TTIC 31220 Final Project Proposal
February 21, 2019
Authors: Kabir Sawhney and Joseph Denby

Motivation: Return variation among financial assets stems from a few dominant factors – inherent sources of risk that are independent from each other. We aim to use dimensionality reduction to see if we can improve on theoretically-constructed factor portfolios in explaining equity market performance.

Background: In financial theory, the rate of return of any given asset can be represented as a linear combination of returns on risk factors plus an error term. However, a great deal of research has gone into studying what the appropriate set of risk factors are from which asset returns are constructed. Hundreds of different models have been proposed – among the most popular is the Fama-French three-factor model, which proposes that returns can be explained by a combination of market risk, outperformance of small vs. big companies (the size effect), and the outperformance of low price-to-book vs. high price-to-book companies (the value effect).

Though there continues to be significant debate on which factors explain returns, most researchers agree that the number of key factors is relatively small. Some researchers have applied linear dimensionality reduction techniques to large factor universes, with a focus on linear PCA, to statistically derive a proposed latent factor representation. Most prominently in Kozak, Nagel, and Santosh (2017), purely statistical reduced-form factors constructed from anomaly portfolios do as well as theoretically constructed portfolios. This research has identified three and five-factor representations that perform similarly to the Fama-French models.

Data description: Novy-Marx and Velikov (2015) identified monthly returns to 32 long/short strategies, examining the returns to each strategy gross and net of trading costs. The dataset covers 486 months of returns, from July 1973 to December 2013, so our total data size is 486 x 32. These strategies were factor strategies, designed to either provide exposure to a specific underlying financial risk or exploit a pricing anomaly to earn excess returns. Their data set also includes returns on a more limited set of 24 strategies going back to July 1963 – part of our project will be determining which of these two subsets of the data to use. We also will use the Fama-French factor models and returns on the S&P 500 equity index over the same time period.

Project plan:

1. Apply a range of non-linear and graph-based dimensionality reduction techniques to the Novy-Mark and Velikov data, including kernel PCA, Laplacian eigenmap, and Isomap. We will also apply linear PCA to have a reference. We will focus on three and five-component latent representations.
2. Use our learned representations in a supervised learning model with the monthly returns of the S&P 500 equity index as the response variable. Choose the representation that generates the lowest squared loss.
3. Use the Fama-French factor models as the covariates in the same supervised learning model, and compare our loss to that from the Fama-French model.
4. Split our data into training and validation sets and repeat the analysis, to determine if in-sample results can be replicated out of sample. We will explore two ways of splitting the data: in time (with older data as the training set) and randomly.