MS&E448 Project Statistical Arbitrage

Carolyn Soo(csjy), Zhengyi Lian(zylian), Hang Yang(hyang63), Jiayu Lou(jiayul)

June 11, 2017

Abstract

This project sought to study and improve the statistical arbitrage strategy put forth by Marco Avellaneda & Jeong-Hyun Lee in their 2008 paper, "Statistical Arbitrage in the U.S. Equities Market"[1]. Like Avellaneda & Lee, we began with the hypothesis that a stock's price (and return) series are driven by market or sector factors represented by Exchange-Traded Fund(ETF)s, and that deviations from some equilibrium level can be modeled as a mean-reverting stationary process known as the spread. Hence, if we can:

- find cointegrated stock-ETF groups whose spread over an x-day trailing window can be estimated and tested statistically for stationarity
- fit a mean-reverting model to the spread process of each stock, to determine their reversion speeds and equilibrium levels
- construct systematic trading signals that quantify how far the spread has deviated from equilibrium
- enter a beta-neutral group of long-short positions in a stock and the relevant ETFs
- close these positions when the spread reverts sufficiently close to the equilibrium

then by carefully timing our exit from these positions, we can profit from the reversion of the spread.

By adapting Avellaneda & Lee's work, we achieved strong returns over the 2003-2008 period, but stagnated thereafter. Subsequently, our research focused on improving performance in the aftermath of the 2008 financial crisis. Ultimately, our project culminated in a strategy which attained the following results in the out-of-sample test set of 2013-2016: Annual return = 2.2%, Sharpe ratio = 0.54, maximum drawdown = -6.6%.

The mediocre risk/reward ratio leads us to suspect that the model adjusts poorly to changing market conditions. Nonetheless, our inclusion of a stationarity test for the spreads, as well as our efforts to deal with incomplete order fills and other execution challenges, enabled us to effectively deal with volatility and drawdown. We also propose several reasons for the post-2012 issues, in addition to recommended avenues for further exploration.

Introduction

In this paper we shall present a statistical arbitrage strategy which executes trades daily at market open, at the market price, based on a trailing window of market data. Our final strategy has a average holding period of 11 days, with an expected Sharpe ratio of 0.54.

Due to limited availability of reliable historical data [1], the strategies discussed in this paper deal exclusively with U.S. equities. Fortuitously, the data has been pre-adjusted for splits and market holidays [5], although

[1] Quantopian has public minute-by-minute data for all U.S. stocks and ETFs traded since 2002; futures data is incomplete and other instruments are not yet covered[2].

care must be taken to avoid bias resulting from forward-filled data [2] or from failures in data processing [3]. Given 14 full years of data (2003-2016), we have elected to reserve the most recent 4 years for out-of-sample testing; our model will be fitted and trained on at most the years 2002-2012.

The investment universe [4] is taken to be the top 500 stocks of Quantopian's contemporary Q1500US; note that performance may vary widely in other universes. Transaction cost is set at 0.05 basis points per transaction. We base signal computations on closing prices of the last trading day, and make trading decisions for each stock only when its signal meets predefined cutoffs. Desired trades (if any) are executed at opening price on the very next trading day [5].

Intuition Behind Statistical Arbitrage

As statistical arbitrage is a familiar concept to many, this discussion on its underlying intuition shall be brief.

Consider a simple universe where there is only 1 stock with price S available for investment. Let this stock be offered by an oil and gas company, such that S is closely related to exactly 1 factor F, oil prices. When oil prices rise due to greater global demand, investors anticipate that shareholder equity will increase, driving up the perceived worth of a single share. Thus stock price S should rise.

Suppose, however, that S does not rise, or rises much less - or more - than predicted based on the oil price increase. This is unsurprising since prices are not perfectly correlated to quantifiable factors[6]; they also contain drivers that are independent of any exogenous factors[7]. So if we remove the effects of common observable factors, and focus only on the idiosyncratic residuals of stock returns, we can identify periods where the share price seems to differ from its mean for no compelling reason.

If we believe that stock prices eventually move towards some average[8], then we expect that over time, the market will adjust to close the gap. Thus the gap represents a profit opportunity: If we long (short) shares of a stock with a higher(lower)-than-expected price, then the narrowing of the gap will let us earn this **spread**[9].

Obviously, real stock prices depend complicatedly on multiple factors, some of which may not be directly observable or quantifiable. Still, the basic idea remains, regardless of how many factors there are: If we can study how or when a stock's price reverts, and what its fair value should be, then we can place our trades to earn the spread.

In a multi-stock universe, we retain the goal of earning the spread. This paper shall consider a "pair" to comprise a stock with its relevant factors, and shall seek to maximize earnings on the spread between each stock's price and the price forecasted by its factors.

To choose what tickers to trade, we first compute an indicator of fast mean reversion time for each ticker to apply as an initial filter, followed by computing a dimensionless "signal" that scales with deviations from the mean; a signal with large magnitude means that the stock price has deviated significantly from its expected value, and (based on the reversion philosophy) should eventually return to its mean.

We use signal cutoffs to determine when to either enter / exit a long position ("buy to open", denoted *bo* / "sell to close" *sc* resp.) or else enter / exit a short position (*so* / *bc* resp.):

[2] Using Quantopian's built-in asof_date capabilities that allow verification of when the values were last updated.

[3] Such failures are deliberately persisted [5] to avoid lookahead bias stemming from retrospective rectifications.

[4] The collection of all stocks that we are permitted to trade in.

[5] Clearly, the forecast horizon is 1 day

[6] If they were, then using empirical data, future stock prices should be perfectly predictable!

[7] For instance, when investors become skittish about a company's stock during a leadership transition, only its share price (and perhaps those of a few close rivals) are noticeably impacted.

[8] One might reasonably believe this average to be literally the stock's historical mean, or else some other pre-defined measure such as growth-adjusted 30-day price projection

[9] The spread is the difference between what the stock price is and what it reverts to.

$$s_{\bar{b}}o = s_{\bar{s}}o = 1.25, s_{\bar{b}}c = 0.75, s_{\bar{s}}c = 0.50$$

We enter a position if and only if either $s_{\bar{s}}o$ or $s_{\bar{b}}o$ is met, and exit a currently-held stock if and only if its signal meets either $s_{\bar{s}}c$ or $s_{\bar{b}}c$. Also, we do not "reuse" signals; that is, once we have entered (exited) a position based on a signal, we do not trade in that stock again until we get the exit (entrance) signal, even if the original signal persists.[10].

Modeling

Studying the Spread Process

In line with our belief that prices (and log returns) will eventually revert to their means, we can profit if we correctly identify periods where an empirical price deviates from its forecast, then take the appropriate position in the stock such that we will profit when the spread goes to zero. Note that for each stock, the forecast is made *not* for the purpose of profiting from its price in and of itself, but rather to compare the actual price to "what it should be" and thusly profit off the gap.

