Data Science in Systematic Investment Management

Opportunities, Challenges and Future Directions

Bio & Background

- Focus on systematic portfolio management at Two Sigma Investments.
- Research and development experience throughout the entire investment process.
- Background in economics & finance, applied mathematics and computing.

Systematic Investment Management

- Provide risk-adjusted returns through use of rules-based and quantitative processes.
- Scalable, diversified, repeatable and historically back-testable.
- Canonical example is Markowitz mean-variance optimization ("MVO").
  - Maximize expected portfolio return while minimizing expected portfolio risk over a suitable large set of financial assets.
  - Optimization problem.
  - Requires return estimates, risk estimates and asset correlation estimates.

Systematic Investment Management

In practice, there may be more constraints and considerations:

- Trading costs: market impact and trading commissions
- Liquidity
- Position limits
- Financing costs
- Common risk factor exposure management
- Counterparty wallet share relationship management for prime brokerage balances and trade execution

-> More complex optimization problem with each incremental consideration requiring quantitative research.

Why Data Science in Systematic Investment Management?

Wikipedia defines data science as an interdisciplinary field that leverages scientific methods, algorithms, processes and systems to draw actionable insights from data across a broad range of domains.

Systematic portfolio management aims to deliver risk-adjusted returns to investors in the data-rich domain of capital markets through use of statistics, optimization, algorithms and scalable processes.

-> Strong match between data science and systematic investment management.

Application Areas in Systematic Investment Management

Systematic investment management solves a maximization problem:

Maximize utility function: expected return – risk – trading costs – other costs

subject to all expressed constraints

Data science can come into play in:

- Portfolio optimization
- Return forecasting (i.e. quantitative strategy development)
- Risk modeling such as volatility and correlation forecasting
- Tail risk modeling
- Parallelized cloud-based computation infrastructure for achieving scalability

Case Study I: Portfolio Optimization

The first example will highlight the algorithmic foundations of data science.

- Numerical convex optimization tends to rely on the differentiability of the objective function and therefore requires careful design of utility function terms and especially constraints.
- With the addition of more complex terms to the utility function that are not linear-quadratic, solving the utility function defined over 5,000 to 15,000 financial assets might require high performance numerical convex optimization algorithms.
  - Gradient Descent & Stochastic Gradient Descent methods
  - Quasi-Newton method: BFGS, L-BFGS,
  - Interior-point methods

Case Study II: Parallelized Compute Infrastructure

- Most research in systematic investment management is predictive modeling: statistical investigation of the relationship between some transformation of historical data and a response.
- The more historical data, the better.
- Assess the hypothesis over a suitable search space of hyper-parameters to maximize a quality metric.

-> Each parameterization might be assessed over millions of data points arising from thousands of financial securities over multiple decades of historical data.

-> Quick iteration requires a parallelized compute infrastructure where large computations can be split up and run in parallel and high performance storage.

Case Study III: Strategy Design

Momentum (Jegadeesh and Titman): past returns predict future returns.

- How to compute past returns? Mean / median return over a configurable rolling window?
- What's the response? Go-forward close-to-close return or a shorter / longer horizon?
- What quality metric to maximize? Correlation, regression beta or t-statistic?

In this example, there is a medium-sized search space linking a central tendency measure of returns to future returns that need to evaluated over millions of asset – timestamp data points.

Opportunities

There is an exponentially growing set of machine readable / ingestible data available for modeling consumption.

Parallelized compute is cheaper and more accessible than ever before via cloud-based solutions.

Publicly available machine learning libraries, adapted to parallelized computation paradigms, have democratized access to sophisticated predictive modeling tools.

Algorithmic advances in data science and machine learning and rapid digitization have opened up previously untapped unstructured video, text and speech-based data for modeling.

Challenges

Garbage in garbage out principle.

Increasing amounts of data does not necessarily mean predictive or relevant data.

Sophisticated techniques and computing power enable data mining and overfitting to historical data in a way that might not generalize.

Priors and observation are critical to successful research.

Future Evolution

- There is an increasing interest and experimentation with unstructured data such as speech snippets, social media posts, financial press & article comprehension, mobile data, video analysis.
- Algorithmic advances in analyzing unstructured data such as natural language processing, deep learning on images and reinforcement learning as applied to trade execution have enjoyed strong growth.
- Acquisition & onboarding, cleaning, storage and rapid automated scoring and evaluation of data sets have become critical.

Thank you!