CIB CAPITAL INVESTMENTS AT BERKELEY
www.ciberkeley.com

Multivariate Pairs Trading: Identification and Modeling of Highly Cointegrated Equity Bundles

Vincent Liu
Capital Investments at Berkeley
University of California, Berkeley
Email: vincent.liu33@berkeley.edu

Tejas Thvar
Capital Investments at Berkeley
University of California, Berkeley
Email: tthvar@berkeley.edu

Abhishek Kumar
Capital Investments at Berkeley
University of California, Berkeley
Email: abhishek.kumar@berkeley.edu

Pairs trading is a market neutral statistical arbitrage technique profiting from the correction of divergences in highly correlated pairs of assets. We found that trading single pairs of equities resulted in erratic and fleeting divergences that would be unable to trade on. We maximized our divergence time and optimized correctional behavior through identifying and trading on highly correlated bundles of equities with existing market indices.

1 Introduction

Statistical arbitrage pairs trading consists of two main stages: selection of securities who have historically moved together during a formation period, and modelling/monitoring the spread [difference in prices] between them during a trading period. (Krauss, 2015). If and when the spread widens beyond a certain identified threshold, you expect the spread to revert back to its long term mean. In our case, going short on our ETF and long on our bundle is equivalent to a short position on the spread which results in a profit as the spread narrows back towards its mean. Pairs trading can also be extended to multivariate settings, trading one asset against a bundle/portfolio of securities (Vidyamurthy, 2004).

Identification of comoving securities has been accomplished through a variety of approaches, including leveraging distance metrics and cointegration testing. Similarly, modelling spread can be implemented via stochastic control theory, and a time series approach. (Jurek and Yang 2007, Gatev et. al 2006).

In this project, we utilized cointegration testing to determine our initial set of comoving securities. Cointegration is a statistical property of non-stationary (i.e. securities) time series variables whose spread tends towards some long-term equilibrium. In other words, a cointegrated relationship between two time series variables implies that a linear combination of these variables is stationary and mean-reverting. As a whole, cointegration is an effective method of ignoring the effect of spurious correlation when regressing several variables.

We also applied a time series approach to model the spread as a mean-reverting process by use of Kalman Filtering. Kalman filtering is an effective state estimation technique, in broader terms combining approximations of a certain variable to improve existing approximations. By dynamically updating predicted states, error covariance will constantly decrease implying a higher degree of certainty with future estimates.

2 Algorithm Overview

Our pairs trading algorithm consists of two modules: pair selection and pairs trading. Pair selection consists of identifying (and in this case constructing) a highly cointegrated pair of assets to ensure a mean reverting and stationary spread to trade upon. The constructed spread signal is then fed into our pair trading module, which identifies and executes optimal trades.

2.1 Pair Selection

Multiple frameworks were tested in order to create a robust pipeline for constructing and identifying a highly cointegrated asset with the QQQ market index. Traditional pairs trading compares single equities, but diversification by bundling allowed for a much stronger and highly co-moving relationship. There are many methods to assess spread-stationarity / a cointegrated relationship, but this was accomplished via both the Johansen and Engle-Granger cointegra-

tion tests which both demonstrate key benefits.

2.1.1 Methods of Cointegration Testing

One method of cointegration testing that was evaluated was the Johansen Test. The Johansen test takes in a set of time series and iteratively checks if some amount of these time series are cointegrating. It is a more robust cointegration than the Engle-Granger Test in the case of multiple time series and lends itself well to the case of determining cointegrated equity bundles. There are two possible versions of the Johansen Test based off of what you are using to measure the test statistic. The two tests are based off the trace and eigenvalues of a generated matrix and the largest difference between the two tests is the alternative hypothesis. With the trace test, the alternative hypothesis states that, at iteration k, the number of cointegrating time series is equal to k. In the case of the eigenvalue test, it states that the number of cointegrating time series is equal to k+1. We used the trace statistic test when applying the Johansen test.

