(IJTBM) 2020, Vol. No.10, Issue No. III, Jul-Sep **e-ISSN: 2231-6868, p-ISSN: 2454-468X**

PAIRS TRADING BASED STATISTICAL ARBITRAGE USING COINTEGRATION APPROACH AND KALMAN FILTER

Lavaneesh Sharma

ABSTRACT

In this paper we explore the pairs trading based statistical arbitrage technique. The proposed pairs trading methodologies was employed to equity trading systems to find the stocks and their underlying ETF's and was able to identify the relative statistical mispricing between the prices of stock-ETF pairs, using regression residuals, and to modeled them as natural mean-reversion processes with a short holding period in the U.S. equities market under any market cycle conditions. The strategy follows a twostep process. First in the ***formation phase*** *we select and identify securities that have long term mean reverting spread; the spread having high standard deviation to allow a profitable strategy. In the actual* ***trading phase,*** *we define rules for trading entry and exit points as prices diverge and converge. Finally, we the back test the strategy using backtrader.*

1. INTRODUCTION

Statistical arbitrage is set of procedures employing an underlying model or method – which while maintaining a level of market neutrality – benefit from relative mispricing of assets. Statistical arbitrage depends heavily on the ability of market prices to return to a historical or predicted mean. The idea is to make many bets with positive expected returns and to produce a low-volatility investment strategy (Avellaneda & Lee, 2010), thereby taking advantage of diversification across assets. Pairs Trading is a relative simple Statistical Arbitrage that aims to finds two assets that have moved together in the past, keep a track of their spread and if this given spread widens, long the asset that has fallen below the long term equilibrium and short its corresponding pair. If the so called relationship persists, the either legs will deliver profits as prices converge and positions are closed. As such selection of pairs of securities of basket of securities becomes imperative.

2. CORRELATION AND COINTEGRATION

Correlation and Cointegration in statistical arbitrage are somewhat associated, but they highlight different theories. Cointegration differs from correlation in the sense that two series maybe highly correlated but need not be cointegrated. For example, if two series are multiples of each other, correlation will be high, but so will the linear combination grow rather than revert to a mean. Correlation implies how two assets move with respect to each other – or strength of movement of two variables – but is significantly unstable over time. If two assets happen to be correlated to each other, then any changes in correlation may allow the asset pairs to mean revert. High correlation is not the only criteria to ensure that the hedges will perform in the long term as well. Correlations based on hedge strategies commonly require frequent rebalance (Alexander, 1999). On the other hand, cointegration measures long-term co-movements in

prices even through a period when correlation appears low. As such, cointegration based on hedge strategies may be more effective in long-term running and short-term dynamic trends. (Miao, 2014). A formal mathematical definition and strategy are given below:

2.1 *CORRELATION*:

Calculating correlation coefficient is the oldest yet simplest method to select and trade commoving assets. A distance based heuristic – that aims to minimize the sum of squared deviations and uses non-parametric thresholds like Bollinger Bands to trigger entry and exit trades. For two assets X & Y, a correlation coefficient defines association between them and how much they are closely related in a linear fashion. The coefficient ρ is given by:

$$\rho = \frac{\sum_i^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^N (X_i - \bar{X})^2 \sum_i^N (Y_i - \bar{Y})^2}}$$

$$where, \bar{X} \; and \; \bar{Y} \; are \; mean \; values \; given \; by$$

$$\bar{X} = \frac{\sum_i^N X_i}{N}$$

$$\bar{Y} = \frac{\sum_i^N Y_i}{N}$$

Trading strategies relying purely on correlation of returns and cannot guarantee long-term performance. It does not guarantee the reversion of the hedge to the underlying, and essentially powerless to prevent tracking error from behaving in the unpredictable manner of a random walk.

Cointegration becomes necessary over and above correlation, as it provides a sufficient evidence to ensure a longer-term evaluation of hedges, and that provides us with necessary confidence that the risk modelling will ensure a longer term trend in prices.

2.2 *COINTEGRATION:*

The idea behind cointegration is to consider a pair of non-stationary time series and form a linear combination of each series to produce a stationary series which has fixed mean and variance. The said stationary series will have short deviations from this fixed mean and we precisely have to take advantage of this, as the series due to its stationarity will revert to its long term mean. Formally speaking,

"if two series $\{x_t\}$ and $\{y_t\}$ be non-stationary series and $a, b \; \epsilon \; R$ are constants, and if the combined series $a\{x_t\} + b\{y_t\}$ turns out to be stationary then we say series $\{x_t\}$ and $\{y_t\}$ are stationary"

2.3 *STATIONARITY:*

A stationary process is one in which when arbitrarily shifted in time doesn't change its joint probability distribution i.e. its mean and variance are time-independent. For a first order stationary process the time series signal independent of time the mean μ(t) and variance $\sigma^2(t)$ are constant over time as:

$\mu(\mathrm{t}) = \mu \ ; \ \sigma^2(t) = \ \sigma$

Similarly, if for second order function that depended only time difference $t_2 - t_1$ rather than on individual times $t_2 \ or \ t_1$.