But what should that forecasted stock price be? To answer this question, on every day T we run a regression for each stock i using a trailing 60-day data window:

$$R_{t,i} = \alpha dt + \sum_{j=1}^{m} \beta_{i,j} \cdot F_{t,j} + dX_t \qquad \forall t \in [T-59, T]$$

where $R_t \triangleq$ the simple day-t return of the stock and $F_{t,j} \triangleq$ the simple day-t return of the jth ETF, computed with respect to the day-$(t-1)$ prices. The β_j's are the coefficients for each ETF, representing the relative strength and direction of the relationship between the stock price and that ETF [11].

This regression decomposes a stock return into a drift term αdt, some components correlated with various ETF returns, as well as a residual dX_t. Empirically, the returns of most stocks seem to have negligible drift [12], corroborating our belief in mean-stationarity and justifying our inattention towards αdt. It is the residual which interests us, since it captures the stock's idiosyncratic fluctuations. Hence, if we study dX_t's behavior, we can better understand the evolution of X_t, the stock's spread away from its equilibrium value.

We hypothesize, as many earlier researchers have[13], that for most stocks, the spread X_t indeed fluctuates around some equilibrium level, and any deviations are bounded. These are the key characteristics of stationary processes, so one natural model choice is the stationary, mean-reverting Ornstein-Uhlembeck (OU) process, which has differential form:

$$dX_t = \kappa\left(m - X_t\right)dt + \sigma dW_t$$

By considering a transformation $f(X_t, t) = X_t e^{\kappa t}$ and applying Ito's Lemma, the above stochastic differential equation can be transformed into:

$$\begin{aligned} df(X_t,t) &= \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial X_t}[\kappa(m - X_t)] + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma^2\right)dt + \frac{\partial f}{\partial X_t}\sigma dW_t \\ &= \left(\kappa X_t e^{\kappa t} + e^{\kappa t}[\kappa(m - X_t)] + \frac{1}{2}(0)\sigma^2\right)dt + e^{\kappa t}\sigma dW_t \\ &= \Big(\kappa m e^{\kappa t}\Big)dt + \Big(\sigma e^{\kappa t}\Big)dW_t \end{aligned}$$

[10] See Figure 10 and 11 in appendix.

[11] Note: The α and β_j's are assumed to be roughly constant over the trailing window, even though we re-run this regression every day with a different window!

[12] See Figure 12 in appendix.

[13] See Statistical Arbitrage in the U.S. Equities Market (Avellaneda & Lee 2008[1]), Pairs trading (Elliott et al.) 2005 etc.

Given an initial condition $f(X_{t_0}, t_0) = X_{t_0} e^{\kappa t_0}$, the above stochastic differential equation for df can be integrated from $t = t_0$ to T to arrive at an analytical expression for $f(X_T, T) = X_T e^{\kappa T}$ at any time T:

$$X_T e^{\kappa T} = X_{t_0} e^{\kappa t_0} + \int_{t_0}^{T} \kappa m e^{\kappa t} dt + \int_{t_0}^{T} \sigma e^{\kappa t} dW_t$$

$$\implies X_T = e^{-\kappa(T-t_0)} X_{t_0} + m(1 - e^{-\kappa(T-t_0)}) + \underbrace{\sigma \int_{t_0}^{T} e^{-\kappa(T-t)} dW_t}_{\sim N\left(0, \sigma^2 \int_{t_0}^{T} e^{-2\kappa(T-t)}\right)}$$

Setting $T = t_0 + \Delta t$, we arrive at

$$X_{t_0+\Delta t} = a + b X_{t_0} + \zeta$$

where $a = m(1 - e^{-\kappa \Delta t}), b = e^{-\kappa \Delta t}$, and $\zeta \sim Norm\left(0, \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa \Delta t}\right)\right)$.

Clearly, this is an order-1 autoregression[14] with parameters a and b. We can estimate a, b by regressing X_t onto its lag-1 series X_{t-1}[15]. Then, using a, b, we can estimate the parameters of our mean-reverting spread model:

$$\kappa = -ln(b)/\Delta t$$
$$m = \frac{a}{1-b}$$
$$\sigma = \sqrt{\frac{\mathbb{V}(\zeta).2\kappa}{1-b^2}}$$

After we estimate κ, we selected stocks with $\kappa > 252/30$ (indicating stocks whose mean reversion time constant $\tau = 1/\kappa$ is predicted to not exceed 30 days) to estimate their signal at the current time-step (i.e. at day n = 60 of the estimation window) as a normalized deviation of the spread from its long-run mean

$$s = \frac{X_n - \mathbb{E}(X)_{eqm}}{\sqrt{\mathbb{V}(X)_{eqm}}} = \frac{X_n - m}{\sigma/\sqrt{2\kappa}}$$

The mean reversion time filter ensures that we avoid having to hold positions for lengthy periods, which serves to manage the risk of our OU model assumptions and parameter estimation procedure becoming invalid before we close our positions. We also adopted Avellaneda's [1] approach to correct for the finite-sample bias in the estimated long-run mean, m_i, of each stock i by subtracting from each m_i the cross-sectional average of m over all stocks fitted to the OU model.

Choosing Signal Cutoffs

Without full data access, and constricted by slow back-test speeds, the best we could do was to conduct a brief (and heuristic) sensitivity test for our Sharpe ratio as a function of the cutoffs. Figure 1 presents some of our trials.

Ultimately, we chose to set both the "long open" and "short open" cutoffs (s_{bo}, s_{so}) to 1.2, and both the "long close" and "short close" cutoffs (s_{bc}, s_{sc}) to 0.6. This decision was made because it shows decent performance across several contexts (Figure 2):

[14] Verified in Figure 13 in appendix.
[15] Here, lags and times are presented in terms of days, in line with the rest of our investigation.

Transaction Cost	Factors	Open Cutoff	Close Cutoff	Sharpe Ratio
On	SPY	1.25	0.3	1.04
...	...	...	...	...
On	SPY	1.2	0.5	1.01
On	SPY	$1 + 1/3 * \sqrt{\sigma}$	$0.6 - 1/3 * \sqrt{\sigma}$	0.96
On	SPY	$1.2 + 1/5 * \sigma$	$0.55 - 1/5 * \sigma$	1.04
On	SPY	$1 + \frac{1}{30} * \frac{1}{\sqrt{\sigma}}$	$0.6 - \frac{1}{30} * \frac{1}{\sqrt{\sigma}}$	0.97
On	**SPY with Weights Optimization**	**1.2**	**0.6**	**1.21**
On	All ETFs with p values smaller than 0.05	1.2	0.6	0.63
On	Single ETF with smallest p value	1.2	0.6	0.73
...	...	...	...	...

Figure 1: Sharpe ratio sensitivity to varying cutoff values and forms (2003-2012). Here, $\sigma \triangleq$ the trailing 60-day volatility of the wider market (The contemporary S&P500 is used as the proxy).

