DATA 612: Final Project Proposal

Derek G. Nokes

2019-07-08

Introduction

Applications of matrix factorization are ubiquitous in modern data-driven science. The need for tools that can map a high-dimensional space into a lower-dimensional space continually grows as the amount of data we collect increases.

A service like Spotify has subscribers that provide indications of interest in particular content in the catalog by either listening to particular items, clicking on items to read descriptions, or adding items to a playlist. Such implicit ratings data allows recommender systems developers to create the classic user by item by implicit rating matrix and apply collaborative filtering. Similarly, trading systems researchers potentially have a very large set of distinct systems that select instruments from a particular instrument universe. The portfolio of positions - which varies over time - provides an indication of the kind of positions each system 'likes'. Many systems traders run ensembles comprised of multiple instances of the same system and instrument universe with different parameter sets. The more diverse the systems within the ensemble (i.e., the lower the co-movement between portfolio components) the faster the return compounding. For the systems researcher, despite knowing the underlying mechanism for selecting instruments, it is not always easy to group systems in such a way as to maximize the diversity within the total portfolio.

Objectives

The dual objectives of this project are to 1) develop robust metrics that can characterize the time-varying level of diversity between trading systems and instruments, and 2) employ matrix factorization techniques to classify trading systems into groups and recommend combinations of systems that used together can enhance the performance of a simple systematic trading strategy.

Motivation

Although the public equity markets are highly accessible for nearly all classes of global investors, these markets pose some significant challenges. In particular, there is a significant degree of co-movement across single stocks, making the construction of a well-diversified portfolio difficult. The high degree of co-movement makes an investor vulnerable to broad-based declines in equity markets.

One of the simplest and most effective strategies employed by active investors to control the risk associated with broad-based declines and enhance performance when markets are rising, involves exploiting a well-known stylized fact of equity markets, namely that stocks that are moving strongly in a particular direction tend to continue to move in the same direction (i.e., they possess 'momentum'). Momentum investing systems focus on identifying stocks that are moving persistently in a particular direction and taking a position to benefit from that directional movement. The long-only version of such strategies buys stocks that are rising most persistently and exits long positions when markets reverse.

To reduce the volatility of such a strategy, stocks that move together are typically grouped and bet as though they represent a single latent 'factor'. Indeed, the most challenging aspect of developing a so-called momentum system is not the identification of momentum stocks, but rather the selection of diverse groups of stocks that - when held together - provide portfolio return smoothing and accelerate the speed of compounding.

In this project, we seek to first develop metrics that can quantify the evolution of the state of diversity in a particular universe of trading systems, then develop a matrix factorization based analytical framework to help researchers improve the diversity within their portfolios.

Data Sources

To perform the proposed analysis, we require a list of the current constituents of the S&P1500 index, along with corresponding instrument master and price data.

The first part of the data set - the instrument master for our universe under study - is to be scraped from the Blackrock iShares website. We will use the holdings of three iShares index-tracking exchange traded funds (ETFs) as proxies for the S&P500, the S&P400, and the S&P600 indices. These three indices comprise the S&P1500. The second part of our data set - corresponding prices for each instrument - are to be obtained from CSI.

Work Plan

The rough work plan for this project is as follows:

1. Create a docker environment to facilitate reproducibility.
2. Implement a systematic trading strategy backtest simulator in PyTorch. Note that the use of PyTorch allows each strategy instance to be backtested in parallel using the GPU (rather than in series using the CPU).
3. Use the simulator to generate transaction data for different trading systems defined by distinct instrument universe and parameter set combinations.

4. Convert time-varying position output into implicit ratings information. Simulated positions of each distinct system represent indications of interest in particular instruments at specific moments in time. Create a system ID by instrument ID by implicit rating matrix for each instant in time.

5. Use matrix factorization to represent the high-dimensional system by instrument by implicit rating space as a lower-dimensional space.

6. Develop a metric to quantify diversity.

7. Combine instrument master data (namely equity name, sector, and capitalization category) with implicit ratings data and explore the relationship between latent factors and system performance.

8. Generate recommendations to improve the performance of an ensemble of trading systems.

Project Risks

As far as we are aware, there is no similar publicly available work. This may be an indication that the proposed approach will not produce useful results.

The scope of the project is also quite large. In project 4, a similar simulator was used to generate transaction data. This simulator, however, was not implemented using PyTorch and did not scale well. As a result, we were not able to generate simulated transactions over a broad set of parameters in a timely manner. We did however test a method to transform position data into a form of implicit ratings data.