TTIC 31220 Final Project Update: Toward a Reduced-Form Factor Portfolio
March 7, 2019
Kabir Sawhney and Joseph Denby

Clarifications to project as a result of feedback

- We had to choose between two subsets of our data: one that went back to 1973 and had more features, or one that went back to 1963 but had fewer features. Based on feedback we received, we are focusing on the subset of our data that has fewer features. Using monthly returns from 1973 to 2013 would give us about 480 data points – to get a robust split into training, validation, and test sets, this data set size would not be sufficient.
- We were unsure about whether to use a time or random split of our data into training and test sets. Based on feedback, we will implement both methods and compare the results.
- In final report and presentation, we will better explain and clarify the financial terminology from our proposal.

Progress in analyzing our research question

- Downloaded and cleaned the Novy-Marx factor strategies, the Fama-French three and five-factor models, monthly returns on the S&P 500, and monthly total returns on the five longest-tenured stocks in the Dow Jones Industrial Average – ExxonMobil, Procter & Gamble, United Technologies, 3M, and IBM.
- Implemented a range of dimensionality reduction techniques to analyze principal components of the Novy-Marx factor set, including linear PCA, kernel PCA, Isomap, and Laplacian eigenmaps.
- Implemented a correction for autocorrelation as a tuning feature, to be used when analyzing the data sequentially in time.

Results obtained

- Our initial analysis was in-sample only. We trained our representation learning methods on the entire data set and did regression analysis on that data. The next step of our analysis will be to split the data into different sets to see if we can get good out-of-sample results.
- Using five individual stocks as response variables, we found that linear PCA has the best R^2 out of the methods we tested – a table of our preliminary results below. We couldn't find a representation that performs as well as the Fama-French models in explaining stock performance. After linear PCA, kernel PCA using the RBF kernel and Isomap had similar performance. We also implemented Laplacian eigenmaps, but did not get good performance from this method.
- Autocorrelation was not as much of a problem as we initially thought it might be. Returns from month to month don't appear to be dependent on the previous months, so using an autocorrelation correction didn't have a meaningful impact on observed results.
- A surprisingly large number of components were needed to get decent performance when reducing the Novy-Marx strategies. We plotted the R^2 value at various numbers of components, from 2 to 24 (the original dimensionality of the data), and found that performance leveled off at 20 components.

Table of results (all values are R^2 obtained using stock returns for the given company as the response)

Company	**Fama-French 3 Factor**	**Linear PCA (k = 20)**	**Kernel PCA (k = 20, kernel = RBF)**
IBM	0.34	0.29	0.27
3M	0.37	0.22	0.21
Procter & Gamble	0.25	0.20	0.20
United Technologies	0.40	0.24	0.25
ExxonMobil	0.36	0.24	0.22

Setbacks or changes necessary in plan

- We had to change the response variable from the monthly returns on the S&P 500 to five individual stocks, for comparing our reduced-dimension analysis to the Fama-French factor models. Initial analysis using the S&P 500 as the response resulted in R^2 values of nearly 1 using the Fama-French factor models. Looking closer at their methodology, the broad US equity market return is one of their factors, so the return on the S&P 500 can almost be perfectly predicted from that factor.

Next steps

- Analyze the S&P 500 monthly returns with reduced-dimension representations of the Novy-Marx factors. Even if we can't compare our results against the Fama-French model, part of the original motivation was to see if non-linear and graph-based dimensionality reduction methods can achieve better performance than linear PCA – we can still see if we can find such a representation for the overall stock market.
- Split data into training, tune, and test sets. Tune hyperparameters for the dimensionality reduction methods, and see which set of hyperparameters and which method generates the best out-of-sample performance (measured by R^2).
- Select the best dimensionality reduction method and compare it to the F-F models, using R^2 as the criteria for comparison.
- Analyze the results and present our conclusions, in the in-class presentation and final report.

Split of Responsibilities

- Kabir: Gathered and cleaned data from various sources, made changes to approach and techniques used as we progressed in the analysis, implemented the autocorrelation correction for time series analysis, and wrote the project proposal and this project update.
- Joseph: Implemented the representation learning and regression methods. Wrote the code to import most of the data into our Jupyter notebook, determine R^2, tune hyperparameters, and graph the results of the tuning.