Project 3: Idea Generation for Price Prediction Signal for a Macro Asset or ETF

There has been a wealth of research on how to construct features using volume and price data. In contrast, we are more curious about how much introducing other data dimensions can help.

Here we propose a possible **high-frequency** ETF price prediction signal based on the co-trading and self-exciting mechanism in cross-stocks under a specific ETF.

Underlying asset: SPY(The SPDR S&P 500 ETF) and its components

Idea generation:

- There is significant co-occurrence-trading between stocks under many different mechanisms Such as
  - Insider traders will split a large order into smaller ones to mitigate the market impact.
  - Institutional traders will place orders on a basket of stocks at the same time.
  - High-frequency traders identify counterparty orders and track the orders placed.
  - Arbitrageurs will hedge by placing a basket of orders and ETFs at the same time.

  > [2209.10334] Trade Co-occurrence, Trade Flow Decomposition, and Conditional Order Imbalance in Equity Markets (arxiv.org)

- Previously I have done ample research about the first 3 mechanism and found that
  - there is significant and time-stable co-trading pattern between stocks in the same industry or ETF.
  - such patterns imply a strong ability to predict future returns.
- For the last mechanism, however, the only step I took stopped at interest. Although we did not have the time to delve deeper into the study and faced significant barriers to obtaining open source high frequency data, this did not prevent us from hypothesizing and designing experimental steps to facilitate subsequent research.
- **BASIC HYPOTHESIS**:
  - Arbitrage behavior contain additional information to predict the future price of ETF.
  - There are 2 aspects: Arbitrage inside the ETF and between ETFs.

Experiment Steps:

- Data
  - Collect and clean high-frequency trading record and price data for SPY and its top 50 holdings, such as Apple, Microsoft, Amazon, NVIDIA, etc.
  - Acquire similar data for ETFs of indexes highly correlated but with few common components, such as SOXS, SOXL, QQQ, etc.
  - Data clean work such as sanity check and infinity check. Also notice that the feature calculation and definition might only make sense when away from near up/down stops.
- Feature Engineering

- Refer to the previously mentioned paper for definition of co-trading and the following paper for the definition of co-trading network.

  $$A_{ij} = \text{co\_trading\_score}(symbol_i, symbol_j)$$

  where

  $$\text{co\_trading\_score}(i, j) = \frac{\text{volume of cotrading from i to j}}{\text{normalized daily trade volume}}$$

  > [2302.09382] Co-trading networks for modeling dynamic interdependency structures and estimating high-dimensional covariances in US equity markets (arxiv.org)

  - Note that,
    - The parameter to define co-occurrence trading might differ a lot under different markets and different mechanisms. For arbitrage trading we can consider to search for co-trading in a 500ms-1s period.
    - **The definition of a pairwise co-trading matrix in the paper is too narrow**. We make further improvement such that the co-trading behavior between stock i and ETF j is not symmetric such that

      $$a_{ij} \neq a_{ji}$$

    - Also, **the definition of same-side trading is not applicable** since arbitrage must be long something and short something. Which means there should be a buy-ETF short-stock matrix as well as a short-ETF long-stock matrix.
    - Further more, the 2-sided co-trading search is under-considerate. Arbitrageurs might have pattern to trade ETFs earlier than stocks, and vise versa. **Which means we can use bias search in the model, like search forward or backward in time.**
  - After the network construction, we can choose the most relevant stocks/ETFs as additional information to predict the movement of ETF. Such features will enter into the train set of our target, SPY.

- Model Training & Prediction:
  - There will be no data size limit so we can apply time-rolling train-test mechanism over time.
  - As we have seen in project 1, such mechanism can help to capture short-term market style transition.
  - **Baseline**: any regression model with simple feature set, for example the ETF itself. For example OLS(simple linear regression) or a simple tree model XGBoost will work, *because we want to measure the additional prediction power we gain.*
  - **Improved**: the same regression model with full feature set, including the ETF-cross-stock and ETF-cross-ETF signals.
- Evaluation:
  - general metrics: RMSE, out-of-sample R^2, pearson, spearman, stratification monotonicity and variance, etc (just as what we have done in project 1).
  - time-rolling metric: coefficient stability and significance.
  - Note that we need to observe the additional prediction power gained.

For the reason that obtaining open source high frequency data will be difficult, this project currently exists only in theory and hopefully I will implement soon.