Another method of cointegration testing utilized was the Engle-Granger test. This test constructs residuals between time series based on static regression. In order for a cointegrated relationship to exist, these residuals must be essentially stationary. This stationarity is evaluated through the Augmented Dicky-Fuller (ADF) test. The ADF test consists of identifying unit roots, or in a time series with value $\alpha = 1$ in the following equation, where Y_t is the value of the time series instantaneously at a given time t and X_e is a exogenous time series:

$$Y_t = \alpha * Y_{t-1} + \beta * X_e + \varepsilon$$

The ADF tests that α is 1 in a modified version of the above model, containing additional lagged difference terms. It is a robust stationarity test, able to handle more complex time series than regular Dickey Fuller testing.

Engle-Granger is thought to be a more robust identifier of cointegrated relationships than the Johansen test and was thus included its own synthetic basket (Gonzalo & Lee 1997)

2.1.2 Optimal Cointegration Testing for Bundles

As they each had proven advantages over the other for cointegrated asset construction, we utilized both the Engle-Granger and Johansen tests to filter equities to construct optimally mean-reverting equity bundles with the QQQ NASDAQ market index. We noticed that baskets of equities had more statistically significant evidence of cointegration than single equities. This behavior suggests tradeable opportunities for arbitrage.

2.2 Optimally Cointegrated Bundle Construction

We first ran cointegration tests on each NASDAQ100 component present in 2018 and 2019 against the QQQ ETF [NASDAQ market index] with a testing period of 2018. Each cointegration test yielded a p-value which we used to filter components, only allowing individual equities with a p-value below 0.2 [i.e. statistically significant cointegrated relationship]. We then created a powerset of this filtered subset and conducted cointegration testing on each of the elements of the powerset and QQQ. We then ranked these cointegration results by p-value, with the lowest p-value and so most statistically significant evidence of cointegration being the bundles HAS, TTWO, IDXX, SBUX, CTAS, ALXN with the Johansen test and HAS, AAPL, TTWO, SBUX, CTAS, ALXN, ALGN, PAYX with the Engle-Granger test. This resulted in highly QQQ-cointegrated assets with high potential for arbitrage.

2.3 Pair Trading

2.3.1 Kalman Filtering

Kalman Filters are linear state space estimators built on the principles of linear regression that account for noise in a stochastic processes by continuously updating the parameters of the filter at every time step. At a high level, the filter begins by predicting a value at a certain time step, then takes in the observed value at that time and adjusts its parameters for future iterations. The various update equations adjust the values such as the variance of the noise of the observed value, variance among iterations, among other parameters. For more details on Kalman Filters, the derivation of its equations, and applications, refer to Walrand's *Probability in Electrical Engineering and Computer Science: An Application-Driven Course*.

The decision to use a Kalman Filter is not novel as it is one of the most frequently used state space models in finance. As seen in Drakos' work, a Kalman Filter is an effective dynamic estimator for stock modeling. Modeling time series with a Kalman Filter was a common technique across much of the literature we read and found it to be a better estimator than an ordinary rolling linear regression. We utilized a Kalman Filter to model the spread between our equity bundle and ETF to determine when to enter and exit positions. At every timestep, the Kalman Filter outputs estimates for the hedge ratios of each component in our synthetic basket as well as an intercept term. In other words, it gave us all the tools to calculate an estimate for the price of QQQ based on the price of each component in our basket. The residuals of this estimate form a mean-reverting and stationary process that we can take advantage of.

2.3.2 Modeling Entries and Exits

We modeled entries and exits by using a Bollinger-band based strategy. We utilized both static and dynamic thresholds when identifying opportunities for arbitrage.

The dynamic strategy considers volatility (standard deviation) of the spread when constructing entries and exits. A lookback period is chosen and an exponential moving average is computed for that lookback period. A constant is also chosen that represents how many standard deviations the bands should encompass. When the spread exceeds the upper band, a short signal is generated. When the spread is below the lower band, a long signal is generated. The signal

ends when the spread crosses the exponential moving average again. Positions are taken in the minute immediately after the signal is generated. Positions are also closed by the end of the day and no overnight positions are ever taken.

Minute level data from 2018 is used to tune the parameters of the dynamic entries and exits. These parameters include a lookback period and how wide the bands should be.

To optimize our entries and exits, we determine the distribution of spread percentage changes from minute to minute in our in-sample period. The bottom 20 % threshold will be used as a condition for entering positions. A graph of dynamic thresholds is included in Figure 2.