A characteristic equation which is formed as a function of backward shifts of its own time series is set to zero. Solving this equation, we get its roots. All its roots must be greater than 1 for the series to be stationary. Therefore, in order to detect stationarity (and hence cointegration) we have statistical hypothesis testing procedure for presence of unit roots.

There are two approaches used for our paper:

1. The Engle-Granger two step method
2. The Johnsen test

We shall discuss these two before we can actually leverage them for our strategy.

2.4 *THE ENGLE-GRANGER TWO STEP METHOD:*

It is used to detect cointegration in time series. It involves the following:

- Regression of two series with respect to each other to understand their relationship
- Using an Augmented Dicky Fuller test on regression residuals.

In the original Dicky Fuller test, the series $z_t = \ \alpha * z_{t-1} + \ w_t$ was used to find the presence of a unit root, where w_t is discrete white noise. The hypothesis test is given as:

$$H_0: \alpha = 1$$
$$H_a: \alpha < 1$$

If we can reject the null hypothesis, then we can assume that the residuals are stationary and this series is cointegrated.

The advantage of this approach is that the regression coefficient β gives us the resultant combination that renders the given series stationary. However, the results differ, according to the choice of which series do we consider independent. We will correspondingly have to compare both p-values in both the cases. The main disadvantage however, is that it can only be applied to a pair of assets. For incorporating more assets into the mix we need the complex Johansen procedure, we can help in analysis of multiple time series at once.

2.5 *JOHANSEN TEST:*

To overcome shortcoming of Engle-Garner method – which can only be applied to two separate time series – we move towards Johansen test. It uses Vector Autoregressive Model (VAR) which is a multi-dimensional extension of Auto Regressive Model AR(ρ). The general form of the model with drift is given as:

$$x_t = \ \mu + A_1 x_{t-1} + \cdots + \ A_p x_{t-p} + \ w_t$$

Where μ is vector valued mean of series, A_i are coefficient matrices for each lag and w_t is Gaussian noise term. Using this we can form a Vector Error Correction Model (VECM):

$$\Delta x_t = \mu + \Pi x_{t-1} + \Gamma_1 \Delta x_{t-1} + \cdots + \Gamma_p \Delta x_{t-p} + w_t$$

Where Δ is the differencing operation: $\Delta x_t = x_t - x_{t-1}$; Γ_1 is the matrix for each lag and Π is the coefficient of the first lag.
The nature of cointegration depends on the rank of coefficient matrix Π . Situation of no cointegration occurs at $\Pi = 0$.

The Johansen test automatically checks for multiple linear combinations to form a stationary portfolio. It does this by *Eigenvalue Decomposition* of Π. The rank of matrix Π is given by r and Johansen test *sequentially* tests weather this rank is equal to zero, one and all through
$r = n - 1$. The hypothesis test then is given as:

$$H_0: r = 0$$
$$H_a: r > 0$$

Where null signifies no cointegration at all, while Alternate suggests that there is a cointegration possible.

2.6 *COINTEGRATION APPROACH:*
If the prices of two assets A and B are given as P_t^A and P_t^B and suppose there is a parameter δ such that P_t^A - δP_t^B is stationary, then a cointegration model for these assets can be represented as:

$$P_t^A - \delta P_t^B = \mu + \epsilon_t$$

with μ as mean of the model (or the prices where they will converge in the long run) and ϵ_t as the cointegration residuals. The amount of profit/loss Q per trade or an investment is given as:

$$\begin{aligned} Q &= (P_t^A - P_{t+1}^A) - \gamma(P_t^B - P_{t+1}^B) \\ &= (P_t^A - \gamma P_t^B) - (P_{t+1}^A - \gamma P_{t+1}^B) \\ &= (\mu + \varepsilon_t) - (\mu + \varepsilon_{t+1}) \\ &= \varepsilon_t - \varepsilon_{t+1}. \end{aligned}$$

As such there are three possible outcomes depending upon the sign of the outcome $\epsilon_t - \epsilon_{t-1}$. For a positive $\epsilon_t - \epsilon_{t-1}$.or $Q > 0$ then a profit was made and on the contrary if it is negative means the pairs trade made a loss. A breakeven was attained for $\epsilon_t - \epsilon_{t-1}$.= 0.