Expanding Universe Starting Size	Max # L/S	ETF	Open Cutoff	Close Cutoff	Sharpe Ratio	Cumulative Returns	Annual Returns
60	10	SPY	1.2	0.6	0.69	191.00%	11.27%
80	20	SPY	1.2	0.6	0.45	92.92%	6.79%
100	20	SPY	1.2	0.6	0.80	135.00%	8.92%
100	20	Multiple ETFs	1.2	0.6	0.24	11.00%	1.05%
100	20	SPY with 2nd day ETF adjustment	1.2	0.6	0.95	197.40%	11.52%
100	20	SPY with 2nd day Buy modification	1.2	0.6	0.80	145.90%	9.41%
100	40	SPY	1.2	0.6	0.95	67%	5.26%
200	40	SPY	1.2	0.6	0.72	103.70%	7.37%
200	40	SPY with 2nd day Sell modification	1.2	0.6	0.50	62.00%	4.94%
500	100	SPY	1.2	0.6	0.51	52.00%	4.28%

Figure 2: Context sensitivity results for $s_{bo}, s_{so} = 1.2$, $s_{bc}, s_{sc} = 0.6$ (2003-2012).

Basic Implementation Training Results: 2003-2012

Given the very basic ideas discussed in the preceding pages, which were heavily based on Avellaneda & Lee's work, these were our results:

Context

- Period 2003-2012
- Using U.S. equities with the 100 largest-cap stocks as of 2003
- Using 1 factor, the S&P 500
- Capital base: USD$100,000

Results

- Annual Return: 7.9%
- Net Gain: 115.3% (in excess of initial portfolio value)
- Annual Volatility: 0.08
- Sharpe Ratio: 0.99

Figure 3: Net gain over time of basic strategy (2003-2012).

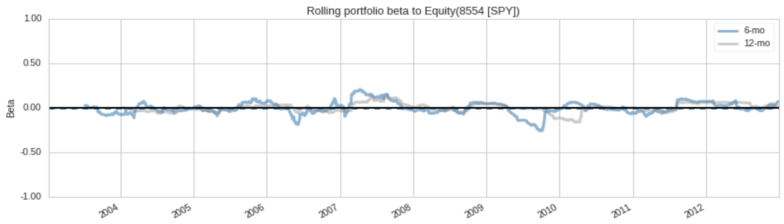


Figure 4: Rolling beta of basic strategy (2003-2012). It remained ≈ 0 across time, as desired.

Note the obvious stagnation of returns after 2008 (Figure 3). Although we succeeded in remaining beta neutral over the course of the strategy (Figure 4), and also achieve a decent Sharpe ratio till the end, we suspect that the 2008 global financial crisis dramatically altered the trading landscape. For instance, financial institutions came under increased scrutiny and tighter regulations, while correlations between asset classes and geographic markets were fundamentally altered[9]. In subsequent sections, we will detail our quest to improve performance during the 2008-2012 period.

In Search of Superior Post-Crisis Results: Portfolio Construction

Assuming an initial capital base of USD$1 million, this section details how we sought a better post-crisis portfolio by implementing several variations of our baseline strategy.

Throughout, we have assumed costless transactions. This is partly because we found that doing costed backtests obfuscated the predictive power of our signals, and partly because in practice hedge funds can lower their own slippage costs. For example, real traders can purchase large quantities of shares from intermediaries out of the public eye, without having to execute block trades that would drive up prices. Thus, ignoring such costs painted a more realistic portrait of our strategy's effectiveness.

Verifying (Weak) Stationarity

There is reason to believe that despite their reversionary tendencies, stock returns might not be truly stationary. For instance, while the finite-horizon behavior of a stock's spread *seems* to demonstrate a

roughly constant mean, said mean might in fact change over longer periods, such that the process eventually reverts to a different equilibrium.

This perspective stems from the fact that if returns and spreads truly had constant averages, then over time, investors would determine the "fair price" for each stock and trade only on deviations. But then since every stock's fair value would be constant and known, there would be little cause for deviations and thus no profit opportunities to incentivize risk-taking. Yet the stock market remains alive and well, hence disproving the notion of true mean-stationarity.

To avoid mistakenly applying our OU model to nonstationary spreads, an augmented Dickey-Fuller test[8]was used to prescreen stocks. This test fits the following model to each stock's spread process X_t:

$$\Delta X_t = \alpha + \gamma X_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta X_{t-j} + \epsilon_t$$

The null hypothesis is that the spread X_t is non-stationary, corresponding to $\gamma = 0$. To prove a stock's spread process is stationary, there must be statistical evidence to reject the null. This test, combined with the earlier filter for fast mean reversion, helps to ensure that we apply our statistical arbitrage model only on stocks which validate its assumptions of mean-stationarity and swift reversion.

Relating Stocks to ETFs

Previously, we mentioned that a stock's expected returns may be forecast by regressing on the market (i.e. the SPY ETF). Yet it makes little sense to indiscriminately regress every stock on an arbitrary set of ETFs; not all stocks are *related* to all ETFs, even if they are *correlated* with the broader market. To more intelligently map stocks to ETFs, we conducted one-off regressions and used F-testing/ANOVA to select which stocks were significantly correlated to which ETFs.[16]

Rather than trading pairs, we trade "groups", each composed of 1 stock together with its related ETF(s). The ratio of a stock position to its jth ETF's is $1 : \beta_j$, where $\beta_j \triangleq$ the corresponding coefficient from the returns regression.

We hedge our portfolio in this manner to achieve beta neutrality; that is, we want to de-correlate our holdings from the broader market, to insulate our alpha from the volatility of individual tickers.[17] This would ordinarily complicate our analysis, but doing many of these group trades cancels out the ETF investments such that the *net* position in any one ETF is approximately 0. Therefore we can treat our hedged portfolio as if it were purely equity. We will carry this simplifying assumption over to subsequent discussion.

Optimizing Weights of New Positions

Consider a typical day where we wish to open m group trades ($i \in [1, m]$), where each group trade consists of a beta-neutral set of long-short positions in a stock together with its ETFs. How do we distribute available funds amongst them?

Initially and naively, we distributed an equal dollar amount to every stock held, with weights normalized such that the sum of absolute weights is always 1. However, this is too crude to be optimal; surely we should place more weight on positions that are more likely to deviate profitably from their expected values?

We believe that there exists a unique, optimal set of weights for these m groups. To determine these weights, we attempted to formulate and solve an optimization problem:

[16]We first regress each stock on the SPY to remove common correlation with the market, then take its residuals and regress on the other 10 pre-selected ETFs (tickers: XLF, HHH, IYR, RKH, SMH, UTH, XLE, XLI, XLP, XLK). This will highlight which ETFs are statistically significant (small p-value) to the response.

