Statistical Arbitrage Pair Trading Using Cointegration, Mean Reversion, and Bayesian Optimization

Hiu Chun (Sunny) Chan, Saurabh Kulkarni, Shaoxiong Yuan, Weiyi Wang

December 8, 2025

1 Introduction and Project Overview

This project develops and evaluates a statistical arbitrage strategy in the equity pairs trading framework, based on mean reversion and cointegration. Compared with directional strategies that attempt to predict the price path of a single asset, pairs trading emphasizes extracting excess returns from *relative mispricing* between two assets while hedging as much systematic risk as possible. The focus therefore shifts from asking whether an individual asset is over- or undervalued in absolute terms to asking whether the spread between two assets exhibits a significant deviation from its long-run equilibrium relationship.

On the asset selection side, the project first constructs a relatively stable equity universe based on the S&P 500 index. Using historical index constituents from 2016 to 2020, we retain only those firms that remain in the index throughout the entire period, in order to ensure persistent liquidity and representativeness of the underlying assets. We then rank these firms by market capitalization and select the top 50 stocks as the base universe for subsequent pair construction and modeling. This design preserves sufficient cross-sectional breadth while controlling the number of candidate combinations, which is essential for systematic screening and backtesting.

Building on this universe, we implement an end-to-end statistical arbitrage pipeline comprising the following key modules:

1. **Candidate pair construction and statistical screening.** From the top-50 universe, we generate all possible bivariate stock pairs and compute, for each pair, the correlation coefficient, the spread series, and its residual process. We then apply the Engle–Granger cointegration test and the augmented Dickey–Fuller (ADF) unit-root test sequentially, in order to identify candidate trading pairs that jointly exhibit high correlation, statistically significant cointegration, and stationarity of the spread. The goal of this step is to reduce the initially large combination space to a subset of asset pairs with a well-defined long-run equilibrium relationship that is suitable for mean-reversion trading.

2. **Characterizing mean reversion and estimating half-life.** For spread series passing the above filters, we model their mean-reversion dynamics using an Ornstein–Uhlenbeck (OU) process and obtain discrete-time estimates of the speed-of-mean-reversion parameter and the associated half-life. The half-life provides a time scale

with clear economic interpretation, which we use to discipline the maximum holding period of individual trades and to control the turnover of the strategy, thereby balancing expected returns against trading costs.

3. **Z-score based trading mechanism and backtesting framework.** After standardizing the spread and converting it into a z-score series, we design a transparent entry and exit mechanism. When the z-score exceeds pre-specified positive or negative entry thresholds, the strategy initiates long–short positions in the two legs with approximately dollar-neutral nominal exposure. Positions are closed when the z-score reverts towards its mean or when the holding period reaches an upper bound determined by a multiple of the half-life. The strategy is implemented in the Backtrader framework, explicitly incorporating a fixed-form trading cost assumption, and records for each trade the execution prices, holding horizon, and profit-and-loss profile to facilitate ex-post analysis and performance evaluation.

4. **Bayesian optimization of strategy parameters.** To avoid ad hoc choices of key parameters such as lookback window length, entry thresholds, stop-loss multipliers, and holding-time multipliers, we formulate the parameter selection problem as a Bayesian optimization task. Within a predefined parameter space, we maximize the Sharpe ratio over the sample period as the objective function and use sequential updates of a surrogate model to efficiently explore the space of parameter combinations. This yields strategy configurations that are more favorable from a risk-adjusted return perspective than naive hand-tuned settings.

In terms of sample splitting, the period from 2016 to 2020 is primarily used to identify cointegration relationships and estimate mean-reversion characteristics, and serves as the basis for pair screening and model specification. The period from 2021 to 2025 is reserved as the main out-of-sample backtesting window, used to assess whether the selected pairs and the strategy parameters obtained via Bayesian optimization can persistently exhibit statistical arbitrage properties under realized market conditions. The subsequent sections of the report focus on the optimized results and backtest performance of representative pairs, such as COST–NEE, and complement these with additional experiments on other pairs to discuss the robustness and limitations of the strategy.

2 Data and Sample Construction

2.1 Data Source and Sample Split

The empirical analysis in this project is based on daily closing prices for individual equities obtained from Refinitiv terminals and related data interfaces. The overall sample period spans from January 2016 to December 2025, with the following split:

- The years 2016–2020 are primarily used to identify cointegration relationships, estimate mean-reversion properties of spreads, and serve as the basis for pair screening

and model specification.