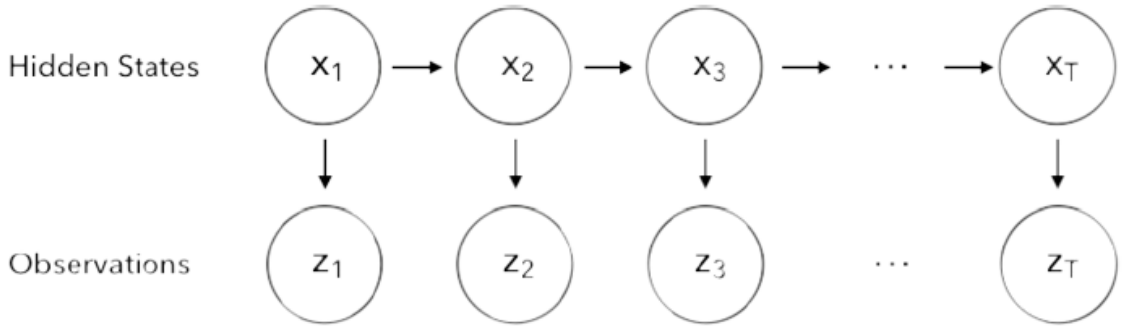
3. KALMAN FILTER

Noise in signal has always been an issue. The Kalman filter is a very popular engineering algorithm for processing noisy data (like market microstructure noise) and allows for an accurate representation of the

true signal (true state). It is essentially a state space model that infers information about the states, given the observation, as data arrives.
For pairs trading it is used to find a dynamic hedge ratio between assets. We shall briefly introduce its working and implementation using Python's $pykalman$ library.

3.1 *WORKING:*
The Kalman filter is particularly useful for rolling estimates of data values or model parameters that change over time. This is because it adapts its estimates at every time step based on new observations and tends to weigh recent observations more heavily. It gives us the flexibility to use a moving average instead of a rigid fixed size window.

Kalan filter therefore is a probabilistic model for a sequential observation $z_i, i \epsilon \{1,2 \dots, T\}$ and a corresponding sequence of hidden states $x_i, i \epsilon \{1,2 \dots, T\}$ represented as follows:



The algorithm iterates between two steps:

- Prediction Step: Estimate current state of the process
- Measurement Step: Use noisy data to update its estimate by giving more weightage to recent observation.

The basic idea behind the algorithm is as follows: certain assumptions about a dynamic system and a history of corresponding measurements will allow us to estimate the system's
state in a way that maximizes the probability of the previous measurements. To achieve the aim of recovering hidden states it assumes model behaves in linear fashion, hidden state is a Markov chain and measurements are subject to Gaussian uncorrelated noise. So essentially it is similar to a Markov chain, with hidden and observed variable having normal distributions.

For python implementation following are key components of our model:

- Initial Hidden State is normally distributed $x_0 \sim N(\mu, \sigma)$ with *initial_state_mean* as μ, and *initial_state_covariance* σ.
- The hidden state X_{t+1} is an affine transformation of $X_t$with *transition_matrix* A, *transition_offset* b, and added Gaussian noise with *transition_covariance* Q:
  $x_{t+1} = A_t x_t + b_t + \epsilon_{t+1}^1, \ \epsilon_t^1 \sim \mathcal{N}(0, Q)$

- The observation z_t is an affine transformation of the hidden state X_t with *observation_matrix* C, *observation_offset* d, and added Gaussian noise with *observation_covariance* R:

$$z_t = C_t x_t + d_t + \epsilon_t^2, \quad \epsilon_t^2 \sim \mathcal{N}(0, R).$$

For conducting a successful pairs trade, we use linear regression between two assets to find out the optimal hedging ratios between them, which in turn tells us how much of each asset to go long or short. The problem however arrives that, this hedging ratio is dynamic and must be readjusted according to the time period in the strategy. One solution is to use rolling linear regression with a "dynamic" lookback window. But this window introduces another free parameter that requires optimization. The proposed solution is to use a "State Space Model" that treats "true" hedge ratio as an unobserved hidden variable and attempts to estimate with a "noisy" observations.

4. PAIRS TRADING STRATEGY:

4.1 *IDENTIFICATION OF COMOVING ASSETS:*

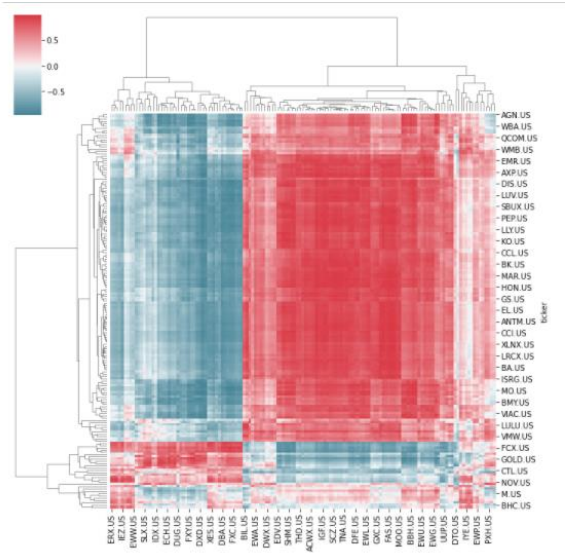
For this paper, we apply a distance based approach to first identify possible pairs for pairs trading. A distance based approach i.e. calculating a correlation between asset prices of their returns is used, as it is simpler and less computationally expensive than that of other statistical models. The speed advantage is particularly useful especially when there is a combination of say example 100 stocks and 100 ETFs that requires 10,000 tests. However, they may not accurately represent the most profitable pairs.

