Statistical Mean Reversion of Airline and Oil Company Stock Prices

CSC 265 Final Project

Dominick Harasimiuk - 30702462

Source Code Available At:
https://github.com/DominickH20/statistical-mean-reversion

May 7, 2021

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

The intent of this project is to build a reliable statistical mean reversion trading strategy on correlated baskets of stocks, specifically airline and oil company stocks. Airline companies are exposed to very similar pools of risk, and the same can be said for oil companies. Because of this, one would expect their stock prices to move together when these underlying risk factors change. Because airlines are large consumers of oil, one might expect the airline and oil sectors to exhibit some relationship as well.

1.1 Statistical Mean Reversion

This project aims to develop a model for how these companies behave in relation to one another. This model can be thought of as the "mean" in this mean reversion concept. I utilize this model in the context of a trading strategy by computing the expected stock return at time t, call it e_t. The expected return can be compared to the actual return, a_t at that same time period t. This generates a residual, $r_t = a_t - e_t$. If this residual is sufficiently positive, then the strategy concludes that the stock in question has grown too much over the period, so it would sell. Likewise, if the r_t is sufficiently negative, then the strategy judges that stock has underperformed over the period relative to its peers, so it would buy. In both cases the strategy is betting on convergence to the statistical relationship between the securities. In this paper, I outline such strategies on pairs of securities as well as baskets of securities, using principal component analysis to handle the multicollinearity problem for regression models implemented on such data.

1.2 Data

1.2.1 Scope

The data used in the subsequent analyses were obtained from the Alpaca Markets historical market data API. The securities examined in this study are the following: American Airlines (AAL), Delta Airlines (DAL), Southwest Airlines (LUV), Allegiant Airlines (ALGT), United Airlines (UAL), Chevron (CVX), Marathon Oil (MRO), Murphy Oil (MUR), Exxon Mobil (XOM), Devon Energy (DVN), SPDR S&P 500 ETF, and iShares Treasury Bond ETF (TLT). There are 5 airline companies, 5 oil companies, and 2 market factors (SPY and TLT). The motivation behind including the market factors in this study is that they help parameterize what is going on in equity and debt markets at any given point in time.

1.2.2 Data Format

The data range from the start of 2016 to May 1st 2021. The years 2020 and 2021 are held out as test data while the rest of the data is used to train the models. Given the occurrence of the pandemic in 2020, this is a particularly challenging test set. The data is organized in a *bar* format. Bars are a common way of aggregating financial data and are structured as follows:

$$B_S = \begin{bmatrix} \text{bar}_0, & \text{bar}_1, & ..., & \text{bar}_n \end{bmatrix}$$

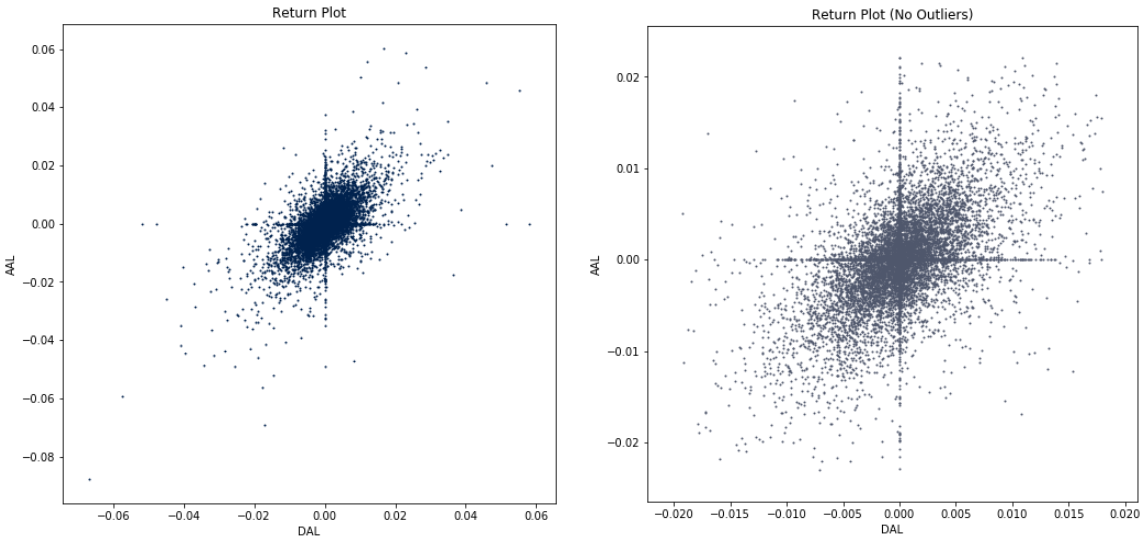
$$\text{bar}_i = \begin{bmatrix} \text{datetime}_i & \text{open}_i & \text{high}_i & \text{low}_i & \text{close}_i & \text{volume}_i \end{bmatrix}$$

Where S is some security, and B_S is the bar set for that security. Each bar summarizes a section of trading activity. The open and close are the first and last trades occurring within the timeframe of the bar, while the high and low are the highest and lowest trade prices observed within the timeframe of the bar. The volume represents the number of shares that traded within the bar. For this analysis, only the closing prices are used.

1.2.3 Data Cleaning

The data set was cleaned and imputed before use in any models. There were missing bars for several securities, so the most abundant time series (SPY) was used as a baseline for bar availability. For every bar in SPY, if there was not an analagous bar present in each other security, S, then the most recent bar in S was backfilled in its place.

There were also several outliers present in the data that would obstruct the fitting of linear models. In order for the models to be able to learn the usual relationship between the securities, I used the



(a) Returns with Outliers

(b) Returns with Outliers Removed

Figure 1: Impact of Outliers on Data

middle 99% of data *for training only*. The models were still evaluated on train and test data that included outliers. The outlier removal was solely done to improve the fit of the linear models.

