Pairs Trading with Orthogonal Distance Regression

Weiyi Chen[1]*

Abstract
In this paper, we will introduce a statistical pairs trading strategy that is a relative value arbitrage on two equities and based on the premise there is a long-run equilibrium between the prices of the stock composing the pairs. We will use orthogonal distance regression to define the degree of deviation from the long-run equilibrium and represent the extent of mutual mis-pricing. Any deviation from the long-run equilibrium is compensated for in subsequent movements of the time series and this pairs trading involves trading on the oscillations about the equilibrium value. We will give example step by step and experimental backtest report of this strategy after introducing the theoretical model based on the econometric paradigm of cointegration and error correction central to the analysis of this pairs-trading strategy.

Keywords
Pairs trading — Orthogonal Distance Regression — Long-run equilibrium

[1] *Yao Class 00, Tsinghua University, Beijing, China*
*__Corresponding author__: weiyi.alan.chen@gmail.com

Contents

Background and Related Work

In order to build a market neutral portfolio, we find pairs that are strongly correlated with one another will remain correlated for some finite period in the future. An easy to see example of a mean-reverting pair of stocks is given in Fig. 1. Using their correlation, we can perform a linear regression[5] on their returns to determine when they have made a significant departure from their expected relationship (spread) and take mean reverting positions profiting on their return to the expected difference.

In this paper we will outline the process for one such pairs trading strategy that will have the following pipeline:

1. Identify pairs of highly correlated equities
2. Select pairs based off of their potential returns
3. Simulate performance of these pairs with back testing

Figure 1. A typical pair found using Linear Regression, we see that entrances to positions occur when the pair of returns has diverged (prices have moved apart from one another) and exits occur when the returns have once again converged (underlying chart from Google Finance).

A more thorough introduction to pairs trading and additional investigations into the underlying theory can be found in Ganapathy Vidyamurthy's book on the same topic (see References).

1. Theory

1.1 Principle

The theme for investing is to sell overvalued and buy the undervalued equities. However, it is possible to determine that a security is overvalued only if we know the true value. But this is hard to do. Pairs trading attempts to resolve this using relative pricing. The specific price of the security will be not of importance. It is only important that the normalized prices of the two securities be the same. If the prices is different, it could be that one of the securities is overpriced, the other security is underpriced, or the mis-pricing[1] is a combination of both.

