EPFL

Financial Big Data

FIN-525

Optimal Causal Path Statistical Arbitrage on SP100

Project Report

by Matthias Wyss (SCIPER 329884)
Lina Sadgal (SCIPER 342075)
Yassine Mustapha Wahidy (SCIPER 345354)

Master in Data Science
Minor in Financial Engineering
MA3

3 ECTS

Prof. Challet Damien

Autumn Semester - 2025

Contents

1 Introduction

Statistical arbitrage pairs trading is a market-neutral strategy that aims to exploit temporary price anomalies between related financial instruments. Since its inception in the 1980s, the traditional approach has relied on identifying pairs of stocks that exhibit strong historical co-movements. When a divergence occurs, the arbitrageur takes a long position on the undervalued security and a short position on the overvalued one, profiting as prices converge back to their long-term equilibrium.
However, classical similarity measures, such as Pearson correlation or Euclidean distance, suffer from a significant drawback: they are highly sensitive to time shifts and misalignments. In high-frequency environments, lead-lag structures between stocks are often non-linear and time-varying, making traditional fixed-time distance measures inadequate for capturing the true underlying dynamics.
To address these limitations, this project implements an integrated statistical arbitrage framework based on the **Optimal Causal Path (OCP)** algorithm. This method utilizes a dynamic programming approach to identify the most suitable non-linear mapping between two time series, allowing for an elastic adjustment of the time axis. By efficiently determining the lead-lag structure at a minute-by-minute resolution, the strategy identifies "leader-follower" relationships where the information contained in the leading stock's returns can be used to predict the future movements of the following stock.
In this report, we apply this sophisticated framework to high-frequency data of the S&P 100 constituents for the period 2015–2017. Our objective is to reproduce the methodology that has generated statistically and economically significant returns in the reference literature. We evaluate the performance of the OCP-based strategy against classical benchmarks to assess its value-add in capturing modern market inefficiencies.
To ensure the reproducibility of our results and provide a clear overview of the implementation, the complete source code, including the OCP engine and the backtesting framework, is available on our GitHub repository: https://github.com/matthias-wyss/OCP-StatArb-SP100.

2 Related Work

The methodology implemented in this project is primarily inspired by the work of Stübinger (2018) [1], who introduced a comprehensive statistical arbitrage framework based on the **Optimal Causal Path (OCP)** algorithm. Stübinger's research addresses a significant limitation in classical pairs trading: the sensitivity of traditional similarity measures, such as Euclidean distance or Pearson correlation, to time shifts and misalignments.

The OCP Algorithm and Lead-Lag Structures The core innovation of the reference paper is the development of a non-parametric, three-step algorithm designed to identify the optimal non-linear mapping and lead-lag structure between two high-frequency time series.

- **Step A** determines an initial optimal lag by assuming a constant lead-lag structure and minimizing a global cost measure across a defined range of lags.
- **Step B** utilizes dynamic programming to permit an elastic adjustment of the time axis, effectively capturing time-varying lead-lag relationships by finding the path of lowest total cost.
- **Step C** extracts the average lag and the fluctuation (standard deviation) of the path to characterize the stability of the relationship.

Stübinger demonstrates through simulation that this approach is robust against noise and efficient in detecting true causal dependencies.

High-Frequency Statistical Arbitrage Unlike lower-frequency strategies, Stübinger applies the OCP framework to minute-by-minute data of the S&P 500 constituents. The strategy logic rests on the premise of **information leadership**: the algorithm exploits information about the leading stock to predict the future returns of the following stock.

The reference paper establishes a rigorous back-testing protocol:

- **Formation Period:** A 1-day window used to identify and select the top $s = 10$ pairs exhibiting the most stable lead-lag structure (i.e., the lowest standard deviation σ_l).
- **Trading Period:** A subsequent 1-day window where trading signals are generated for the following stock based on the movement of the leader.
- **Execution Logic:** Signals are triggered using Bollinger Bands (with parameters $d = 20$ and $k = 2.5$) and an economic threshold r to ensure returns cover transaction costs.