1.3 Returns and Markouts

1.3.1 Returns

Asset returns represent a scaled first difference in the data, so they eliminate the serial component in the time series data along with the complications associated with this. The actual return as mentioned above can be defined as follows:

$$a_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Normal approximations have been used for asset returns on the monthly time scale, however, over short intervals this assumption breaks down. The hourly returns used in this study are abundant close to 0 and extreme returns occur with greater frequency than under normality assumptions. Despite this fact, OLS is a *BLUE* estimator, so we can still confidently estimate a line despite the distribution not being normal.

1.3.2 Markouts

Markouts are a way of computing hypothetical trade profits or losses. At every point along the time series, we hypothetically enter a trade and hold for some predefined number of time periods, after which we exit the trade. Note that the entry can be a selling (short selling) or buying trade, while

the exit is simply the opposite side of the entry transaction. To specify this more rigorously, we can write:

$$M_{t,k} = \begin{cases} P_{t+k} - P_t \text{ if Buy} \\ P_t - P_{t+k} \text{ if Sell} \end{cases}$$

This is the markout at time t, held for k periods. Notice that $M^B_{t,k} = -M^S_{t,k}$, meaning that the profit or loss of a buying and selling trade at a given point in time are exact opposites (since the asset can only move one direction). A good trading signal is one that is able to separate positive from negative markouts.

2 Pairwise Mean Reversion

As mentioned in the introduction, airline companies are exposed to similar pools of risk. Oil companies are also exposed to similar pools of risk. Airline and oil companies are both exposed to the price risk of oil itself, so it would stand to reason that the stock prices of companies in this sector would move together. This assumption underpins the pairwise mean reversion models presented in this section.

2.1 Pairwise Correlations

To get an idea of the relationships present among the companies in the data set, a pairwise correlation analysis was conducted on the hourly returns of each of the securities. We can see from Figure 1 above

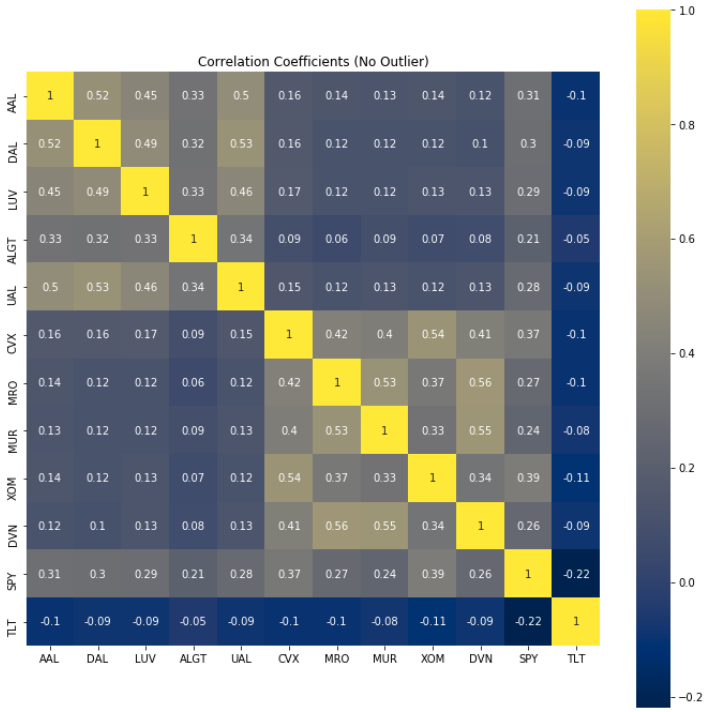


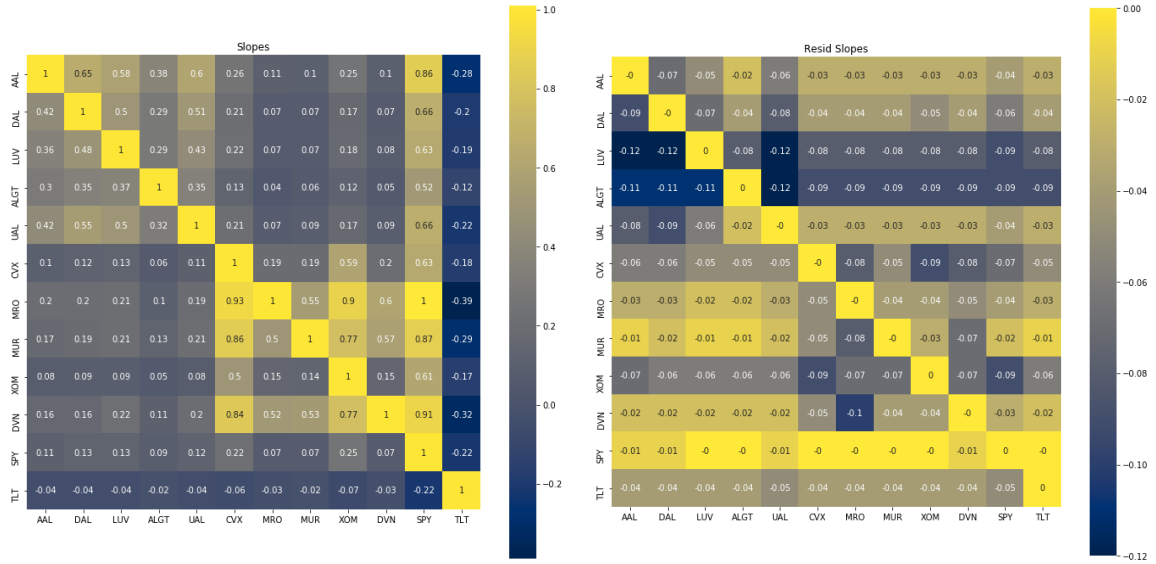
Figure 2: Pairwise Correlation - Outliers Removed

and section 6.3 in the appendix that the outliers in the data set had a noticeable positive impact on the correlations between the securities, so for training purposes, it was appropriate to remove them. There are clear sections of correlatedness in Figure 2 above. We can see that the 5 airline companies are together correlated, while the same can be said for the 5 oil companies in the data set. There is

also some positive correlation between the companies in the data set and the S&P 500 as well as some negative correlation with price of bonds.

