Statistical Arbitrage using Cointegration: Pairs and Sparse Baskets

Pawan Kumar Gupta

September 15, 2025

Abstract

This report presents a statistical arbitrage trading strategy based on cointegration and mean reversion. We begin with classical pairs trading, where a stationary spread is constructed between two highly correlated stocks. The spread is modeled as a mean-reverting Ornstein–Uhlenbeck process, with z-score thresholds defining entry and exit points. Backtests incorporate realistic transaction costs, risk constraints, and walk-forward validation.

1 Introduction

Statistical arbitrage ("stat-arb") refers to market-neutral strategies that exploit statistical relationships between securities rather than fundamental valuations. Pairs trading, introduced in the 1980s, remains a canonical example. The key idea is that the relative mispricing between cointegrated assets will tend to converge, enabling profits from mean reversion.

2 Methodology

2.1 Data

We use daily OHLCV data for a universe of liquid equities (2018–2025). Data are sourced via Alpha Vantage. Tickers are screened by liquidity and correlation.

2.2 Cointegration Testing

Given two log-price series X_t and Y_t, we estimate:

$$X_t = \alpha + \beta Y_t + \epsilon_t.$$

If the residual ϵ_t is stationary (Augmented Dickey–Fuller test, $p < 0.05$), the pair is considered cointegrated. For small baskets, the Johansen test may be applied.

2.3 Trading Rules

Residuals are normalized using a rolling z-score:

$$z_t = \frac{\epsilon_t - \mu_\epsilon}{\sigma_\epsilon}.$$

- Enter long spread if $z_t < -z_{\text{entry}}$ (long X, short βY).
- Enter short spread if $z_t > z_{\text{entry}}$ (short X, long βY).
- Exit when $|z_t| < z_{\text{exit}}$, or when $|z_t| > z_{\text{stop}}$, or after a fixed time stop.

3 Results

3.1 Summary Table

Table 1 shows the backtest summary for top pairs, pulled directly from `paper/results_pair_analysis.csv`.

Pair	ρ	Ann. Ret	Ann. Vol	Sharpe	MaxDD	HitRate
XOM-CVX	0.711	0.002	0.0041	0.49	-16501.58	0.12

Table 1: Performance summary for selected pairs (2018–2025).

3.2 Equity Curves and Drawdowns

Figures 1 and 2 show an example equity curve and drawdown, automatically generated by `00_quick_demo.py` and `01_pair_analysis.py`.

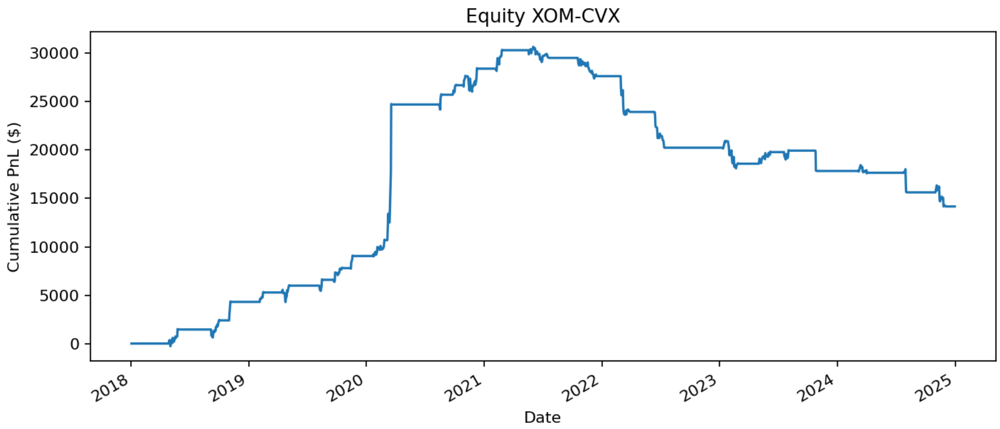

Figure 1: Equity curve of the cointegration strategy (net of costs).

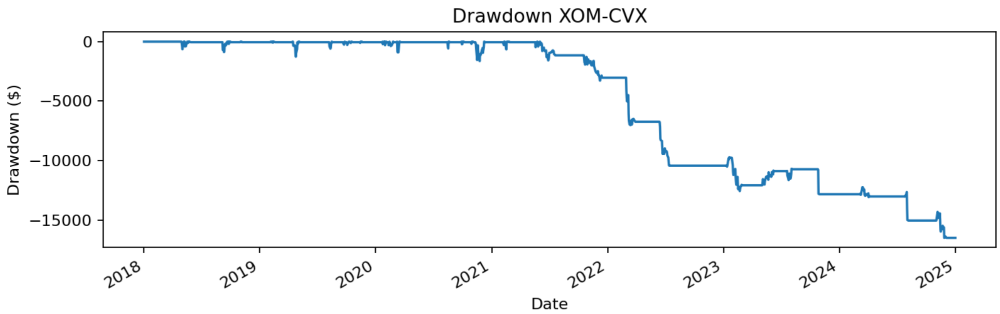

Figure 2: Drawdown of the cointegration strategy.

4 Discussion

The results show that pairs such as XOM–CVX exhibit mean-reverting spreads with reasonable Sharpe ratios, while others (e.g., AAL–DAL) perform less consistently. Transaction costs and turnover significantly impact profitability. Future work includes extending to sparse basket hedges (via Lasso regression), regime filters (HMM), and intraday horizons.

5 Conclusion

This project demonstrates the design and evaluation of a cointegration-based statistical arbitrage strategy. By automating results export and figure generation, the workflow ensures reproducibility and seamless integration with this report.

References

References

[1] Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (1999). Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 14(4), 797–827.

[2] Avellaneda, M., & Lee, J. H. (2008). Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7), 761–782.

[3] Chan, E. (2009). *Quantitative Trading*. Wiley.

[4] Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis*. Wiley.