[17]This means higher rewards for lower risk.

$$\max_{w_1,\ldots,w_m} \frac{\sum_{i=1}^{m} w_i.(-s_i)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}w_i w_j}}$$

where s_i is some signal indicative of potential expected return from mean reversion in the group trade position i, and σ_{ij} is the covariance between signal s_i from group i and signal s_j from group j. As our algorithm takes a long position in the spread ($w_i > 0$) when $s_i < 0$ and a short position in the spread ($w_i < 0$) when $s_i > 0$, we chose to maximize the function $\sum_{i=1}^{m} w_i.(-s_i)$, which is indicatve of the expected return from the set of new positions we intend to open long or short. The first order condition (FOC) for this optimization problem can be written as a linear system of m equations:

$$\sum_{j=1}^{m} \sigma_{ij} \cdot c \cdot w_i = -s_i$$

where c is some unknown constant for each i from 1 to m. Substituting $v_i = cw_i$, the FOC is rewritable as

$$\sum_{j=1}^{m} \sigma_{ij} v_i = -s_i$$

We solve this for v_i, which can then be normalized to obtain w_i:

$$\mathbf{v} = \mathbf{\Sigma^{-1}}(-\mathbf{s})$$
$$w_i = \frac{v_i}{\sum_{i=1}^{n} v_i}$$

Here, $\mathbf{\Sigma}$ is the signal covariance matrix of the m group trade positions, to be estimated from the trailing 60-day values of $s_i : i \in [1, m]$:

$$\sigma_{ij} = \frac{1}{n-1}\sum_{t=1}^{n} \left(s_i(t) - \bar{s_i}\right)\left(s_j(t) - \bar{s_j}\right)$$
$$\bar{s_i} = \frac{1}{n}\sum_{t=1}^{n} s_i$$

Here, we cannot use Markowitz-style objective functions (expected return, Sharpe ratio)[3] because the long-run expectated return of a spread is zero, so the conventional notion of a return does not make sense. Furthermore, the estimation of the required covariance matrix $\mathbf{\Sigma}$ is error-prone in light of limited and/or noisy financial data, so a theoretical derivation was simpler[18].

Hence, the above optimization can be simplified to just assigning weights in proportion to signal magnitudes. In other words, we used signal strength as a proxy for the short-run spread return, and maximized the weighted sum of signal strengths after normalizing it by the portfolio's standard deviation. This corresponds to allocating more funds to stocks which are predicted to revert faster and more profitably than the others. Backtests showed that this intuitive approach yielded slight improvements in our backtests versus just allocating funds equally to every new position.

Risk Management Philosophy

Recall that we do not "reuse" signals; we keep our holdings of a stock unchanged even when its initial signal to buy (or sell) persists over several days, and will change our position only when we see the appropriate

[18] Assuming that the spread process $X_{i,t}$ for each stock i is idiosyncratic and hence independent, their signals, which are normalized spreads, are expected to have variance $\sigma_{ii} = 1$ and covariance $\sigma_{ij} = 0$.

signal. Although this does cost us potential profit[19], this is an important point of risk management: A very strong signal might stem from a stock that is sinking quickly due to a corporate scandal, and which may not recover for a long time (if at all) - In such cases, doubling down would be very undesirable despite the clear deviation from the stock's historical mean. To prevent the algorithm from buying into a fast-sinking ship, it makes sense to hold off until we can verify the transience of this deviation by studying the rest of the market.

The above consideration highlights a fundamental weakness of statistical arbitrage: Its belief in, and completely dependence on, mean reversion. The strategy's profits (and losses) depend on the ability of prices to return to equilibrium[7] in a reasonable time frame. In times of crisis, investor panic and pessimistic sentiment can depress prices for lengthy periods, which prevents spreads from closing and also starves the algorithm of short-selling opportunities. In such difficult circumstances, the strategy will perform very poorly[20].

To ameliorate this weakness, we scale stock and ETF returns with a volume adjustment factor equal to $\frac{\text{avg daily volume over 60-day trailing window}}{\text{previous day's volume}}$ as in [1]. This scaling factor serves to downweight mean-reversion signals for stocks experiencing larger than average trading volumes, while emphasizing signals under small volumes. This is desirable because sudden increases in trading volume may indicate that the stock has been significantly affected by exogenous factors, so its spread process is likely to have experienced a long-lasting shift rather than a temporary deviation. Given these circumstances, we should logically place less trust (and trade fewer or smaller positions) in signals associated with stocks experiencing abnormally large volume changes, as they are less likely to revert within a reasonable time frame (if ever). Along the same vein, we also reject stocks whose predicted mean reversion time constant exceeds 30 days[21].

Capital constraints are another threat: Given a fixed budget, as well as the need for a buffer to ensure we can close bad shorts at will, there is a natural constraint on the number and dollar value of positions we may enter. If we run out of funds to complete a "group" trade, then we fail to become beta neutral, thus unknowingly exposing our portfolio to market risk. To forestall these difficulties, we chose to restrict the number of longs and shorts that we could hold at any instant, as well as avoided holding extremely large positions in any one ticker. This not only ensured diversification[22], but also prevented us from being hit by margin calls (which might force us to relinquish positions that we should have held onto at unfavorable moments). Of course, these decisions come with their own difficulties. For instance, we surely miss profit opportunities when we skip some of the recommended trades. Nonetheless, being able to control which opportunities we give up is infinitely better than being backed into a corner by margin calls and insolvency.

Lastly, market illiquidity is a critical roadblock: If we cannot fill a planned order[23], then we cannot fully open or close positions on both sides of the spread. If we cannot enter all the intended positions, not only does our portfolio fail to achieve beta-neutrality (thus exposing us to unplanned market risk), we also lose some profit opportunity when the stock mean-reverts, since we failed to fully take the other side of the spread. This may also be an issue when trying to exit our positions: If the spread re-widens before we can close out, then we are forced to take smaller profits at an inferior timing. On a single transaction this cost might be written off, but it adds up significantly over many group trades and long horizons.

Unfortunately, illiquidity is a risk inherent to trading low-volume tickers. We rejected the idea of pre-screening our universe for tickers that would be extremely prone to fill failure, since any *effective* pre-screening would have to condone lookahead bias. Instead, we experimented with three courses of action:

(1) Persisting incompletely-filled orders until they were completed. This was straightforward to implement, albeit suboptimal since it affects the portfolio's beta neutrality and increases the "bad fill" rate (since the tendency is to 'win' contrarian trades).

[19] It is easy to envision cases where doubling down would have paid off... but always in hindsight!

[20] This is especially since we did not implement stop losses, which contradict the spirit of statistical arbitrage by preventing us from profiting off reversions. If stop losses are desired, they must be chosen carefully via testing to ensure they lie far outside the spread's usual range.

[21] As inferred from the parameter κ in their individual OU processes. Based on Avellaneda & Lee's discussion[1].

[22] Diversifying allows us to control idiosyncratic downside risk.

[23] Unfortunately this may not be a rare event, especially when trading small-volume tickers

(2) Filling as much of a stock order as possible, then ordering only the proportionate ETF amount that same day.

(3) Filling as much of every order as possible, then trading off any excess ETF positions the next day in response to incompletely-filled equity orders.

Our simulations showed that (3) performed the best by a slight margin across a wide range of contexts (using the S&P500 as the lone factor, using several ETF factors, etc.), so this is what we have chosen to do.