Pairs trading involves with the idea that the mis-pricing

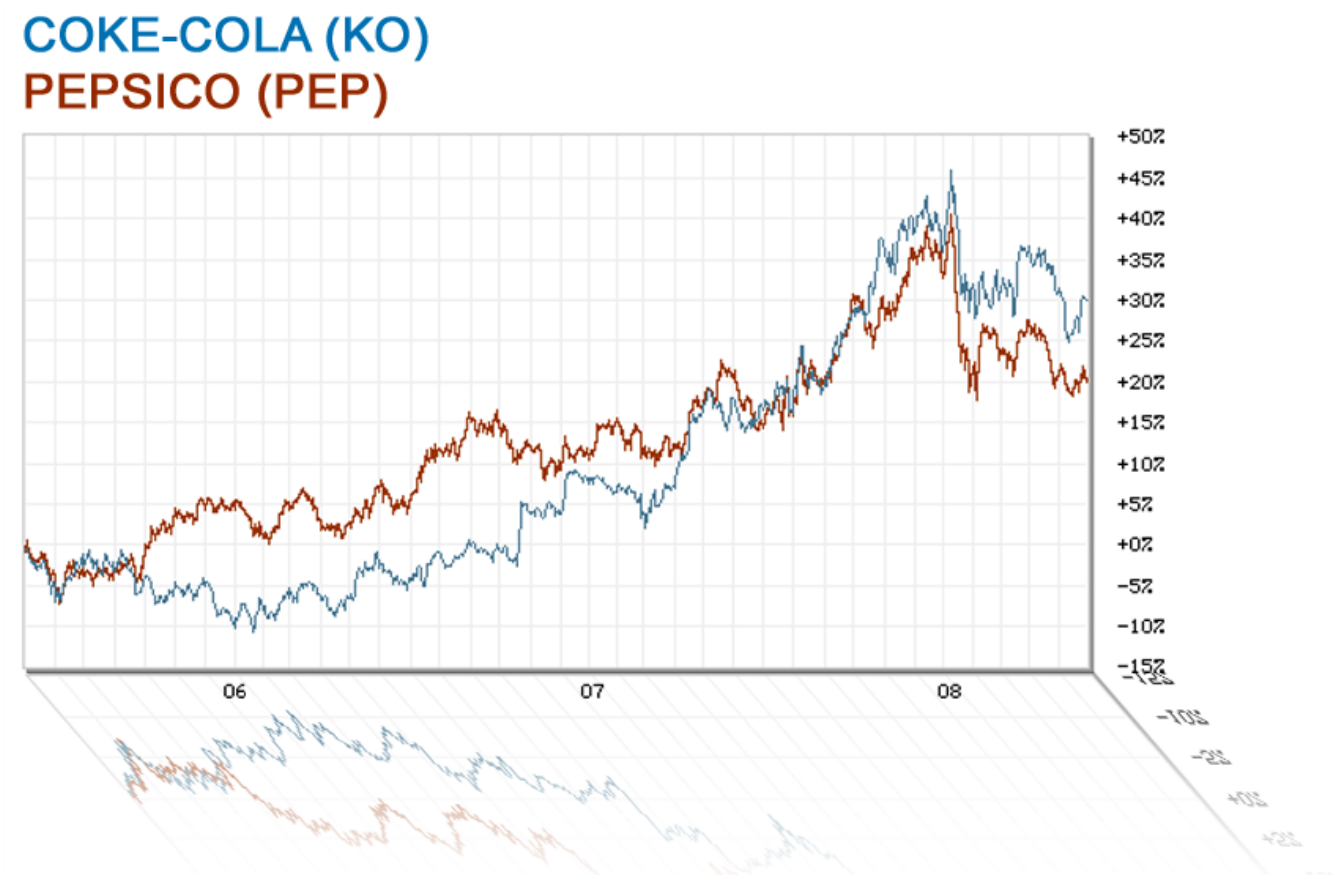


Figure 2. Example: high correlated pair

will correct itself in the future. The mutual mis-pricing between the two securities is captured by the notion of spread. A long-short position is constructed such that it has a beta and therefore minimal exposure to the market. Hence, the returns from the trade are uncorrelated to market returns, a feature typical of market neutral strategies.

Unlike the purely empirical approach, the method that we subscribe to comprises theoretical valuation concepts that are then validated with empirical models and data. We will later show that the theoretical valuation approach helps us to easily identify pairs based on the security historical price. It also leads to the formula used to measure the spread, the degree of mis-pricing between the two securities. According to arbitrage pricing theory[2], if two securities have exactly the same risk factor exposures, then the expected return is the same. The actual return may differ slightly because of different specific returns for the two securities. Let the price of securities A and B at time t be p_t^A and p_t^B, and at time $t+i$ be p_{t+i}^A and p_{t+i}^B, respectively. The return in the time period i for the two securities given as $log(p_t^A) - log(p_{t+i}^A)$ and $log(p_t^B) - log(p_{t+i}^B)$.

Now let us say that we have the prices of both securities at the current time. The return on both securities is expected to be the same in all time frames. In other words, the increment to the logarithm of the prices at the current time must be about the same for both the securities at all time instances in the future. This, of course, means that the time series of the logarithm of the two prices must move together, and the spread calculation formula is therefore based on the difference in the logarithm of the prices.

Having explained our approach, we now need to define in precise terms what we mean when we say that the price series or the log price series of the two securities must move together. The idea of co-movement of two time series has been well developed in the field of econometrics. We discuss it in the following section on cointegration[6].

1.2 Cointegration (correlation)

In time series we are briefly taught the preprocessing step for non-stationary series. The series is typically transformed into a stationary time series by differencing. By extension, when analyzing multivariate time series where each of the component series is non-stationary, it would then make sense to difference each component and then subject them to examination.

Let us now state the idea of cointegration more formally. Let y_t, and x_t be two non-stationary time series. If for a certain value γ, the series $y_t - \gamma x_t$ is stationary, then the two series are said to be co-integrated. Real-life examples of cointegration

abound in economics. In fact, the first demonstrations and tests of cointegration involved economic variable pairs like consumption and income, short-term and long-term rates, the M2 money supply and GDP, and so forth.

The explanation for cointegration is captured by error correction[1]. The idea is that cointegrated systems have a long-run equilibrium. The formal theorem stating that error correction and cointegration are equivalent representations. We shall not attempt to discuss the proof of the theorem, but simply present here for your information.