2.2 Pairwise OLS

With this understanding of the correlations between the data as well as the fact that the outlier free return data exhibits ellipsoidal linear relationships, it became clear that OLS would be a good model for this data. OLS was able to fit lines to the data well, however, because the data fail to satisfy some assumptions of OLS, inference on the OLS parameters was not possible. The coefficients of the pairwise OLS model can be found in Figure 3(a). We see that the the strength of the slope coefficients



(a) Model Slope Coefficient

(b) Residual Slope Coefficient

Figure 3: Pairwise OLS

lines up well with the strength of the pairwise correlation between the assets.

Include regression plot, coefficient plots

2.3 Profit and Loss Analysis

Introduce cumulative markout sums and parameter tuning. include the PNl curves, and some pnl heatmaps comment on allegiant

3 Principal Component Basket Mean Reversion

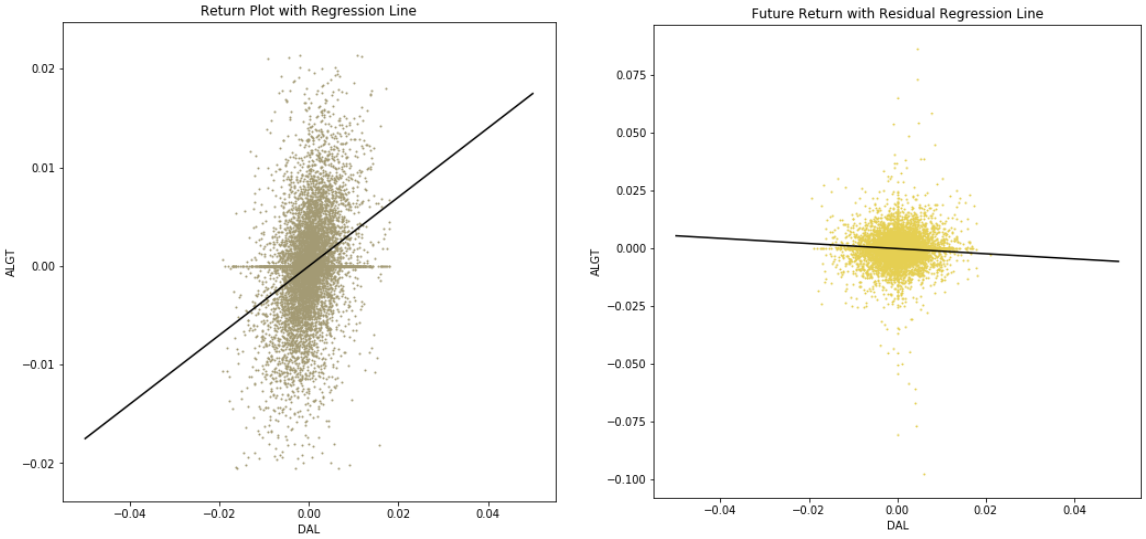
overview of goal and intuition here

3.1 Dimensionality Reduction

variance explained - show plots of first 4 PCs correlation with returns anomalies with coefficients

3.2 Basket OLS

Describe fit and any coefficient plots here describe feature selection



(a) OLS Fit on No Outlier Data

(b) Future Returns Regressed on Model residuals

Figure 4: Return Plots with Regression

3.3 Basket Lasso for Feature Selection

Omit this - mention briefly

3.4 Profit and Loss Analysis

plots of some pnls, also buying and selling matrices Discuss why there is shortfall

4 Discussion

discuss shortfall of full strategy and why this might have occurred dangers of assuming things revert to mean, mention correlation breakdowns during covid mention statistical issues with fit - high Dimensionality and lasso doing a poor job of feature selection because there are so many data points close to the 0,0

5 Conclusion

highlight what was accomplished and what takeaways and next steps are

6 Appendix

6.1 RMSE for PC Regression Models

Table 1: Airline PC Regression RMSE

Ticker	PC1	Full	Lasso
AAL	0.003519	0.003487	0.003487
DAL	0.002762	0.002725	0.002725
LUV	0.002839	0.002819	0.002819
ALGT	0.003561	0.003554	0.003554
UAL	0.002931	0.002902	0.002902

Table 2: Oil PC Regression RMSE

Ticker	PC1	Full	Lasso
CVX	0.002338	0.002136	0.002136
MRO	0.004708	0.004647	0.004647
MUR	0.004573	0.004536	0.004536
XOM	0.002240	0.002047	0.002047
DVN	0.004302	0.004281	0.004284

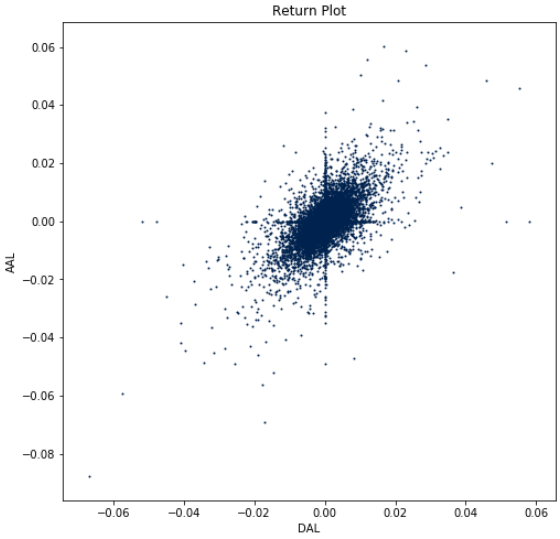
Table 3: Airline and Oil PC Regression RMSE

