Statistical Arbitrage on US Equity

Kunal Rajani (kr386@cornell.edu)

Nov 10, 2012

Abstract

The objective of this analysis is to obtain an alpha that is based on statistically exploring inefficiencies in stock prices. The strategy involves decomposing stock prices in each industry into principal components that explain the most variance and then regress the stock prices on those components to obtain the stock's dependence on them. Next the components are forecasted using GARCH model and hence the forecasted evolution of the stocks is also obtained based on the regression results. Based on these forecasts I will create a long-short neutral arbitrage strategy with the aim of achieving high risk adjusted returns.

1 The Idea

I obtained this idea while I was working on testing cointegration to create an arbitrage opportunity. A big universe of stocks makes it computationaly intensive and so I was trying to think of another strategy that aims to exploit divergence of stock prices from their expectation and trades to close that gap. Instead of cointegration which regresses two pairs of stocks I thought of another statistical technique to obtain stock price expectations and then trade the gap between them and the actuals.

PCA is a technique that works to extract the explainatory structure of the variance between a group of variable. Furthermore, the components are orthogonal and hence perfectly fit the description of the criteria of independent variables for regression. I used these as my building blocks of the analysis: create an expectation of the stock prices based on certain decomposed components that explain the maximum variance. It was equally rewarding in terms of computation as now only the principal components will need to be computed and forecasted which requires quite a bit lower effort as compared to comparing all pairs of stocks to find cointegrated ones. Now that I'll be able to obtain the gap between the expected price and the actual price I aim to bet against that gap and create an arbitrage.

2 The Analysis

2.1 Framework

Since my analysis is statistically based, I plan to reset my models over a horizon of 10 days. This means that every 10 days I will recalibrate the decomposition of stocks, the regression and the GARCH models to forecast the future prices so any new information can be incorporated as it becomes available. I also choose the stocks that are likely to have a high liquidity based on their price and the volume traded. Based on this rule I get rid of low priced and low volume stocks. Since I expect the gap between predicted and actual values to close within this horizon I am going to deal in low volatility stocks that are not expected to diverge from expectations. Next, I conduct the analysis on the prices of the stocks rather than the returns since I will be making my bets on the convergence of prices. To

also have a balanced and industry neutral approach I will create trades for each industry separately based on the analysis. Having choosen the stocks universe and my horizon I proceed below.

2.2 Principal Components Analysis

Once I obtain a group of stocks with the minimum variance I standardize them to ensure that highly priced stocks don't influence the decomposition. Then I run an eigen value decomposition on their covariance matrix to obtain a set of vectors that are able to explain 90% of the variance in those stocks. Since I take the stocks with low variance I am able to get a small number of components to explain as much variance.

2.3 Regression

Now that I have obtained the principal vectors I run a regression on the stock prices onto the components obtained based on the training period. Note that all these vectors are standardized and furthermore the components are orthogonal and hence perfectly match the criteria of linear regression. After the regression I have obtained the coefficients that highlights how the stock prices depend on the components. All the regressions contained some significance and based on a random sample the average R-square was over 85%. Although this is not quite a surprise since the components are a linear combination of the stocks it helps strengthen the belief in the theory.

2.4 GARCH Forecasting

At this point I have the components that explain the variance among a group of stocks and how the actual stock prices depend on those components. To forecast the evolution of those stock prices I forecast the principal components using GARCH(1,1)/ARMA(1,1) models. This is done by first fitting the model to the components lookback period of 300 days after which the specifications are used to forecast for the next horizon which is 25 days. This is the process that takes the most amount of time.

2.5 Taking positions

Once the principal components are forecasted I use the results of the regression to forecast the stock prices themselves, after destandardizing them using their mean and standard deviation. If the actual value of the stock is above a certain spread from the forecasted I short the stock and if it is below a certain spread I long the stock. This process is repeated every day and the stock positions changed depending on whether they are above or below their forecasts.

2.6 Balancing

After the above procedure is carried out we would have obtained the positions of the lowest volatility stocks in each industry. To be market neutral and balance out our exposures I squared off our positions in each industry with the stocks of that particular industry that made the portfolio industry neutral and hence market neutral. This helps us get rid of the exposure to systemic risk and hence focus our efforts into the arbitrage trade that we seek to profit from.