Let ε_{x_t} be the white noise process corresponding to time series $\{x_t\}$. Let ε_{y_t} be the white noise process corresponding to the time series $\{y_t\}$. The error correction representation is

$$y_t - y_{t-1} = \alpha_y(y_{t-1} - \gamma x_{t-1}) + \varepsilon_{y_t}$$
$$x_t - x_{t-1} = \alpha_y(y_{t-1} - \gamma x_{t-1}) + \varepsilon_{x_t})$$

Let us interpret the above equations. The left-hand side is the increment to the time series at each time step. The right-hand side is the sum of two expressions, the error correction part and the white noise part. Let us look at the error correction part $\alpha_y(y_{t-1} - \gamma x_{t-1})$ from the first equation. The term $y_{t-1} - \gamma x_{t-1}$ is representative of the deviation from the long-run equilibrium (equilibrium value is zero in this case), and γ is the coefficient of cointegration. α_y is the error correction rate, indicative of the speed with which the time series corrects itself to maintain equilibrium. Thus, as the two series evolve with time, deviations from the long-run equilibrium are caused by white noise, and these deviations are subsequently corrected in future time steps.

We will now illustrate that the idea of error correction does indeed lead to a stationary time series for the spread. Two independent white noise se- ries with zero mean and unit standard deviation were generated to represent ε_{y_t} and ε_{x_t}, respectively. The other values were set as $\alpha_y = -0.2, \alpha_x = 0.2$ and $\gamma = 1.0$. Note that it is important to have the two coefficients α_y and α_x set to opposite signs for error-correcting behavior. The values for the two time series $\{x_t\}$ and $\{y_t\}$ were then generated using the simulated data and the equations from the error correction representation. A plot of the two series is show as in Fig. 3.

Subsequently, the spread at each time instance was calculated using the known value for γ. A plot of the spread series and its autocorrelation[4] is shown in Fig. 4 and Fig. 5. It is easy to appreciate from the autocorrelation function that the spread series is indeed stationary.

A more direct approach to model cointegration is attributed to Stock and Watson, called the common trends model[3]. The primary idea of the common trends model is that of a time series being expressed as a simple sum of two component time series: a stationary component and a non-stationary component. If two series are cointegrated, then the cointegrating linear composition acts to nullify the non-stationary components, leaving only the stationary components. To see what we mean, consider two time series

$$y_t = n_{y_t} + \varepsilon_{y_t}$$

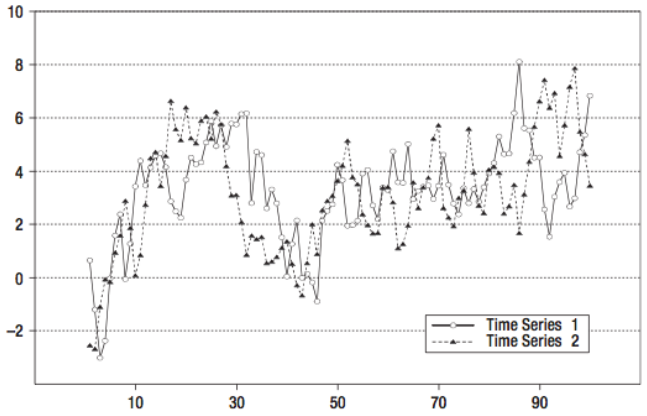


Figure 3

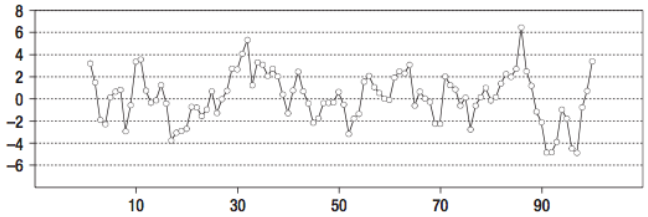


Figure 4

$$z_t = n_{z_t} + \varepsilon_{z_t}$$

where n_{y_t} and n_{z_t} are the random walk (non-stationary) components of the two time series, and ε_{y_t} and ε_{z_t} are the stationary components of the time series. Also, let the linear combination $y_t - \gamma z_t$ be the co-integrating combination that results in a stationary time series. Expanding the linear combination and rearranging some terms, we have

$$y_t - \gamma z_t = (n_{y_t} - \gamma n_{z_t}) + (\varepsilon_{y_t} - \gamma \varepsilon_{z_t})$$

If the combination in above equation must be stationary, the non-stationary component must be zero, implying that $n_{y_t} = \gamma n_{z_t}$, or the trend component of one series must be a scalar multiple of the trend component in the other series. Therefore, for two series to be co-integrated, the trends must be identical up to a scalar. We will rely on the Stock-Watson model[7] to establish cointegration.

