NICHOLAS HAMMOND
BSc (Hons) Economics and Econometrics

Exploring Inefficiencies Uncorrected by Authorised Participants and Other Agents within the Exchange-Traded Fund Market Using Statistical Arbitrage

Submitted May 2024
for the degree of Bachelor of Science
Department of Economics and Related Studies
University of York

Supervisor:

Charalambos (Pambos) Pattichis

Words: 5000

Abstract

Whilst many studies investigate the incidence of ETF mispricing or if there have been opportunities to execute an arbitrage strategy involving ETFs and their underlying assets, there is little public literature that attempts to backtest a potential strategy to ultimately make the market more efficient. We built and backtested a pairs trading strategy based on the concept of statistical arbitrage to trade two ETFs, SPY and QQQ, against two of their mutually dominant underlying equities, AAPL and MSFT, to determine whether APs and other arbitrageurs have maintained an efficient market. The strategy returned -3.93% and a Sharpe ratio of -13.16 during the backtesting period between 2 January 2014 and 31 December 2023, providing evidence that APs successfully maintained an efficient ETF market during this period.

Table of Contents

Table of Contents 3
1. Introduction 4
2. Literature Review 6
   - 2.1. Theoretical Framework 6
   - 2.2. Empirical Studies Evaluating the Efficacy of Basic Pairs Trading and More Sophisticated Statistical Arbitrage Strategies 6
   - 2.3. Empirical Studies Analysing ETF Mispricing and Arbitrage Opportunities 7
   - 2.4. Improvements on Previous Work & Current Gaps in Existing Literature 7
3. Data and Methods 9
   - 3.1. Economic Hypothesis 9
   - 3.2. Descriptive Statistics of Data 9
   - 3.3. Using Cointegration to Find Viable Candidates and Calculating the Rate of Decay 9
   - 3.4. Calculating Stochastic Forecasts of the Spread 11
   - 3.5. Trading Signal Generation and Portfolio Allocation 12
   - 3.6. Updating the Spread 12
   - 3.7. Measuring Strategy Performance 13
4. Analysis & Interpretation of Results 14
   - 4.1. Portfolio Backtesting 14
   - 4.2. Risks 17
   - 4.3. Improvements Towards Future Works 19
5. Conclusion 20

Bibliography 21

Appendix A: Identified Cointegrated Pairs 24

Appendix B: Sample Data Output for AAPL x SPY (Training Data from 2014Q3 and Trading Data from 2014Q4) 25

Appendix C: Identified Trading Signals: 26

1. Introduction

Exchange-Traded Funds (ETFs) were first listed in January 1993 with *SPY*, a fund passively tracking the S&P 500 Index. Since its introduction, the index has been hailed as the best single gauge of large-cap U.S. equities (S&P Dow Jones Indices, 2024). ETFs were designed to operate differently to regular equities, with the power to control the number of shares in circulation shifted onto Authorised Participants (APs), a third party to the ETF issuer. The issuer grants APs the power to control dilution by either exchanging a proportional value of the underlying securities for ETF shares, or by redeeming ETF shares for the underlying assets. APs, therefore, have a unique opportunity to generate a return by correcting a market mispricing, known as the *ETF Arbitrage Mechanism*, shown in Figure 1 below.

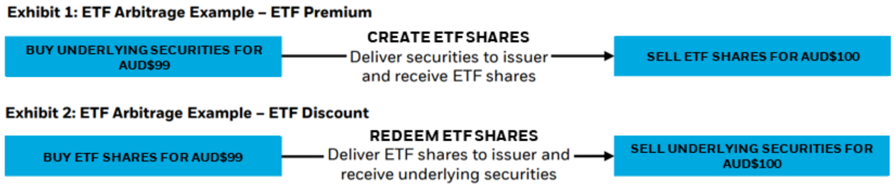

Figure 1: The ETF Arbitrage Mechanism in action (Blackrock, n.d.).

Whilst market makers have a contractual obligation to the exchange to maintain liquidity of certain products, APs have a natural informational (and contractual) advantage as they are the sole party who can control the liquidity of ETFs. Therefore, they should be able to consistently maintain an efficient market. This should prevent other economic agents from generating a profit by creating a portfolio that mimics this process, holding ETF shares hedged against the underlying securities.

This dissertation aimed to test the efficiency of APs by constructing and backtesting a trading strategy that utilised the concept of statistical arbitrage applied to the ETF market and its underlying securities. Statistical arbitrage is a bottom-up methodology developed by a team within Morgan Stanley in the 1980s which emerged from basic pairs trading strategies (i.e., long/short equity based upon fundamental competitor analysis). It uses a contrarian mean reversion principle that generates long/short signals for the assets within the investable universe and has grown to become a popular method amongst hedge funds to generate alpha.

Our conjecture within this research is that an arbitrage opening emerges due to market volatility, which may generate a profit if one invests in a group of assets and waits for mean reversion to return its spread back to equilibrium.

For the purpose of this research, the investable universe consists of two ETFs, *SPY*[1] and *QQQ*[2], and two equities, *AAPL*[3] and *MSFT*[4]. All four assets are dollar-denominated assets listed on U.S. exchanges; SPY is listed on the New York Stock Exchange (NYSE) Arca, and the others are listed on the Nasdaq. By focusing on dollar-denominated assets, foreign exchange (FX) noise was eliminated when searching for cointegration. Each of the equities chosen are among the largest underlying assets of both ETFs, acting as a vector for the strategy to exploit the ETF Arbitrage Mechanism as an external economic agent.

The strategy was backtested using daily market close pricing data for the assets, which we programmatically sourced from Yahoo Finance using a Python script. The dataset covered the period 2 January 2014 to 31 December 2023, inclusive.

[1] SPDR S&P 500 ETF Trust. An ETF that passively tracks the S&P 500 index

[2] Invesco QQQ Trust Series 1. An ETF that passively tracks the Nasdaq-100 index

[3] Apple Inc

[4] Microsoft Corp

2. Literature Review

2.1. Theoretical Framework

Fama (1965b) described three different types of the Efficient Market Hypothesis (EMH): the weak form stated that all past information is reflected in an asset's market price; the semi-strong form stated that all publicly available information is reflected in the current market price; and the strong form went further by stating that all information about an asset that is known by *any* economic agent is reflected in its market price. Fama later expanded this with three conditions that facilitate EMH: perfect information, absence of transaction costs, and aggregate price agreement (Fama, 1970).

The premise of our strategy aligns with both the weak and semi-strong forms of EMH, but *not* the strong form. Should there be no opportunity to successfully arbitrage the ETF market using our strategy, we would effectively provide supportive evidence of the strong form of EMH, particularly because it would create favourable information asymmetry within the market. Although realistic transactions incur costs such as exchange commissions and fees related to holding short positions, the model assumes that these are negligible such that this condition for EMH is not violated.

Arbitrage pricing theory (Ross, 1976) stated that any deviation from an identified equilibrium that lacked a clear fundamental cause, such as an earnings call, is a deviation from the fair value of some (or all) of the assets within the group. This concept supports statistical arbitrage mean reversion principles as they identify and exploit deviations in the spread.

Both EMH and arbitrage pricing theory are imperative for creating a foundation for the economic hypothesis of this dissertation outlined in Section 3.1.

2.2. Empirical Studies Evaluating the Efficacy of Basic Pairs Trading and More Sophisticated Statistical Arbitrage Strategies

A baseline distance approach for pairs trading (Gatev, Goetzmann and Rouwenhorst, 2006) yielded an average annualised excess return of 11% in a backtest between 1962 – 2002, utilising a 12-month formation period and a six-month trading period. The strategy showed that, even with simplistic methodology when compared to many modern quantitative strategies, there have been historical opportunities to generate alpha with profits uncorrelated to S&P 500 returns. However, these opportunities have dwindled in recent years as increased hedge fund activity has improved the overall efficiency of financial markets.