Empirical Benchmark and Performance In the original study covering 1998 to 2015, the OCP strategy generated an annualized return of **54.98%** after transaction costs, with an annualized Sharpe ratio of **3.57**. These results significantly outperformed traditional benchmarks such as correlation (COR), Manhattan distance (MAN), and lagged cross-correlation (LCC). Furthermore, the author found that these returns do not load on common systematic risk factors, suggesting that the OCP algorithm captures a unique form of market inefficiency.

By implementing this framework on the S&P 100 universe, our project seeks to validate these findings and assess the persistence of these lead-lag anomalies in more recent market environments.

3 Method Implemented

3.1 Data

The dataset used in this project consists of historical Best Bid and Offer (BBO) quotes for all constituents of the S&P 100 index. Prof. Challet Damien provided the data as 309 compressed `.tar` archives, corresponding to 102 stocks and the SPY index over the years 2015, 2016, and 2017. Each yearly `.tar` archive contains daily parquet files with all BBO updates recorded for that specific ticker. Although 102 tickers were expected, only 93 contain usable data, as some `.tar` archives are empty. Each parquet file corresponds to a single trading day and is named following the convention `YYYY-MM-DD-TICKER-bbo.parquet`. Since BBO data is generated at high frequency, a single day may contain several hundred thousand events.
Each row corresponds to a market event where at least one of the best bid or best ask levels has changed. The main fields include:

- `xltime`: fractional days since 1899-12-30, representing the event timestamp.
- `bid-price`, `ask-price`: best bid and ask quotes.
- `bid-size`, `ask-size`: quote depths (not used in subsequent analysis).

Since the raw data is event-based with irregular timestamps, preprocessing is required to produce uniform price series suitable for analysis.

Loading and extraction. Each yearly archive is extracted programmatically. Daily parquet files are loaded sequentially using Polars to allow efficient memory usage. Invalid or incomplete rows are discarded to ensure data consistency.

Timestamp conversion and trading hours filtering. The `xltime` column is converted into proper datetime objects localized to the `America/New_York` timezone, as tickers belong to NASDAQ and NYSE markets. A separate `date` column is added to facilitate day-by-day processing. Each day's data is filtered to include only regular trading hours (9:30–16:00). The NASDAQ trading calendar is used via `pandas_market_calendars` to ensure that all tickers are aligned on the exact same trading days, which is critical for computing consistent lead-lag relationships and returns in the OCP algorithm.

Resampling and mid-price construction. For each trading day, the mid-price is computed as

$$m_t = \frac{\texttt{bid-price}_t + \texttt{ask-price}_t}{2}.$$

The series is then resampled at a 1-minute frequency restricted to regular trading hours. Missing values are forward-filled and backward-filled to ensure a complete series. This procedure guarantees that all 390 minutes per trading day are present, which is essential for consistent return computation and for the stability of lag estimation in the OCP analysis.
Figure 1 shows the 1-minute mid-prices for AAPL on October 23rd, 2015.

Return computation. Minute-by-minute returns are computed from the resampled mid-price series as

$$r_t = \frac{m_t - m_{t-1}}{m_{t-1}}.$$

Each trading day therefore contains 389 returns, aligned with the 390 mid-price values. Figure 2 shows the corresponding 1-minute returns for the same day.

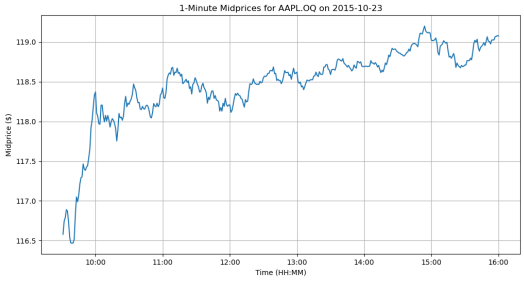

Figure 1: 1-minute mid-prices for AAPL on October 23, 2015.