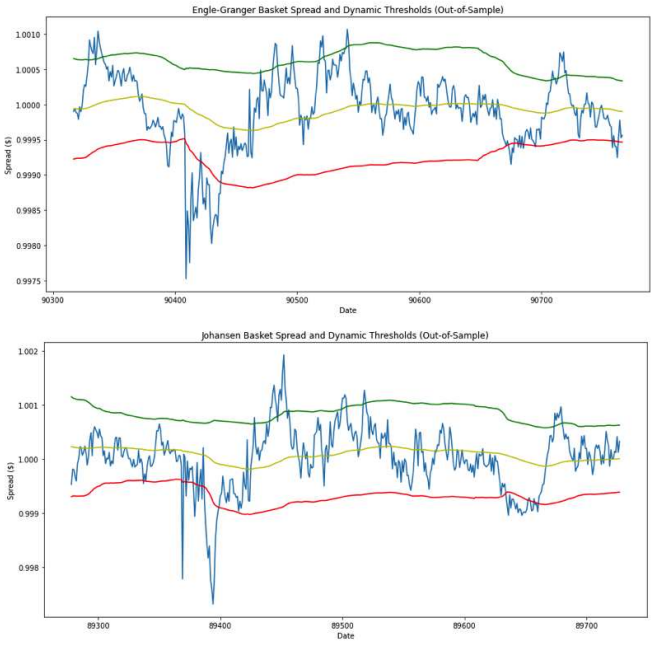

Figure 1. Dynamic Thresholds for Cointegrated Bundles

The static strategy uses minute level data from our in-sample period (2018) to compute the mean and standard deviation of the spread. We then used a constant multiple of the standard deviation as the upper and lower thresholds for our mean reversion strategy. Signals are generated during our trading period when our spread is above/below our upper/lower threshold. Note, the actual positions we take need to be lagged by 1 minute from our signals. Without this shift, our results would be reliant on high-frequency trading technology and latency. A graph of static thresholds is included in Figure 3.

We implemented various methodologies to further optimize entry and exit points. We first attempted to implement the OU model framework, which fits the Ohrstein-Ulbeck process to tested time series to output estimates of long term mean level, mean-reversion speed, and volatility at optimal levels. This can be treated as an optimal stopping problem to then return optimal entry and liquidation levels for a portfolio consisting of a long and short of the two co integrated

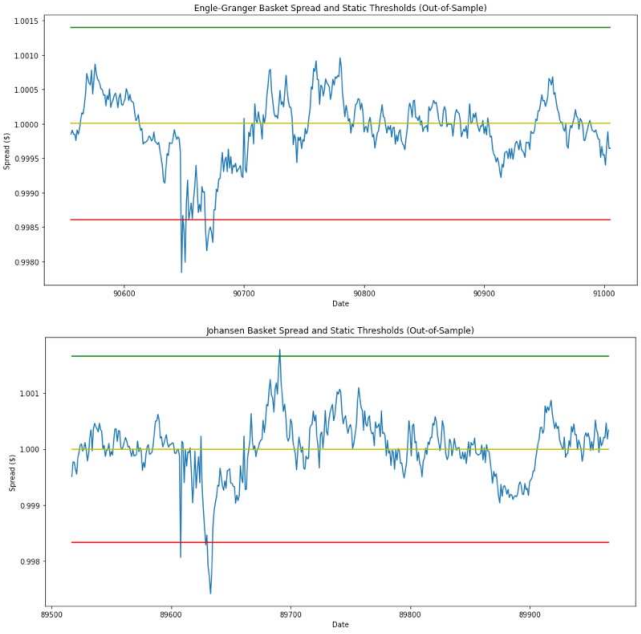

Figure 2. Static Thresholds for Cointegrated Bundles

assets.

Another method we had tried involved predicting the next value of the spread. For example, if we knew the spread up to time step k, referred to the value at k as s_k, we estimated that $s_{k+1} = s_k + (s_k - s_{k-1})$. This formula assumes that the value at the next time step is equal to the sum of current time step and the difference between the current time step and the one prior, forming some sort of basic regression. If the estimated value was above or below our designated z-threshold, we would enter a position.