To balance this tradeoff between computational cost and quality of resulting pairs, Krauss (2017) states:

- Select pairs with a stable spread that shows little drift to reduce the number of candidates
- Test the remaining pairs with the highest spread variance for cointegration

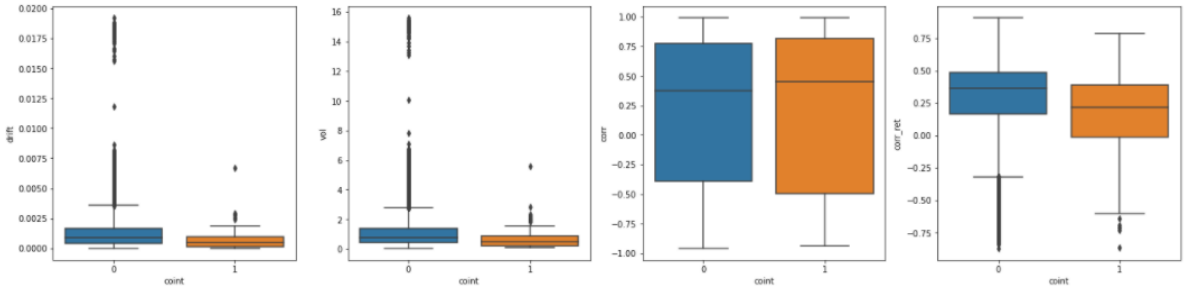
This process aims to select cointegrated pairs with lower divergence risk while ensuring more volatile spreads that, in turn, generate higher profit opportunities. However, including multiple samples for testing for cointegration may introduce data snooping bias, that may increase the number of false positives.

For our case we sample of 172 stocks and 138 ETFs traded on the NYSE and NASDAQ, with daily data from 2010 – 2019 provided by Stooq. These securities are largest average dollar volume over same period in their class. We begin by pre-processing the data and set a 0.99 cut-off for correlation so as to remove the highly correlated assets and remove stationary time series at a 95 percent confidence interval. The correlation cluster is visualized as follows:



We then go on to find the distance-based heuristics to recognize the cointegrated pairs between around 23,000 pairs of stocks and ETFs. Low volatility and drift (linear regression of time trend on spread) and high correlations between normalized price series and between their returns are simple proxies for cointegration. For each pair we evaluate Engle-Granger and Johansen cointegration tests. For both tests, we assume that the cointegrated series (the spread) may have an intercept different from zero but no trend. To check for the significance of the cointegration tests, we compare the Johansen trace statistic for rank 0 and 1 to their

respective critical values and obtain the Engle-Granger p-value. We apply and select the pairs where both test statistics agree. For the over 46,000 pairs across both sample periods, the Johansen test considers 3.2 percent of the relationships as significant, while the Engle-Granger considers 6.5 percent. They agree on 366 pairs (0.79 percent). The charts below show a comparison of heuristics for series that are cointegrated versus those that are not:

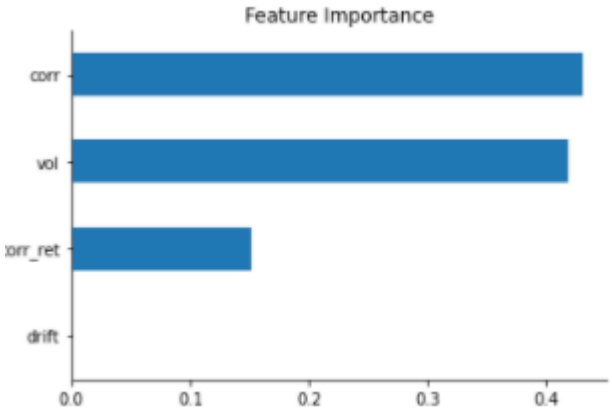


		25%	50%	75%	count	max	mean	min	std
	coint								
corr	0	-38.88%	37.10%	77.53%	2374700.00%	99.17%	20.61%	-95.88%	60.70%
	1	-49.28%	45.31%	81.63%	16100.00%	99.07%	20.13%	-93.50%	66.96%
drift	0	0.04%	0.09%	0.17%	2374700.00%	1.92%	0.13%	0.00%	0.18%
	1	0.01%	0.05%	0.10%	16100.00%	0.67%	0.07%	0.00%	0.08%
vol	0	46.08%	80.12%	139.56%	2374700.00%	1560.03%	117.39%	5.77%	145.64%
	1	24.79%	51.40%	87.97%	16100.00%	561.25%	70.22%	8.10%	68.08%

For evaluation of predictive accuracy of the heuristics we run logistic regression with these given features to predict significant cointegration and decision tree based classifier that gives us relative importance of these features.