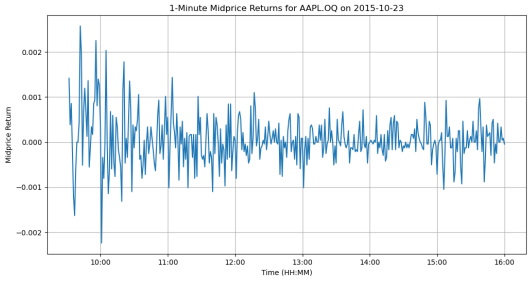

Figure 2: 1-minute mid-price returns for AAPL on October 23, 2015.

Quality control and data selection. The pipeline ensures that only trading days between `2015-01-01` and `2017-03-31` are included, as most tickers have no data beyond March 31, 2017. Missing days are detected using the NASDAQ calendar, and exception days can be removed if necessary. One notable exception is November 27, 2015, when the U.S. stock market experienced an unusually early closure due to Thanksgiving. Allowing this exception increases coverage from 17 complete tickers to 90 out of the 93 usable tickers. Only tickers with complete and validated series are copied to a dedicated folder for downstream analysis, producing a final dataset with fully aligned trading days. The SPY index data was fully available and also included in this dataset.

Output format. The cleaned and processed data for each selected ticker is stored as a single Parquet file, resulting in 90 files in total and 1 file for the SPY index data. Each file contains exactly 565 trading days with 389 minute-by-minute returns per day, for a total of 219,785 rows, including the timestamp and mid-price return. Parquet format provides efficient, compressed, and reproducible storage suitable for analytical workloads, forming the foundation for all subsequent steps of the project, including causal path analysis and backtesting of statistical arbitrage strategies. Initially, the raw dataset occupied approximately 55 GB, but after preprocessing and selection, we obtain 91 Parquet files of roughly 2 MB each.

3.2 Optimal Causal Path (OCP) Algorithm

The Optimal Causal Path (OCP) algorithm is employed to identify statistically stable intraday lead–lag relationships between pairs of stocks. Unlike traditional correlation-based approaches, OCP explicitly allows for time deformation between two return series, enabling the detection of both constant

and time-varying lags without assuming stationarity or cointegration. In this project, the OCP algorithm is used exclusively as a *pair selection mechanism* during a daily formation period, providing directional information that is subsequently exploited by the trading strategy.

Input Structure and Pair Universe The input to the OCP algorithm consists of minute-by-minute return matrices constructed from the preprocessed data described in Section 3.1. For each trading day D, a matrix

$$R_D \in \mathbb{R}^{T \times N_D}$$

is built, where $T = 389$ corresponds to the number of 1-minute returns during regular trading hours, and N_D denotes the number of stocks with complete data available on that day. Due to occasional missing trading days for individual stocks, N_D may vary across days but typically remains close to the full universe of 90 selected tickers.
For each day D, all unordered stock pairs are generated from the available tickers, yielding

$$\binom{N_D}{2}$$

candidate pairs. In the typical case where $N_D = 90$, this corresponds to approximately 4,005 pairs per day. The OCP algorithm is applied independently to each of these pairs using only data from the formation day.
Each stock pair is represented by two aligned return series

$$x = (x_1, \ldots, x_T), \qquad y = (y_1, \ldots, y_T),$$

where x_t and y_t denote the minute-by-minute returns of the two stocks at time t.

Step A: Constant Lag Estimation As an initial approximation, a constant lag between the two return series is estimated. For a candidate lag $l \geq 0$, the following cost function is defined:

$$c(l) = \sum_{i=1}^{T-l} |x_{i+l} - y_i| .$$

The optimal initial lag $\hat{l}_{\text{init}}$ is obtained by minimizing this cost over a bounded lag window:

$$\hat{l}_{\text{init}} = \arg \min_{l \in \{0, \ldots, L_{\max}\}} c(l),$$

where $L_{\max} = 30$ minutes. This bound reflects the assumption that exploitable intraday lead–lag effects occur at short time horizons and also serves to reduce sensitivity to noise. The estimated constant lag provides a robust initialization for the subsequent dynamic programming stage and significantly reduces computational complexity.

