EPFL

Financial Big Data

FIN-525

Optimal Causal Path Statistical Arbitrage on SP100

Project Proposal

by Matthias Wyss (SCIPER 329884)
Lina Sadgal (SCIPER XXXXXX)
Yassine Mustapha Wahidy (SCIPER 345354)

Master in Data Science
Minor in Financial Engineering
MA3

Prof. Challet Damien

Contents

1 Introduction

This project aims to implement a statistical arbitrage trading strategy based on the Optimal Causal Path (OCP) algorithm. We will focus on the S&P 100 constituents over the period 20??-2017, using tick-level Best Bid/Offer (BBO) data. The objective is to replicate the OCP framework proposed in [1], identify high-quality pairs with stable lead-lag relationships, and evaluate the profitability of the strategy against a naive buy-and-hold S&P100 benchmark.

2 Data Preparation

For each stock i and timestamp t, we compute the midprice:

$$P_{\text{mid}}^{(i)}(t) = \frac{P_{\text{bid}}^{(i)}(t) + P_{\text{ask}}^{(i)}(t)}{2}.$$

These midprices are then resampled to 1-minute intervals:

$$P_{\text{mid}}^{(i)}(m) = P_{\text{mid}}^{(i)}(t_m),$$

and minute-by-minute returns are computed:

$$r^{(i)}(m) = \frac{P_{\text{mid}}^{(i)}(m) - P_{\text{mid}}^{(i)}(m-1)}{P_{\text{mid}}^{(i)}(m-1)}.$$

All possible stock pairs (x, y) are formed, yielding $C_{100}^2 = 4950$ pairs per day.

3 Optimal Causal Path (OCP) Algorithm

Given return series $r_x = (x_1, \ldots, x_N)$ and $r_y = (y_1, \ldots, y_M)$ $(N \geq M)$, the OCP algorithm is applied in three steps:

3.1 Step A: Constant Lag Estimation

For each candidate lag $l \in \{0, \ldots, M-1\}$:

$$c_l = \sum_{i=1}^{M} |x_{i+l} - y_i|,$$

and the initial lag is chosen as

$$\hat{l}_{\text{initial}} = \arg\min_l c_l.$$

This lag provides the starting point for the next step.

3.2 Step B: Optimal Causal Path with Variable Lag

The total cost of a warping path $p = \{(n_i, m_i)\}_{i=1}^{I}$ is:

$$c_p(r_x, r_y) = \sum_{i=1}^{I} |x_{n_i} - y_{m_i}|,$$

subject to:

1. Boundary: $(n_1, m_1) = (1, 1), (n_I, m_I) = (N, M)$
2. Monotonicity: $n_1 \leq \cdots \leq n_I, m_1 \leq \cdots \leq m_I$

3. Step size: $(n_{i+1} - n_i, m_{i+1} - m_i) \in \{(1,0),(0,1),(1,1)\}$

Step B starts from the diagonal path shifted by $\hat{l}_{\text{initial}}$ and iteratively adjusts the path locally to minimize c_p, until convergence.

3.3 Step C: Final Lag and Fluctuation

The estimated lag and its fluctuation are:

$$\hat{l} = \frac{1}{I}\sum_{i=1}^{I}(n_i - m_i), \quad \sigma_l = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left[(n_i - m_i) - \hat{l}\right]^2}.$$

The top 10 pairs with non-zero lag and lowest σ_l are selected for trading.

4 Trading and Backtesting

4.1 Formation Period

Each trading day is split into:

- Formation period: 391 minutes (1 trading day)
- Trading period: following 391 minutes

During the formation period, OCP estimates $\hat{l}$ and σ_l for all pairs. Top 10 pairs with non-zero lag and stable fluctuations are selected.

4.2 Entry Signals

For leader x and follower y, compute rolling mean and standard deviation over past d minutes:

$$\mu(t) = \frac{1}{d}\sum_{i=t-d}^{t-1} r_x(i), \quad \sigma(t) = \sqrt{\frac{1}{d}\sum_{i=t-d}^{t-1}(r_x(i) - \mu(t))^2}.$$

Define Bollinger bands:

$$\text{Upper}(t) = \mu(t) + k\sigma(t), \quad \text{Lower}(t) = \mu(t) - k\sigma(t).$$

Trade signals:

$$\begin{cases} r_x(t) > r_{\text{cost}} \text{ and } r_x(t) > \text{Upper}(t) & \Rightarrow \text{long } y \\ r_x(t) < -r_{\text{cost}} \text{ and } r_x(t) < \text{Lower}(t) & \Rightarrow \text{short } y \\ \text{otherwise} & \Rightarrow \text{no trade} \end{cases}$$

4.3 Portfolio Construction

- Dollar-neutral: hedge with S&P100 index. - Close positions at target return, end of trading day, or stock delisting. - Portfolio consists of top 10 active pairs.

4.4 Performance Evaluation

- Portfolio return:

$$R_{\text{portfolio}} = \sum_{t \in \text{active positions}} r_y^{\text{position}}(t)$$

- Benchmark: S&P100 buy-and-hold

$$R_{\text{MKT}} = \frac{P_{\text{S\&P100}}(T_{\text{end}}) - P_{\text{S\&P100}}(T_{\text{start}})}{P_{\text{S\&P100}}(T_{\text{start}})}$$

The project will compare the OCP strategy to this naive benchmark.

5 Conclusion

This project will replicate the OCP statistical arbitrage framework on high-frequency SP100 data, evaluate its profitability, and benchmark it against a passive market strategy.

References

[1] Johannes Stübinger. *Statistical arbitrage with optimal causal paths on high-frequencydata of the SP 500*. eng. FAU Discussion Papers in Economics 01/2018. Nürnberg, 2018. URL: https://hdl.handle.net/10419/173658.