1.3 Residual spread (event definition)

In this section, we fit the cointegration model to the logarithm of stock prices. For the cointegration model to apply, we would require the logarithm of stock prices to be a non-stationary series. The assumption that the logarithm of stock prices is a random walk (read as non-stationary) is a rather standard one.

Let us say that two stocks A and B are co-integrated with the non-stationary time series corresponding to them being $\{log(p_t^A)\}$ and $\{log(p_t^B)\}$ respectively. Applying the error

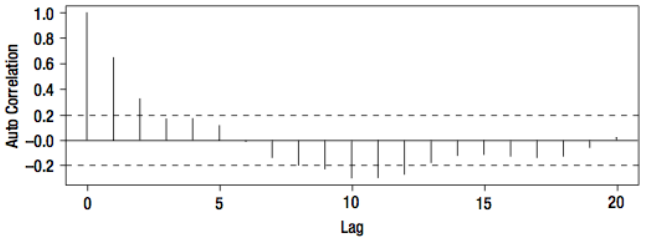


Figure 5

correction representation described here, we have

$$log(p_t^A) - log(p_{t-1}^A) = \alpha_A log(p_{t-1}^A) - \gamma log(p_{t-1}^B) + \varepsilon_A$$

$$log(p_t^B) - log(p_{t-1}^B) = \alpha_B log(p_{t-1}^A) - \gamma log(p_{t-1}^B) + \varepsilon_B$$

The parameters that uniquely determine the model are the cointegration co- efficient γ and the two error correction constants α_A and α_B. Therefore, estimating the model involves determining the appropriate values for α_A, α_B, and γ. The left-hand side of the above equation is the return of the stocks in the current time period. On the right-hand side, note the expression for the long-run equilibrium, $log(p_{t-1}^A) - \gamma log(p_{t-1}^B)$, in both the equations. In words, it is the scaled difference of the logarithm of price. Incidentally, this coincides with what we termed the spread in our earlier discussion. Also notice that the subscripts for stock prices in the expression for the long-run equilibrium is $t-1$. The past deviation from equilibrium plays a role in deciding the next point in the time series. Therefore, knowledge of the past realizations may be used to give us an edge in predicting the increments to the logarithm of prices; that is, returns.

Consider a portfolio with long one share of A and short γ shares of B. The return of the portfolio for a given time period is given as

$$[log(p_{t+i}^A) - log(p_t^A)] - [log(p_{t+i}^B) - log(p_t^B)]$$

. Rearranging the terms a little bit, we have the above equation equals to

$$spread_{t+i} - spread_t$$

Therefore, the return on the portfolio is the increment to the spread value in the time period i. We have successfully associated a portfolio with a stationary time series. The one thing that remains is providing an interpretation for γ, the cointegration coefficient.

2. Methodology

The discussion so far briefly outlines how we might trade once we know two stocks are cointegrated. We do concede that the course of the discussion so far has brought up more questions on the details. How do we identify candidate stock pairs? Can we verify that they are indeed cointegrated? How do we determine the cointegration coefficient? What is the most appropriate value for delta? We explore the questions and issues involved in the subsequent chapters. To that end, we provide a road map for the design and analysis of the pairs trading strategy.

The steps involved are as follows:

1. Identify stock pairs that could potentially be cointegrated. This process can be based on the stock fundamentals or alternately on a pure statistical approach based on historical data. Our preferred approach is to make the stock pair guesses using historical price information.
2. Once the potential pairs are identified, we verify the proposed hypothesis that the stock pairs are indeed cointegrated based on statistical evidence from historical data. This involves determining the cointegration coefficient and examining the spread time series to ensure that it is stationary and mean reverting.
3. We then examine the cointegrated pairs to determine the delta. A feasible delta that can be traded on will be substantially greater than the slippage encountered due to the bid-ask spreads in the stocks. We also indicate methods to compute holding periods.

2.1 Finding and Selecting Pairs

The first step in our pairs trading strategy is to identify pairs of highly correlated equities. This is done by calculating the correlation between all pairs in a given basket over a look-back period prior to the date of interest. For this paper the S&P 100 was used. Once the correlation matrix of all pairs has been calculated, we use a threshold to give us only pairs with a high degree of correlation (around 0.97). Table 1 contains 10 highly correlated pairs from our trial typical of those found on some day.

