ORIE 4741 Course Project Proposal
Application of Machine Learning Techniques in Statistical Arbitrage

Zicheng Men, Sheng Zhang, Shan He

September 23, 2016

Executive Summary

In this project, we propose to apply machine learning techniques we learned from the course ORIE 4741 to the topic of statistical arbitrage. The central question we would like to answer is: *how to decide if an arbitrage opportunity appears in the stock market?* To answer this question, we would like to generate trading signals to alert investors of potential arbitrage opportunities by using supervised learning method and Principal Component Analysis.

Statistical arbitrage is defined as the automated trading strategy that identifies statistical mispricing of one or more assets (stock in our case) according to its historical expected value. It relies heavily on statistical analysis and data mining techniques. Since arbitrage opporunity is considered as riskless profit for investors or traders, topic of statistical arbitrage can be very important for trading and investment companyies, especially algorithmic trading company. Statistical arbitrage coupled with high-speed automated trading systems can generate enormous profits for algo-trading companies. It also helps to reinforce the efficiency of the market, since it can lay pressure upon the stock price to return to its "true" value by continously capturing stocks mispricing. Moreover, it is also valuable for non-profit purpose. The trading signals generated from our model can be used as benchmark for investors' own stock analysis.

Since we focus on stock arbitrage opportunities in this project, the dataset we plan to use is the stock closing price of the companies listed in Russel 2000 index over a 10-year period. Since companies from Russel 2000 are identified as small-cap companies, which, generally speaking, receive less attention from the public, therefore are more likely to be mispriced. An alternative approach to determine the dataset is to select stocks with low volumes.