Expanding upon this, Vidyamurthy (2004) popularised using cointegration for various strategies, including pairs and n-ary group strategies, and employed both parametric and non-parametric methods which have since been adapted for a variety of approaches. For example, a multivariate approach was used as an advanced indexing strategy, with the returns of underlying assets of the EuroStoxx 50 (Dunis and Ho, 2005) compared to the returns of the index itself.

Finally, a more sophisticated active statistical arbitrage strategy outperformed traditional “buy-and-hold” strategies of ETFs whilst simultaneously generating a positive return between September 2008 and March 2009 (Galenka, Popova and Popova, 2012), the worst period for financial markets, during the Global Financial Crisis (Dwyer, 2009).

These papers all illustrate how various cointegration-based statistical arbitrage strategies can outperform the market, generating positive returns even in times of crisis. Their key differentiating factor is that they all generate returns using a unique contrarian approach with cointegration as a baseline, hence a strategy deployed to one investable universe may not be applicable for another. Such techniques, however, are all at risk of over-fitting and may not continue to be profitable when forward tested out-of-sample.

2.3. Empirical Studies Analysing ETF Mispricing and Arbitrage Opportunities

In a study that assessed intraday arbitrage opportunities between 423 passively managed U.S. equity ETFs and their underlying, Box et al. (2021) found insufficient evidence to suggest that ETF arbitrage opportunities preceded trading within the underlying. This means that other economic agents did not react to potential arbitrage opportunities due to insufficient capital, transaction costs, or insufficient returns given the risk of opening a position. It could be argued that insufficient liquidity was a reason for this, however, this was unlikely.

Contrastingly, empirical evidence was found to support the incidence of liquidity mismatch when the underlying was relatively illiquid to the ETF (Pan and Zeng, 2017); the illiquidity ultimately resulted in increased volatility of the ETF and a reduction in market efficiency. An explanation for this is that the study assessed corporate bond ETFs, where the underlying assets are traded over-the-counter, which makes the fragility easier to observe. Pan and Zeng attributed the conflict to APs being both bond dealers and arbitrageurs, which caused a larger relative mispricing when the conflict worsened.

Despite these papers disagreeing with one another, their results can likely be attributed to them analysing different underlying asset classes (bonds and equities), however, it still shows that arbitrage opportunities may exist within an investable universe of ETFs.

2.4. Improvements on Previous Work & Current Gaps in Existing Literature

This dissertation is an expansion and improvement upon the KAL-EGY strategy (Hammond, Munday and Cammish, 2024), which was submitted to the Southeastern Hedge Fund Competition run in association with Georgia State University, applied to a different investable universe. Compared to KAL-EGY, the strategy described in this dissertation generates more accurate hedge ratios, which are imperative within statistical arbitrage to maintain a market-neutral strategy. Additionally, we applied the Kalman Filter to the hedge ratio, rather than the logarithmic price ratio between the two assets. Furthermore, the hedge calculated was more accurately through the repeated use of the Johansen maximum eigenvalue test at multiple stages in the lead-up to executing trades, rather than a singular use of the Engel and Granger test.

Research papers in the public domain predominantly test whether there is a risk of a mispricing, or for the incidence of external ETF arbitrage, rather than attempting to exploit market inefficiencies to generate a profit. Successful research in this area is unlikely to be published as quantitative researchers would prefer to seek institutional investment and therefore use their work for personal gain. Any quality research in this field would immediately be digested by the market, causing them to adapt and improve efficiency; this in turn removes the opportunity to generate a profit using such research. Moreover, this dissertation contributes to existing research through its unique research methodologies that are applied to a select investable universe.

3. Data and Methods

3.1. Economic Hypothesis

Due to data availability restrictions, the strategy operates using asset daily closing price data, which differs to APs who operate on a high frequency basis and look for mean-reverting price deviations during periods when there are no fundamental changes to the assets being traded. Therefore, we hypothesise: (1) Some individual trades may prove to be profitable, satisfying arbitrage pricing theory (Ross, 1976), although these will be unable to generate alpha in the long-run; and (2) successful opportunities to arbitrage ETFs against their underlying are limited, supporting the strong form of EMH (Fama, 1965b).

3.2. Descriptive Statistics of Data

We programmatically retrieved daily price data for each of the assets at market close from Yahoo Finance, using a Python script. We then adjusted this data for dividends and stock splits so backtesting would remain consistent and unbiased. A summary of this dataset is shown in Table 1 below.

	Equities		**ETFs**		**Investable Universe**
	AAPL	MSFT	SPY	QQQ	
Min	16.96168	31.78906	151.489	77.75071	16.96168
Max	194.8406	377.9111	460.8	410.7568	460.8
Mean	80.57476	146.7882	307.3299	203.6536	184.58662
Median	56.58998	88.92873	296.7595	172.9602	130.94447
Std. Dev.	54.54845	106.6166	94.88674	105.894	108.82334
No. Obs.	2515	2515	2515	2515	10060

Table 1: Descriptive statistics of the assets within the investable universe.

3.3. Using Cointegration to Find Viable Candidates and Calculating the Rate of Decay

Within the investable universe of two equities and two ETFs, there existed four potential paired combinations, as the equities and ETFs cannot be paired against one another. Each of these combinations form what are referred to as a *candidate*. These were then split and grouped by quarter to create possible training data for the model. Potential candidates are shown in Figure 2.

Figure 2: Possible ticker pair combinations.

For each candidate, we then derive the natural logarithm of the price series for the two assets to form the following linear combination:

$$Y = \beta_1 X_1 + \beta_2 X_2$$

where the values $\beta_1 \ldots \beta_2 \in \mathbb{R}$ are unrestricting cointegrating parameters, hence $Y \in I(0)$ implying that the linear combination is stationary.

Using the Johansen maximum eigenvalue test (Johansen, 1988), the hypothetical linear combinations were then placed into vector error form and the cointegrating matrix was tested for its maximum eigenvalue. If the eigenvalue were greater than zero at 95% confidence for the quarter of trading being tested, we deemed the group to be cointegrated. This test expands upon the Engle and Granger (1987) cointegration test, which checks whether two time-series variables satisfy the following equation:

$$\alpha + \varepsilon = X_1 - \beta_2 X_2$$

Where, given that this is a pairs trading strategy, both tests are logically identical, although the equation from Engle and Granger easily illustrates our goal from each input.

We chose the Johansen test for two reasons: (1) the Engle and Granger test cannot be scaled beyond two variables; and (2) it requires a secondary test to compute the hedge ratio, reducing its computational efficiency. Reason (1) is not relevant at present; however, it makes the code more scalable for future improvements.

The hedge ratio, or the *spread*, was then calculable as follows:

$$\Delta = \frac{eigenvalue[1][0]}{eigenvalue[0][0]}$$

where Δ is a negative scalar value and was used to set parameters $\beta_1 \ldots \beta_2$ such that for every X_1 shares traded, $\beta_2 X_2$ should be traded in the opposite direction similar to the Engle and Granger cointegration test. For example, should the spread be computed to be -2; $\beta_1 = 1$ and $\beta_2 = -2$, the following linear combination would emerge:

$$Y = X_1 - 2X_2$$

Furthermore, to mitigate the risk of finding spurious cointegrating relationships, we use the weightings $\beta_1 \ldots \beta_2$ calculated by Δ to check whether the identified relationship existed during the previous quarter. We use the Augmented Dickey-Fuller (ADF) Test (Dickey and Fuller, 1979) for stationarity and test Y using the following equation under $H_0: \gamma = 0$ and $H_1: \gamma < 0$ at 95% confidence:

$$\Delta Y_t = \alpha + \omega t + \gamma Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-i} + \varepsilon_t$$

If we reject H_0, the value of γ is used to calculate the half-life of Y to ensure that mean reversion will occur within a feasible time period, reducing the probability that fundamental changes will occur. Choosing a time period is an arbitrary choice which differs amongst statistical arbitrage strategies; for the purpose of this dissertation, we chose 15.75 days, the average length of 25% of U.S. trading days in a quarter (63).

