[Working paper] Statistical Arbitrage in U.S. Equities: Baseline and Robustness Study

Anirudh Krovi[1]

[1]PhD, Northwestern University; MBA, NYU Stern; Formerly at McKinsey & Company, anirudh.krovi@stern.nyu.edu

August 8, 2025

Abstract

We present a conventional statistical arbitrage framework applied to a broad U.S. equities universe, combining sector-relative residual signals with tail-based selection, sector momentum confirmation, and equal-weight portfolio construction. The framework is modular, allowing transparent substitution of signal, execution, and risk-control components. In a 726-day out-of-sample period, the strategy delivers an annualized return of 23.3% with 18.8% volatility, a Sharpe ratio of 1.21, and a maximum drawdown of 17.9%. Robustness checks across entry/exit thresholds and maximum holding periods show performance stability and consistent trade win rates, supporting the reliability of the design. Extensions to broader sector coverage, finer market-cap stratification, and alternative cost and execution models present natural avenues for further development.

1 Introduction

Statistical arbitrage (stat-arb) is a class of trading strategies that aim to profit from temporary mispricings between related securities (Gatev et al., 2006; Avellaneda and Lee, 2010). These strategies are usually market-neutral and assume that certain cross-sectional price relationships—based on fundamentals, factor exposures, or statistical patterns—will revert toward an equilibrium.

Implementations vary widely. Some focus on simple stock pairs (Gatev et al., 2006), while others use large cross-sectional models with multiple predictive signals (Avellaneda and Lee, 2010). Results in practice depend on choices about transaction cost modelling, execution assumptions, and the definition of the trading universe (Lehmann, 1990; Moskowitz et al., 2012).

In this study, the scope is kept narrow for clarity and speed of testing. We exclude sectors with very distinct risk profiles, such as large-cap technology and market-making firms. We also remove the largest index constituents to avoid benchmark effects and crowding. For the prototype, we use ten liquid stocks from different sectors. This keeps calculations fast and allows for rapid iteration before scaling to a larger universe.

A central element of any mean-reversion strategy is how "fair value" is defined. Approaches in the literature include moving averages, factor-model residuals, and cointegration-based estimates. Here, we use the rolling mean of the relevant sector ETF. This simple and transparent measure fits the sector-neutral design, smooths idiosyncratic noise, and provides a clear baseline for later comparisons.

The rest of the paper is organised as follows. Section 2 reviews related work on statistical arbitrage, mean-reversion modelling, and portfolio design. Section 3 describes the dataset and preprocessing. Section 4 explains the modelling framework, including signals, portfolio formation, and cost treatment. Section 5 presents results and robustness checks. Section 6 offers closing remarks.

2 Related Literature

Early work on relative-value strategies in equities focused on pairs of stocks. Gatev et al. (2006) showed that matching stocks with similar historical price patterns and trading when their prices diverged could generate abnormal returns over long periods. This approach is easy to understand but limited in scale because only a small number of pairs are suitable at any time.

Later studies extended the idea to larger portfolios. Avellaneda and Lee (2010) proposed a cross-sectional framework that measures each stock's return after removing market, sector, and style effects. Portfolios are then built from these residuals, subject to risk and neutrality constraints. This work highlights the main trade-offs in stat-arb: stronger signals can mean less diversification, while larger universes can increase transaction costs.

Research on return predictability also feeds into stat-arb design. Lehmann (1990) found that short-term return reversals are common, supporting mean-reversion strategies. Moskowitz et al. (2012) documented momentum effects over longer horizons, showing that both reversal and trend-following can exist in the same market depending on the time frame. In practice, many stat-arb strategies combine different types of predictors and control risk at the portfolio level.

A key choice in any mean-reversion strategy is how to define "fair value." Common methods include:

- Moving averages or other statistical filters (Poterba and Summers, 1988; Lehmann, 1990).
- Residuals from cross-sectional regressions (Avellaneda and Lee, 2010).
- Cointegration-based price relationships (Elliott et al., 2005; Gatev et al., 2006).
- Microstructure-based estimates in high-frequency trading.