Table 1. list of top 10 highest correlated pairs on Dec 10th

Index	Equity 1	Equity 2	Correlation
1	LMT	RTN	0.994625
2	LMT	SBUX	0.982215
3	GD	LMT	0.982196
4	MET	WFC	0.979455
5	MMM	TXN	0.978430
6	FOXA	TWX	0.978277
7	BA	SBYX	0.978092
8	GD	RTN	0.978003
9	TXN	UTX	0.977723
10	HON	MMM	0.977600

Once we have a set of highly correlated pairs, we then calculate their cumulative returns over the look back period. We then plot the returns for each pair and find the best fit line for these returns (see Fig. 6). This best fit line is the baseline that we will compare all future return pairs to. For all of the past returns in our preceding period, we determine the

standard deviation of their distance from the best-fit line that we will use as our signal to determine when to enter a position in this pair. The best-fit line will be used to determine when to exit positions.

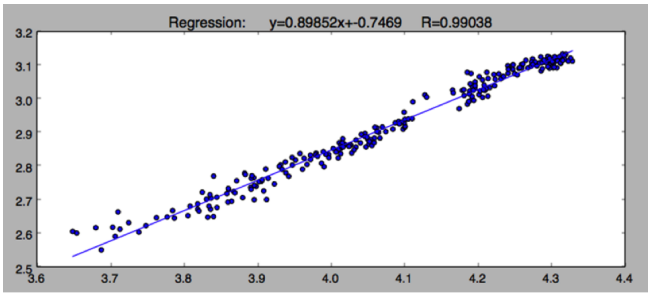


Figure 6. A typical set of returns for a pair (APD, CSCO here) of highly correlated equities with a best-fit line plotted.

Equipped with the best fit line and standard deviation of the distance of return pairs from the best fit line (this distance is called the residual of the pair), we will enter positions when the residual is more than one standard deviation away from the best fit line, expecting mean reversion to lead this pair?s returns back across the best fit line. When the residual crosses the best-fit line, we then exit the position having made a profit. See Fig. 7 for an example of the residual for a pair.

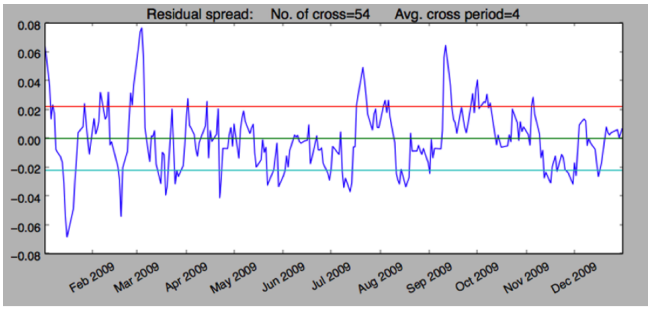


Figure 7. A typical residual spread (the distance each pair of returns is from the best fit line) for with standard deviation lines marked above and below the central line (which is representative of the best fit line).

To determine which pairs will be most profitable, note that profit is made based off of two factors: (1) distance of the residual from the best fit line when we enter a position, and (2) the frequency at which the residual crosses the best fit line. The product of these two factors yields the profitability of a given pair, thus we look for those pairs that are historically more profitable in this sense, and will use these pairs in our back test trades. See Table 2 for a list of the top 10 most profitable pairs for a given period in our back test.

2.2 Back Tests

We implemented a trading strategy in simulation as follows:

1. On each trading day, we generate the most profitable pairs for each two stocks in S&P 100 (as described above). We enter an equally weighted position when the residual is greater than 1 standard deviation away from the best-fit line (e.g., long in A, short in B if above +1 sigma; short in A, long in B if below -1 sigma).

2. When the residual crosses the best-fit line or position is held a month, exit the position.

3. We add transaction costs by modeling commissions, slippage penalties and short interests. Commission are set at $0.0035 per share with a $2.95 minimum per trade, slippage is estimated at 5bps for each transaction (including that the price will move against the trader by 0.05% when entering and exiting a position), and short interest rate is estimated at 3.5% annually.