Ticker	PC1	Full	Lasso
AAL	0.004306	0.003475	0.003478
DAL	0.003450	0.002721	0.002722
LUV	0.003363	0.002809	0.002809
ALGT	0.003865	0.003552	0.003553
UAL	0.003602	0.002896	0.002896
CVX	0.002320	0.002124	0.002125
MRO	0.004851	0.004640	0.004641
MUR	0.004661	0.004530	0.004530
XOM	0.002231	0.002044	0.002044
DVN	0.004414	0.004275	0.004278

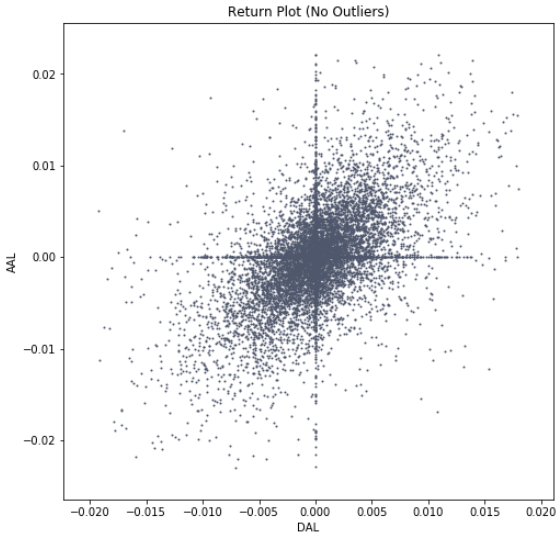
Table 4: Full PC Regression RMSE

Ticker	PC1	Full	Lasso
AAL	0.004301	0.003458	0.003458
DAL	0.003446	0.002713	0.002713
LUV	0.003359	0.002801	0.002801
ALGT	0.003863	0.003543	0.003545
UAL	0.003598	0.002891	0.002891
CVX	0.002316	0.002104	0.002105
MRO	0.004854	0.004634	0.004635
MUR	0.004666	0.004530	0.004531
XOM	0.002227	0.001996	0.001996
DVN	0.004418	0.004271	0.004275
SPY	0.001469	0.001329	0.001330
TLT	0.001574	0.001546	0.001546