Step B: Optimal Causal Path with Variable Lag While Step A assumes a constant lag, intraday lead–lag relationships are often time-varying. To capture this behavior, Step B computes an optimal causal alignment path between the two return series.
A causal path is defined as a sequence of index pairs

$$p = \{(n_i, m_i)\}_{i=1}^{I},$$

which aligns x_{n_i} with y_{m_i} subject to the following constraints:

- **Boundary conditions:** $(n_1, m_1) = (1, 1)$ and $(n_I, m_I) = (T, T)$,
- **Monotonicity:** $n_{i+1} \geq n_i$ and $m_{i+1} \geq m_i$,
- **Causality and step size:** $(n_{i+1} - n_i, m_{i+1} - m_i) \in \{(1, 0), (0, 1), (1, 1)\}$.

The total cost associated with a path is given by

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

The optimal causal path is obtained via dynamic programming by minimizing $c_p(x, y)$ subject to the above constraints. To ensure scalability across thousands of daily stock pairs, the search space is restricted to a band around the diagonal shifted by the initial lag estimate $\hat{l}_{\text{init}}$. This banded dynamic programming approach allows moderate deviations from the constant lag while preventing spurious alignments and ensuring computational feasibility.

Step C: Lag Estimation and Stability Criterion Once the optimal causal path has been identified, the effective lag along the path is summarized by its empirical mean and dispersion. The average lag is defined as

$$\hat{l} = \frac{1}{I} \sum_{i=1}^{I} (n_i - m_i),$$

and the lag fluctuation is quantified by

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

The estimated lag $\hat{l}$ determines the leader–follower relationship within the pair: a positive value indicates that stock x tends to lead stock y, while a negative value indicates the opposite. The fluctuation σ_l serves as a stability measure: low values correspond to persistent lead–lag relationships, whereas high values indicate unstable or noisy alignments. Only pairs with a non-zero estimated lag and sufficiently low σ_l are retained for trading consideration.

Daily Pair Selection and Output Format For each formation day, the OCP algorithm is applied to all candidate stock pairs. The resulting pairs are ranked according to increasing lag fluctuation σ_l, and the ten most stable pairs are selected. The output of the OCP stage for each trading day is a structured table containing, for each selected pair:

- the leader stock,
- the follower stock,
- the estimated lag $\hat{l}$,
- the lag fluctuation σ_l.

This daily output constitutes the sole input to the trading strategy described in Section 3.3. Importantly, the OCP algorithm itself does not generate trading signals; it only identifies directional relationships that are subsequently exploited by the strategy.

3.3 Distribution of Causal Lags

To validate the dynamic nature of the OCP algorithm, we analyze the distribution of the estimated optimal lags ($\hat{l}$) across all traded pairs (Figure 3).
As shown in the histogram, the distribution is heavily concentrated in the 1-to-2 minute range. This finding suggests that while the OCP algorithm allows for flexible time warping, the most statistically significant arbitrage opportunities in the S&P 100 are extremely short-lived. The rapid decay of these lags reinforces the need for a high-frequency execution infrastructure, as the echo of the leader's movement dissipates almost immediately. The presence of a thin tail extending beyond 5 minutes confirms that OCP occasionally identifies longer-term structural delays, but these are rare compared to the dominant immediate reactions.