In this paper, fair value is the rolling mean of the relevant sector ETF. This simple choice matches the sector-neutral design, smooths short-term noise, and is easy to update.

Portfolio construction choices also affect results. Large-cap technology stocks and index heavyweights can dominate risk and are heavily influenced by benchmark flows (Grinold and Kahn, 2019; Khandani and Lo, 2007). Market-making firms behave differently from typical equities and are excluded. As in many academic (e.g., Jegadeesh and Titman, 1993) and practitioner studies, we start with a small, liquid, and diversified universe to allow faster testing before scaling up.

3 Data

The trading universe is constructed to balance liquidity, sectoral diversity, and tractability during model development. Six sector ETFs from the S&P 500 are selected as anchors: Industrials (XLI), Energy (XLE), Utilities (XLU), Financials (XLF), Health Care (XLV), and Materials (XLB). These sectors are chosen to avoid idiosyncratic risk concentrations, such as those present in large-cap technology or market-making entities, while still representing a broad cross-section of the economy. Highly visible mega-cap stocks are excluded to reduce the influence of index membership effects and crowding from benchmark-driven flows. This choice also limits portfolio risk from single-name dominance and focuses the strategy on more idiosyncratic cross-sectional relationships.

From each sector ETF, the top ten constituents by contribution to the index are selected, yielding a 60-stock universe. This approach ensures that the chosen names are both liquid and representative of sector performance, while still allowing for a manageable dataset during early-stage modelling. The final universe is:

- **XLI – Industrials:** GE, RTX, CAT, UBER, GEV, BA, ETN, HON, UNP, DE
- **XLE – Energy:** XOM, CVX, COP, WMB, EOG, KMI, MPC, SLB, PSX, OKE
- **XLU – Utilities:** NEE, CEG, SO, DUK, VST, AEP, SRE, D, EXC, PEG
- **XLF – Financials:** BRK-B, JPM, V, MA, BAC, WFC, GS, MS, C, SPGI
- **XLV – Health Care:** LLY, JNJ, ABBV, UNH, ABT, MRK, TMO, ISRG, AMGN, BSX
- **XLB – Materials:** LIN, SHW, NEM, ECL, APD, MLM, VMC, NUE, CTVA, FCX

Daily adjusted close prices for both stocks and their corresponding sector ETFs are sourced primarily from Stooq, using the `pandas_datareader` interface. For redundancy and validation, equivalent series are also obtained via the `yfinance` API. Vendor-provided adjustment factors are applied to account for splits and dividends. Small gaps (up to two consecutive trading days) are forward-filled to preserve continuity; securities with more substantial missing data over the sample are excluded entirely.

Each stock is mapped to its sector ETF, which serves two purposes: (i) to de-mean stock prices by sector in signal construction, and (ii) to define the "fair value" benchmark as the rolling mean of the sector ETF's price. This sector-relative framing ensures that the residual captures idiosyncratic, rather than broad sectoral or market-wide, movements.

Rationale for Top-10-by-Contribution Selection

Selecting the top contributors to each sector ETF index serves several goals:

1. **Liquidity and tradability:** High index-weight names are typically among the most liquid stocks in their sectors, ensuring minimal slippage and realistic cost modelling.
2. **Sector representativeness:** These names collectively explain most of the sector ETF's return and variance, making them natural candidates for sector-relative strategies.
3. **Signal purity:** By anchoring residuals to sector leaders, the strategy focuses on deviations within the sector's most influential names—where mean-reversion pressures are more likely to reflect idiosyncratic factors rather than noise from thinly traded constituents.
4. **Balanced exposure:** Equal selection across sectors prevents over-representation from particularly volatile or crowded sectors, maintaining the intended cross-sectional neutrality.

This construction ensures that the strategy tests its logic on a relevant yet computationally tractable subset of the market, with the flexibility to scale to the full sector universe once the modelling framework is validated.