Table 2. list of top 10 most profitable pairs on Dec 10th, 2013

Index	Equity 1	Equity 2	Correlation	Profitablity
1	DD	GILD	0.971380	69.62
2	MDT	MET	0.971099	63.47
3	FOXA	GILD	0.971349	61.75
4	GD	RTN	0.978003	56.64
5	LMT	SBUX	0.982215	49.76
6	FOXA	TWX	0.978277	46.04
7	RTN	SBUX	0.972998	39.98
8	GD	LMT	0.982196	39.90
9	COP	MA	0.974329	36.99
10	TXN	UTX	0.977723	32.91

The first back test as showed in Fig. 8 shows very positive results, with a consistent average Sharpe ratio of more than 1.0 each year.

In our second back test we remove the process to assess which pairs will be most profitable, and solely use the most correlated pairs. As you can see in Fig. 9, performance is still positive but reduced in comparison with the back test considering profitability process.

Finally, as a control, we ran a back test in which the pairs were selected randomly from the selected correlated pairs each day. The purpose of a randomly selected portfolio is to illustrate the performance we would expect if there were no predictive information. Performance of that test is illustrated in Fig. 10 in the last page.

As you can see, performance for the random correlated pairs is significantly worse than for our test cases. More detailed back test reports, including specific transactions are available in attachments.

3. Discussion and Further study

3.1 Discussion

We propose a pairs selecting and corresponded event triggering criteria based on relative pricing and the idea that

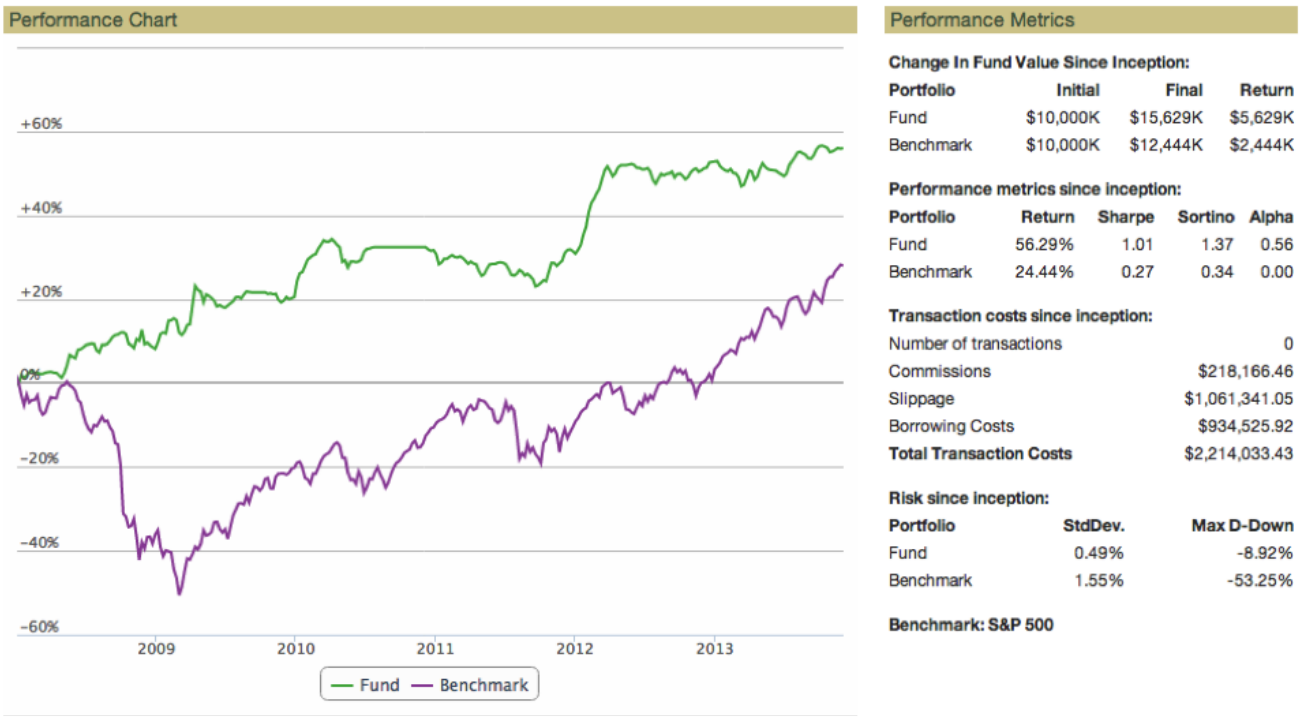


Figure 8. Performance of the ODR pairs trading strategy for 2008 to 2013 with most profitable pairs.

mis-pricing will correct itself in the future. The event predicts higher returns for equities whose recent performance has had low returns than for equities whose recent performance have over-performance.