Training Results and Discussion: 2008-2012

To determine which trading schemes were superior, as well as to tune our parameters, we experimented with the following algorithmic variants in a costless environment where the trading universe was updated annually to be the 500 largest-cap stocks:

(1) Baseline (SPY): Spread estimated from a pair comprising stock and the S&P500 ETF (i.e. SPY) — this is the same strategy implemented on page 5

(2) Baseline (SPY) + Stationary: Combining (1) with the A.D. Fuller stationarity test to pre-screen stocks

(3) SPY + Multiple ETF + Stationary: Similar to (2), but now each stock is regressed on the SPY and multiple sector ETFs

(4) SPY + Best Sector ETF + Stationary: Similar to (3) except a stock's spread is regressed on the SPY and its most-correlated sector ETF

Universe	Max Stocks	Strategy	Performance Metrics		Risk Metrics		
			Annual Return	Sharpe	Beta	Volatility	Max Drawdown
100	20	Baseline (SPY)	9.5%	0.78	0.01	0.13	-14.4%
		Baseline (SPY) + stationary	4.0%	0.57	-0.01	0.07	-7.7%
		SPY + multiple ETF + stationary	4.1%	0.65	-0.01	0.07	-9.0%
		SPY + best sector ETF + stationary	5.7%	0.86	0	0.07	-10.2%
500	**100**	Baseline (SPY)	6.3%	0.67	0.01	0.1	-11.3%
		Baseline (SPY) + stationary	**7.3%**	**1.23**	**0**	**0.06**	**-8.4%**
		SPY + multiple ETF + stationary	2.6%	0.54	0.01	0.05	-12.2%
		SPY + best sector ETF + stationary	4.2%	0.84	0.01	0.05	-6.6%
1000	200	Baseline (SPY)	8.1%	0.92	-0.01	0.09	-14.6%
		Baseline (SPY) + stationary	3.1%	0.58	-0.01	0.06	-10.4%
		SPY + multiple ETF + stationary	3.6%	0.81	-0.01	0.04	-5.9%
		SPY + best sector ETF + stationary	3.4%	0.77	-0.01	0.04	-5.8%

Figure 5: Performance and risk summary for strategy variants under different contexts (2008-2012). Note: Value at Risk (VaR) is calculated as the 5th percentile of the portfolio's daily return.

As seen in Figures 5 and 6, backtests demonstrate that all variants achieved positive annual returns with generally low systematic risk exposure[24]. However, with a 500-ticker investment universe and a maximum of 100 long and 100 short holdings at any one time, variant (2) attained the highest Sharpe ratio at 1.23. Thus, compared to the baseline, the stationarity check seems to enable the algorithm to achieve strong returns while reducing volatility and drawdown.

Variants (3) and (4) struggled to perform well. Possible reasons for their poor results include:

[24] i.e. low overall beta, rolling beta

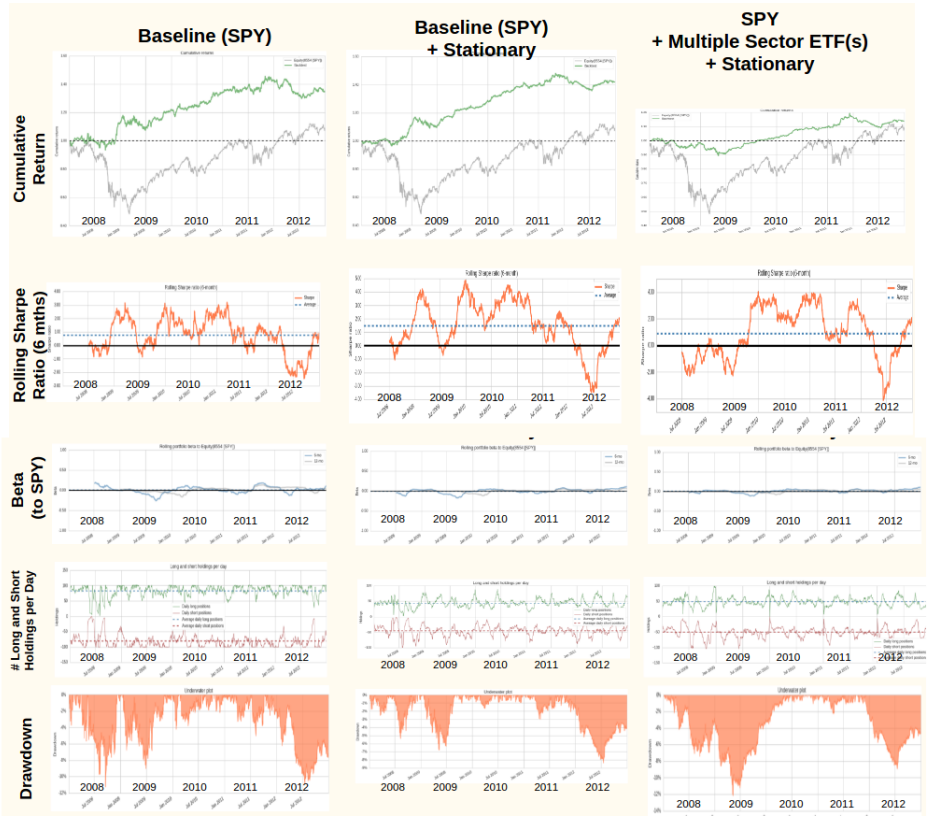


Figure 6: Time series of portfolio performance and risk metrics during training period (2008-2012).

- The use of sector ETFs "regressed away" too many of the fluctuations in the stock spread processes, thus preventing the algorithm from profiting off reversionary deviations
- The models assume that stock-factor correlations remain fairly constant over time, which may not be true; stocks that were initially mapped to highly-correlated ETFs may have these relationships significantly altered by market changes. Given the frequent and extreme volatility during 2008-2012 (e.g. the financial crisis, quantitative easing, Eurozone woes), the high volatility and maximum drawdowns of these two variants are unsurprising.

Based on its superior performance, we chose (2) as our strategy for out-of-sample testing.

Out-of-Sample Results and Discussion: 2013-2016

Recall that we set aside the period 2013-2016 as our out-of-sample data, having fitted our model and tuned our parameters without it. We now present results and analysis revolving around this set.[25]

Performance

Figures 7 and 8 show that the chosen "best model" (strategy (2)) achieved a much lower Sharpe ratio of 0.54 over the test period (2013-2016), compared to its training result of 1.23. Nevertheless, it still outperforms the baseline model over the same period; again, the stationarity check seems to significantly reduce volatility and maximum drawdown.

[25]Not using this data until now avoids overfitting, a phenomenon where researchers succumb to temptation and tweak their model to perfect their claimed performance.