We then calculate the half-life (Baviera and Santagostino Baldi, 2019):

$$\lambda = \frac{-\log(2)}{\gamma}$$

using the value of γ calculated from the ADF test. λ is recalculated using the same process every 30 days to reevaluate whether we wish to continue to hold a position [Appendix A].

3.4. Calculating Stochastic Forecasts of the Spread

Now that a suitable candidate had been found, we proceeded to generate training data to forecast future values of the spread. We ran an iterative loopback algorithm to calculate the spread from the previous 20 trading days everyday between day 20 and the final day of the cointegrated quarter for our candidate. Given an average of 63 trading days in a quarter, this equates to approximately 43 calculated values of the spread, providing a suitably large sample size for an accurate stochastic forecasting engine.

We calculated the candidate's forecasted spread for the next quarter using the Kalman Filter (Kalman, 1960), using the training data as its input. The aim of this was to expand upon a strategy used by Renaissance Technologies' Medallion Fund which utilised hidden Markov models (Lipton, 2021) to compute stochastic forecasts in discrete time. The Kalman Filter expands upon hidden Markov models by generating stochastic forecasts in continuous time. Renaissance's strategy was originally used to assign bull and bearish states to singular equities, whereas this strategy utilises a novel approach – applying the forecasting engine to the spread.

Within this strategy, the system was assumed to be stable, such that its transition matrix, A, evolves as a random walk, acting as a suitable model for stock prices as they follow this same pattern in the short and medium-run (Fama, 1965a). It is also consistent with satisfying arbitrage pricing theory, as we assume in our model that there are no fundamental changes to our assets, hence significant price deviations signal a mean-reverting arbitrage opportunity. Furthermore, we assume that system stability means that the covariance matrices are also constant, hence do not need updating, and that no other control factors are required.

Each timestep (days) is treated as a continuous process with the current state of the system estimated using the transition matrix and its input taken as a new measurement. The system proceeded to take the conditional probability of the values given the system state to update its current state. The algorithm loops iteratively for each timestep, calculating forecasted values of the spread for the entirety of the next quarter.

3.5. Trading Signal Generation and Portfolio Allocation

Using a combination of the training data and the forecasted data, z-scores are calculated for the candidate's dataset. A z-score is how many standard deviations above or below the calculated or forecasted spread is from its mean and is calculated as follows:

$$z = \frac{x - E(Y)}{\sigma(Y)}$$

We set the z-score tolerance to one standard deviation from the mean. When $z \leq -1$ we long the spread with weightings β_1, β_2 and, conversely, when $z \geq 1$, we short the spread with the weighting $-(\beta_1, \beta_2)$. The model is also prevented from generating multiple sequential buy/sell signals such that a position is only opened once and does not continue to compound.

Positions were closed if any of the criteria in Table 2 were met.

TAKE PROFIT	**STOP LOSS**	**UNCLOSED POSITION**
The z-score crosses zero and mean reversion has taken place	The P&L reaches -20% of its position	The trade had not been closed by the final day of the forecasted quarter

Table 2: Criteria for closing positions.

3.6. Updating the Spread

As the forecasted spread poses potential inaccuracies and could have overly drifted from its true value, we used these calculations merely to forecast a date in the future to open a position. Instead, the true hedge ratio was calculated on that date using the previous 20 trading days of observed pricing data leading up to the execution date, giving a more representative value [Appendix B] [Appendix C]. This is a novel concept and, although it adds some inaccuracy to the model as the value is calculated at the same time of execution, given the size of our positions and the number of trades that were executed the risk can be deemed to be minimal. This method is better than alternative existing methods because it aims to mitigate the risk of errors arising from the stochastic forecasting process breaking the market neutrality of this strategy.

This method to open a position assumed that price variation leading up to the end of the trading day is often small and the trade could still execute during electronic trading hours (provided sufficient liquidity). Furthermore, the strategy assumed that the sum of the variation of all long and short trades satisfies the Law of Large Numbers with zero expectation:

$$\sum_{i=1}^{n} \beta_{1i}\varepsilon_{1i} + \beta_{2i}\varepsilon_{2i} \xrightarrow{p} E[\beta_1\varepsilon_1 + \beta_2\varepsilon_2]$$

where

$$E[\beta_1\varepsilon_1 + \beta_2\varepsilon_2] = 0$$

and ε denotes the residual error arising from executing trades in this manner.

3.7. Measuring Strategy Performance

Once we had found suitable candidates, we ran a portfolio backtest and the strategy's performance was measured using the following five metrics: (1) nominal returns; (2) Compound Annual Growth Rate (CAGR); (3) Sharpe ratio (Sharpe, 1994); (4) daily drawdown; and (5) maximum daily drawdown.

These metrics were calculated using the following formulae:

$$Nominal\ Returns = Terminal\ Portfolio\ Value - Initial\ Portfolio\ Value$$

$$CAGR = \left(\left(\frac{Terminal\ Portfolio\ Value}{Initial\ Portfolio\ Value}\right)^{\left(\frac{1}{length\ of\ backtesting, years}\right)} - 1\right) * 100$$

$$Sharpe\ ratio = \frac{R_{Portfolio} - R_{Risk-free\ asset}}{\sigma_{Portfolio}}$$ [5]

$$Daily\ Drawdown = \left(\left(\frac{Portfolio\ Value}{\max Portfolio\ Value_{252\ day\ rolling\ window}}\right) - 1\right) * 100$$

$$Maximum\ Daily\ Drawdown = \min Daily\ Drawdown_{252\ day\ rolling\ window}$$

Here, daily drawdown and maximum daily drawdown calculate how much the Net Asset Value (NAV) declines each day, relative to its absolute worst performance, rolling on a 252-day window.[6]

[5] Calculated using an average 10-year U.S. T-Bill yield of 2.31% and a portfolio standard deviation of 1.62%.
[6] On average, there are 252 trading days in the U.S. each year.

4. Analysis & Interpretation of Results

4.1. Portfolio Backtesting

4.1.1. Portfolio Construction

We gave the portfolio $100,000,000 of initial capital investment and backtested the strategy between 2 January 2014 and 31 December 2023, providing it with the capacity to generate signals over a ten-year period. The portfolio allocated 30% of available cash to any position to open, hence opening multiple positions would result in diminishing returns for those opened later. All positions were assumed to be unleveraged, and only available capital was risked with each trade, hence the cash generated from opening short positions remained in the brokerage account ready to cover a margin call should one have occurred.

Statistical arbitrage strategies benefit most when the market experiences increased volatility, hence we expected the strategy to be most active during 2018 and 2020. The S&P 500 declined by ~10% in 2018, fuelled by market uncertainty; tariffs imposed by the Republican administration; and rate hikes by the U.S. Federal Reserve (Frazee, 2018), which would have created an opportunity for the strategy to generate a return. Furthermore, the Covid-19 pandemic caused the S&P 500 to fall by an extraordinary 12% on 16 March 2020 (Pisani, 2021), resulting in the invocation of a circuit-breaker by the stock-exchanges themselves to try to contain the market volatility; meanwhile, analysis of 8K and DEF14A filings found that technology firms earned high positive returns (Mazur, Dang and Vega, 2020). The contrarian returns between the S&P 500 and the equities in our investable universe therefore posed a prime opportunity for the strategy to generate a significant return.