4 Methodology

Our statistical arbitrage framework is built as a sequence of modular stages. Each stage is self-contained, auditable, and easily replaced, so that alternative specifications can be tested without re-engineering the entire pipeline. We begin with signal construction, move to candidate selection, and then integrate these with portfolio construction, transaction cost modelling, and performance evaluation.

4.1 Residual-Based Signals and Daily Pick Lists

Universe and mapping. The universe comprises the ten largest constituents by index weight in each of six liquid S&P 500 sector ETFs: Industrials (XLI), Energy (XLE), Utilities (XLU), Financials (XLF), Health Care (XLV), and Materials (XLB). Each stock i is mapped to its sector ETF $s(i)$. All series are aligned to a common trading calendar to ensure time-consistent observations.

Rolling sector-relative residuals. Let $P_{i,t}$ be the adjusted close of stock i at date t, and $E_{s,t}$ the adjusted close of its sector ETF. Over a rolling window of L trading days, we estimate a local linear fit:

$$\beta_{i,t} = \frac{\mathrm{Cov}_L(P_{i,\cdot}, E_{s,\cdot})}{\mathrm{Var}_L(E_{s,\cdot})}, \quad \alpha_{i,t} = \overline{P}_{i,t}^{(L)} - \beta_{i,t}\,\overline{E}_{s,t}^{(L)},$$

yielding the sector-implied fair value and residual:

$$\widehat{P}_{i,t} = \alpha_{i,t} + \beta_{i,t}E_{s,t}, \quad R_{i,t} = P_{i,t} - \widehat{P}_{i,t}.$$

Rationale. A short lookback (here $L = 10$) keeps the regression parameters responsive to recent market conditions. Using regression residuals rather than raw spreads removes both level and beta effects, so that deviations are purely idiosyncratic within the sector.

Time-series standardisation. To compare residual magnitudes across stocks, we standardise each stock's residuals over the same rolling window:

$$Z_{i,t} = \frac{R_{i,t} - \overline{R}_{i,t}^{(L)}}{\sigma_{i,t}^{(L)}}, \quad \sigma_{i,t}^{(L)} = \mathrm{sd}_L(R_{i,\cdot}).$$

This ensures the scale reflects each stock's own recent volatility, avoiding cross-sectional biases from high-volatility names.

Daily tail selection. On each day t, we form the cross-section $\{Z_{i,t}\}$ across the full universe and select the most extreme observations by absolute value:

$$K_t = \lceil \tau N_t \rceil,$$

with $\tau = 0.10$ denoting the tail proportion. The K_t most negative $Z_{i,t}$ form the long candidate set $\mathcal{L}_t$, and the K_t most positive form the short candidate set $\mathcal{S}_t$.

Rationale. This tail filter focuses trading on the most pronounced dislocations each day, controlling capacity and reducing false positives from ordinary noise.

Persistence. We save the full residual panel $\{R_{i,t}\}$, the standardised panel $\{Z_{i,t}\}$, and the daily candidate lists to disk. This separation between feature generation and portfolio simulation makes the pipeline auditable and allows downstream experiments to use exactly the same inputs.

Key design choices.

- *Short lookback:* prioritises responsiveness at the cost of noisier estimates, mitigated by the tail filter.
- *Time-series standardisation:* preserves idiosyncratic scale; cross-sectional ranking is introduced only at the selection stage.
- *Universe-wide tails:* allows capital to move freely to sectors with the strongest daily mispricings, while sector effects are already neutralised in residual construction.

4.2 Trading Strategy: Entry, Exit, Execution, and P&L

The trading engine converts the daily z-scores into positions through three filters: (1) magnitude threshold, (2) tail rank, and (3) sector-momentum confirmation. Signals are generated at the close of t and executed at the open of $t+1$ to prevent look-ahead bias.

