# ORIE 4741 Course Project Midterm Report
# Application of Machine Learning Techniques in Statistical Arbitrage

**Zicheng Men, Sheng Zhang, Shan He**

October 27 2016

## 1 Project Framework

The purpose of this project is to develop a trading strategy that tracks arbitrage opportunities in a stock index. The net present value (NPV) of a stock index is computed from the weighted average of the prices of selected stocks. On the other hand, the actual trading price of the ETF (exchange traded fund) of this index is determined by the relationship between supply and demand of the market. Therefore the discrepancy between its NPV and actual trading price creates arbitrage opportunities. It is commonly practiced to use constituents of an index to predict the index fluctuation. According to empirical studies, the fluctuation of each stock price can be decomposed into systematic fluctuation, which is due to the momentum of the industry, and idiosyncratic fluctuation, which is irrelevent of the industry.

We will employ linear regression and support vector regression onto the constituents of the index. In order to prevent overfitting, we use Principal Component Analysis (PCA) to reduce the dimensions of the feature space.

We want to model the price of the index such that it accounts for a drift, which measures systematic deviations from the market, and a price fluctuation that is a mean-reverting Ornstein Uhlenbeck process to the overall industry level. We represent the idiosyncratic fluctuation by the generalization errors from regression on principal components to generate trading signals.

## 2 Data Selection

S&P 100 index and the adjusted close prices of its constituents are our primary data in this project. We download the data from Yahoo Finance using R. S&P 100 index is a stock

market index which is composed of 102 leading US stocks, including Apple, Microsoft, etc. We use adjusted close prices of 101 out of 102 stocks in S&P 100 index as our features. We exclude one of the stocks because it has missing data from April to June. We only use 102 days from April. 11, 2016 to Sep. 2, 2016 as our samples because given 101 feature variables, we need at least 102 observations to train the 102 parameters (for adding offset model) in the linear regression model. Our output is adjusted close price of iShares S&P 100 ETF from the corresponding period.

To begin with, we do a basic linear regression using least squares (LS) fitting. The empirical error on the 101 constituents over 102 training days is around $10^{-3}$. Then we improve our model by adding offset and standardizing features to input data.

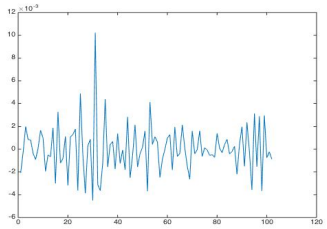

**Figure 1:** Empirical error without offset

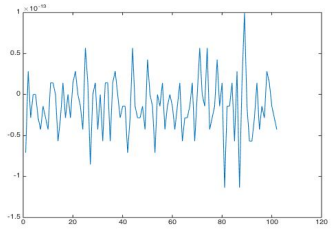

**Figure 2:** Empirical error with offset

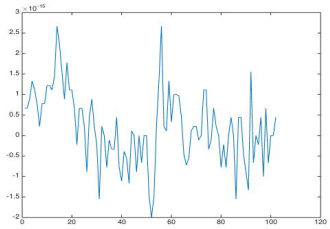

**Figure 3:** Empirical error standardization and offset

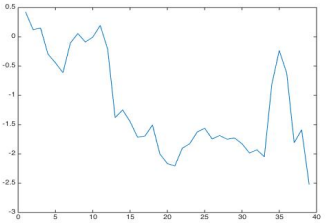

**Figure 4:** Generalization error without offset

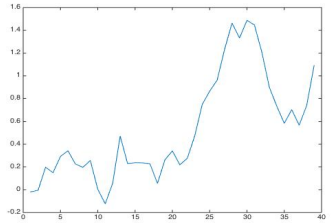

**Figure 5:** Generalization error with offset

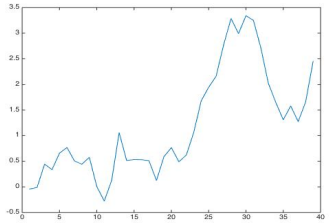

**Figure 6:** Gen. error standardization and offset

From Figures 1 2 and 3, we observe that the empirical errors of three models are acceptable (at the levels of $10^{-3}$, $10^{-13}$, $10^{-15}$ respectively). However, when we test this linear model using 39 examples that are not in the training set, i.e. adjusted close prices of each stock from Sep. 5, 2016 to Oct. 27, 2016, the generalization errors of the three models are far from satisfactory, as shown in Figures 4 5 and 6.

This implies that we are facing an overfitting problem, even though weve used the smallest possible training set. There are too many parameters in the linear model, which gets us into this problem. Hence, we apply Principal Component Analysis (PCA) to reduce the dimension of the model on the basis of the feature-standardized model, which has smallest empirical and generalization errors.

# 3 Principal Component Analysis

We apply PCA to reduce the feature dimensions. We use the previous 102 training days to estimate the correlation matrix of 101 features. Since PCA exports the principal components in the descending significance order, we compute the eigenvalues of the correlation matrix (Figure 7) and observe that the first 20 eigenvalues reveal almost all information of the correlation matrix.

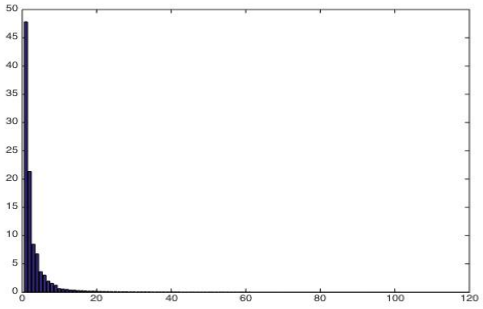


**Figure 7:** Eigenvalues of correlation matrix

By comparing the empirical errors of model with first one principal component (PC), first two PC, , and first 15 PC, we choose first 14 PC as the features of the new model since the first 14 has the least errors from both training set and testing set. And we will use the errors of the regression of the first 14 PC to generate trading signals.

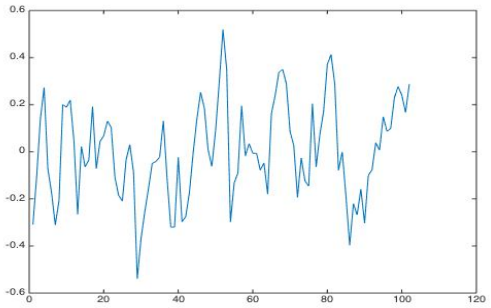


**Figure 8:** Empirical errors of first 14 PC

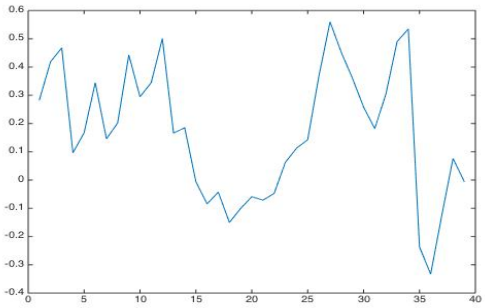


**Figure 9:** Gen. error of first 14 PC

# 4 Future Plan

We will try different methods to construct a model for the errors, which is interpreted as the idiosyncratic fluctuations in the prices, to determine the patterns of the error. And we will generate a statistics of errors as the signal to start or end trade. Eventually, the accuracy of model and the efficiency of the arbitrage strategy will be tested.