Universe	Max Stocks	Strategy	Performance Metrics			Risk Metrics			
			Annual Return	Sharpe	Sortino	Beta	Volatility	Max Drawdown	VaR*
500	100	Baseline (SPY)	2.2%	0.31	0.44	0.07	0.08	-18.7%	-0.9%
		Baseline (SPY) + stationary	2.2%	**0.54**	0.75	0.01	0.04	**-6.6%**	-0.47%
		SPY + Multiple ETF + Stationary	2.4%	0.59	0.84	0.01	0.04	-6.9%	-0.53%
		SPY + Best Sector ETF + stationary	2.5%	0.59	0.84	0.01	0.04	-6.9%	-0.53%

Figure 7: Comparison of strategy variants over test period (2013-2016). Investment universe of 500 largest cap stocks. Maximum 100 long and 100 short positions in selected stocks. No transaction costs.

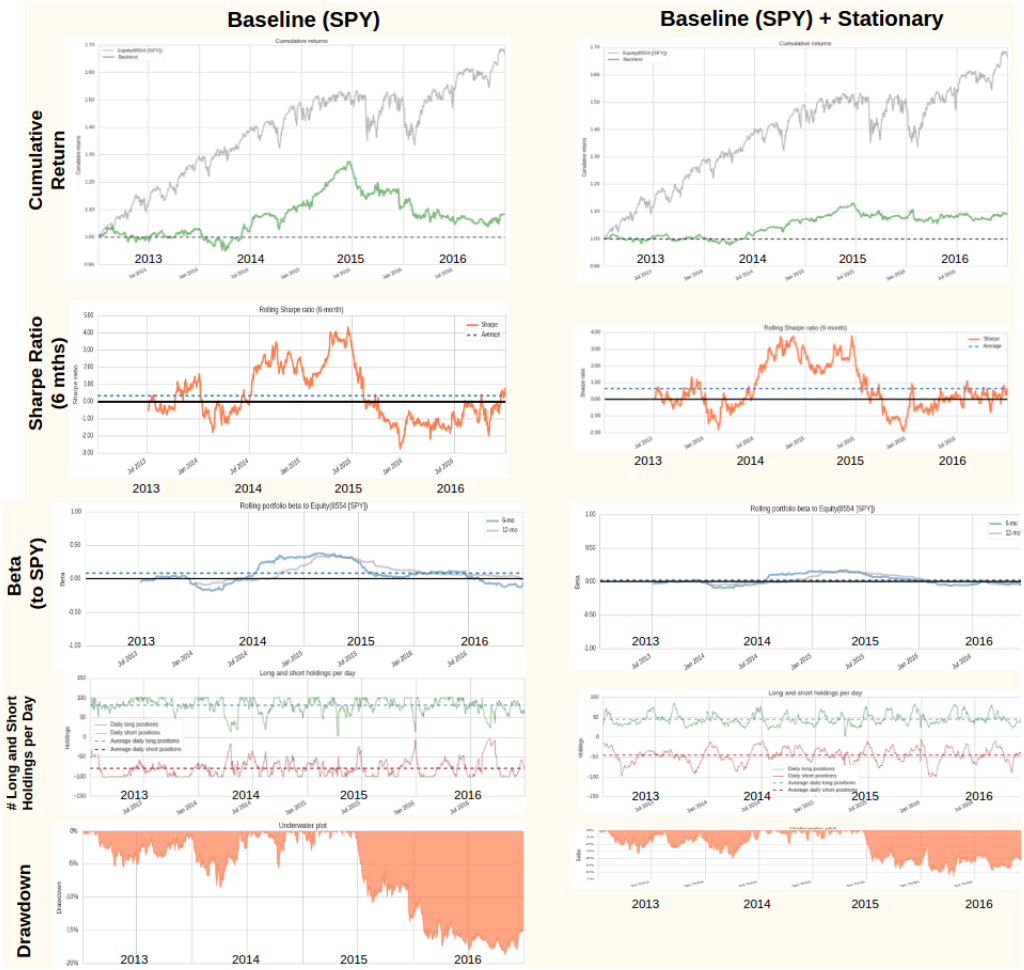


Figure 8: Time series of portfolio performance and risk metrics during test period (2013-2016). Cases involving "Multiple ETF" and "Best Sector ETF" are similar to "Baseline (SPY) + Stationary". Investment universe of 500 largest cap stocks. Maximum 100 long and 100 short positions in selected stocks.

Nonetheless, this is a disappointing performance, especially in light of (2)'s strong training results. We suspect that the abysmal test results could be attributed to several factors, including but not limited to:

- Avenalleda & Lee concluded their study before the 2008 global financial crisis, having run their model only on preceding years. Perhaps their research was overfitted to earlier years, when the market had been more predictable and less exposed to global risks than it has been in recent years. In the

aftermath of the crisis, it is also reasonable to suspect that tighter regulations and altered investing attitudes resulted in a fundamentally different trading environment post-2008, which may have erased the arbitrage opportunities that they had been exploiting.

- The publication of Avellaneda & Lee's research meant that their ideas — and their optimized parameters — were now publicly known. This enabled other traders to begin replicating or countering regression-based strategies in live markets, which would surely eliminate any profit potential that they might have had.
- In recent years, trading has also been impacted by unprecedented Fed interventions. For instance, the 2012 quantitative easing increased market liquidity and drove up stock prices across the board. This would have killed off prospective short opportunities and thus prevented us from taking both sides of a group trade, while also distorting how our algorithm saw stock spreads. The anti-inflationary measures that were implemented later, such as the 2015 rate hikes, would have further warped the markets. Our algorithm was not capable of adjusting for these events, and would have certainly lost its predictive powers thereafter.

One consolation was that our stationarity check noticeably improved the out-of-sample risk metrics. For instance, our model (2) saw a worst-case maximum drawdown of just 6.5%, compared to the baseline (1)'s drawdown of 18%. Our chosen model (2) also ensured that the portfolio was more consistently beta-neutral, which meant that (2) faced less systematic risk than the baseline (1). Tail risk and downside risk, as captured by Value-at-Risk (VaR) and Sortino ratio respectively, also improved under model (2).

Out of curiosity, we also investigated the performance of models (3) and (4) in the 2013-2016 markets. As noted in figures 7 and 8, (3) and (4) both performed marginally better than our chosen strategy (2). However, their VaR values showed a slight deterioration indicating larger tail risk. We suspect that (3) and (4) were better able to capture the dependence of stock returns on multiple factors, which improved returns but increased their vulnerability to extreme events. However, this is only speculation and will necessitate further investigation.

Summary and Conclusion

This project sought to replicate and extend Avellaneda & Lee [1]'s statistical arbitrage strategy, by making many market-neutral group trades within the US equity market and profiting from mean reversion. Though we found success during 2003-2008, our returns stagnated in later years, potentially due to market changes, overexploitation by competing traders and other unpredictable events. It seemed that implementing a statistical test to verify spread stationarity played a key role in limiting volatility and maximum drawdown, which helped to preserve our returns even in difficult conditions. However, poor test results show that our strategy fails to adapt to sudden changes in the economic or trading environment, which implies that more work needs to be done to build a dynamic, self-adjusting strategy that can survive a changing world.