4.1.2. Risk-Adjusted Performance

The portfolio had a terminal value of $96,067,990.04, resulting in a nominal return of -3.93% and a CAGR of -0.40%, resulting in a Sharpe ratio of -13.16. The unprofitable nature of the strategy is indicative of the strong form of EMH holding (Fama, 1965b) as it was unsuccessful in detecting pricing efficiencies spanning multiple days. Contrarily, had it proved to be profitable, a wider risk of the strategy could be attributed to the spread being skewed towards a higher notional value of one asset, hence its performance could have simply been credited to the rapid growth of the wider market.

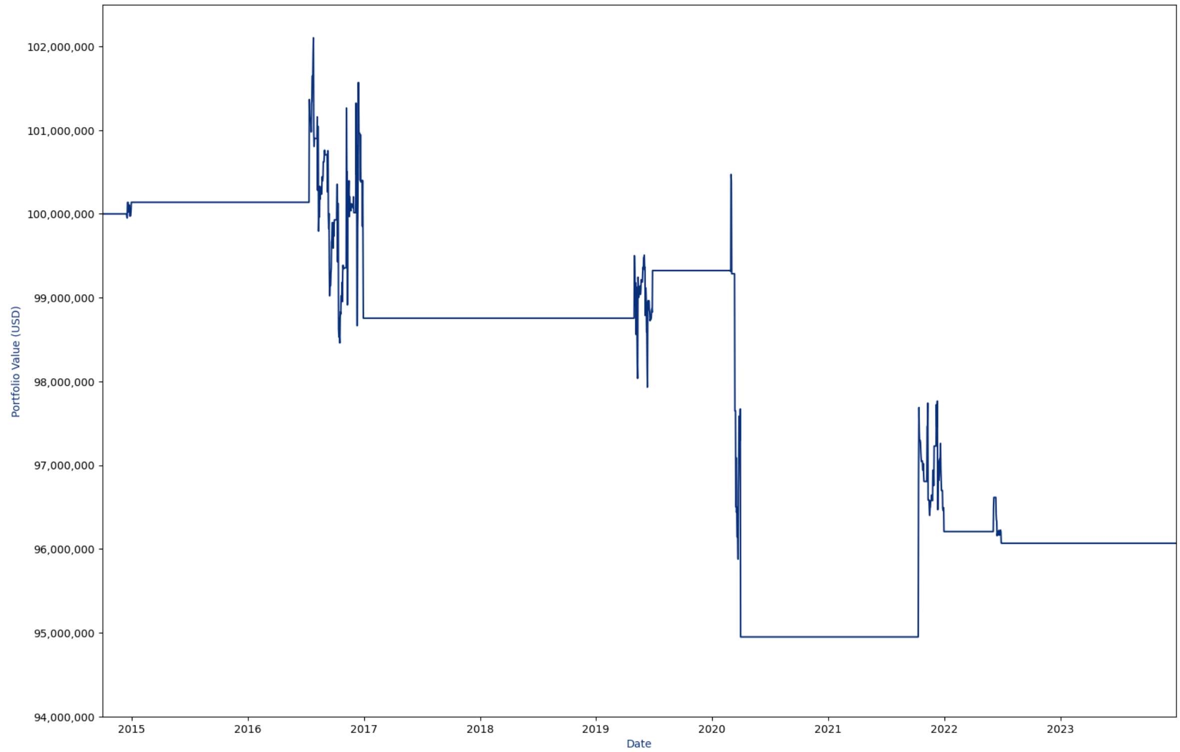

Figure 3: Portfolio NAV during the backtesting period of 2 January 2024 – 31 December 2023.

30 trading signals were generated throughout the backtesting period, 13 of which were to long the spread, 2 to short the spread, and 15 to close out positions. 40% of the trades identified were Type-II Errors (false-positives), with the largest loss incurred from a trade between *AAPL* and *SPY* between 28 February 2020 and 31 March 2020. This trade risked $22,014,891.27, and lost $1,867,189.05 (-8.48%). Here, *AAPL* fell by 6.98% and *QQQ* fell by 12.49%. This result is in line with arbitrage pricing theory as, during this time, the Covid-19 pandemic caused the stock market to enter freefall and was yet to recover, hence a fundamental change within these assets had occurred that was unobservable by this model.

Moreover, this result is not contrarian to Mazur, Dang and Vega's research (2020) as technology firms did not have the opportunity to recover in this time period. The most profitable trade stemmed from a pairs trade between *MSFT* and *QQQ* between 13 December 2016 and 14 December 2016, during which *MSFT* fell by 0.48% and *QQQ* fell by 0.21%. As Δ was heavily skewed to form a linear combination, shorting *MSFT* by a greater than proportional amount than *QQQ*, the trade still yielded a positive return via mean-reversion.

Amongst the remaining trades, seven yielded a profit or loss no greater than $100,000 which, assuming an average of $30,000,000 risked per trade, implies a $\pm 0.33\%$ return; this contributed negligibly to the portfolio's overall performance.

Despite expectations, a disproportionate number of trades were executed between 10 July 2016 and 30 December 2016 relative to the rest of the backtesting period. During this time, the algorithm executed six trading opportunities between *MSFT* and *QQQ*, and one between *AAPL* and *QQQ*. This activity shown in Figure 4 shows how, whilst these were occasionally profitable, the strategy was unable to time the market accurately and missed opportunities to accurately capture mean reversion.

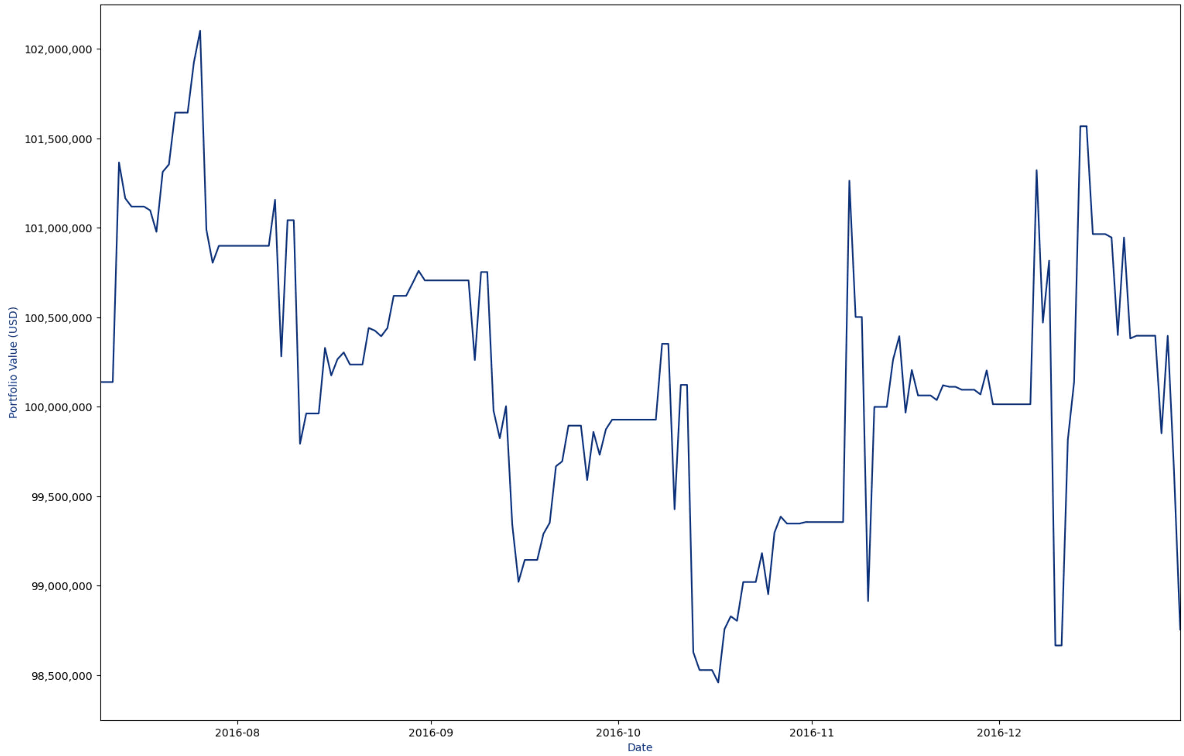

Figure 4: Snapshot of backtest between the period 10 July 2016 – 30 December 2016.