3 The Results and Conclusions

To briefly illustrate the entire approach we can look at this flow: Industry classification -> Select low volatility stocks -> standardize to bring all prices to same level -> perform principal components analysis -> regress stock

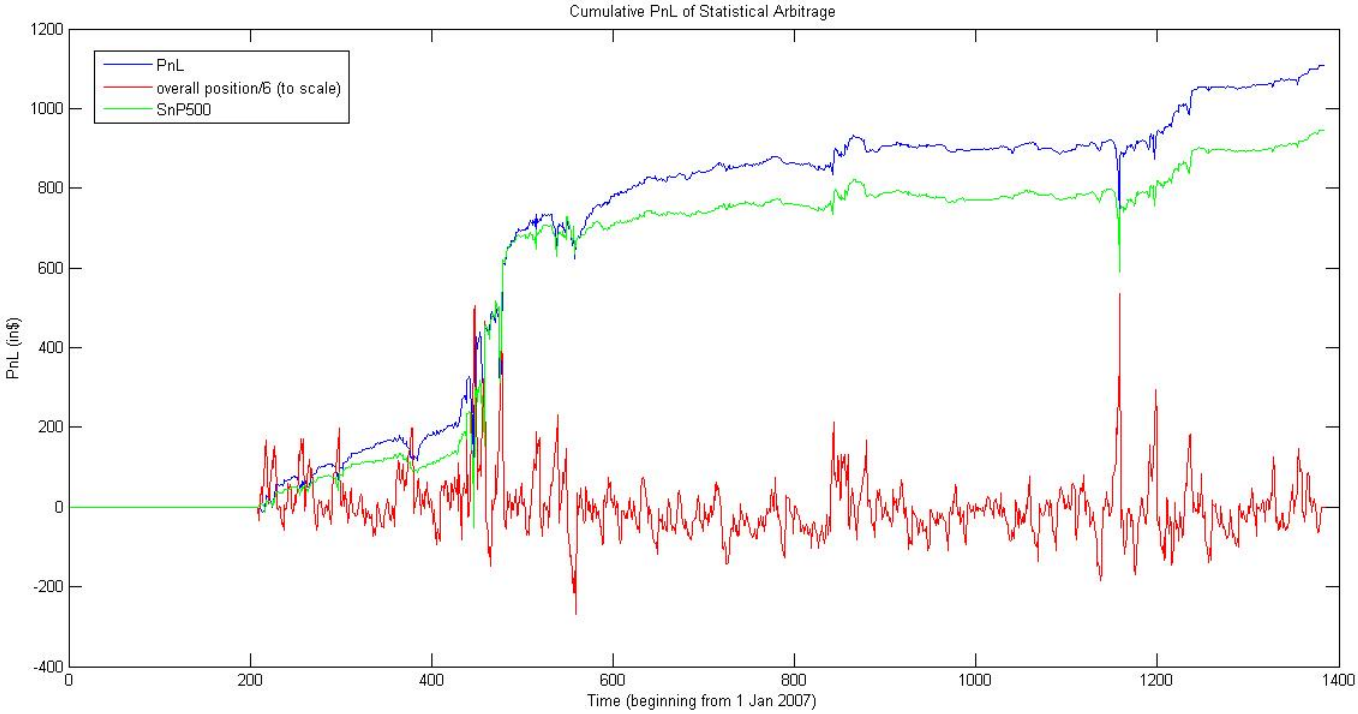

Figure 1: Evolution of PnL of the Stat. Arb. trades

prices on these components -> forecast the components based on GARCH modeling -> predict the stock prices based on these forecasts and regression results -> destandardize -> obtain the difference between these expectations and actual stock price -> trade on the gap to close down.

When the above analysis is run on the given stocks universe it generates the following PnL. We can see how the statistical approach was able to extract the inefficiencies in the stock prices and trade on them. We also see that our positions fluctuated around zero as the trading strategy took long-short positions in different stocks. This strategy generated a return of over 50% from 2007-2012 at a Sharpe Ratio of 4.6%. We can also see that the statistical arbitrage portfolio beat the SnP500 index consistently throughout the investment period.