6.2 Return Plots

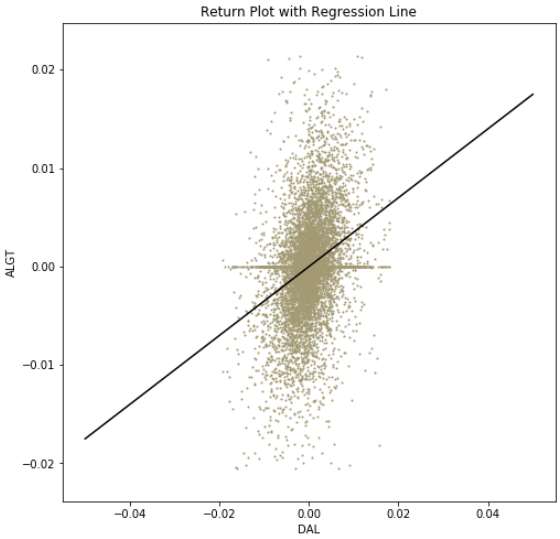


(a) Returns with Outliers

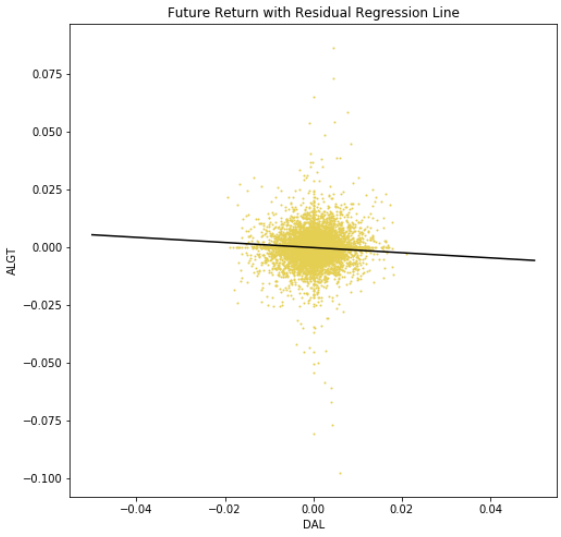


(b) Returns with Outliers Removed

Figure 5: Impact of Outliers on Data



(a) OLS Fit on No Outlier Data



(b) Future Returns Regressed on Model residuals

Figure 6: Return Plots with Regression

6.3 Correlations

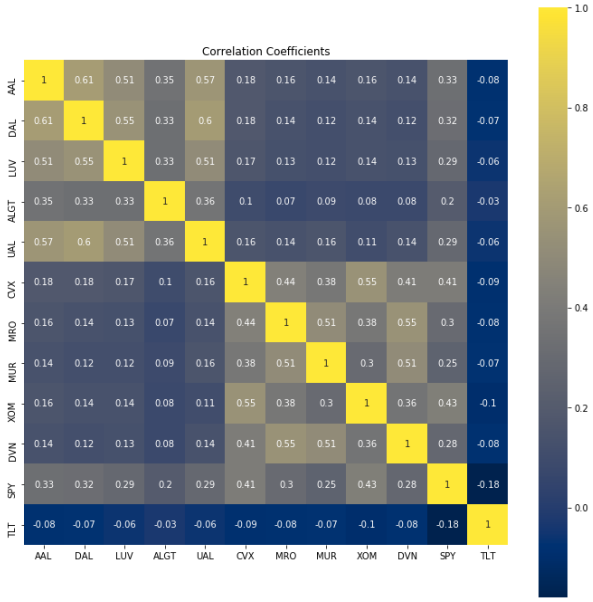


Figure 7: Pairwise Correlation - Outliers Present

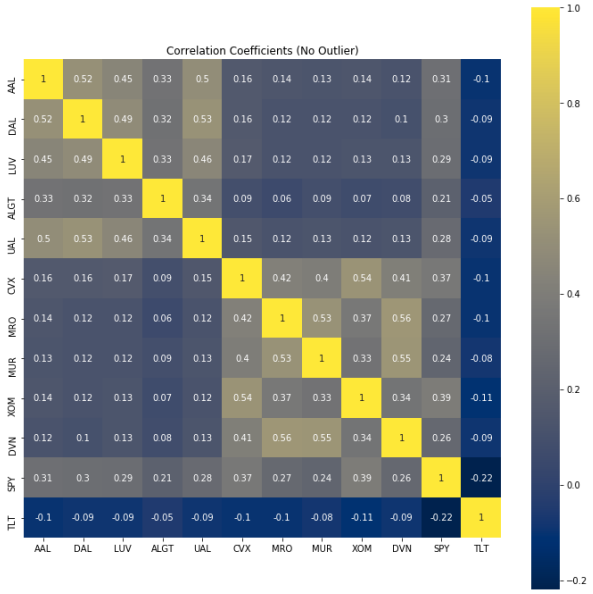


Figure 8: Pairwise Correlation - Outliers Removed

6.4 Principal Component Analysis

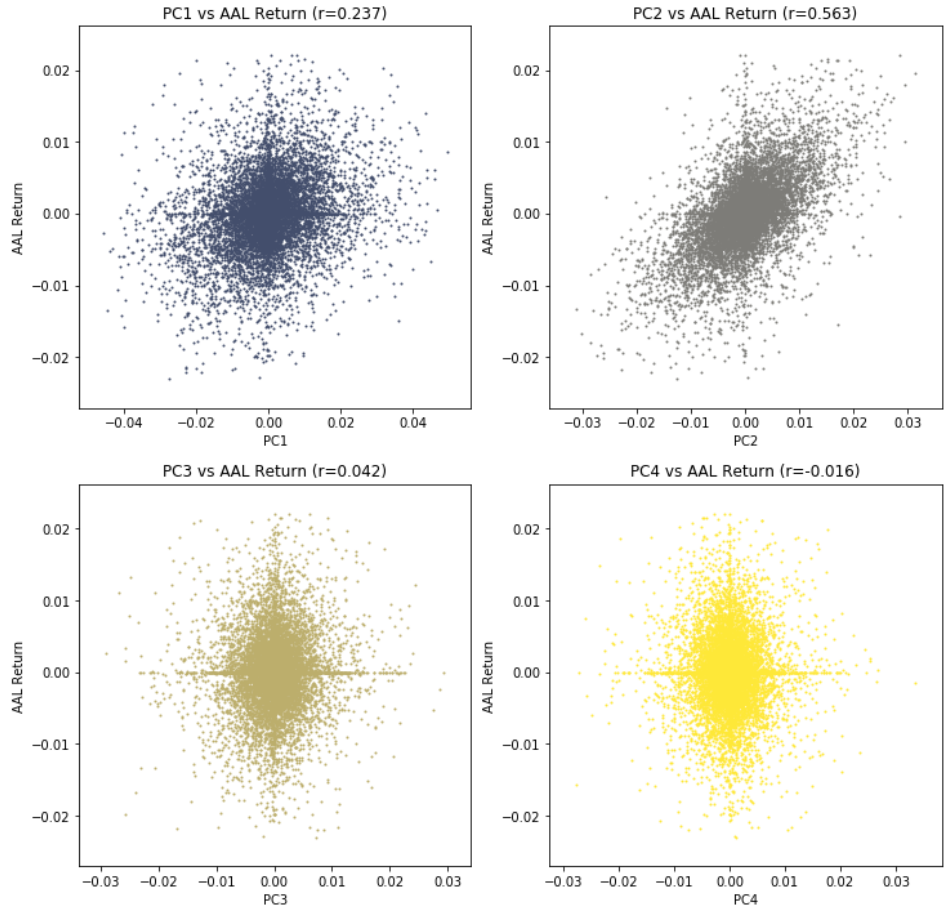
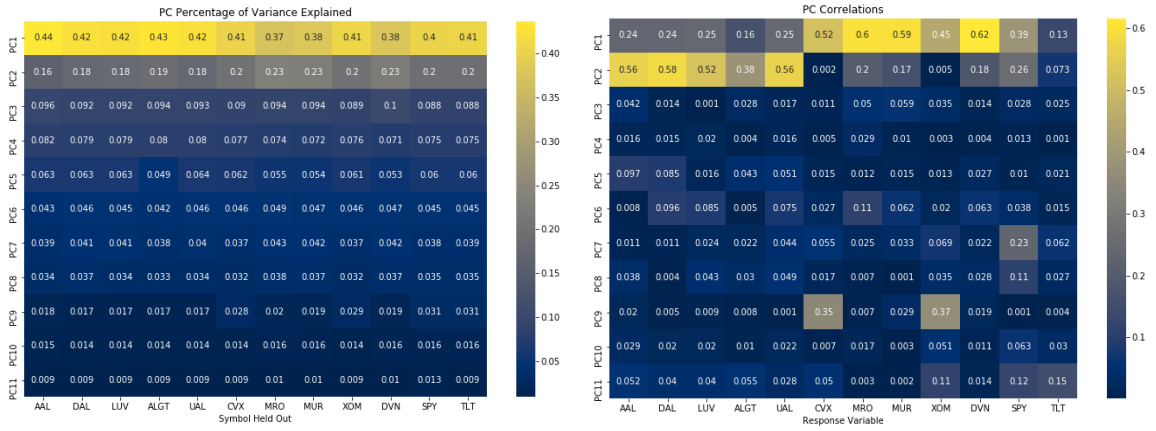