One separate performance metric of the strategy was how accurately it chose the correct asset to long and short. Whilst many trades resulted in both assets moving in the same direction (as expected with ETF arbitrage strategies), thus the hedge ratio and decision to long/short the spread based upon its z-score was more important, incorrect signals that resulted in each asset moving in the opposite direction could be more detrimental. Of the 15 signals, three involved trades where the equity and ETF moved in opposing directions, of which two were predicted correctly. Of these, however, all of them were still profitable trades, implying that the hedge ratio had mitigated this incorrect prediction. This cannot be trusted with high certainty given such a low sample size, although it does provide some insight that the model was successful in identifying the correct position and maintaining a market-neutral strategy.

The final metric that was used to measure the portfolio's performance was its drawdown, shown in Figure 5. The portfolio's worst return with respect to its daily drawdown and maximum drawdown occurs during the pairs trade between *AAPL* and *SPY* during the Covid-19 outbreak, where the portfolio fell by 5.50%. Other than this, the daily drawdown peaks at -3.40%, indicating that the portfolio experienced overall stability throughout the backtesting period and was never at risk of sudden implosion, even during black swan events.

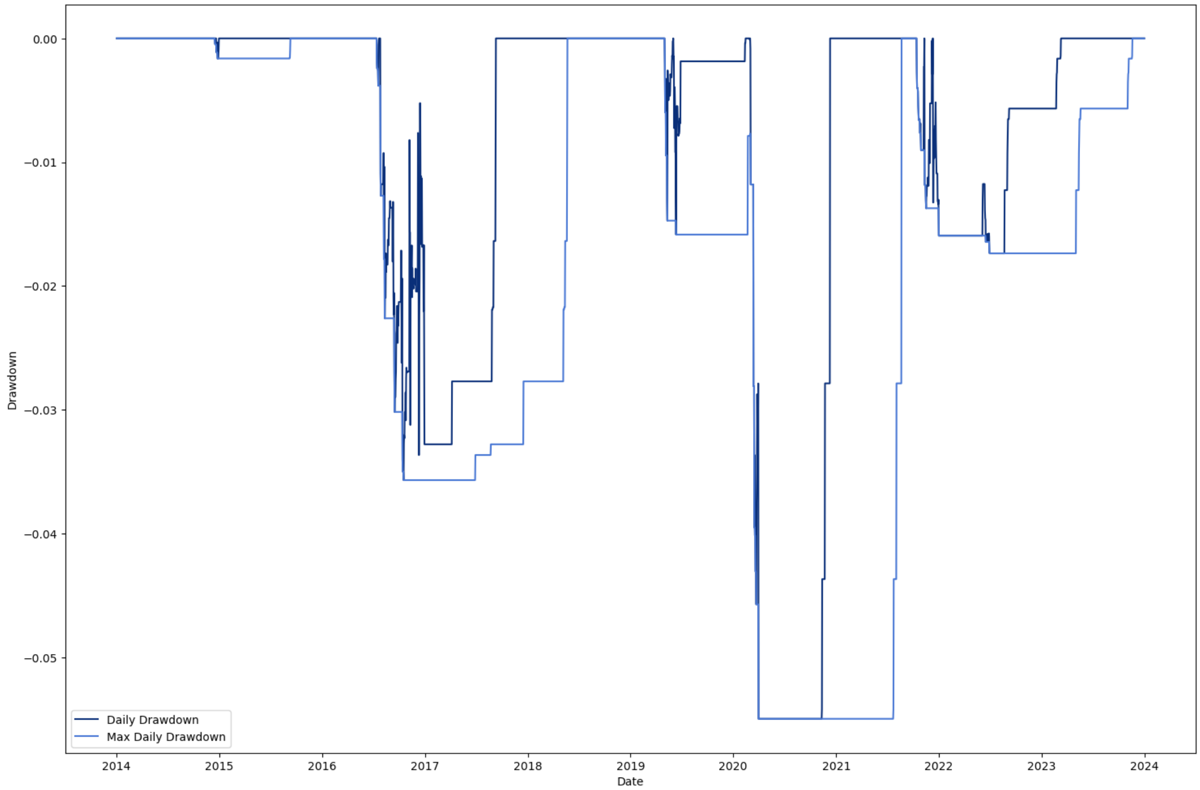

Figure 5: Daily drawdown and maximum daily drawdown of the portfolio.

4.2. Risks

4.2.1. Idiosyncratic Risks

Fundamental shocks to the assets pose a threat to the strategy if they impact the price ratios between cointegrated pairs. These may arise for several reasons, including (but not limited to): (1) *AAPL* or *MSFT* engaging in M&A activity, or (2) adjusted forward-outlook on company performance. Such shocks would instantly invalidate arbitrage pricing theory (Ross, 1976) and any positive outcome from them would be spurious rather than the product of successful arbitrage. Naturally, this exposes the strategy to increased downside risk and the model does not have a specific termination clause that executes if a cointegrating relationship breaks down, however, we attempt to mitigate this with the hard stop loss mechanism.

Furthermore, opening a short position has limited upside potential and unlimited downside potential, a risk that may be compounded by a decomposition of the relationship between our cointegrated pairs. This risk is mitigated by holding the cash allocated to short trades in the brokerage account such that it can always be covered in the event of a margin call.

Finally, the investable universe itself is small and restricted which constrains the model by how many potential candidates that can be identified. This constraint is amplified by the fact that the strategy assumes that the chosen underlying assets of the ETFs will continue to be dominant into the future. As observed with Apple losing its position as the largest company globally (by market capitalisation) at the start of this year due to slowing growth relative to its competitors (Soni, 2024), with Microsoft continuing to rally amidst the growth in artificial intelligence, the assumption regarding the underlying assets in our investable universe may be weak. A mitigation for this would be to review the relative performance of the equities within the investable universe on an annual basis, replacing them with their counterparts where appropriate.

4.2.2. Model Inaccuracies and Lower than Expected Returns

Profit-generating trades (Type-I errors) may arise as the model misses the opportunity to identify and execute a trade. Moreover, other inefficiencies and errors may result in it executing unwanted trades (Type-II errors). The hard stop loss was used to minimise Type-II errors, although given the low volatility of this strategy, this was never invoked and could have potentially been restricted further.

By the nature of identifying market inefficiencies, executing trades uses market forces to push the assets back towards their true market value, which may cause slippage. This is because we assume that the model can execute every trade at the quoted market price, when in reality trades may have to “walk-the-book”. Moreover, the model does not account for any brokerage fees or commissions, although these should be minimal as the strategy only executed 60 trades (30 signals of pairs trades).