We also attempted to find the peak or dip in the spread and take a position there. Assuming the derivative of the time series at time step k was $s_k - s_{k-1}$ and using threshold t, we took a position any time $s_k - s_{k-1} < t$ and s_k was outside of our z-threshold. This idea was based off the principle that a maximum or minimum is when the first derivative was equal to 0. We generated the threshold t by looking at the spread changes across the previous trading year and used the 20th percentile, to ensure that we only took positions when we were most confident that the derivative was closest to 0.

3 Backtest Results

Our backtest's trading period begins on January 1, 2019 and finishes on December 31, 2019. All the parameters for each strategy are determined using an in-sample period beginning on January 1, 2018 and ending on December 31, 2018.

When constructing cointegrated bundles, we used two methods of cointegration. These were the Engle-Granger test and the Johansen test. Both are discussed earlier and both tests yielded different bundles. As mentioned previously, the Engle-Granger cointegration test was evaluated based on the

strength of cointegration between QQQ and the sum of the stocks in our bundle. The Johansen cointegration test was evaluated based on the cointegration score of all the stocks in our bundle and QQQ.

Two different strategies are evaluated for each of these bundles. The difference in these strategies lies in how entries and exits are determined. For both strategies, we use the in-sample period (2018) to tune parameters.

In the static strategy, we find the mean and standard deviation of the spread during our in-sample period. Our entry threshold to open a long position is 1.25 standard deviations below the spread mean. Our entry threshold to open a short position is 1.25 standard deviations above the spread mean. In both cases, we exit our trade when the spread returns back to the mean.

In the dynamic strategy, we simulate trading during the in-sample period with entries and exits determined using a variety of parameters. The most successful set of parameters are then adopted for our out-of-sample period. For this strategy, we tested every combination of a lookback period in the range of 30 to 300 minutes (30 minute intervals) and a z-score threshold in the range of 1 to 2.5 (intervals of 0.25).

To evaluate the performance of the various methods we used to calculate entries and exits, we created a function to create a trade log. Each row of the trade log is a distinct trade executed by the strategy and contains information about the trade including start and end time, holding period, whether we were long or short, position sizes, entry and exit prices and profit data.

In order to avoid using fractional shares in the construction of our synthetic bundle, a lot size of 1000 shares was used. In every trade in the backtest, 1000 shares of QQQ were traded against the corresponding hedge ratios of each component of the bundle. To make sure that each component of the synthetic bundle was a whole number of shares, hedge ratios were rounded to the nearest thousandth.

While the spread data generated by our Kalman Filter on our bundle's price data has hedge ratios that update every minute, the hedge ratios at the minute of entry determine the position sizes we take for the entirety of our trade. This means that the profit of our strategy cannot be directly computed from our spread data.

Below are summary tables of our results. The first table contains the results of the strategies using the cointegrated bundle generated with the Engle-Granger cointegration test. The second table contains the results of the strategies using the cointegrated bundles generated with the Johansen cointegration test. Both cointegrated bundles are back-tested with static and dynamic entries and exits, with and without the position optimization function.

Table 1. Top Cointegrated Bundles

Engle-Granger	**Johansen**
1. HAS, AAPL, TTWO, SBUX, CTAS, ALXN, ALGN, PAYX	1. HAS, TTWO, IDXX, SBUX, CTAS, ALXN
2. HAS, AAPL, TTWO, SBUX, CTAS, ALXN, ALGN	2, HAS, TTWO, SBUX, CTAS, ALXN
3. HAS, AAPL, TTWO, SBUX, CTAS, ALXN, ALGN, BMRN	3. HAS, TTWO, IDXX, SBUX, CTAS, ALXN, PAYX
4. HAS, AAPL, SBUX, CTAS, ALXN, ALGN, BMRN, PAYX	4. HAS, TTWO, SBUX, CTAS, ALXN, PAYX
5. HAS, AAPL, TTWO, SBUX, CTAS, ALXN, ALGN, BMRN, PAYX	5. HAS, TTWO, IDXX, SBUX, ALXN

Table 2. Optimized Static Thresholds

	Lower Bar Threshold (mean - 1.25 * std)	**Mean**	**Upper Bar Threshold (mean + 1.25 * stddev)**
Engle-Granger Bundle	0.9985	1.0000	1.0015
Johansen Bundle	0.9982	0.9999	1.0018