All in all, a key takeaway was the importance of risk management throughout the entire process, from building a model to optimizing parameters and executing trades. We have also been led to believe that the linear regression-based approach employed by Avellaneda & Lee, as well as our team, is too simplistic to have any competitive edge in current markets. Given what we know now, we would strongly recommend using more sophisticated methods — such as principal component analysis, or clustering — to identify and use better predictive features for stock spreads. We also suggest implementing mechanisms that detect market events which might invalidate key assumptions, so that parallel algorithms may adjust our strategy to account for different conditions.

The following list details other avenues for future exploration:

Alternative Factor Selection

Our research showed that a model which includes sector ETFs performs only slightly better than a model which regresses stocks on the S&P500 alone. This result held even when we "regressed out" the market's effect on returns before mapping the stocks to ETFs[26]. One possible explanation is that simple linear regression is too crude to accurately assess the relationships between stock returns and ETFs, and is returning factors which are irrelevant or spuriously correlated; perhaps other machine learning methods like LASSO or nonparametric regressions would improve performance.

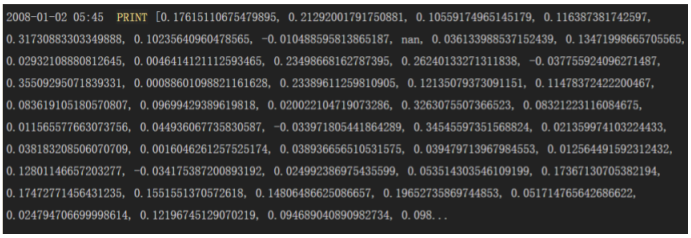


Figure 9: Adjusted R^2 values of various tickers, when regressed on the 10 sector ETFs

We verified this concern by considering the adjusted R^2 values of regressing the market-independent components of returns on all 10 non-market ETFs (Figure 9). The abysmally low R^2 values suggest that as suspected, the sector ETFs are poor predictors of stock returns. Therefore, we propose that further investigations should consider non-ETF factors. For instance, qualitative attributes such as market capitalization or industry classification could be studied as prospective predictors of a stock's reversionary tendencies:

More Robust Parameter Optimization Procedures

One area that can be improved within the quantopian platform is a more robust weights optimization scheme that accounts for other meaningful constraints, such as an upper limit on the order amount for stocks as a percentage of their daily volumes, explicit constraints on our cash position or leverage ratio. Unfortunately, many of our attempts at parameter optimization were conducted using trial and error, rather than intelligent search. This was due to Quantopian's strict limitations on data accessibility: As we could not export any values to external optimizers, nor were we able to invoke private methods hidden by the site's black-box operations, brute-force sensitivity testing seemed the easiest approach to tune our parameters[27]. As Quantopian improves the integration between its algorithmic and research capabilities, perhaps they will also enable different optimization methods e.g. gradient descent, walk-forward optimization.

Conditionally-Heteroskedastic Models

In this paper we applied linear regression models at crucial junctures, assuming that their residuals were Gaussian and characterized stationary spread processes. However, it is known that stock returns exhibit volatility clustering and heteroskedasticity, as well as fatter tails. This makes returns series good candidates for GARCH models. However, the conditional non-stationarity of the residuals would fundamentally change how the spread process should be modeled, and is therefore not compatible with the preceding investigation.

Non-stationary Models

In 'Portfolio Construction', we argued for the necessity of verifying the stationarity of a stock's returns using the augmented DickeyFuller test. Stocks which are statistically determined to be non-stationary are then removed from consideration, as the assumptions of our model are invalid in such cases. However, if we could somehow adapt the OU process to account for non-stationary tickers, then we might be able to develop a portfolio that can profit off even more stocks within our universe.

[26] Because we were concerned that the market effect might mask the smaller sector effects

[27] Online references only demonstrate similar trial-and-error methods.

References

[1] Avellaneda, Marco, and Jeong-Hyun Lee. "Statistical Arbitrage in the U.S. Equities Market." SSRN Electronic Journal (2008): 1-47. Web.

[2] "Basic Usage." Quantopian Help. N.p., n.d. Web. 22 Apr. 2017. <https://www.quantopian.com/help>.

[3] "Markowitz Mean-Variance Portfolio Theory." <https://sites.math.washington.edu/∼burke/crs/408/fin-proj/mark1.pdf>

[4] M.H. Moore, QuantStart. "Basics of Statistical Mean Reversion Testing". 23 Oct 2013. <https://www.quantstart.com/articles/Basics-of-Statistical-Mean-Reversion-Testing>.

[5] Payne, John. "Re: Questions from Students." Message to Lisa Marina Borland. 21 April 2017. E-mail.

[6] "Quantopian." Quantopian. N.p., n.d. Web. 16 Apr. 2017. <https://www.quantopian.com>.

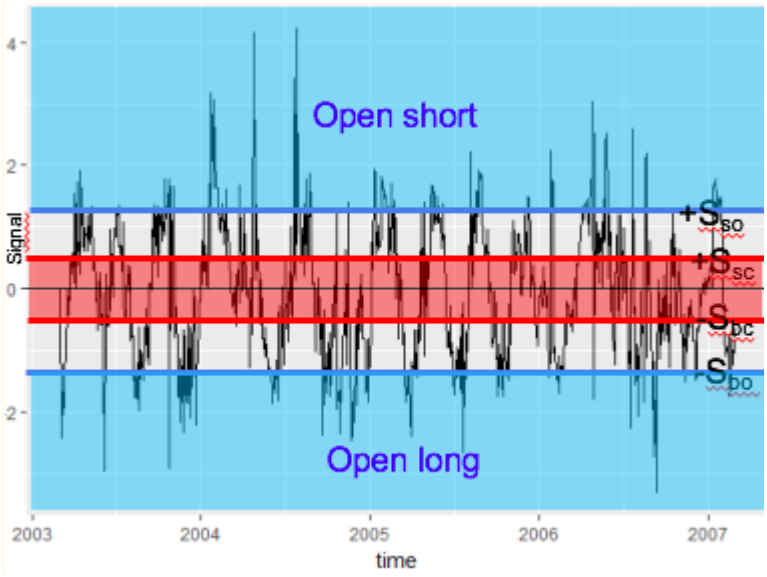
[7] Staff, Investopedia. "Statistical Arbitrage." Investopedia. N.p., 26 Nov. 2003. Web. 13 May 2017. <http://www.investopedia.com/terms/s/statisticalarbitrage.asp>.

[8] StatsModels in Python. "Augmented Dickney Fuller Unit-Root Test". <http://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.adfuller.html>

[9] Zhang, Bing, Xindan Li, and Honghai Yu. "Has recent financial crisis changed permanently the correlations between BRICS and developed stock markets?" The North American Journal of Economics and Finance 26 (2013): 725-38. Web.

Appendix A Supporting Figures

Figure 10: The evolution of a ticker's signal over time, together with the relevant signal cutoffs.



***

Figure 11: A sample spread process over time, together with its signal cutoffs.