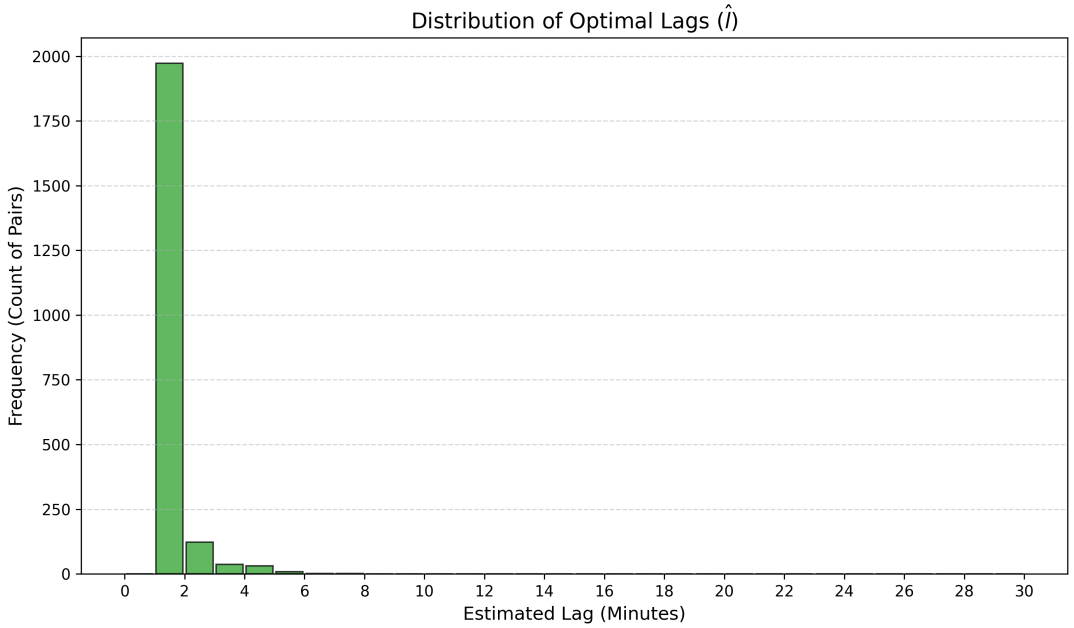

Figure 3: Distribution of Optimal Lags ($\hat{l}$) identified by OCP. The distribution is heavily right-skewed, indicating that most exploitable lead-lag relationships occur at very short horizons (1–2 minutes).

3.4 Trading Strategy Implementation

To evaluate the economic value of the causal relationships identified by the OCP algorithm, we implement a statistical arbitrage strategy. The strategy operates on a rolling daily basis: for every trading day D, the universe of tradable pairs is defined exclusively using data from the formation day $D-1$. This strict separation prevents look-ahead bias in the pair selection process.
The trading logic relies on a "Leader-Follower" mechanism. We hypothesize that significant idiosyncratic shocks to the "Leader" stock will eventually be reflected in the "Follower" stock after a time lag $\hat{l}$ estimated by the OCP algorithm.

3.4.1 Signal Generation

For each minute t on trading day D, we monitor the returns of the Leader stock, denoted as r_t^L. We compute a rolling mean μ_t^L and standard deviation σ_t^L over a lookback window of $w = 20$ minutes. A trading signal is triggered if the Leader's return exhibits a significant anomaly, defined as a deviation exceeding $k = 2.5$ standard deviations from the rolling mean:

$$|r_t^L - \mu_t^L| > k \cdot \sigma_t^L \tag{1}$$

Additionally, to filter out negligible price movements that would likely be eroded by transaction costs, we impose a minimum economic threshold constraint: $|r_t^L| > r_{\min}$, where $r_{\min}$ is set to 4 basis points (0.04%).

3.4.2 Market-Neutral Execution

Upon detecting a positive shock in the Leader ($r_t^L > \mu_t^L + k\sigma_t^L$), we assume the Follower stock is temporarily undervalued due to the response latency. To exploit this while neutralizing broad market risk, we enter a market-neutral pair trade:

- **Long** the Follower stock (r^F).
- **Short** the Market Index (r^I, represented by SPY).

Conversely, if the Leader suffers a negative shock, we Short the Follower and Long the Index. To ensure realistic execution, trades are entered at the open of the subsequent minute $(t+1)$, acknowledging that the signal calculated at time t uses close-to-close returns and is only actionable afterwards.

3.4.3 Dynamic Reaction Window

A critical innovation of this framework is the use of OCP metrics to define a dynamic exit schedule. Unlike fixed-horizon strategies, the holding period is tailored to the specific stability of the pair. The expected reaction window W is defined as:

$$W = \left[\hat{l} - z\sigma_l, \quad \hat{l} + z\sigma_l\right] \tag{2}$$

where $\hat{l}$ is the estimated lag and σ_l is the lag fluctuation derived from the OCP path. We use $z \approx 2.8$ (corresponding to a 99.5% confidence interval).
The position is closed when either of the following conditions is met:

1. **Profit Take:** The cumulative return of the hedged position $(r^F - r^I)$ exceeds the target threshold $r_{\min}$ within the reaction window.
2. **Time Stop:** The time limit $(\hat{l} + z\sigma_l)$ expires without the target return being met. In this case, the position is liquidated immediately to free capital.

