Alejandro Suarez

Daniel Muniz

<https://users.cis.fiu.edu/~giri/teach/Capstone/F18/mentors.pdf>

**Capstone Project Proposal**

**Proposal**

* Acquire data from S&P 500 Stocks
* Map data into 2 or 3 dimensions using the t-distributed Stochastic Neighbor Embedding (t-SNE) visualization technique
* Identify clusters
* Compute summary statistics, create histograms and hypothesize distributions for these clusters
* Using t-SNE and PCA techniques, create various portfolios from the resulting clusters (k= 5, 7, 9)
* Compare to control portfolios (Capital Asset Pricing Models)
* Compare how the shape of clusters change across time

**Details**

In many financial funds across the world, fund managers try to attain the highest average yearly they can by constructing a well-balanced portfolio of securities that maximize gains while minimizing risks. Portfolios may comprise of stocks and bonds; bonds, stocks, and cash; stocks, options, and mortgages, etc. Grouping stocks with uncorrelated or negative correlations between their returns is a common risk management strategy amongst fund managers to offset potential losses.

We propose to leverage the power of the t-SNE algorithm to generate visualizations and clusters that can be used as the basis for high-performance portfolios. We will take S&P 500 daily and weekly stock returns from the past decade, embed them in a low dimensional space, build 6 portfolios from the resulting clusters with the highest returns (3 using t-SNE, 3 using PCA), and compare them to 2 control portfolios. These control portfolios will come from the Capital Asset Pricing Model, as well as an index of the top 50 performing stocks from the last 10 years.

For each cluster portfolio, we’ll graph side by side boxplots of the returns over the years (2009-2018), as well as a distribution of those returns. Then we’ll ultimately compare the performance of all 8 portfolios on the first 3 quarters of 2019 (1/1 - 9/30).

We also will perform an exploration of the data using the t-SNE algorithm. We’ll build the clusters starting from the 2009 data, and then compare how the clusters behave across time. Do they hold their shape well over time? How much do the centroids change and move, if at all? How do the mappings change from daily to weekly data? Do stocks with low daily volatilities have periods of high weekly volatility? Do stocks belonging to similar industries necessarily group together? What is the distribution of dollar gains and losses within each cluster?

**Data sources**

Quantopia (<https://www.quantopian.com/algorithms/5d83f0c090b3850047f50d04>)

Kaggle (<https://www.kaggle.com/camnugent/sandp500/downloads/sandp500.zip/4>)

Google datasets

Finance department at FIU

CRSP

QUANDL.COM

<https://www.tickdata.com/equity-data/>

Alpha vantage

**Tools**

Python, R, t-SNE, tableau, excel, SQL

**Hypothesis**

The CAPM model will perform the best on 2019 data. For our t-SNE exploration, cluster shapes will generally hold across time, with few stocks switching clusters.