Figure 9: Decomposition Plot



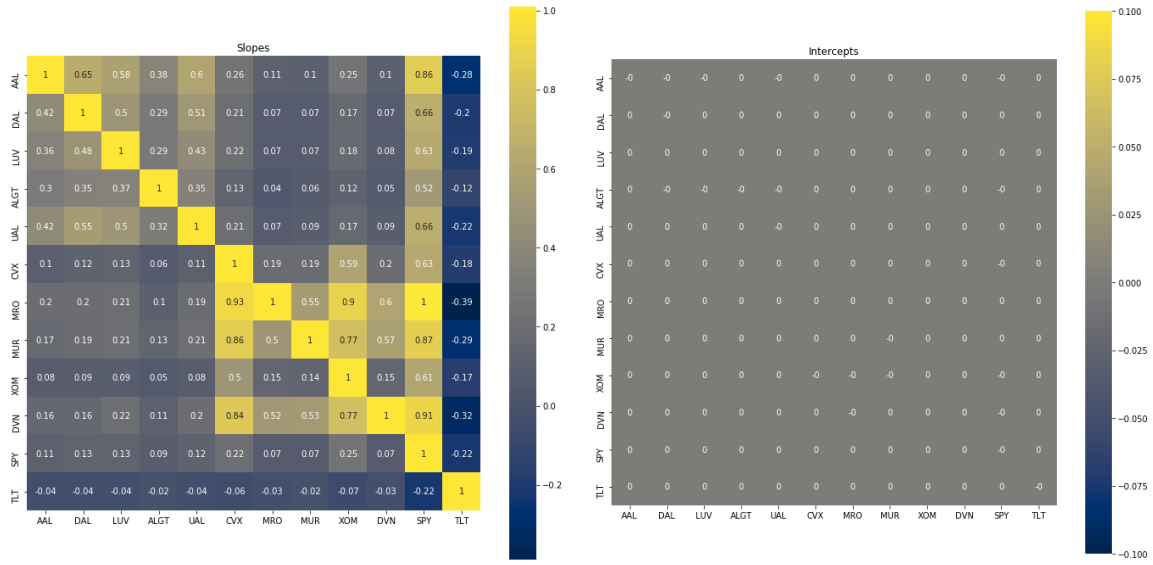
(a) PCA Percentage of Variance Explained