This mechanism ensures that the strategy allows more time for "looser" causal links (high σ_l) to converge, while quickly discarding failing signals for "tight" links.

4 Results

We evaluate the performance of the OCP-based statistical arbitrage strategy over the period from January 2015 to March 2017. To isolate the predictive power of the causal paths identified by the OCP algorithm, we first analyze the strategy's performance on a gross basis (excluding transaction costs).

4.1 Aggregate Performance

Table 1 summarizes the key performance indicators of the strategy. Over the backtesting period, the strategy generated a cumulative return of 34.49%, significantly outperforming a market-neutral baseline. The Annualized Sharpe Ratio of 1.73 indicates a strong risk-adjusted return, suggesting that the "Leader-Follower" signals provide a genuine statistical edge.

Table 1: Strategy Performance Metrics (Gross, Jan 2015 – Mar 2017)

Metric	**Value**
Total Return	34.49%
Annualized Sharpe Ratio	1.73
Max Drawdown	-5.39%
Daily Win Rate	46.85%
Total Trades	9,969

An interesting observation is the Daily Win Rate of 46.85%. Although the strategy wins less than half the time, it remains highly profitable. This positive skew implies that the average gain from a successful OCP convergence is significantly larger than the average loss from a failed signal. This validates that the OCP algorithm is capturing substantial mean-reversion opportunities rather than random noise.

4.2 Equity Curve and Risk Analysis

Figure 4 illustrates the cumulative profit and loss (PnL) of the strategy. The equity curve demonstrates consistent growth with limited volatility, validating the effectiveness of the market-neutral hedging approach (Long Follower / Short SPY). The steady upward trend confirms that the alpha is not driven by a few lucky outliers but by a persistent edge.

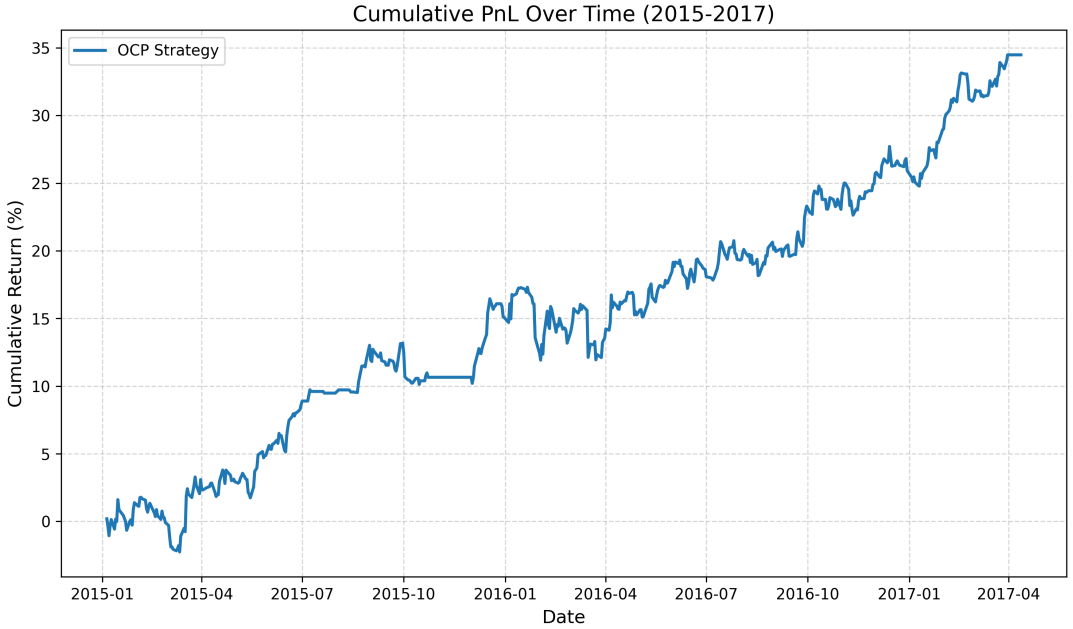

Figure 4: Cumulative returns of the OCP Strategy (Gross). The upward trajectory indicates consistent alpha generation over the 2-year period.