- The years 2021–2025 are reserved as the core backtesting window, in which the strategy's performance and risk characteristics are evaluated on realized historical price paths.

This split balances the need for a sufficiently long history with the requirement of a clear temporal separation between the "model design" phase and the "strategy evaluation" phase, thereby mitigating look-ahead bias and reducing the risk of overfitting.

2.2 Equity Universe and Pair Space

For the selection of underlying assets, the project constructs a base equity universe from the S&P 500 index. Specifically, we aggregate index constituents over the 2016–2020 period and retain only those firms that remain in the index throughout the entire five-year interval. This filter ensures that sample firms exhibit relatively stable listing status and liquidity conditions over the research horizon, and reduces structural distortions caused by frequent entries and exits from the index.

Given this stable constituent set, we then rank firms by market capitalization and select the top 50 stocks as the base universe for subsequent pair construction and strategy implementation. On the one hand, this size provides a sufficiently rich cross-sectional set of candidates, ensuring that the search for cointegrated pairs is statistically meaningful. On the other hand, keeping the universe at a moderate scale helps control computational burden. With 50 stocks, we obtain $\binom{50}{2} = 1225$ unique stock pairs, which is small enough to make large-scale pair screening, backtesting, and parameter optimization operationally feasible while still allowing meaningful cross-sectional exploration.

All possible bivariate stock pairs generated from this top-50 universe form the initial candidate pair set. The subsequent steps—correlation-based filtering, cointegration testing, and mean-reversion diagnostics—are applied within this set and progressively shrink the universe of tradable pairs to those with more refined statistical and economic justification.

2.3 Data Preprocessing and Quality Control

The preprocessing stage involves several data quality and consistency checks:

1. **Trading day alignment and cross-sectional synchronization.** All stock price series are aligned to a common trading calendar. Non-trading days are removed, and missing observations across the cross-section are handled in a consistent manner, ensuring that prices for the two legs of a pair are directly comparable at each time point.

2. **Handling missing values and outliers.** Isolated missing records arising from suspensions, technical issues, or data-interface errors are treated via forward filling or, when necessary, by excluding the affected dates from the analysis. Extreme price

jumps or evidently erroneous entries are identified and corrected or removed, in order to reduce their impact on cointegration tests and mean-reversion estimation.

3. **Consistency checks on price series.** Prior to constructing log returns and spread series, we conduct basic consistency checks on raw closing prices to ensure the absence of duplicate records, date misalignment, or currency mismatches. This step is essential for the reliability of subsequent statistical tests and backtesting results.

After these preprocessing and quality-control steps, we obtain a data set covering 2016–2025 with high continuity and internal consistency. This data set forms the basis for the pair screening, mean-reversion modeling, and strategy backtesting analyses conducted in the remainder of the project.

3 Methodology and Model Construction

3.1 Candidate Pair Generation and Statistical Screening

After constructing the equity universe and the underlying data set, the first step is to generate the pair space from the top-50 stock universe. Concretely, we enumerate all possible bivariate combinations, yielding 1225 candidate pairs. For each pair, we work with log closing prices to construct a spread series and compute the within-sample linear correlation coefficient, which allows us to discard pairs that exhibit little or no co-movement in their price dynamics.

Following this preliminary correlation filter, we subject the remaining pairs to cointegration and stationarity tests. Cointegration is assessed using the Engle–Granger two-step procedure: we regress the log price of one stock on the log price of the other, treat the regression residuals as an estimate of the equilibrium spread, and then apply an ADF unit-root test to these residuals. Rejection of the unit-root null is interpreted as evidence that the residuals are stationary and that the two price series are cointegrated. Stationarity of the constructed spread itself is further evaluated by applying the augmented Dickey–Fuller (ADF) test directly to the spread series as a robustness check. Only those pairs that simultaneously satisfy three conditions—sufficiently high correlation, statistically significant cointegration, and stationarity of the spread—are retained as candidates that are statistically suitable for mean-reversion trading.

This sequence of statistical filters compresses the original pair space into a smaller but higher-quality set of tradable pairs, which provides the foundation for the subsequent mean-reversion modeling and strategy design.