The criteria were utilized in a pairs trading strategy in which we enter equally weighted long and short positions in event-triggered pairs. The strategy was evaluated in three back tests:

1. With most profitable pairs;
2. With most correlated pairs;
3. With randomly correlated pairs.

As we remove requirements for profitability assessing (Fig. 9) and correlation ranking (Fig. 10), the back tests significantly underperform the former test that includes these factors (Fig. 8). This suggests that there is actionable information in our pairs trading strategy.

3.2 Further study

There are some important factors to consider in validation, like time travel, market behavior and so on. For the time travel, we can look at specifically one pair to track record how or whether they successfully converge. By doing this study, we will be more successfully modify the pairs selection to choose those potential ones. For market behavior, based on current research the pairs trading strategy did a great job when the market is bearish since this is a market neutral strategy. But on the other hand, like the market in 2013 is really bullish, it?s hard for pairs trading to catch up with the benchmark with super high returns this year.

Also, we have other important factors to consider in optimization, which are itemized as follows.

1. pairs universe (S&P 500 vs. S&P 100 ? larger universe is better, but has much larger computational expense for backtests ? O(N2); also, need to account for any stocks that could have limitations on shorting)
2. look-back period: # pairs meeting correlation threshold
3. ranking period: potential profitability scores
4. correlation (cointegration) threshold: sensitivity to different thresh- olds (# events, potential profitability measure, and back-tested return and Sharpe)
5. number of crossings / std dev for ranking: sensitivity of backtested return and Sharpe to different thresholds for number of crossings and standard deviation score

Acknowledgments

So long and thanks for course advisor Zhaoguo Zhan's teaching notes on statistics and time series analysis.

References

[1] Engle, Robert F. and C. W. Granger. (1987) *Co-integration and Error Correction: Repre- sentation, Estimation and Testing.*, econometrica 55, no. 2: 251?253.

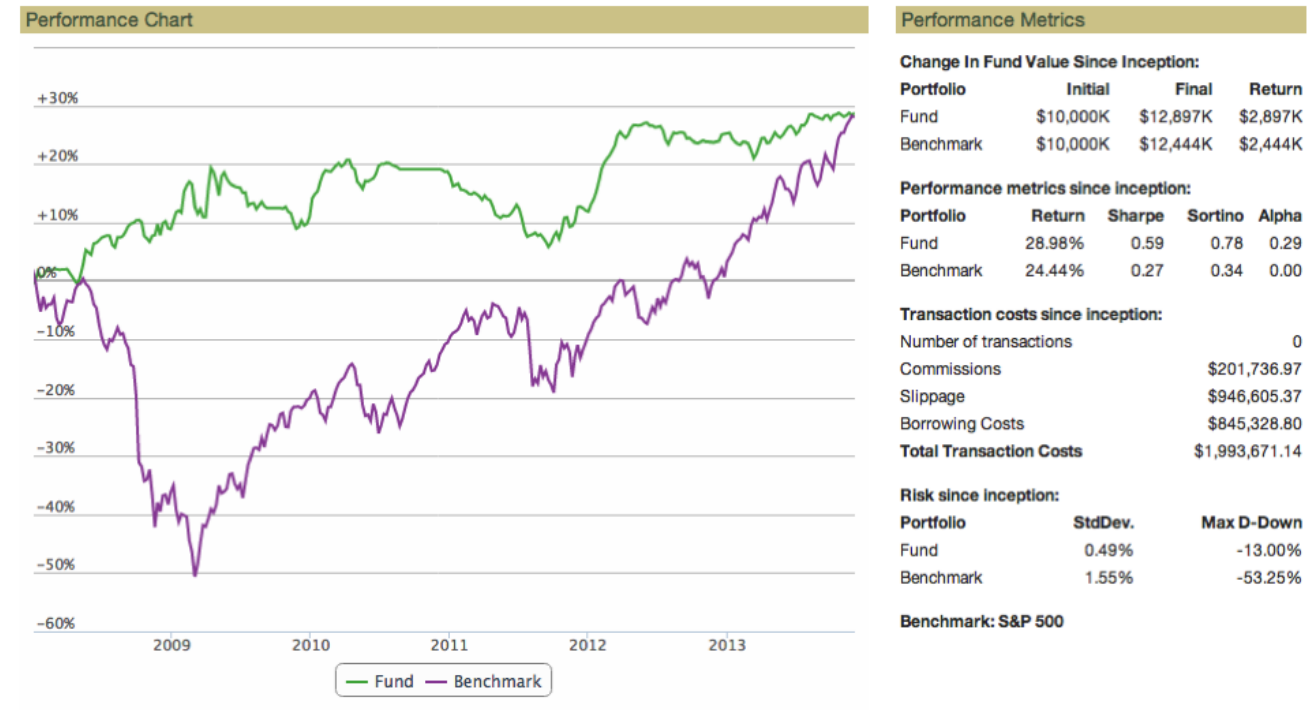


Figure 9. Performance of the ODR pairs trading strategy with most correlated pairs.