The risk profile, shown in Figure 5, remains stable throughout the period, with a maximum drawdown of only -5.39%. This low drawdown is a direct result of the dynamic reaction window mechanism, which quickly liquidates positions that fail to converge within the expected OCP lag time ($\hat{l} + z\sigma_l$).

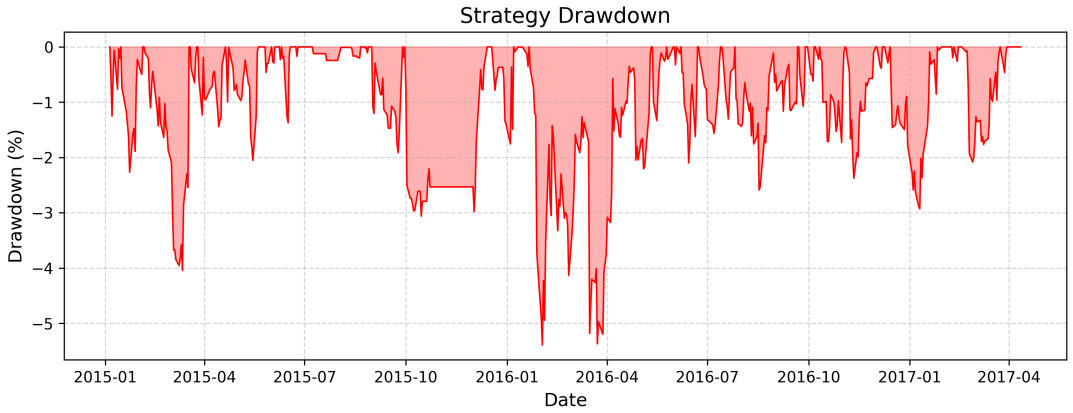

Figure 5: Drawdown profile over the trading period. The strategy avoids deep losses, highlighting effective risk management.

5 Discussion

5.1 Signal Efficacy vs. Transaction Costs

While the gross performance confirms the predictive power of the OCP algorithm, the practical implementation faces challenges due to execution friction. Our backtest recorded a total of 9,969 trades, implying an average turnover of approximately 18 trades per day.
When applying a realistic transaction cost model of 4 basis points per trade, the net performance drops significantly to -364.27%. This disparity highlights a classic trade-off in high-frequency statistical arbitrage: the signal is strong (Sharpe 1.73), but the current entry threshold ($r_{\min} = 4$ bps) is too sensitive. The strategy captures many small, profitable moves that are ultimately drowned out by the cost of crossing the bid-ask spread.

5.2 Sensitivity Analysis: Optimizing Signal Quality

To explore the trade-off between signal capture and execution volume, we performed a sensitivity analysis by varying the entry threshold ($r_{\min}$) from 4 to 20 basis points (Figure 6).