3.2 Mean-Reversion Modeling and Half-Life Estimation

For the spread series that pass the statistical screening stage, we model their mean-reversion behavior using an Ornstein–Uhlenbeck (OU) process. In continuous time, the dynamics of

the spread X_t can be written as

$$dX_t = \kappa(\mu - X_t)\,dt + \sigma\,dW_t,$$

where μ denotes the long-run equilibrium level, $\kappa > 0$ is the speed of mean reversion, σ is the volatility parameter, and W_t is a standard Brownian motion.

In the empirical implementation, we first construct the log price spread $s_t = \log P_t^{(1)} - \log P_t^{(2)}$ from the two assets' closing prices, aligned on common trading dates. We then approximate the OU dynamics by running a discrete-time regression of spread changes on the lagged spread,

$$\Delta s_t = \alpha + \beta s_{t-1} + \varepsilon_t,$$

and interpret the slope coefficient as $\beta \approx -\kappa$. In practice we set $\kappa = -\beta$ as the estimated speed of mean reversion. Given κ, we define the half-life of mean reversion as the time it takes for a deviation from equilibrium to decay by half,

$$t_{1/2} = \frac{\ln 2}{\kappa} = -\frac{\ln 2}{\beta}.$$

The half-life parameter admits a direct economic interpretation: it characterizes how quickly deviations of the spread from its equilibrium are expected to dissipate. As such, it provides an objective benchmark for setting the holding horizon of individual trades. A very short half-life suggests rapid mean reversion, which may induce excessively high trading frequency and amplify transaction costs; a very long half-life implies sluggish mean reversion, tying up capital for extended periods and reducing capital efficiency. In practice, we use the estimated half-life as a key input when specifying the maximum holding period in the trading rules and, through subsequent parameter optimization, seek a balance between mean-reversion strength and trading costs.

Once the mean-reversion parameters have been estimated, the spread series is standardized into a z-score series to serve as the basis for generating trading signals. This normalization facilitates consistent comparison across different pairs and assigns a clear numerical interpretation to entry and exit thresholds.

3.3 Trading Rules and Backtesting Framework

At the signal level, the strategy employs a z-score threshold mechanism to define trading rules. When the absolute value of the z-score for a given pair exceeds a pre-specified entry threshold, the strategy initiates offsetting long and short positions in the two stocks, maintaining an approximately dollar-neutral exposure to hedge out broad market movements. The underlying economic premise is that the spread is likely to revert towards its long-run equilibrium, so that profits can be realized as the spread narrows. Positions are closed either when the z-score reverts to a neighborhood around zero or when the holding period reaches an upper bound determined by a multiple of the estimated half-life, at which point the trade's realized return and risk profile are recorded.

On the implementation side, we build a unified backtesting engine on top of the Backtrader framework. The engine is responsible for loading the preprocessed price data, converting them into time series compatible with the backtesting platform, injecting the custom pairs trading strategy, and setting the initial capital and transaction cost parameters. After the strategy run, the engine produces a collection of performance metrics, including final portfolio value, total return, annualized Sharpe ratio, maximum drawdown, number of trades, win rate, and the distribution of daily returns. It also generates visual diagnostics such as the equity curve and drawdown curve to provide an intuitive assessment of strategy behavior.

On top of this infrastructure, we introduce a Bayesian optimization module that treats key hyperparameters—such as lookback window length, z-score entry thresholds, stop-loss multipliers, and holding-time multipliers—as decision variables to be tuned. Using the Sharpe ratio over the in-sample window as the objective function, the optimizer searches the parameter space in an automated manner. This procedure yields strategy configurations that exhibit more robust performance for a given pair and forms the basis for the empirical results reported for representative pairs in the subsequent sections.

4 Empirical Results

This section presents the empirical findings from our full pipeline, beginning with a universe-level diagnostic of all stock pairs generated from the top-50 constituents of the S&P 500. After ranking pairs by their cointegration significance, correlation strength, and stationarity properties, we identify a subset of statistically suitable candidate pairs. We then highlight the strongest pairs and subsequently perform a detailed optimization and backtesting study on the most promising one.

4.1 Summary of Top Candidate Pairs

Table 1 reports the top ten statistically suitable pairs based on low cointegration p-values, high correlation, and strong ADF stationarity. These pairs satisfy all filtering criteria and collectively provide a diverse set of mean-reverting spread relationships for further investigation.