4.2.3. Liquidity Shortages

The strategy may experience a strain on liquidity, particularly if other statistical arbitrage funds are operating within the same investable universe. This is likely a certainty as the equities have been within the top five of the S&P 500 for over a decade. The strategy risks being crowded out if another fund deploys a larger amount of capital to its positions, particularly if they are leveraged, and may result in the trades executing at an unfavourable rate. Decreasing liquidity paired with increasing leverage has previously been found to be a factor that has caused many statistical arbitrage funds to be forced into liquidating their portfolios in August 2007 (Khandani and Lo, 2011), making this an ever-present risk.

Furthermore, crowding out becomes a greater issue because although the exchange operates using a first in/first out (FIFO) algorithm on the basis of Price/Time priority (CME Group, n.d.), it gives order execution priority to orders of larger volume. Therefore, a fund with more deployable capital could gain a preferrable entry price, even if our strategy reaches the exchange earlier.

Fortunately, the investable universe involves two of the most liquid equities. As of 13 May 2024, the 30-day average daily volume was 62.62m shares for Apple (YCharts, 2024a), and 19.16m shares for Microsoft (YCharts, 2024b). APs and market makers also have a duty to maintain ETF liquidity within the market, hence trading these on the exchange should not pose a significant issue.

4.3. Improvements Towards Future Works

The first main improvement for this strategy would be for it to utilise high-frequency data, as we only had access to daily closing prices for this dissertation. This would give the strategy the opportunity to search for market inefficiencies at a micro level, placing it one step closer to being able to compete with High-Frequency Trading (HFT) firms and APs more effectively.

The strategy assumed that only market orders would be used. This could be improved by using a group of All-Or-None (AON) orders around the bid-ask spread ready for market open. This poses two key benefits. Firstly, each trade would be entered almost an entire trading day earlier and, secondly, as the limit orders are around the bid-ask spread, the exchange may offer a rebate for providing market liquidity. This mitigates the risk of exchange fees subsuming some potential profit, whilst also providing a profit cushion.

Finally, a larger investable universe would improve this strategy greatly, particularly as it would mean that not only pairs trades could be executed. For example, if this methodology was to be expanded to have 20 equities that formed their respective ETFs and/or other ETFs were included, the probability that a cointegrated linear combination could be found would vastly increase, particularly if it were to be expanded for n-ary group statistical arbitrage. The strategy would therefore not be constrained to one discrete length of linear combination.

5. Conclusion

Ultimately, this strategy generated a -3.93% return during its ten-year backtesting period. This provides evidence that APs effectively maintain market efficiency with the ETF Arbitrage Mechanism but also more broadly that, over multiple days, the strong form of EMH holds. However, if this strategy had access to more intraday pricing data and had its threshold to open trades reduced, this strategy may be more successful. The unsuccessful result of this strategy may further be attributed to errors in the stochastic forecasts [Appendix C] where the z-score output seemed to act as a random walk in one direction, hence skewing the validity of signal generation.

The use of high-frequency data would further prove to be especially important because it would illustrate whether this strategy struggled to generate accurate predictions from the use of the Kalman Filter. Should this be repeated with daily pricing data again, reverting to stochastic forecasting in discrete time (using hidden Markov Models) may prove to be beneficial.

Furthermore, other ETF arbitrage strategies utilising a different methodology may prove to be more successful. For example, in a seminal paper that applied deep learning techniques to statistical arbitrage, a Convolutional Neural Network (CNN) model outperformed other parametric models by four times, and other alternative deep learning models by two times with an average annualised return of 20% (Guijarro-Ordonez, Pelger and Zanotti, 2021). Applying such a concept to create a more sophisticated forecasting engine may yield improved returns in a more complicated strategy, whilst also reducing the number of Type-II errors that it causes.

We look forward to seeing how market-neutral strategies continue to evolve into the future and how such strategies can continue to maintain an efficient market.

Bibliography

Baviera, R. and Santagostino Baldi, T. (2019). Stop-loss and leverage in optimal statistical arbitrage with an application to energy market. *Energy economics*, 79, pp.130–143. doi:https://doi.org/10.1016/j.eneco.2018.03.024.

BlackRock. (n.d.). *Authorized participants and market makers*. [online] Available at: https://www.blackrock.com/au/intermediaries/ishares/authorized-participants-and-market-makers [Accessed 28 Apr. 2024].

Box, T., Davis, R., Evans, R. and Lynch, A. (2021). Intraday arbitrage between ETFs and their underlying portfolios. *Journal of Financial Economics*, 141(3). doi:https://doi.org/10.1016/j.jfineco.2021.04.023.

CME Group. (n.d.). *Supported Matching Algorithms - Electronic Platform Information Console - Confluence*. [online] Available at: https://www.cmegroup.com/confluence/display/EPICSANDBOX/Supported+Matching+Algorithms [Accessed 14 May 2024].

Dickey, D.A. and Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74(366), pp.427–431. doi:https://doi.org/10.2307/2286348.

Dunis, C.L. and Ho, R. (2005). Cointegration portfolios of European equities for index tracking and market neutral strategies. *Journal of Asset Management*, 6(1), pp.33–52. doi:https://doi.org/10.1057/palgrave.jam.2240164.

Dwyer, G. (2009). *Federal Reserve Bank of Atlanta*. [online] www.atlantafed.org. Available at: https://www.atlantafed.org/cenfis/publications/notesfromthevault/0909 [Accessed 24 May 2024].

Fama, E.F. (1965a). Random Walks in Stock Market Prices. *Financial Analysts Journal*, [online] 21(5), pp.55–59. Available at: https://www.jstor.org/stable/4469865 [Accessed 8 May 2024].

Fama, E.F. (1965b). The Behavior of Stock-Market Prices. *The Journal of Business*, 38(1), pp.34–105. doi:https://doi.org/10.1086/294743.

Fama, E.F. (1970). Efficient Capital Markets: a Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), pp.383–417. doi:https://doi.org/10.2307/2325486.

Frazee, G. (2018). *6 factors that fueled the stock market dive in 2018*. [online] PBS NewsHour. Available at: https://www.pbs.org/newshour/economy/making-sense/6-factors-that-fueled-the-stock-market-dive-in-2018 [Accessed 14 May 2024].

Galenko, A., Popova, E. and Popova, I. (2012). Trading in the Presence of Cointegration. *The Journal of Alternative Investments*, 15(1), pp.85–97. doi:https://doi.org/10.3905/jai.2012.15.1.085.

Gatev, E., Goetzmann, W.N. and Rouwenhorst, K.G. (2006). Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *Review of Financial Studies*, [online] 19(3), pp.797–827. doi:https://doi.org/10.1093/rfs/hhj020.

Guijarro-Ordonez, J., Pelger, M. and Zanotti, G. (2021). Deep Learning Statistical Arbitrage. *SSRN Electronic Journal*. [online] doi:https://doi.org/10.2139/ssrn.3862004.

Hammond, N., Munday, O. and Cammish, A. (2024). *From Black to Green Gold: Arbitraging to Net Zero with the Kalman Filter*. [Southeastern Hedge Fund Competition] Available at: https://github.com/NickHammond0/Portfolio-of-Work/blob/6c4759fc2a33ff4349e436517feff9be0d8c53df/Southeastern%20Hedge%20Fund%20Competition/KAL-EGY.pdf [Accessed 23 Apr. 2024].

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3), pp.231–254.

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, [online] 82(1), p.35. doi:https://doi.org/10.1115/1.3662552.

Khandani, A.E. and Lo, A.W. (2011). What happened to the quants in August 2007? Evidence from factors and transactions data. *Journal of Financial Markets*, 14(1), pp.1–46. doi:https://doi.org/10.1016/j.finmar.2010.07.005.