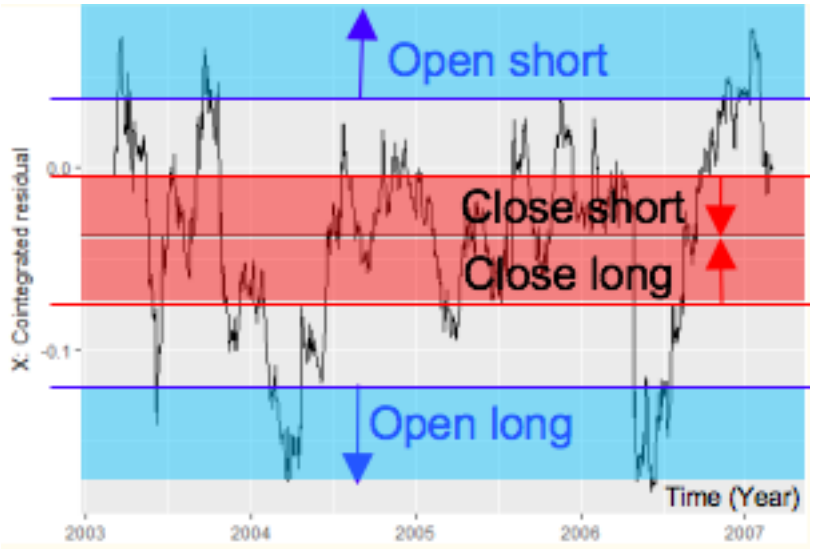
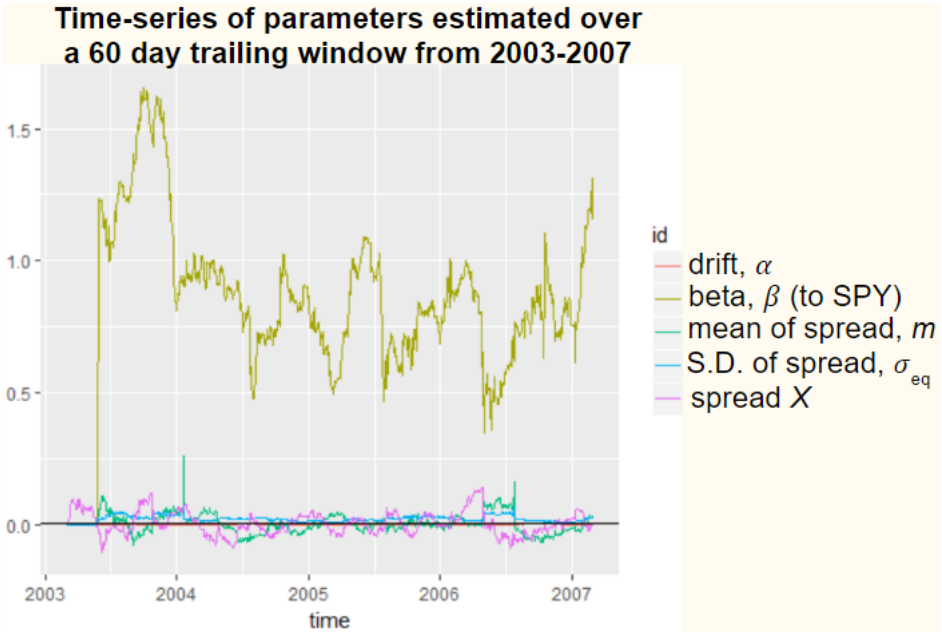
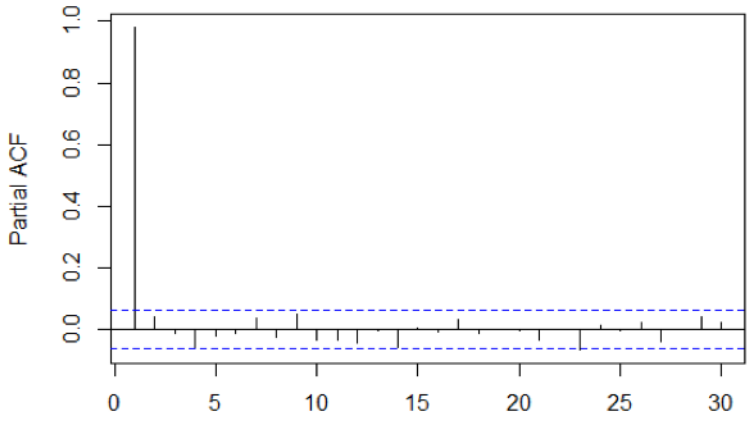


Figure 12: The drift of this ticker (red) appears negligibly close to zero at all times, an observation that holds true for most tickers.



***

Figure 13: A sample partial autocorrelation function plot from the market data of an actual ticker. Note the presence of a uniquely significant PACF at lag 1 and the sharp cutoff thereafter, which identify this as an AR(1) process.



Appendix B Execution Discussion: Real-Life Challenges

Our paper, in relying on a simulated environment to run backtests, has simplified away some execution challenges that would have to be addressed in the live markets. Some of them are noted below.

In our implementation, we assumed a costless trading environment. In real life, however, trading costs must be successfully optimized away[28], which could be very difficult. Because the average holding time of our strategy is around 11 to 12 days, this suggest a rather high portfolio turnover, which would - without careful cost management - add up to significant slippage.

Another challenge we would face in practice, but not in a simulated environment, are operational challenges. For instance, in Quantopian we were easily able to correct bugs in our code and re-run backtests, and trades were guaranteed to execute after the default fifteen-minute paper trading delay; in real life, programming mistakes might be irreversible or even fatal, while delays might be wildly unpredictable. High-frequency rivals or human error might also hobble our execution process, and create whole host of issues that we did not have to account for within our simulations.

Lastly, though Quantopian's restricted data usage policies prevented us from making live adjustments to our algorithm or performing custom analysis[29], real markets would similarly fail to provide us with perfect information. In fact, prompt access to clean and complete data is not guaranteed in practice, and may be worse than in our simulations.

[28] For example, avoid block trading by buying or selling with intermediaries such as banks or brokerages

[29] Much energy was expended trying to obtain a histogram of our holding times, with no success

Appendix C Miscellaneous Thoughts

- We began by implementing Avellaneda & Lee's paper[1] using a universe of the top 100 largest-cap stocks in Quantopian's QS1500 collection, and limiting our portfolio to at most 20 long and 20 short positions at any one time. However, we quickly realized that these holding limits were usually unmet, suggesting that we were trading on too few opportunities. We thus decided to expand our universe to 500 stocks, as some quick tests showed that 1000 stocks included too many low-liquidity tickers to earn a solid profit.

- We pre-selected the following 11 factors as candidates for mapping our stocks onto: SPY, XLF, HHH, IYR, RKH, SMH, UTH, XLE, XLI, XLP, XLK. The inclusion of SPY has clear motivations, and the rest are chosen as representatives of a range of common sectors (adapted from the list proposed by Avellaneda & Lee[1]).