Table 1: Top 10 Suitable Stock Pairs Based on Cointegration, Correlation, and Stationarity

Pair	Cointegration p-value	Correlation	ADF p-value	Mean Z-score	Std Z-score
COST.OQ–NEE.N	0.000033	0.981836	0.000037	1.35×10^{-16}	1.0
V.N–ABT.N	0.017675	0.979157	0.008362	4.97×10^{-16}	1.0
HD.N–ACN.N	0.012285	0.974604	0.001419	-2.82×10^{-16}	1.0
TMO.N–COST.OQ	0.042744	0.973373	0.030303	0.00	1.0
ABT.N–COST.OQ	0.014473	0.972814	0.014070	-1.35×10^{-15}	1.0
GOOG.OQ–ACN.N	0.024877	0.968758	0.011432	2.12×10^{-15}	1.0
WMT.N–NEE.N	0.006351	0.968013	0.024297	0.00	1.0
GOOG.OQ–HD.N	0.004690	0.966925	0.001005	1.06×10^{-15}	1.0
GOOGL.OQ–ACN.N	0.030324	0.966247	0.012596	-1.54×10^{-15}	1.0
CRM.N–ABT.N	0.017416	0.966218	0.010661	1.81×10^{-16}	1.0

Among these candidates, the COST–NEE pair stands out clearly: it exhibits the strongest cointegration signal (lowest p-value) and the highest correlation among all screened pairs. This combination indicates a robust mean-reversion structure, making it an ideal benchmark for demonstrating our full workflow of parameter optimization and out-of-sample strategy evaluation.

4.2 Bayesian Optimization Results for the COST–NEE Pair

Within the universe constructed from the top-50 S&P 500 constituents, the COST–NEE pair exhibits a pronounced and relatively stable mean-reversion structure over the 2016–2020 period after passing correlation, cointegration, and stationarity filters. We therefore use this pair as a benchmark to showcase the full pipeline.

On the strategy side, we perform Bayesian optimization for the COST–NEE pair on the 2016–2020 in-sample window, using the Sharpe ratio as the objective function. The optimizer explores the predefined parameter space for key strategy hyperparameters and yields a representative optimal configuration with the following characteristics:

- Lookback window length of approximately 28 trading days;
- z-score entry threshold of approximately 1.5 in absolute value;
- Stop-loss multiplier of approximately 3.0;
- Holding-time multiplier of approximately 1.6, combined with the estimated half-life to cap maximum holding duration;
- Estimated half-life of approximately 20.5 trading days.

Under this parameter configuration, the optimized in-sample Sharpe ratio is around 1.6, consistent with the mean-reversion properties revealed by the earlier statistical tests. A concise summary of the resulting optimized parameters and in-sample performance statistics is reported in Listing 1, while Figure 1 shows the corresponding in-sample equity curve and drawdown over the 2016–2020 training window.

Listing 1: Bayesian optimization results for the COST–NEE pair

```
Stock Pair for Pair Trading:
Stock 1: COST.OQ
Stock 2: NEE.N
Kappa: 0.0338
Half-Life: 20.51

Best parameters found:
lookback: 28
entry_threshold: 1.5011681487
stoploss_factor: 2.9883173389
holding_time_factor: 1.6174815096
half_life: 20.5111349853
Best Sharpe Ratio: 1.5805
```