[2] Gatev, Evan, G., William, N. Goetzmann, and K. Greet Rouwenhorst. (1999) *Pairs Trading: Performance of a Relative Value Arbitrage Rule.*, NBER Working Papers 7032

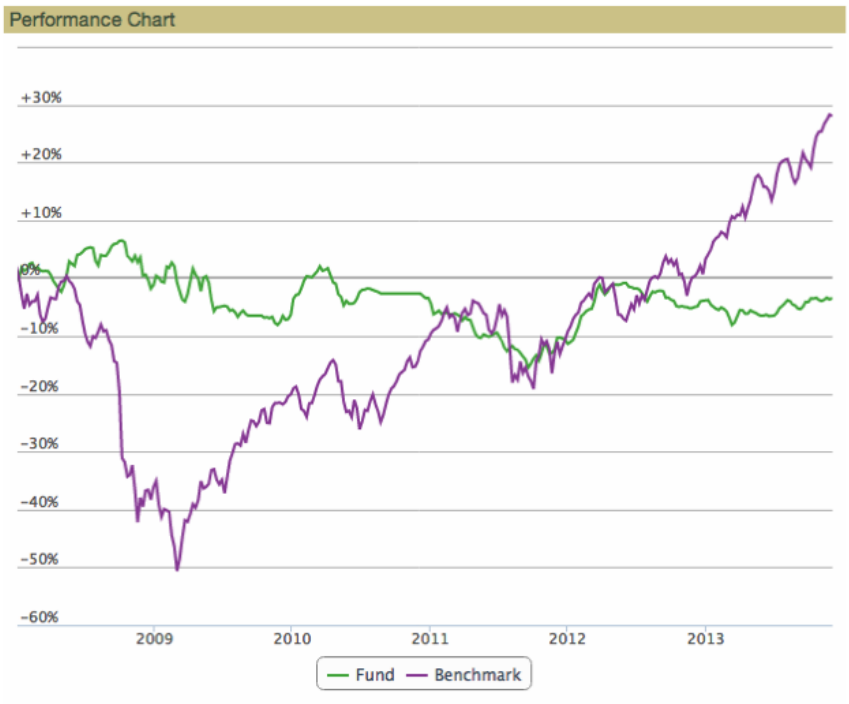
[3] Stock, James H. and Mark W. Watson. (1988) *Testing for Common Trends.*, Journal of the American Statistical Association 83, no. 404: 1097?1099.

[4] Zhaoguo Zhan *Time Series Analysis*, Teaching notes 8: 110-112

[5] Philips, P. C. B. and S. N. Durlauf. *Multiple Time Series Regression with Integrated Processes*, Review of Economic Studies 53: 473.

[6] Park, J. Y. S. Oularis and B. Choi. *Spurious Regressions and Tests for Co-integration*, CAE working paper 88?07

[7] Ganapathy Vidyamurthy *Quantitative Methods and Analysis*, Wiley Finance: 73-80



Performance Metrics

Change In Fund Value Since Inception:

Portfolio	Initial	Final	Return
Fund	$10,000K	$9,673K	($326,185.27)
Benchmark	$10,000K	$12,444K	$2,444K

Performance metrics since inception:

Portfolio	Return	Sharpe	Sortino	Alpha
Fund	-3.26%	-0.03	-0.04	-0.03
Benchmark	24.44%	0.27	0.34	0.00

Transaction costs since inception:

Number of transactions	0
Commissions	$164,410.10
Slippage	$772,381.61
Borrowing Costs	$724,993.64
Total Transaction Costs	$1,661,785.35

Risk since inception:

Portfolio	StdDev.	Max D-Down
Fund	0.50%	-20.88%
Benchmark	1.55%	-53.25%

Benchmark: S&P 500

Figure 10. Performance of the ODR pairs trading strategy in which correlated pairs were selected randomly.