Step 1 — Candidate pool. From the cross-sectional tails, we retain only those meeting the z-score magnitude threshold:

$$\mathcal{L}_t = \{i : Z_{i,t} \leq -Z_{\text{entry}}\}, \quad \mathcal{S}_t = \{i : Z_{i,t} \geq Z_{\text{entry}}\}.$$

Step 2 — Sector momentum filter. For each security, we compute k-day momentum of its sector ETF:

$$M_{s(i),t} = \prod_{u=t-k+1}^{t} \left(1 + \frac{E_{s(i),u} - E_{s(i),u-1}}{E_{s(i),u-1}}\right) - 1.$$

We only go long if $M_{s(i),t} \leq 0$ and short if $M_{s(i),t} \geq 0$, avoiding trades that fight sector trends.

Step 3 — Entry and exit rules. A trade is entered if it meets both the magnitude and momentum conditions. Positions are closed when $|Z_{i,t}| < Z_{\text{exit}}$ or when the maximum holding period $H_{\max}$ is reached.

Step 4 — Position sizing. Capital is equally allocated across all active positions:

$$w_{i,t} = \frac{1}{N_t}.$$

Residualisation and the sector filter contribute partial neutrality without imposing hard constraints.

Step 5 — P&L and aggregation. Long returns are $(P_{i,t}/P_{i,t-1})-1$, short returns are their negatives. The gross portfolio return is:

$$\tilde{r}_t = \sum_{i \in \mathcal{A}_t} w_{i,t} \cdot \sigma_i \cdot R_{i,t},$$

with $\sigma_i = \pm 1$ for long/short. Transaction costs are deducted on entry and exit. Equity is compounded as:

$$E_t = \prod_{u=1}^{t} (1 + r_u),$$

from which all performance metrics are computed.

5 Results

The performance of the strategy reflects the design principles outlined in Section 4. By anchoring each stock's price to its sector ETF through rolling residuals, the model captures short-lived deviations from a dynamic "fair value" benchmark. Daily scanning, combined with modest holding periods and balanced long/short exposure, allows the strategy to exploit mean reversion while limiting exposure to prolonged trends. The following results quantify how these choices translate into risk-adjusted returns over the test period.

5.1 Baseline Performance

Over the 726 trading days in the sample, the baseline configuration delivered:

- **Annualised Return:** 23.32%
- **Annualised Volatility:** 18.79%
- **Sharpe Ratio:** 1.21
- **Sortino Ratio:** 1.58
- **Maximum Drawdown:** −17.90%

Figure 1 shows the cumulative equity curve for the baseline strategy. Performance is positive and persistent, with relatively shallow drawdowns and no prolonged flat periods, consistent with the strategy's high turnover and daily selection process.

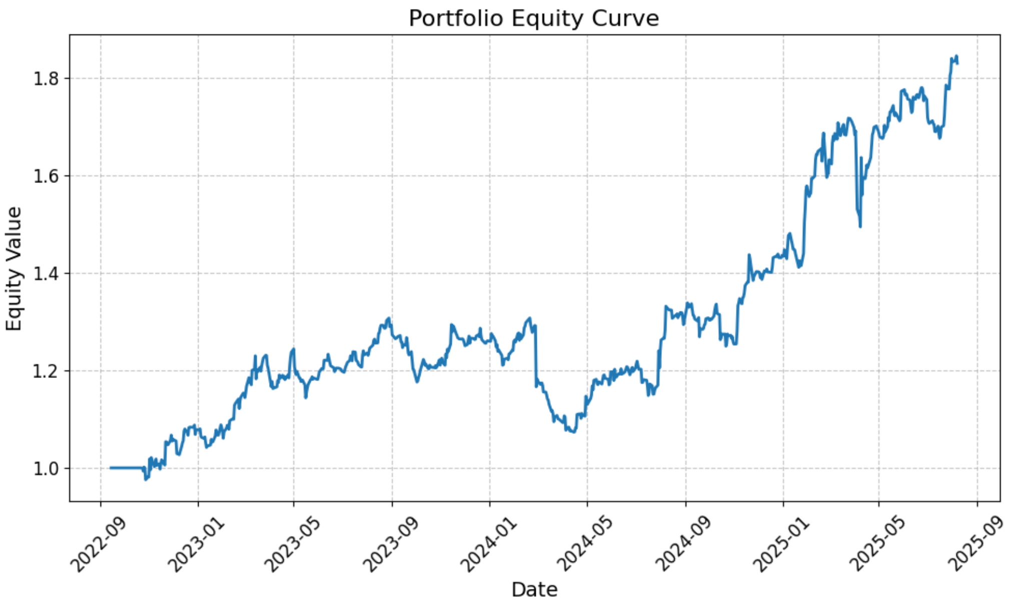

Figure 1: Cumulative equity curve for the baseline configuration.