AUC for Logistic Regression	AUC for Decision Tree Classifier
0.815	0.821

This shows that there are a large number of false positives: correctly identifying 80 percent of the 366 cointegrated pairs implies over 16,500 false positives, but eliminates almost 30,000 of the candidates and calls for more substantial testing to screen the large universe more effectively.



Using the above steps we find the top 10 instances of most common trading pairs. They were as follows:

	s1	s2	n	name1	name2
1352	T.US	VOX.US	6	AT&T	VANGUARD COMMUNICATION SERVICES ETF
384	FXF.US	MDLZ.US	5	INVESCO CURRENCYSHARES SWISS FRANC TRUST	MONDELEZ INT
388	FXF.US	NOV.US	5	INVESCO CURRENCYSHARES SWISS FRANC TRUST	NATIONAL OILWELL VARCO
391	FXF.US	RIG.US	5	INVESCO CURRENCYSHARES SWISS FRANC TRUST	TRANSOCEAN
532	AMJ.US	MDLZ.US	5	JPMORGAN ALERIAN MLP INDEX ETN	MONDELEZ INT
547	DIG.US	MDLZ.US	5	PROSHARES ULTRA OIL & GAS	MONDELEZ INT
549	DJP.US	MDLZ.US	5	IPATH BLOOMBERG COMMODITY INDEX TR ETN	MONDELEZ INT
571	ERX.US	MDLZ.US	5	DIREXION DAILY ENERGY BULL 2X SHARES	MONDELEZ INT
630	FXN.US	MDLZ.US	5	FIRST TRUST ENERGY ALPHADEX FUND	MONDELEZ INT
644	IYE.US	MDLZ.US	5	ISHARES US ENERGY ETF	MONDELEZ INT

4.2 *FINDING HEDGE RATIONS AND KALMAN FILTER APPLICATIONS:*

We now process to compute the spread for each candidate pair in the time period 2017-2019 using a rolling hedge ratio. We will also use Bollinger Band so as to identify spreads that go beyond two standard deviations from our moving average for our long short signals and its crossing of moving average in reverse as exit signals. For this we first smooth the prices and remove noise from the data using Kalman Filter and using rolling hedge ratio using Kalman filter to obtain a dynamic hedge ratio. The linear regression is simply a regression between pairs of assets and gives us the amount of asset to go long/short. The Python's *pykalman* library does most of the heavy lifting for us.

Modelling the spread as mean reverting stochastic process in continuous time we can model it as an *Ornstein-Uhlenbeck* process, which gives the half-life or the approximate time required for the spread to converge. Bollinger bands are then computed using the z-spread that captures deviations from moving average with a window equal to two half-lives in terms of rolling standard deviations. The z-scores give us the trading signals:

- If the z-score is **below** two we enter a **long position**, which implies spread has gone 2 standard deviations **below** the moving average (the asset return/price will **rise** mean revert back and above to long term average)
- If the z-score is **above** two we enter a **short position**, which implies spread has gone 2 standard deviations **above** the moving average (the asset return/price will **fall** mean revert back and below to long term average)
- Exit the trade when spread reaches the same value again.

Rules are derived on a quarterly basis that passed tests in lookback period, but allow pairs to exit during the subsequent 3 months. Dropping pairs not closing in the 6-month period will also be dropped. Here's sample of final result obtained:

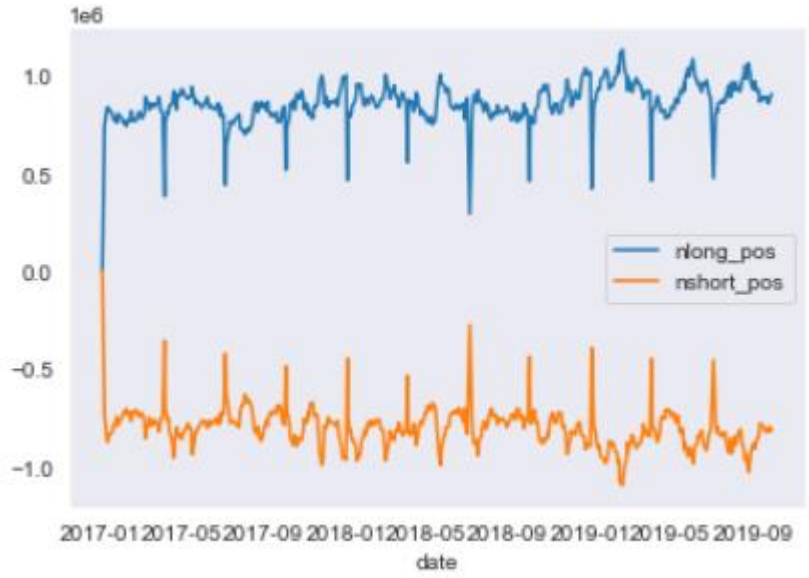
	s1	s2	hedge_ratio	period	pair	side
date						
2017-01-03	AA.US	ACWI.US	-0.533861	1	16	-1
2017-01-12	AA.US	ACWI.US	-0.533861	1	16	0
2017-01-03	AA.US	ACWX.US	-0.799916	1	54	-1
2017-01-12	AA.US	ACWX.US	-0.799916	1	54	0
2017-01-03	AA.US	DEM.US	-0.896395	1	376	-1