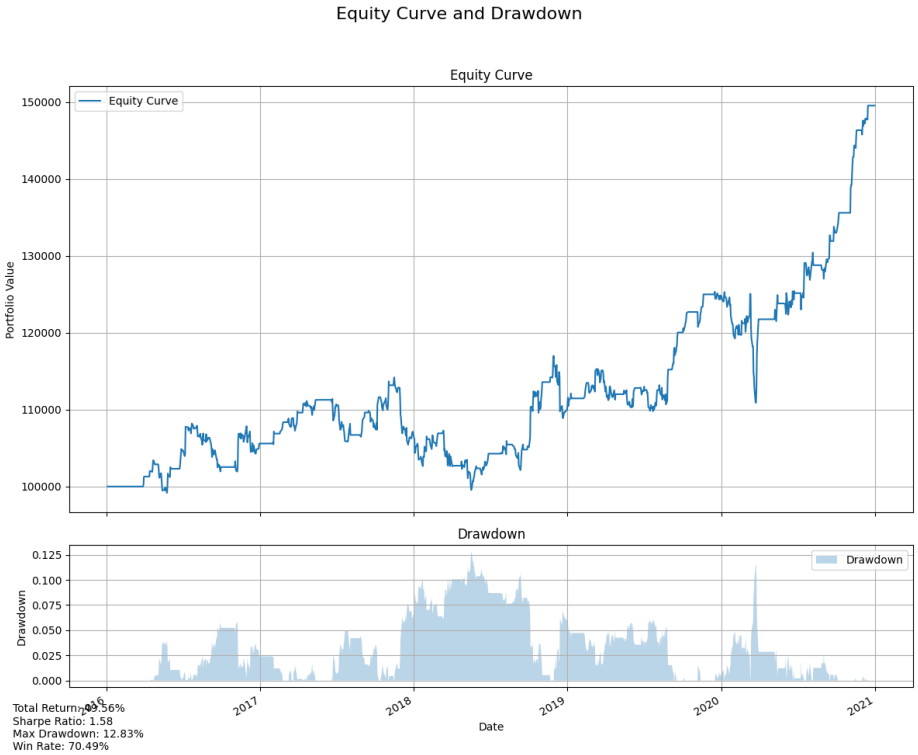

Figure 1: In-sample equity curve and drawdown for the COST–NEE strategy over 2016–2020 under the Bayesian-optimized parameter configuration.

4.3 Out-of-Sample Backtest Performance for COST–NEE

Using the optimized parameter configuration above, we conduct a full out-of-sample backtest of the COST–NEE strategy over the 2021–2025 period. Starting from an initial capital of 100,000, the key performance statistics can be summarized as follows:

- Final portfolio value of approximately 106,277, corresponding to a total return of about 6.28%;

- Annualized Sharpe ratio of approximately 1.69, indicating a moderate level of risk-adjusted performance;
- Maximum drawdown of about 29.3%, reflecting significant equity drawdowns during episodes of persistent spread dislocation;
- A total of 57 trades, of which 37 are profitable and 20 unprofitable, yielding a win rate of roughly 64.9%;
- Mean daily return of approximately 0.09% with a standard deviation of about 0.81%, consistent with the reported Sharpe ratio.

These statistics are reported in Table 2.

Table 2: Out-of-sample performance metrics for the COST–NEE pair (2021–2025). Initial capital is 100,000.

Metric	Value
Initial capital	100,000
Final portfolio value	106,277
Total return	6.28%
Annualized Sharpe ratio	1.69
Maximum drawdown	29.27%
Total number of trades	57
Winning trades	37
Losing trades	20
Win rate	64.91%
Mean daily return	0.09%
Standard deviation of daily returns	0.81%

From the equity curve and drawdown profile, the strategy exhibits a gradual upward trend over the out-of-sample period, punctuated by several episodes of medium-sized drawdowns. The bulk of profits arises from a number of phases in which the spread departs from and then reverts towards its equilibrium level, while the largest drawdowns are concentrated in windows where the spread remains dislocated for an extended period and mean reversion occurs more slowly than anticipated.

At the trade level, most positions close within a horizon consistent with one to several estimated half-lives, which is in line with the OU-based mean-reversion estimates. A small number of trades exhibit substantially longer holding periods, typically associated with persistent spread deviations; these trades are major contributors to the maximum drawdown and highlight aspects of risk management and threshold selection that merit further refinement.

An example of the out-of-sample equity curve and associated drawdown for the COST–NEE strategy is shown in Figure 2. Trade-level examples illustrating the timing and profitability of individual mean-reversion episodes can also be visualized.

Overall, the COST–NEE pair delivers a moderate total return and a reasonably attractive Sharpe ratio in the out-of-sample period, but at the cost of non-negligible tail risk. This