Lipton, A. (2021). The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution. *Quantitative Finance*, 21(7), pp.1077–1081. doi:https://doi.org/10.1080/14697688.2021.1901972.

Mazur, M., Dang, M. and Vega, M. (2020). COVID-19 and the March 2020 Stock Market Crash. Evidence from S&P1500. *Finance Research Letters*, [online] 38(101690), p.101690. doi:https://doi.org/10.1016%2Fj.frl.2020.101690.

Pan, K. and Zeng, Y. (2017). ETF Arbitrage Under Liquidity Mismatch. *SSRN Electronic Journal*, 2017(59). doi:https://doi.org/10.2139/ssrn.3723406.

Pisani, B. (2021). *One year ago stocks dropped 12% in a single day. What investors have learned since then*. [online] CNBC. Available at: https://www.cnbc.com/2021/03/16/one-year-ago-stocks-dropped-12percent-in-a-single-day-what-investors-have-learned-since-then.html [Accessed 14 May 2024].

Ross, S.A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, [online] 13(3), pp.341–360. doi:https://doi.org/10.1016/0022-0531(76)90046-6.

S&P Dow Jones Indices (2024). *S&P 500® - S&P Dow Jones Indices*. [online] www.spglobal.com. Available at: https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview [Accessed 10 May 2024].

Sharpe, W.F. (1994). *The Sharpe Ratio*. [online] Stanford. Available at: https://web.stanford.edu/~wfsharpe/art/sr/sr.htm [Accessed 24 May 2024].

Soni, A. (2024). *Microsoft briefly overtakes Apple as world's most valuable company*. [online] Reuters. Available at: https://www.reuters.com/technology/microsoft-overtakes-apple-worlds-most-valuable-company-2024-01-11/ [Accessed 14 May 2024].

Vidyamurthy, G. (2004). *Pairs trading : quantitative methods and analysis*. Hoboken, N.J.: J. Wiley.

YCharts. (2024a). *Apple 30-Day Average Daily Volume*. [online] Available at: https://ycharts.com/companies/AAPL/average_volume_30 [Accessed 14 May 2024].

YCharts. (2024b). *Microsoft 30-Day Average Daily Volume*. [online] Available at: https://ycharts.com/companies/MSFT/average_volume_30 [Accessed 14 May 2024].

Appendix A: Identified Cointegrated Pairs

	Ticker1	Ticker2	Quarter	Cointegrated	Johansen Trace Statistic	Johansen 95% Critical Value	Johansen Eigenvectors	Hedge Ratio
0	AAPL	QQQ	2014Q3	True	[15.6011591 1.53617886]	[15.4943 3.8415]	[[82.57891398441237, -3.53140187572486], [-111.96083987114818, 54.14869717167276]]	-1.355804
1	AAPL	QQQ	2015Q2	True	[15.52573346 5.36604347]	[15.4943 3.8415]	[[34.71361721893355, -106.00644412160355], [-105.80476869778188, 76.65928371359827]]	-3.047933
2	AAPL	QQQ	2016Q3	True	[21.1826809 3.4671275]	[15.4943 3.8415]	[[31.566558495544193, -34.65729007236719], [-98.53641852621885, 46.19195248770511]]	-3.121545
3	AAPL	QQQ	2018Q3	True	[16.64494104 0.98095774]	[15.4943 3.8415]	[[15.99004270858679, -21.234011125611946], [-96.10120403234426, 34.016070786713364]]	-6.010066
4	AAPL	QQQ	2019Q1	True	[27.36675471 0.95490324]	[15.4943 3.8415]	[[60.85591654777209, -33.038936406865275], [-126.34304631550813, 39.67371612154958]]	-2.076101
5	AAPL	QQQ	2019Q4	True	[15.97376312 1.89400602]	[15.4943 3.8415]	[[102.60644116846828, -21.95873283763803], [-203.24618963889213, 16.580481229258982]]	-1.980833
6	AAPL	SPY	2015Q2	True	[18.27083001 6.48727947]	[15.4943 3.8415]	[[34.752285834485484, -92.64960846760516], [-176.57357449106723, 93.88722022934496]]	-5.080920
7	AAPL	SPY	2016Q2	True	[15.84698376 4.11072855]	[15.4943 3.8415]	[[4.360738272213597, 16.470842314047843], [86.46794999056914, -34.385590835817695]]	19.828741
8	AAPL	SPY	2018Q2	True	[19.35960387 3.55182262]	[15.4943 3.8415]	[[31.51940028508571, -18.929350677744196], [-105.300636747577, 0.5979167694772552]]	-3.340820
9	AAPL	SPY	2019Q1	True	[16.73317604 1.19115473]	[15.4943 3.8415]	[[26.435365245584183, -33.53013861063847], [-80.96062661576893, 53.05328044698807]]	-3.062588
10	AAPL	SPY	2019Q4	True	[15.76119708 1.99185532]	[15.4943 3.8415]	[[104.47215211172882, -18.985461631899437], [-253.7541755794473, 13.175172883532483]]	-2.428917
11	AAPL	SPY	2021Q3	True	[15.54958764 3.54181035]	[15.4943 3.8415]	[[78.690849057346, -15.215896213966243], [-90.6972751652398, 89.25802284012492]]	-1.152577
12	MSFT	QQQ	2014Q2	True	[16.52264234 0.27949354]	[15.4943 3.8415]	[[107.29772335079502, 2.7480264045691727], [-67.91930154346747, 31.079120214542428]]	-0.632999
13	MSFT	QQQ	2016Q3	True	[17.44777869 5.00017762]	[15.4943 3.8415]	[[24.556297028236393, -92.3778747297191], [4.95893935626856, 144.91508198470427]]	0.201942
14	MSFT	QQQ	2017Q1	True	[16.27179428 3.88811974]	[15.4943 3.8415]	[[143.8484606741834, 11.115407565936167], [-77.31328122347122, 29.810757710665158]]	-0.537463
15	MSFT	QQQ	2018Q3	True	[15.80454887 3.25347528]	[15.4943 3.8415]	[[34.45242447614516, -55.946813162546626], [-110.27890790782409, 71.13532607311062]]	-3.200904
16	MSFT	QQQ	2019Q3	True	[18.00756806 2.45691706]	[15.4943 3.8415]	[[101.7785722340933, -55.57766647438713], [-21.696755863904897, 75.09518975696002]]	-0.213176
17	MSFT	QQQ	2021Q4	True	[26.59609095 7.24178331]	[15.4943 3.8415]	[[124.32119841208174, 6.286183428079636], [-178.70388117285444, 18.761802387259674]]	-1.437437
18	MSFT	QQQ	2022Q1	True	[22.60346303 7.42478863]	[15.4943 3.8415]	[[90.89527278841865, -30.69000474233266], [-56.388629373270724, 40.8496968677888]]	-0.620369
19	MSFT	SPY	2014Q2	True	[16.43238223 0.65090981]	[15.4943 3.8415]	[[103.51478800347158, 8.053150551293696], [-97.82021614959535, 40.62038148327115]]	-0.944988
20	MSFT	SPY	2015Q2	True	[16.81022937 5.47326342]	[15.4943 3.8415]	[[10.434625052422465, -21.870821016365618], [-167.7732241987603, 44.43036234458618]]	-16.078510
21	MSFT	SPY	2016Q2	True	[16.89217879 3.42310781]	[15.4943 3.8415]	[[10.624300924218, 26.01623860462239], [76.84998620940192, -53.167806628796505]]	7.233416
22	MSFT	SPY	2016Q3	True	[19.77679484 6.34141197]	[15.4943 3.8415]	[[5.934912013448213, 45.717239994337625], [77.92037334007094, -137.9600297066011]]	13.129154
23	MSFT	SPY	2017Q3	True	[16.95744323 2.10409183]	[15.4943 3.8415]	[[44.04804153211277, -61.29245595910558], [12.212276021485868, 158.30290877255754]]	0.277249
24	MSFT	SPY	2018Q3	True	[17.62863197 6.69732882]	[15.4943 3.8415]	[[88.1266658078555, -7.580962128963967], [-154.07552741956312, 68.09330303721539]]	-1.748342
25	MSFT	SPY	2019Q1	True	[17.3822049 1.58233433]	[15.4943 3.8415]	[[4.642681661034642, -44.26637594587914], [-34.543509315151795, 50.18787329754457]]	-7.440422
26	MSFT	SPY	2019Q3	True	[19.56222782 1.87792188]	[15.4943 3.8415]	[[107.85198802256924, -39.04910173985598], [-33.63070568232383, 75.12730499728862]]	-0.311823
27	MSFT	SPY	2022Q1	True	[20.57769146 6.41947734]	[15.4943 3.8415]	[[90.27192054713139, -39.208193981725245], [-81.2273755740114, 69.20644188882025]]	-0.899808