4.3 *BACKTESTING THE ALGORITHM AND EVALUATION:*

We are now ready to formulate our strategy on a back testing platform, execute it, and evaluate the results. To do so, we need to track our pairs, in addition to individual portfolio positions, and monitor the spread of active and inactive pairs to apply our trading rules. The evaluation of our strategy depends on:

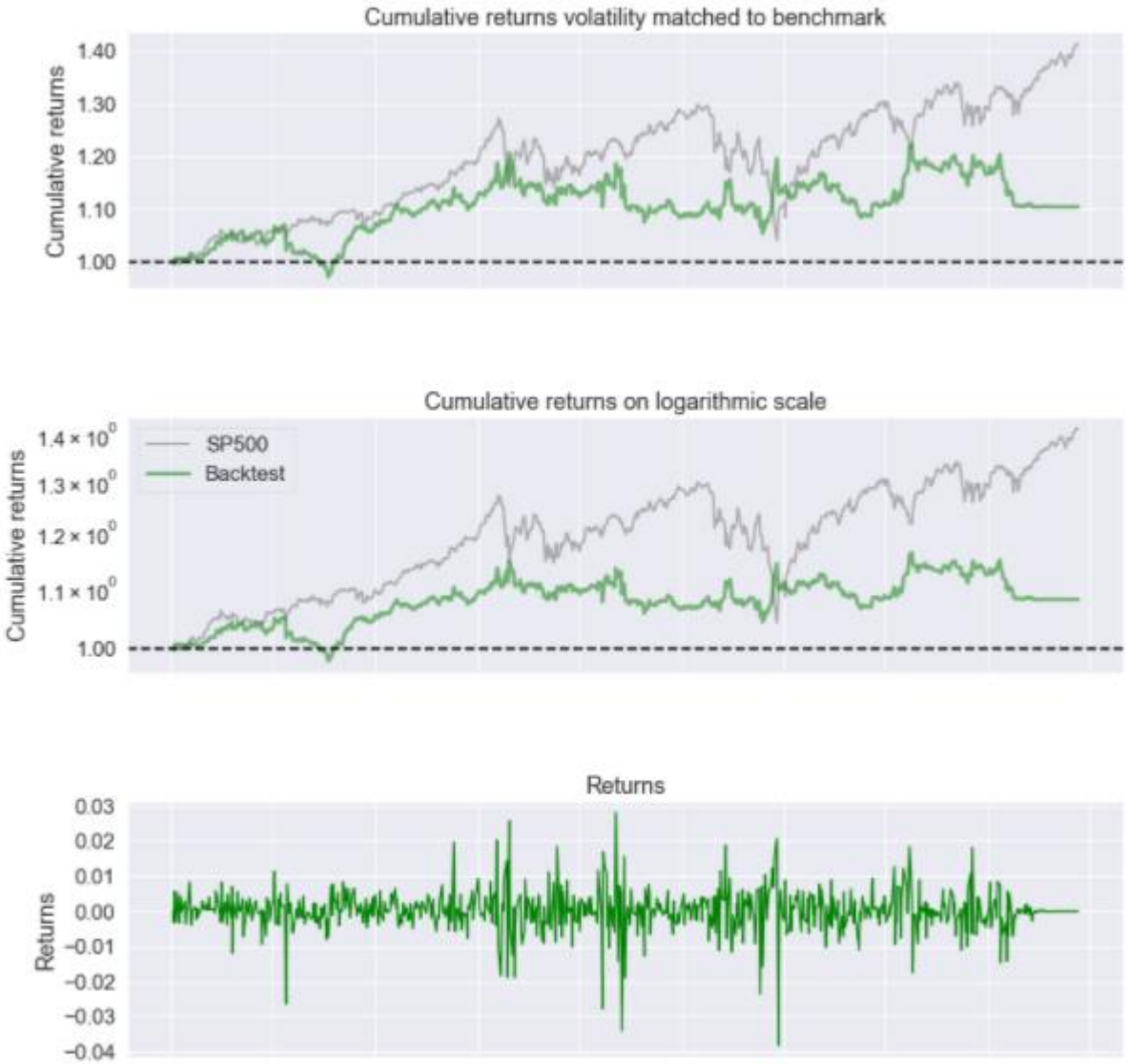
- Daily Exit from rules that have passed the negative return threshold.
- Daily Entry to new positions that have passed the entry signals
- Dynamic hedging for varying number of pairs.