(b) Principal Component Correlation with Response

Figure 10: PCA Metrics

6.5 Regression Coefficients

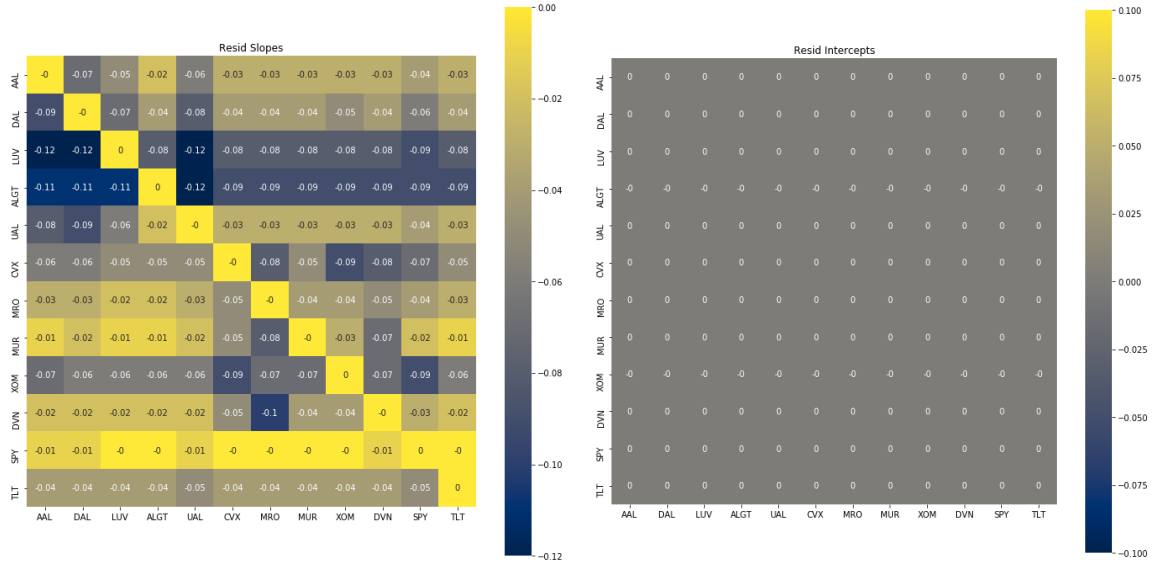
6.5.1 Pairwise Regression



(a) Slope Coefficient (b) Intercept

Figure 11: Pairwise Regression

6.5.2 Pairwise Residual Regression



(a) Slope Coefficient (b) Intercept

Figure 12: Pairwise Residual Regression

6.5.3 Principal Component Regression

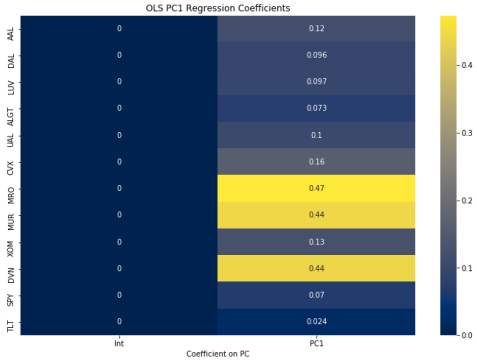


Figure 13: PC1 OLS Regression

Figure 14: Full PC OLS Regression

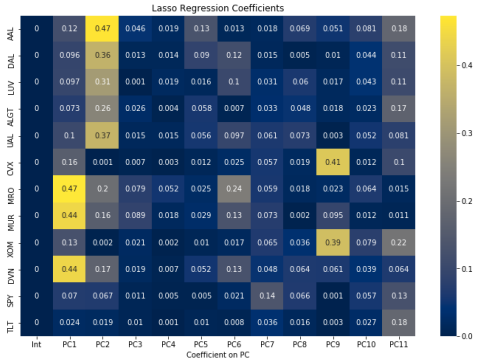
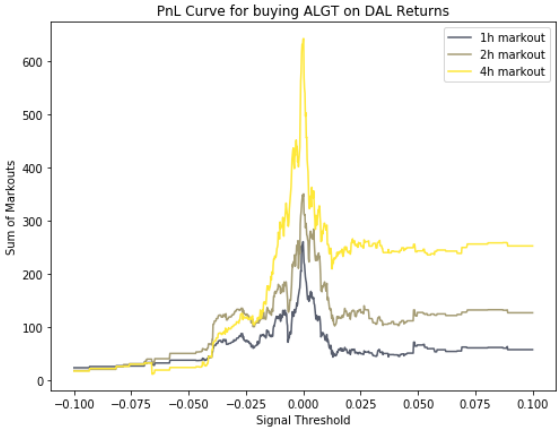
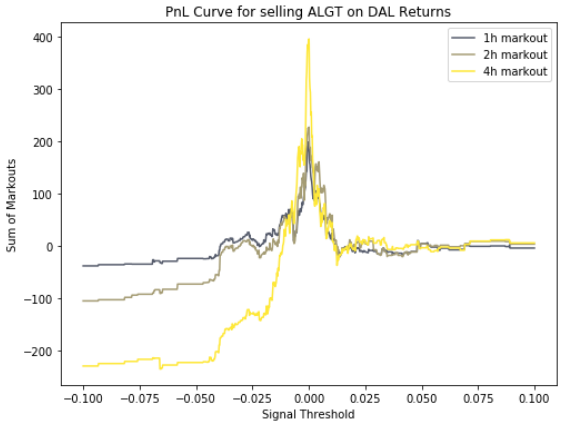