Figure 6: List of identified signals and its hedge ratio on the last day of the quarter of training data generation.

Appendix B: Sample Data Output for AAPL x SPY (Training Data from 2014Q3 and Trading Data from 2014Q4)

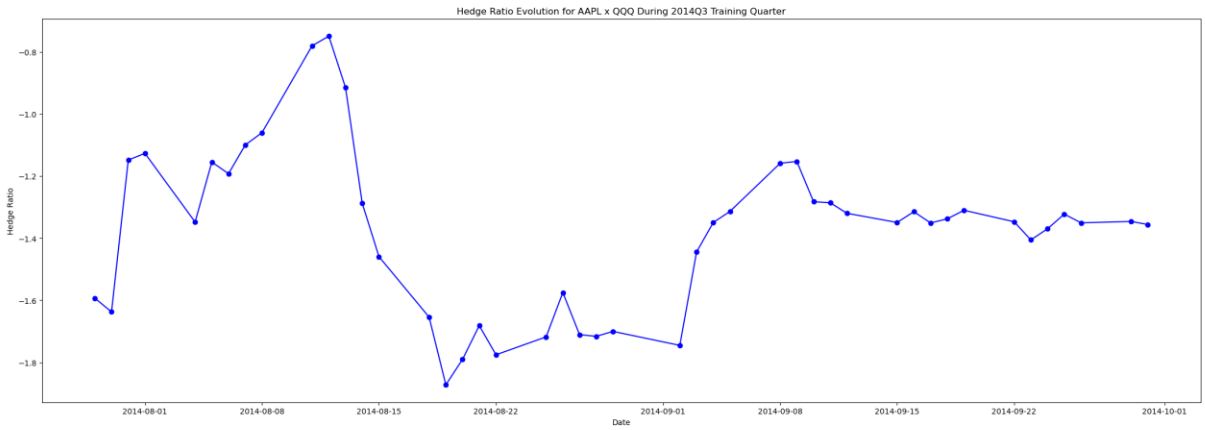

Figure 7: Hedge ratio evolution for AAPL x SPY in 2014Q3.

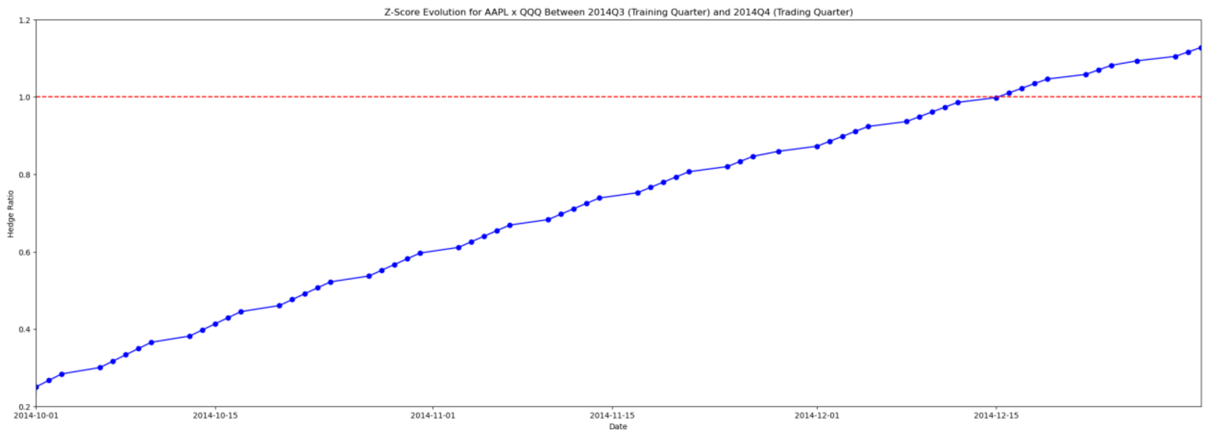

Figure 8: Z-Score evolution for AAPL x SPY in 2014Q4.

Appendix C: Identified Trading Signals:

Date	Ticker1	Ticker2	Quarter	Hedge Ratio	Action
2014-12-16	AAPL	QQQ	2014Q4	-0.675583	Long
2014-12-31	AAPL	QQQ	2014Q4	-0.635019	Close
2016-12-07	AAPL	QQQ	2016Q4	-1.816644	Long
2016-12-13	MSFT	QQQ	2016Q4	3052.946215	Long
2016-12-14	MSFT	QQQ	2016Q4	-3467.081565	Close
2016-12-15	MSFT	QQQ	2016Q4	3937.394810	Long
2016-12-16	MSFT	QQQ	2016Q4	-4471.506539	Close
2016-12-19	MSFT	QQQ	2016Q4	5078.071083	Long
2016-12-20	MSFT	QQQ	2016Q4	-5766.916742	Close
2016-12-21	MSFT	QQQ	2016Q4	6549.205036	Long
2016-12-22	MSFT	QQQ	2016Q4	-7437.611557	Close
2016-12-23	MSFT	QQQ	2016Q4	8446.531352	Long
2016-12-27	MSFT	QQQ	2016Q4	-9592.312174	Close
2016-12-28	MSFT	QQQ	2016Q4	10893.519364	Long
2016-12-29	MSFT	QQQ	2016Q4	-12371.236670	Close
2016-12-30	AAPL	QQQ	2016Q4	-1.586821	Close
2019-04-30	MSFT	SPY	2019Q2	-31.266291	Short
2019-06-06	AAPL	SPY	2019Q2	-2.061259	Long
2019-06-14	AAPL	QQQ	2019Q2	-1.021685	Long
2019-06-28	AAPL	SPY	2019Q2	-1.895195	Close
2019-06-28	AAPL	QQQ	2019Q2	-0.931042	Close
2019-06-28	MSFT	SPY	2019Q2	-72.880810	Close
2020-02-28	AAPL	SPY	2020Q1	-0.868526	Long
2020-02-28	AAPL	QQQ	2020Q1	-0.703703	Long
2020-03-31	AAPL	SPY	2020Q1	-0.631513	Close
2020-03-31	AAPL	QQQ	2020Q1	-0.508751	Close
2021-12-10	AAPL	SPY	2021Q4	-1.231324	Short
2021-12-31	AAPL	SPY	2021Q4	-1.351084	Close
2022-05-06	MSFT	QQQ	2022Q2	-0.262102	Long
2022-06-30	MSFT	QQQ	2022Q2	-0.564479	Close

Figure 9: List of identified trading signals during the backtesting period.