We due an analysis of the portfolio, (pythons *PyFolio* library) and gives us the positions and metrics of our said portfolio. The results from back testing the strategy are as follows:

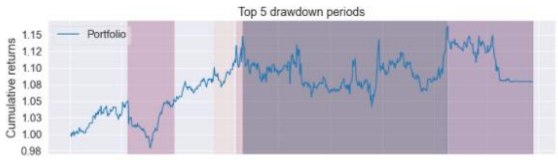


Portfolio Tear sheet for a period of 35 months from 3/1/2017 to 18/12/2019

Annual return	2.62%
Cumulative returns	7.96%
Annual volatility	9.30%
Sharpe ratio	0.32
Calmar ratio	0.29
Stability	0.51
Max drawdown	-9.18%
Omega ratio	1.06
Sortino ratio	0.45
Skew	-0.6
Kurtosis	7.11
Tail ratio	1.03
Daily value at risk	-1.16%
Gross leverage	0
Daily turnover	inf%
Alpha	0.06
Beta	-0.24

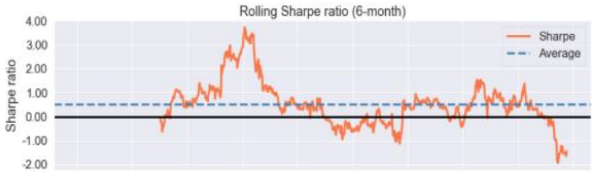

Cumulative returns volatility matched to benchmark
Cumulative returns
1.40
1.30
1.20
1.10
1.00
Cumulative returns on logarithmic scale
SP500
Backtest
Cumulative returns
1.4 × 10^0
1.3 × 10^0
1.2 × 10^0
1.1 × 10^0
1.00
Returns
Returns
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04


