Large Scale Machine Driven Statistical Arbitrage

Introduction

Our hypothesis is that we can treat the markets as an abstract problem, and utilise machine-learning principles and theory-free modelling to predict future relative value. Certain datasets provide a glimpse of the future of relative stock returns. We utilise a suite of algorithms and research processes to refine and harness information from these datasets to predict relative price movements. Our approach thrives on scale, which enhances performance and builds a protective moat against competitors. Furthermore, a fully automated pipeline ensures costly human capital can focus on innovation and refinement.

Datasets

We seek datasets with comprehensive coverage and those with extensive histories. We run search algorithms on samples to assess the likelihood of discovering good models. Once we have acquired a dataset, our goal is to extract all useful information in a dataset, by transforming each dataset into multiple new datasets that contain information orthogonal to the original dataset.

Search Algorithms

Our search algorithms are employed on our datasets. These algorithms derive models from the datasets. Each model is a weak predictor relying on a few data fields to forecast relative future returns. Search algorithms are only allowed to utilise some available historical data to find good models. Good models meet criterias such as working in large universe sizes, low exposure to known risk factors, high sharpe ratios, low correlation to production models, robust to universe sizes, low turnovers and high capacities.

Theory Free Searches

When running search algorithms, the lack of an imposed structure will trade off type II errors for type I errors. It is inevitable that there will be noise introduced to our pool, and this is true whether or not the models are found by a machine or a human. However, the core belief here is that our process can find true positive models that produce useful predictions by utilising datasets that contain information about future returns.

Validation Process

Given the risk of overfitting, our models undergo stringent checks to increase the probability of robustness. A designated hard validation period is kept separate from the search algorithms. Before good models are productionized, they have to pass the validation process. The purpose of a validation process is to validate that the model is able to generalise into an out-of-sample period, and that it value-adds to the production pool of models.

Overcoming Noise

Some models will still be productionized out of sheer statistical chance despite being noisy. The primary insight is that we can apply the law of large numbers to models as well. As long as our process can generate models with positive expectation, we can ensemble a large number of models, and the asymptotic behaviour of false positive models is that noise signals will cancel each other out, and the sum of that is negative PnL equivalent to the transaction costs of turning over the noise models.As long as the PnL generated from true positive models can overcome the turnover cost of the noise models, this strategy will be able to generate positive PnL.

Conclusion

Our approach with machine-driven statistical arbitrage allows us to utilise the law of large numbers at the model level, and extract unintuitive information from datasets about future returns. As the number of datasets and models increase, the number of uncorrelated, true positive signals will increase and so will our performance. Since our process is largely algorithmic, incremental improvements increase marginal benefits and decrease marginal costs of each new dataset.