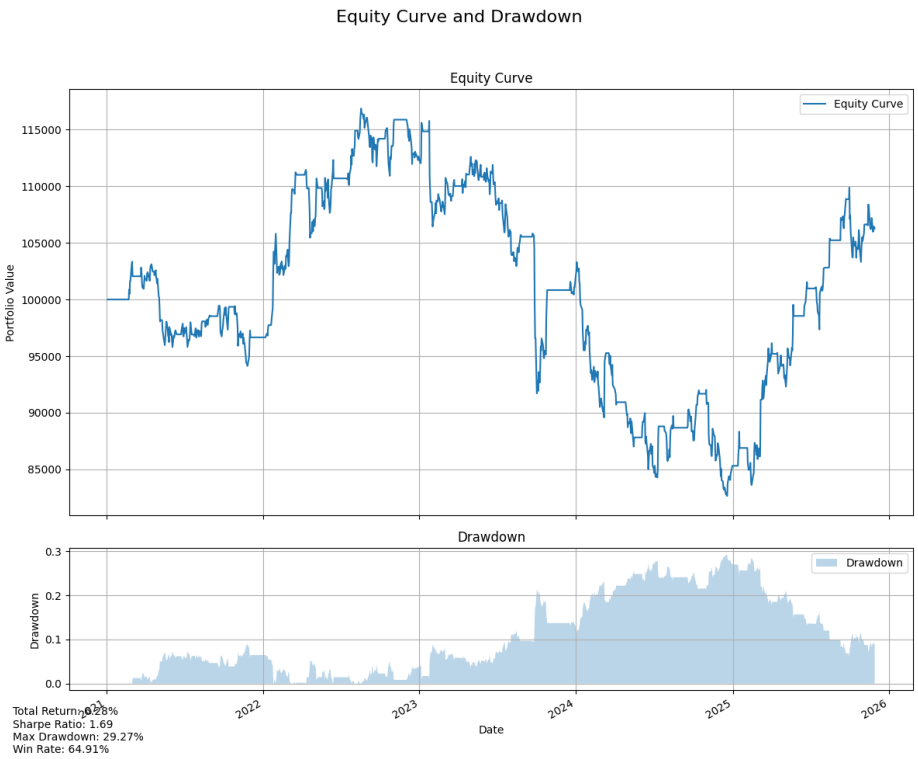

Figure 2: Out-of-sample equity curve and drawdown for the COST–NEE strategy over 2021–2025.

evidence supports the applicability of the cointegration and OU-based mean-reversion framework for this particular pair, while also underscoring the limitations of relying on a single pair as the sole driver of portfolio-level performance.

4.4 Other Pairs and Robustness Analysis

To evaluate the robustness of the strategy across different underlying assets, we apply the same modeling and optimization pipeline to a subset of other pairs selected from the "suitable pairs" set. Each pair is subjected to the same sequence of mean-reversion modeling, Bayesian hyperparameter optimization, and out-of-sample backtesting.

The results reveal substantial heterogeneity in performance across pairs. Some pairs continue to exhibit strong mean-reversion behavior and attractive risk-adjusted returns out of sample, while others show weak or deteriorating performance, or even unfavorable risk–return profiles. This heterogeneity reflects both the time-varying nature of cointegration and mean-reversion relationships and the possibility of overfitting to a specific in-sample period.

As a representative example of a strongly performing pair under the proposed framework, the ACN–TXN pair achieves the following indicative performance under an appropriate parameter configuration:

- Total return of approximately 16.5%;

- Sharpe ratio of about 1.66;
- Win rate close to 90%;
- Maximum drawdown of approximately 8%.

These results demonstrate that, for pairs with more pronounced and stable mean-reversion structures, the same strategy framework can deliver considerably stronger risk-adjusted returns than in the COST–NEE case. At the same time, there are pairs that pass in-sample statistical tests but exhibit noticeably weaker or even failed performance out of sample. This highlights the inherent instability of cointegration and mean-reversion properties and cautions against relying solely on single-period statistical diagnostics when assessing long-run strategy viability.

Table 3 summarizes the performance of several representative pairs, including COST–NEE and ACN–TXN.

Table 3: Summary of out-of-sample performance for representative pairs (2021–2025).

Pair	Total Return (%)	Sharpe Ratio	Win Rate (%)	Max Drawdown (%)
COST–NEE	6.28	1.69	64.91	29.27
ACN–TXN	16.50	1.66	90.00	8.00

A graphical summary of out-of-sample performance across multiple pairs (for example, using bar charts of Sharpe ratios and drawdowns) can further highlight the dispersion in outcomes and the sensitivity of performance to pair selection, as illustrated in Figure 3.