In the next figure we notice how the overall position is zero. This is because it is a balanced portfolio without any market/industry exposure. We see how the PnL, though lower than the Stat Arb trades, is consistently increasing and then increases quite slightly from 2010 onwards. The balanced strategy generated fairly modest returns of 10% over 2007-2012 with a Sharpe Ratio of 4.1%.

4 Further Work

We see that the results of the statistical arbitrage don't seem to be robust when looking at the Sharpe Ratio. Although they do provide a high return a low Sharpe Ratio denotes the amount of instability and risk in achieving

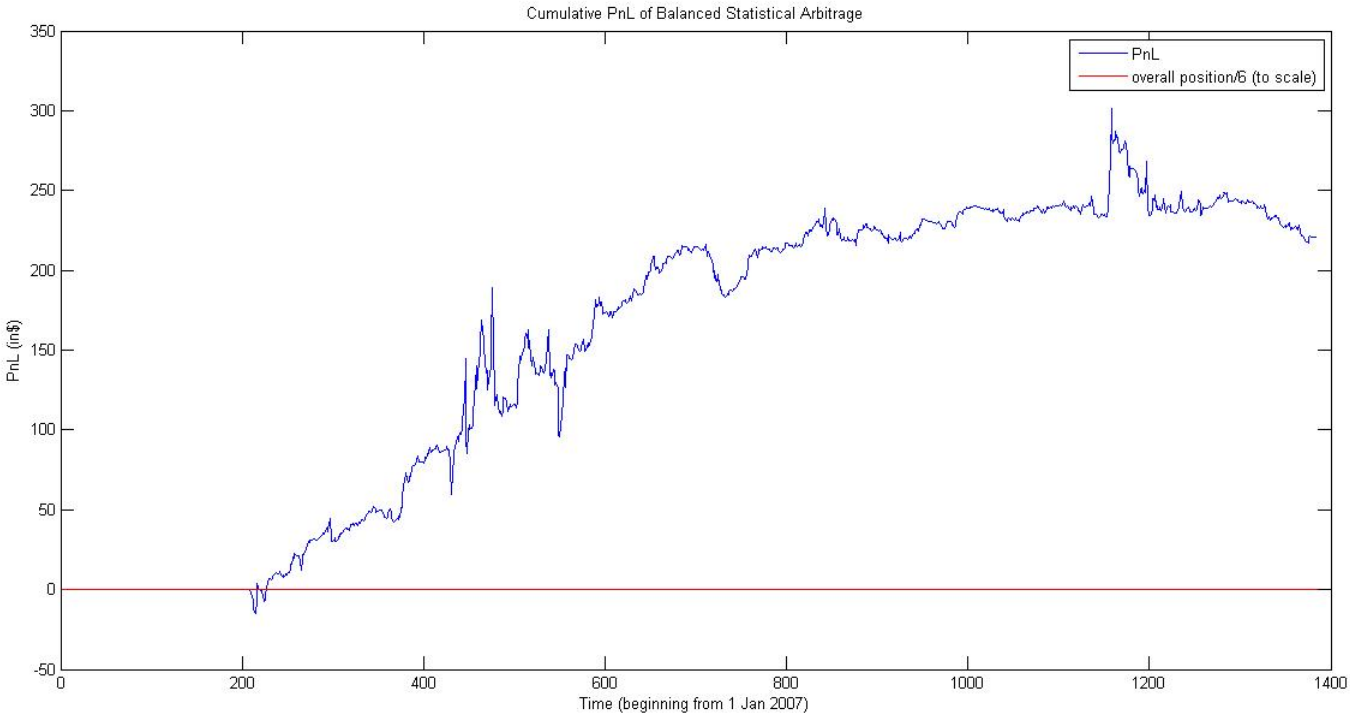

Figure 2: Evolution of PnL of Balanced Strategy

those returns. To make the strategy more robust I would like to look at the fundamental variables to be able to group the stocks together. For ex. calibrating our strategy based on high growth or high profit margin stocks can help us avoiding shorting those stocks but short those that have a high debt to equity rations and exhibit a lower liquidity.

Nevertheless, the Statistical Arbitrage discussed above shows that it is possible to extract inefficiencies in the markets using statistical techniques by simply using the price information of the stocks.