Table 3. Optimized Dynamic Threshold Parameters

	Lookback period	**Z-Score**
Engle-Granger Bundle	240	1.25
Johansen Bundle	240	1.25

4 Conclusions/Further Work

In general, our attempts to optimize the strategy and improve it were successful. The bundles that we construct are a unique and uncommon approach to statistical arbitrage. Because these bundles are created based on various tests of cointegration, they are sometimes nonobvious. Compared to our initial tests with SPY and VOO (Vanguard's version of SPY) where mean reversion happened rapidly at the tick level and there wasn't much opportunity for arbitrage, the bundles that we use in our approach have relationships that are more realistically exploitable.

The results indicate that our strategy is unable to capture a significant amount of profit. There are many reasons for why this could be the case. As mentioned earlier, the spread data generated is based on time adapting hedge ratios that update every minute. However, when we actually trade this strategy, the hedge ratios at the minute of entry determine our position sizes for the entire trade. This disconnect between

Table 4. Engle-Granger Cointegrated Bundle Trading Results

	Position Optimization	Returns (%)	Sharpe Ratio	Average Holding Period (minutes)	Average Trades per Day
Static Entry / Exit	No	-8.023	-2.460	27.202	2.781
	Yes	-1.733	-0.942	29.787	1.662
Dynamic Entry / Exit	No	-5.492	-1.536	18.859	7.171
	Yes	-0.647	-0.282	20.237	3.831

Table 5. Johansen Cointegrated Bundle Trading Results

	Position Optimization	Returns (%)	Sharpe Ratio	Average Holding Period (minutes)	Average Trades per Day
Static Entry / Exit	No	-4.068	-1.165	30.207	2.717
	Yes	-0.869	-0.348	32.983	1.637
Dynamic Entry / Exit	No	-0.841	-0.184	20.159	6.758
	Yes	1.391	0.525	21.140	3.843

our ideal spread data and our actual position is one of the reasons for this strategy's relatively poor performance.

Another possible reason for this strategy's underperformance lies in our entries and exits. While we have successfully applied various methods to optimize our entries and exits, these methods can still be improved further. Improving the modeling of the mean-reversion process would help us set better entries and exits and improve our backtest. Some options for doing so include fitting the Ornstein-Uhlenbeck process to our spread series, using machine learning methods to optimize entries and exits or other mathematical tools and approaches to maximizing the amount of mean reversion captured by our strategy.

Moreover, trade selection could also be optimized by implementing the Kelly Criterion or other bet-sizing or allocation ratio algorithms. This could help us pick our trades and lean into the trades we are more confident about and expose ourselves less to the trades we are less confident about.

Finally, it would also be interesting to apply this framework to products other than indices. While we explored the cointegration relationships between the NASDAQ100 and its components in our approach, another approach could include exploring cointegration relationships between an individual equity and a hand-picked basket of related stocks. For example, applying this framework to AMZN and a bundle of technology, e-commerce and cloud stocks could yield interesting results.

Besides equities, this framework can definitely be applied to any other asset class and even apply to products across multiple asset classes as long as they are comoving or cointegrated in some manner. This framework can be used in a variety of different ways to explore cointegrated relationships within and among all asset classes.

5 Bibliography

1. Avellaneda, Marco, and Jeong-Hyun Lee. Statistical Arbitrage in the U.S. Equities Market. Department of Mathematics, New York University.
2. Chen, Huafeng, et al. Empirical Investigation of an Equity Pairs Trading Strategy. Tsinghua University, PBC School of Finance.
3. Drakos, Stefanos. Statistical Arbitrage in S&P500. Journal of Mathematical Finance.
4. Endres, S, and J Stübinger. "Optimal Trading Strategies for Lévy-Driven Ornstein–Uhlenbeck Processes." Taylor & Francis.
5. Krauss, Christopher. Statistical Arbitrage Pairs Trading Strategies: Review and Outlook. Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics, 2015.
6. Jean Walrand. Probability in Electrical Engineering and Computer Science: An Application-Driven Course. Quorum Books, 2014.