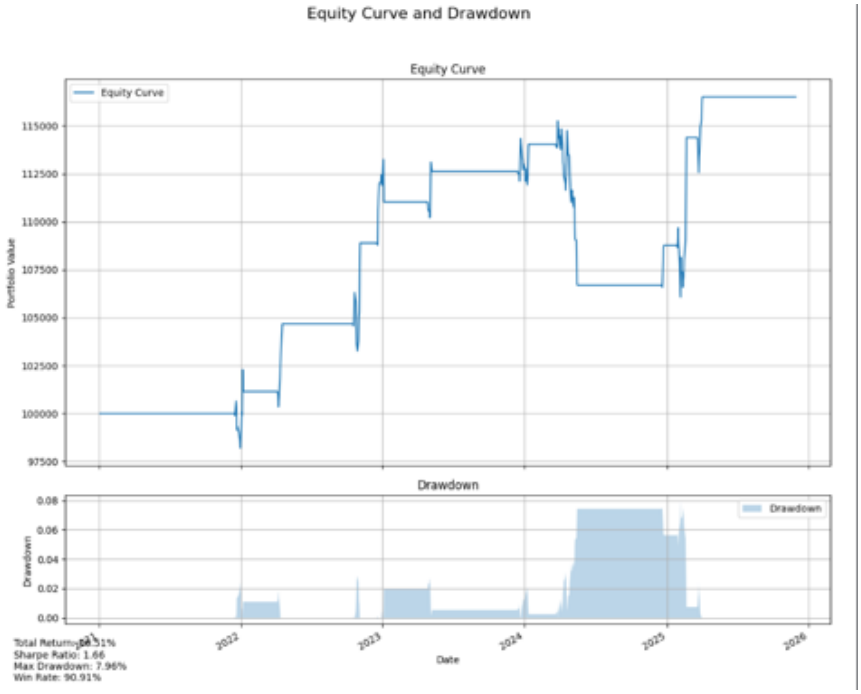

Figure 3: Out-of-sample performance across multiple pairs.

Taken together, the extended experiments across multiple pairs support the following conclusions. First, for assets exhibiting pronounced and relatively stable mean-reversion structures, a pairs trading strategy grounded in cointegration testing, OU modeling, and Bayesian optimization can deliver attractive out-of-sample risk-adjusted returns. Second, strategy performance is highly sensitive to pair selection and parameter configuration, and mean-reversion properties themselves are time-varying. In practical implementations, it is therefore more prudent to trade a diversified portfolio of carefully screened pairs and manage risk at the portfolio level, rather than relying on a single pair whose statistical properties may shift over time.

5 Conclusion

This project develops an end-to-end framework for statistical arbitrage in equity markets based on pairs trading. Starting from a stable subset of S&P 500 constituents over 2016–2025, we construct a pipeline that encompasses data preprocessing, pair selection via correlation and cointegration tests, mean-reversion modeling using Ornstein–Uhlenbeck dynamics, and a Backtrader-based backtesting engine augmented with Bayesian hyperparameter optimization. The empirical analysis demonstrates that, for pairs such as COST–NEE and ACN–TXN that exhibit pronounced and persistent mean-reversion characteristics, the strategy can generate positive out-of-sample risk-adjusted returns, with Sharpe ratios and win rates that are economically meaningful.

At the same time, the results highlight several important limitations. First, performance varies substantially across different pairs: some pairs that pass in-sample statistical diagnostics deliver weak or even adverse performance out of sample, underscoring the time-varying nature of cointegration and mean reversion and the sensitivity of the strategy to pair selection and parameter choices. Second, the current implementation operates at a daily frequency and employs relatively stylized assumptions about transaction costs and execution, without fully capturing liquidity frictions, slippage, and order execution constraints that would arise in a real-world trading environment.

Overall, the main contribution of this project lies in providing a concrete and reproducible empirical framework that integrates cointegration testing, OU-based mean-reversion modeling, and Bayesian optimization into a coherent pairs trading strategy applied to real data. Future extensions could further enhance robustness and practical relevance by: trading a diversified portfolio of multiple pairs with dynamic capital allocation at the portfolio level; incorporating multivariate cointegration techniques such as the Johansen test to capture richer long-run relationships; adopting more detailed models of trading costs, liquidity, and execution; and combining the existing mean-reversion structure with machine learning methods for short-horizon spread forecasting. These directions may help improve both the stability and the implementability of statistical arbitrage strategies in practice, while maintaining a disciplined approach to model risk and overfitting.