Rolling portfolio beta to SP500
Beta
6-mo
12-mo

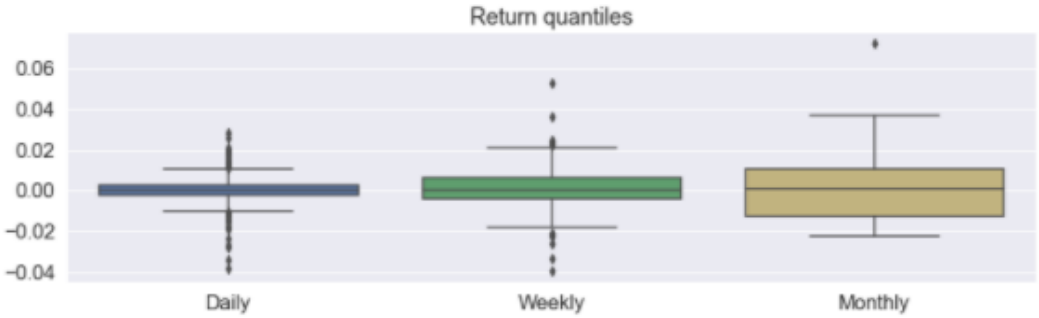

Top 5 drawdown periods
Cumulative returns
Portfolio

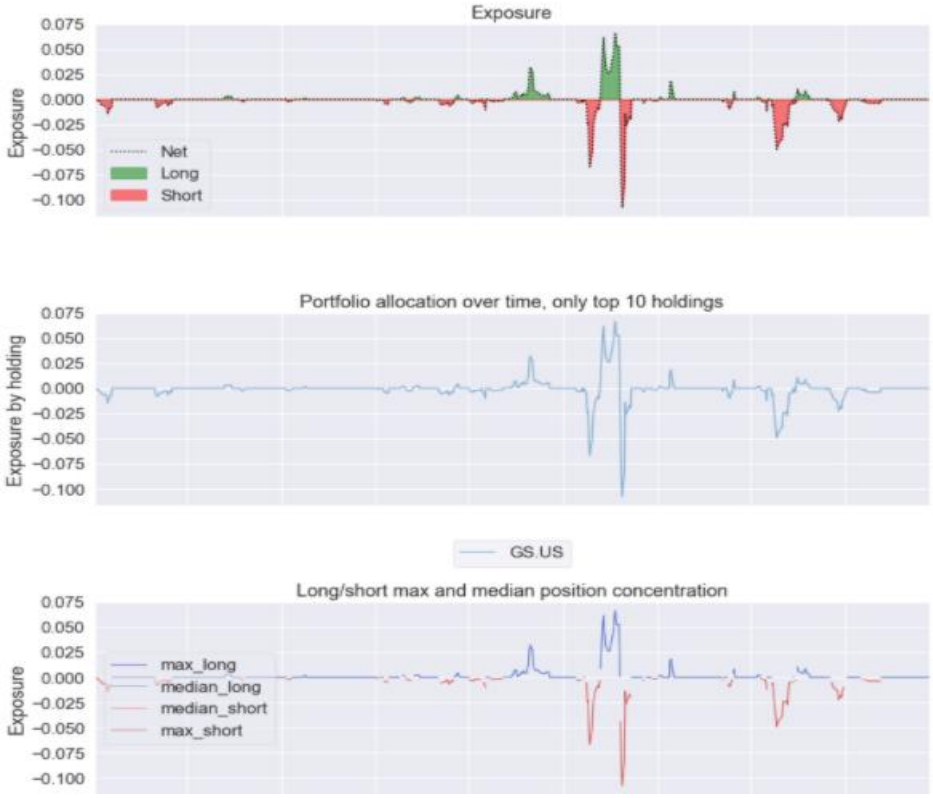

Rolling volatility (6-month)
Volatility
Volatility
Benchmark volatility
Average volatility


Underwater plot
Drawdown


Rolling Sharpe ratio (6-month)
Sharpe ratio
Sharpe
Average


Monthly returns (%)
Year
Month
Annual returns
Year
Mean
Returns
Distribution of monthly returns
Number of months
Mean
Returns


Return quantiles
Daily
Weekly
Monthly



5. CONCLUSION AND FUTURE WORK

The strategy that we developed in this paper demonstrate the anatomy of statistical arbitrage based on cointegration in the form of pairs trading, however, the performance metrics must be taken with the grain of salt. There are few ways to better the performance.

Cointegration is a very useful concept to identify pairs or groups of stocks that tend to move in unison. Compared to the statistical sophistication of cointegration, we used very simple and static trading rules; the computation on a quarterly basis also distorts the strategy, as the patterns of long and short holdings show. We need to screen a larger universe and optimize several parameters including and especially the trading rules. Also for risk management purposes we need to account for concentrated positions that arise when certain assets appear relatively often on same side of traded pair. The implementation can also be extended to basket of securities instead of just individual pairs. Finally, many advanced machine learning models can be used in place that can predict absolute size or the direction of price movements for given investment universe of horizon.

In this paper, a dynamic pairs trading system was developed using cointegration approach and Kalman filter. All of the assets that were used for the proposed pairs trading system belonged to the NYSE and NASDAQ. There were three essential steps in the whole trading process. First, we identified the pairs (time series prices data in particular) that showed cointegration. Highly correlated assets and stationary series were filtered out. The cointegration tests were applied systematically between the stock and ETF pairs and out of the universe of almost 46,000 pairs, a total of 366 pairs were selected. Then, smoothing

or rolling regression was applied to find out the hedge ratio (that tells us how much to go short/long between the pair) using a Kalman Filter. Post that, the entry and exit dates for long and short positions were identified, Bollinger bands to set the threshold values to two standard deviations was computed. Lastly, the strategy was tested using *backtrader* and the portfolio measurements obtained.

REFERENCES

- George J. Miao High Frequency and Dynamic Pairs Trading Based on Statistical Arbitrage Using a Two-Stage Correlation and Cointegration Approach
- Cartea & Penalva, 2012, Where is the value in high frequency trading?
- Gatev, Goetzmann, & Rouwenhorst, 2006, Pairs trading: Performance of a relative-value arbitrage rule. The Review of Financial Studies, 19(3), 797–827.
- Engle and Granger, Co-integration and error correction: Representation, estimation, and testing. Econometrica, 55(2), 251–276.
- Alexander, C. (1999). Optimal hedging using cointegration. Philosophical Transactions of the Royal Society, Series A, 357, 2039–2058. http://dx.doi.org/10.1098/rsta.1999.0416
- Avellaneda, M., & Lee, J. H. (2010). Statistical arbitrage in the U.S. equities market. Quantitative Finance, 10, 1–22. http://dx.doi.org/10.2139/ssrn.1153505
- Caldeira, J. F., & Moura, G. V. (2013). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. Revista Brasileira de Financas (Online), Rio de Janeiro, Brazil, 11(1), 49–80. http://dx.doi.org/10.2139/ssrn.2196391
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimates for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431. http://dx.doi.org/10.2307/2286348
- Perlin, M. S. (2009). Evaluation of Pairs-trading strategy at the Brazilian financial market. Journal of Derivatives & Hedge Funds, 15(2), 122–136. http://dx.doi.org/10.1057/jdhf.2009.4