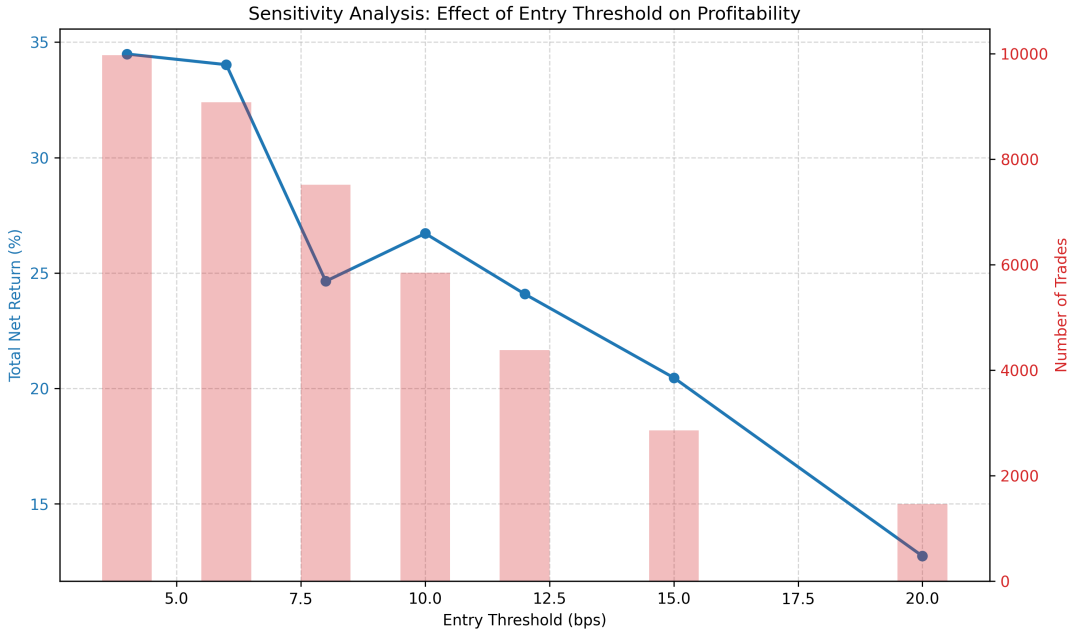

Figure 6: Sensitivity of Gross Return (blue line) and Trade Volume (red bars) to the entry threshold ($r_{\min}$). While higher thresholds reduce the total cumulative return, they reduce trade volume at a significantly faster rate, implying an improvement in average profit per trade.

As illustrated in Figure 6, there is a non-linear relationship between the threshold and strategy performance. Increasing the threshold from 4 bps to 15 bps reduces the trade count by approximately 70% (from ∼10,000 to ∼3,000), whereas the Gross Return only declines by approximately 45%.
This divergence is critical. It indicates that a large portion of the strategy's volume comes from marginal, low-quality signals that contribute disproportionately to transaction costs while adding little to the bottom line. By raising the threshold to 15 bps, the strategy effectively filters out this noise, isolating the high-conviction opportunities that are robust enough to survive transaction fees.

6 Conclusion and Future Work

6.1 Conclusion

This project explores the application of the Optimal Causal Path (OCP) algorithm to intraday statistical arbitrage within the S&P 100 data. By replacing traditional correlation-based metrics with a

dynamic time-warping framework, we aimed to capture non-linear lead-lag relationships that persist despite market efficiency.

Our results provide strong empirical evidence that the OCP algorithm successfully identifies predictive causal links. The strategy achieved a Gross Total Return of 34.49% and an Annualized Sharpe Ratio of 1.73 over the 2015–2017 period, validating the core hypothesis that "Leader" price movements can predict "Follower" reactions. However, the analysis also highlighted the severe constraints of high-frequency execution. The strategy's high turnover (nearly 10,000 trades) resulted in a Net Return of -364.27% after accounting for transaction costs, demonstrating that the current signal sensitivity is economically unviable without optimization.

The distribution of optimal lags, heavily concentrated at the 1-minute mark, confirms that the S&P 100 is highly efficient, with information transmission occurring on timescales likely faster than the 1-minute resolution of our dataset. Nevertheless, the OCP framework proved robust in determining the direction of these causal links, providing a valuable directional filter for pairs trading.

6.2 Future Work

To bridge the gap between theoretical alpha and realizable profit, future research should focus on three key areas:

1. **Data Expansion:** We restricted our analysis to the S&P 100, a universe of highly liquid, efficient stocks where lags are minimal. Applying the OCP framework to small-cap indices or cross-asset pairs (e.g., ETFs vs. component stocks) may reveal longer, more exploitable lags due to lower liquidity and slower information diffusion.

2. **Execution and Resolution:** The current analysis aggregated high-frequency BBO data into 1-minute intervals to ensure computational feasibility. However, this aggregation compresses the sub-minute lead-lag relationships visible in the raw data. Future work should utilize the original tick-level data to resolve these milliseconds-level lags. Additionally, replacing the current market-order assumption with limit orders would capture the bid-ask spread rather than paying it, significantly reducing the transaction costs observed in this study.

References

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