Figure 15: Lasso PC Regression

6.6 Profit and Loss Signal Curves

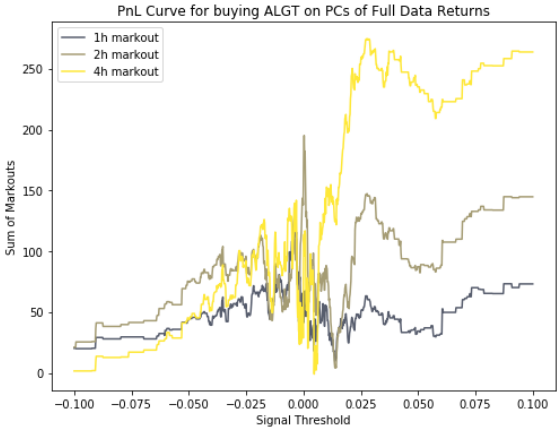


(a) Buy Side PnL

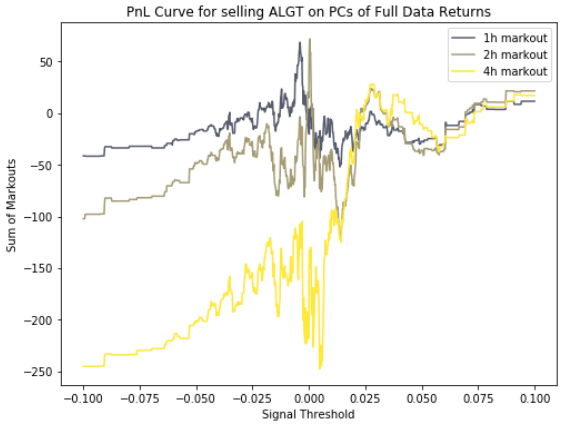


(b) Sell Side PnL

Figure 16: Pairwise Pnl Curves - ALGT



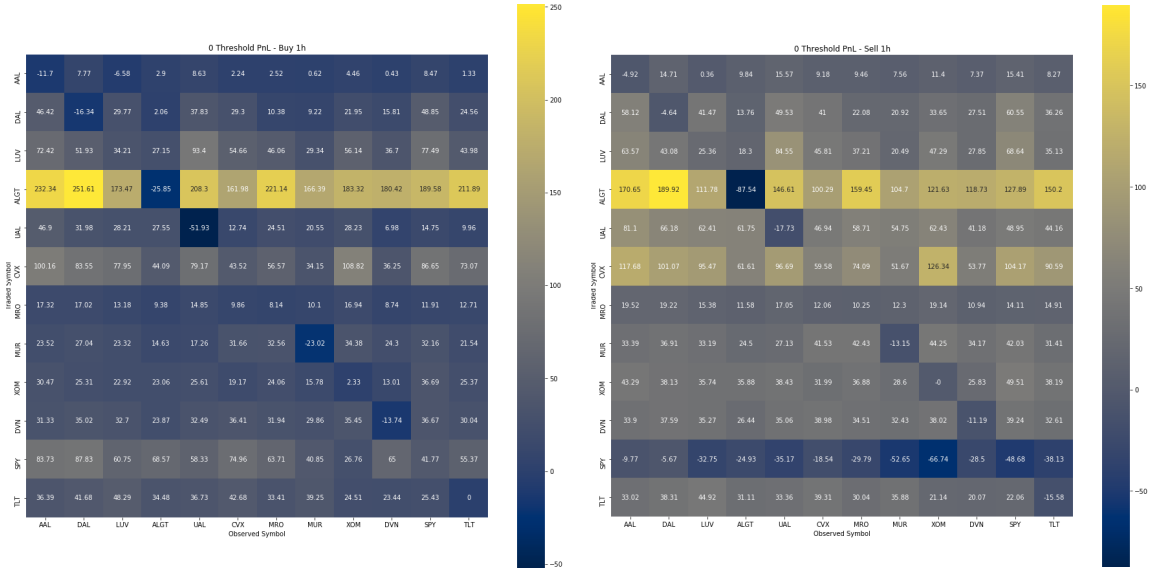
(a) Buy Side PnL



(b) Sell Side PnL

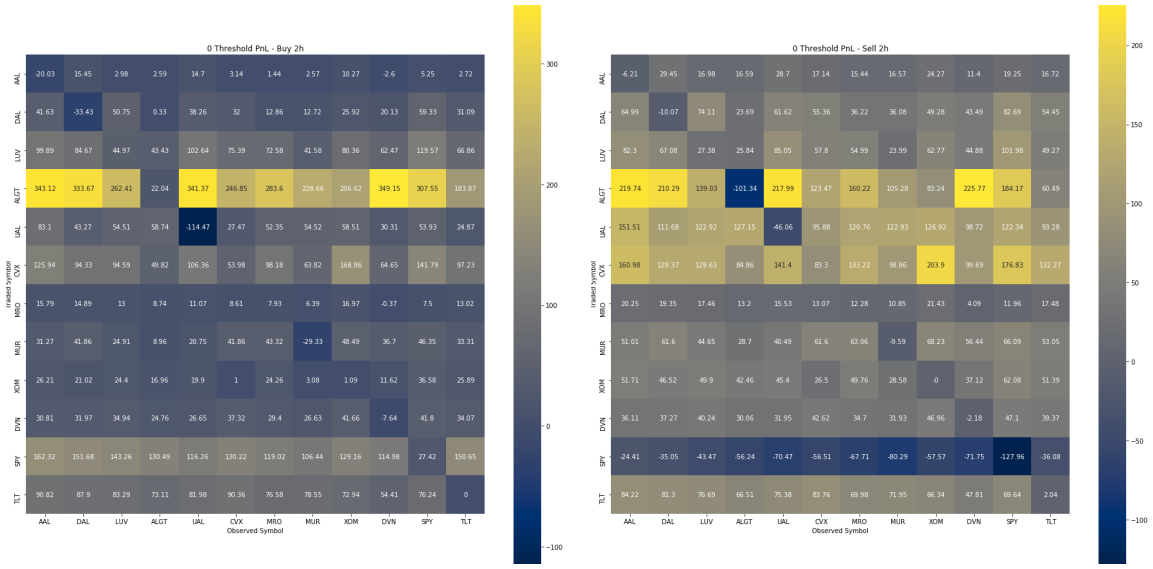
Figure 17: Basket Pnl Curves - ALGT

6.7 Profit and Loss over Markout Horizons



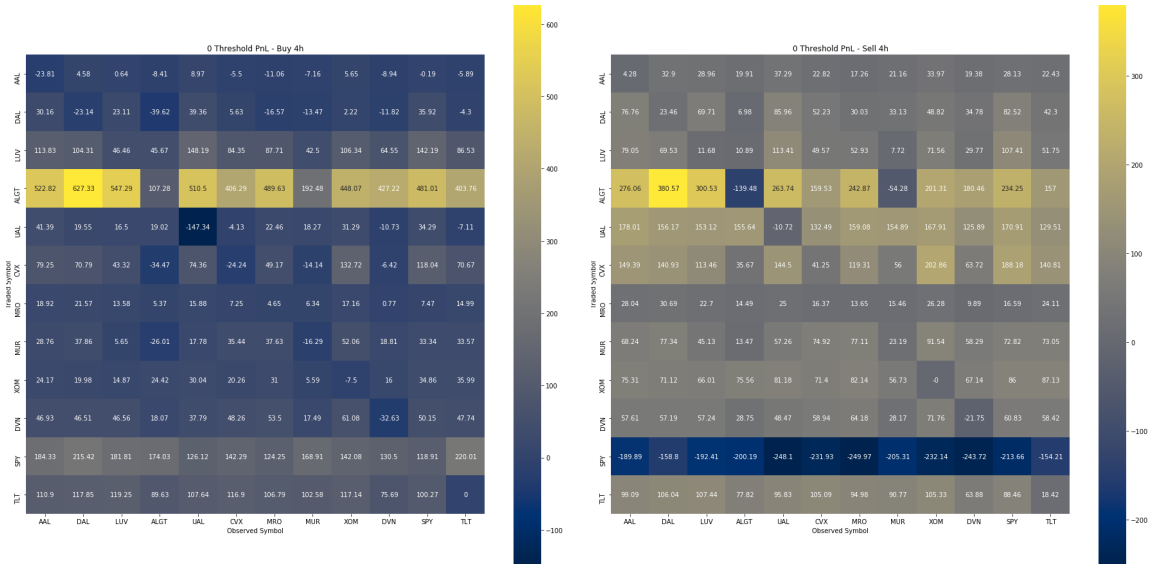
(a) Buy Side PnL (b) Sell Side PnL

Figure 18: 1-Hour Markouts



(a) Buy Side PnL (b) Sell Side PnL

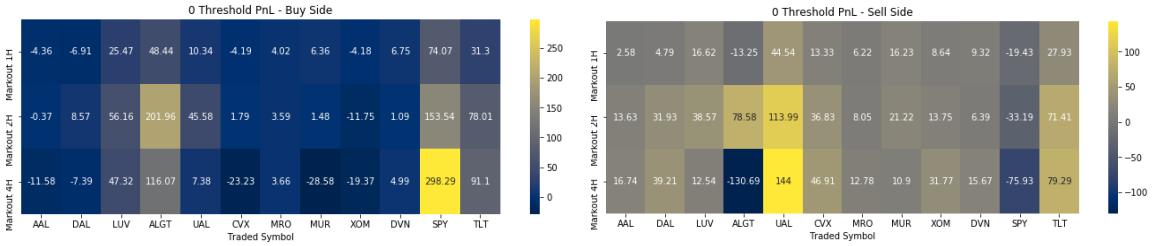
Figure 19: 2-Hour Markouts



(a) Buy Side PnL

(b) Sell Side PnL

Figure 20: 4-Hour Markouts



(a) Buy Side PnL

(b) Sell Side PnL

Figure 21: Basket Model PnL