5.2 Robustness Checks

To test sensitivity to entry/exit rules and holding constraints, we evaluate several parameter variants:

- **Case1:** Baseline configuration.
- **Hold2:** Reduced maximum holding period.
- **Zentry2.0:** Higher entry z-score threshold.
- **Zentry2.1:** Slightly higher entry z-score threshold.
- **Zexit0.45:** Stricter exit z-score threshold.

	AnnRet	AnnVol	Sharpe	Sortino	MaxDD	Trades	WinRate	AvgTradeRet	MedTradeRet	AvgHoldDays	AvgDailyTurnover
Case1	0.2332	0.1879	1.2098	1.5810	-0.1790	1442	0.5298	0.002027	0.001286	3.32	3.98
Hold2	0.1697	0.1959	0.8983	1.1624	-0.1672	1444	0.5173	0.001324	0.000808	2.60	3.98
Zentry2.0	0.2260	0.1914	1.1604	1.5659	-0.1974	1297	0.5312	0.001838	0.001318	3.34	3.58
Zentry2.1	0.2241	0.1946	1.1363	1.5784	-0.1773	1012	0.5326	0.002333	0.001398	3.35	2.80
Zexit0.45	0.2245	0.1875	1.1743	1.5370	-0.1770	1442	0.5305	0.001983	0.001263	3.36	3.98

Table 1: Robustness of performance metrics across strategy parameter variations.

5.3 Interpretation

The baseline (Case1) achieves a Sharpe ratio above 1.2 with a modest maximum drawdown, indicating that the short-horizon mean-reversion edge is consistent through time. Reducing the holding period (*Hold2*) lowers returns and Sharpe, suggesting that some trades require more than two days to converge and that premature exits sacrifice P&L. Increasing the entry threshold (*Zentry2.0, Zentry2.1*) trims trade count and turnover, modestly lowering volatility without significantly impairing returns — a sign that the most extreme dislocations carry much of the strategy's profitability. Tightening the exit threshold (*Zexit0.45*) leaves performance broadly unchanged, implying that the original exit logic already captures most of the mean-reversion benefit.

Overall, the robustness checks confirm that the strategy's edge is not narrowly dependent on one parameter setting: it persists across variations in entry/exit rules and holding constraints, with only moderate trade-offs between turnover, capacity, and risk-adjusted return.

6 Conclusions and Next Steps

6.1 Conclusions

The residual-based statistical arbitrage framework developed here demonstrates that short-horizon, sector-neutral dislocation signals can produce attractive risk-adjusted returns in a liquid large-cap universe. Over the 726-day backtest, the base configuration achieved an annualised return of 23.3% with annualised volatility of 18.8%, yielding a Sharpe ratio of 1.21 and a Sortino ratio of 1.58. Drawdowns were moderate at -17.9%, consistent with the mean-reversion premise and the dispersion of trades across sectors.

Robustness tests indicate that performance is resilient to reasonable variations in entry/exit thresholds and holding period caps. The combination of short lookback windows for residual estimation, time-series standardisation, and daily cross-sectional tail selection effectively isolates idiosyncratic dislocations. These dislocations tend to decay within a few trading days, which explains both the high turnover and the roughly three-day average holding period observed in the strategy.

The methodology's design choices—such as sector-by-sector relative value modelling, a balanced large-cap universe, and avoidance of high-volatility market makers or mega-cap index drivers—contribute to stability by limiting exposure to structural biases or single-name shocks. This setup also keeps computational demands tractable while preserving sector representation, allowing the focus to remain on clean residual signals rather than index or sector beta.

