Funny Indices

1 Setting

In this problem we will explore the world of funny indices in a toy stock universe composed of 100 stocks. You are are given the prices for each stock from time $t : 0 \rightarrow N$ (*data_challenge_stock_prices.csv*). You can assume that the return: $r_t^s = (p_{t+1}^s - p_t^s)/p_t^s$, where p_t^s is the price of stock s at time t, is conditionally independent of r_{t-1}^s.

Additionally you are given 15 indices with their corresponding prices at each timestamp (*data_challenge_index_prices.csv*). Each index I_i is a related to the stock prices. Specifically if $R_t^{I_i}$ is the return of Index I_i at time t then:

$$R_t^{I_i} = f_i(r_t^0, r_t^1, ..., r_t^{99}) + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i)$, is i.i.d. (independent of t) particular to each index.

2 Problem

Analyze and answer the following about this data: (Note that since you will be working with returns which are small numbers, its advisable to **transform the returns to bps to help with fitting**).

1. Each stock belongs to a particular sector S. There are M sectors ($M < 100$). Find the value M.

2. Find the partition of stocks into sectors. You should store the set of stocks for each sector to help speed up the next part.

3. Given an additional constraint (if needed) that **each index I_i is a function of stocks only from a particular sector** and conforms to the functional form given in (1), solve for as many indices as you can. As stated above, the index returns at time t can be predicted using stock returns at time t.
   A solution for an index I_i, involves a model which implements f_i and reports which sector this index pertains to. By construction, these indices have the noise ε_i tuned such that a correct solution should be able to achieve $\geq 40\%$ predictive correlation. You are also encouraged to reason about the possible functional forms for index I_i in a data-driven approach.

4. Collect all the indices you could solve (lets say k) and compute the covariance $\Sigma_{k\times k}$ for your predictions $\mu_{k\times N}$, (you will have k predictions for each timestamp).

5. Construct a trading strategy using μ and Σ from the previous part. (you will have a prediction μ_t for each time t and Σ is a static estimate. You are given the following constraints/conditions:

   - you are able to trade the Index I_i at its price at time $t-1$, using price information from stocks at time t. Effectively this will allow you to use the predictions on the return $\hat{R_t^{I_i}}$ and the estimated covariance Σ from above.
   - You need to decide an allocation $A^t_{I_i}$ for each index at each time using your predictions. Additionally $A^t_{I_i}$ is constrained as $-1 \leq A^t_{I_i} \leq 1$, or your max long/short position is constrained to \$1.
   - Your long-short positions should net out completely, i.e. $\sum_i A^t_{I_i} = 0, \forall t$.

6. Finally, report the mean, stdev and sharpe of your trading strategy. Write a small note discussing assumptions in the above execution scheme.

3 Guidelines

It would help to keep the following points in mind while attempting the problem:

- If you get stuck trying to figure out the sector partition, feel free to skip ahead to index return prediction and just put in all stock returns as indicators.

- We encourage trying a vast variety of fitting techniques but still having some valid reasoning behind them. While some of these indicies are inherently "funny", most of them should be easy to fit with relatively less data and simple ideas. Training time for a model should be a few minutes in the worst case.

- When working with returns which are inherently small numbers, a lot of fitting issues might arise if these are not converted to bps or standardized in some other way.

- In the trading-strategy construction you may try different allocation schemes and see how they impact mean-pnl and sharpe.

- If your allocation scheme is complicated (like an optimization problem) you may only want to run it for the last 5-10k points in the dataset. Your mean-pnl and sharpe results should be stable with this amount of data since the underlying data is homogenous.

- Finally, we hope you enjoy discovering the many small *eureka-voila!* moments while exploring the dataset even if you're not able to complete the problem. Please reach out to us with any suggestions!