Table 2: Current Universe Composition

Sector	**# Stocks**	**Selection Basis**
Industrials (XLI)	10	Top contributors to index by weight
Financials (XLF)	10	Top contributors to index by weight
Technology (XLK)	10	Top contributors to index by weight
Consumer Discretionary (XLY)	10	Top contributors to index by weight
Energy (XLE)	10	Top contributors to index by weight
Health Care (XLV)	10	Top contributors to index by weight
Total	**60**	All large-cap, liquid constituents

6.2 Limitations

While the backtest results are promising, several limitations should be noted:

- **Transaction costs and slippage:** The current framework assumes frictionless execution. Real-world implementation would require incorporating realistic bid–ask spreads, partial fills, and market impact.
- **Universe restriction:** The current stock set is limited to 60 large-cap US equities across six sectors. Results may not generalise to mid- or small-cap stocks, or to international markets without modification.
- **Lookback and holding period choices:** The chosen 10-day residual window and average holding period of ∼3 days are optimised for this dataset and may require adjustment under different market regimes.
- **Signal scope:** The residual z-score signal is univariate and does not yet integrate other potentially informative microstructure or macro variables.
- **Data quality:** Backtest accuracy depends on the quality of historical price and ETF membership data; survivorship bias and corporate action handling could influence results.

6.3 Next Steps

Several extensions and refinements could further strengthen and broaden the scope of this framework:

1. **Execution and cost modelling:** Incorporate realistic bid–ask spreads, market impact models, and intraday fill logic to convert gross returns into more accurate net-of-cost estimates.
2. **Universe expansion:** Increase the breadth of the strategy by
   (a) Adding more sectors beyond the current six,
   (b) Expanding from the top-10 contributors per sector to a wider set (e.g., top 20 or full sector membership), and
   (c) Including cross-market universes such as European or Asian large caps.
3. **Market capitalisation classification:** Partition the stock universe into large-, mid-, and small-cap buckets to study whether mean-reversion strength, decay rates, and turnover dynamics differ across market capitalisations.
4. **Signal enrichment:** Complement residual z-scores with other short-horizon features (e.g., realised volatility shifts, abnormal volume, order book imbalance) to improve entry precision and reduce false positives.
5. **Dynamic risk overlays:** Apply volatility-scaling, sector balancing, or beta-neutral weighting schemes to further stabilise risk-adjusted returns while controlling concentration.
6. **Regime adaptation:** Incorporate macro or volatility regime detection to dynamically modulate exposure, potentially improving drawdown control during adverse conditions.
7. **Live monitoring:** Transition from backtesting to a controlled live or paper-trading environment to track stability, slippage, and latency-sensitive behaviour in real market conditions.

These steps are designed to move the strategy from a research-grade prototype toward a production-ready statistical arbitrage model, while retaining the transparency, interpretability, and modularity of the current pipeline.

References

Marco Avellaneda and Jeong-Hyun Lee. Statistical arbitrage in the us equities market. *Quantitative Finance*, 10(7):761–782, 2010. doi: 10.1080/14697680903124632.

Robert J. Elliott, John van der Hoek, and William P. Malcolm. Pairs trading. *Quantitative Finance*, 5 (3):271–276, 2005. doi: 10.1080/14697680500149370.

Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.

Richard C. Grinold and Ronald N. Kahn. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill Education, 3rd edition, 2019.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993. doi: 10.1111/j.1540-6261.1993.tb04702.x.

Amir E. Khandani and Andrew W. Lo. Lo, has quantitative equity investing run its course? *Financial Analysts Journal*, 63(2):13–23, 2007. doi: 10.2469/faj.v63.n2.4525.

Bruce N. Lehmann. Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, 105 (1):1–28, 1990. doi: 10.2307/2937816.

Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, 2012. doi: 10.1016/j.jfineco.2011.11.003.

James M. Poterba and Lawrence H. Summers. Mean reversion in stock prices: Evidence and implications. *Journal of Financial Economics*, 22(1):27–59, 1988. doi: 10.1016/0304-405X(88)90021-9.