Stationary, But Not Profitable?
A Critical Look at Pairs Trading

Davide Pandini, PhD, CMT, MFTA, CFTe, CSTA
SIAT (Societa' Italiana Analisi Tecnica)
Corso Magenta 56, 20123 Milano (MI), Italy
davide.pandini@hotmail.com

"*Virtue is the golden mean between two vices, the one of excess and the other of deficiency*"
Aristotle (384–322 BCE), Greek philosopher and polymath

Abstract

This paper provides a comprehensive re-examination of cointegration-based pairs trading and multi-asset statistical arbitrage in the context of modern, liquid ETFs and equity markets. While statistical arbitrage is a well-established field, much of the foundational literature relies on outdated pre COVID-19 sample periods, narrow asset universes, or simplified assumptions that ignore the impact of transaction costs and parameter instability. The primary objective of this study is to determine whether traditional measures of long-run equilibrium – specifically stationarity and cointegration – remain reliable predictors of out-of-sample alpha for active investment managers in a regime characterized by high volatility clusters and structural shifts. By utilizing a diverse universe of assets including MSCI country funds, U.S. sector benchmarks, and commodity-linked vehicles, this research bridges the gap between theoretical econometrics and practical investing and trading execution.

Methodological Framework – The study introduces a fully reproducible, end-to-end framework for spread-based trading. A core pillar of this methodology is a strict train/test protocol: models are identified during an *in-sample* training window (2007–2017) and then frozen and evaluated during an extensive *out-of-sample* testing window (2018–2025). This separation is vital for active managers to ensure that reported performance is not inflated by look-ahead bias or in-sample overfitting. The paper details two primary statistical approaches for spread construction: the Engle–Granger[1] two-step procedure for asset pairs and the Johansen system-based procedure[2] for multi-asset combinations. The methodology constructs two-asset spreads via OLS (Ordinary Least Squares linear regression) hedge ratios and multi-asset spreads via Johansen eigenvectors, applies Engle–Granger and Johansen cointegration diagnostics, and implements an ex-ante 21-day rolling z-score normalization computed without self-inclusion. A significant technical innovation discussed in this work is the use of a zero-intercept OLS specification for hedge-ratio estimation. Including a static intercept can lead to "stale" bias in new price regimes. In contrast, by forcing the intercept to zero, the model delegates level-correction to an adaptive rolling window, thereby improving out-of-sample robustness and ensuring the hedge ratio reflects only the proportional relationship between asset prices.

Normalization and Signal Generation – To ensure disciplined rule-based execution, the raw spread is transformed into a rolling z-score. The research advocates for a conservative normalization method that excludes self-inclusion, where the current day's spread is compared to a benchmark mean and standard deviation fixed *before* the current move. This causally clean approach avoids the "self-dampening" of z-scores that occurs when the current observation is used to set its own benchmark, providing a more honest assessment of market shocks.

Trading signals follow a symmetrical state-machine logic: long positions are entered when the z-score drops below -2 and exited at -1, while short positions are entered above +2 and exited at +1. This stepwise exit strategy is designed to capture meaningful retracements while avoiding premature closure on minor fluctuations.

Empirical Case Studies and The Cointegration Paradox – The results present what we call the "Cointegration Paradox": while cointegration is a standard for risk management, it is neither a sufficient nor a strictly necessary condition for profitability. Among the most relevant case studies analyzed in this work there are:

- *Robust successes*: The CAN-RSA (Canada and South Africa) and AUS-CAN (Australia and Canada) spreads demonstrated high risk-adjusted performance, leveraging the shared resource-dependence of these economies. The multi-asset AUS-CAN-RSA spread further illustrated how combining three assets can filter

out idiosyncratic risks, achieving a positive net Sharpe ratio even after transaction costs.

- *Regime failures*: The GLD-SLV (Gold/Silver) case study reveals the fragility of historical relationships. Although gold and silver are traditionally viewed as a cointegrated pair, they failed formal cointegration tests and decoupled significantly during the 2020 COVID-19 shock. Similarly, the EWN-EWQ (Netherlands/France) spread demonstrated that structural economic divergence – such as the heavy weighting of semiconductors in the Netherlands vs. luxury goods in France – can cause a historically stationary relationship to break down.
- *Profitable anomalies*: Conversely, the SPY-AAPL-MSFT multi-asset spread outperformed buy-and-hold benchmarks despite a lack of in-sample cointegration. The strategy succeeded by dynamically capturing short-horizon relative-value premiums that static allocations could not realize, suggesting that transient convergence can be a valid source of alpha for active managers even without long-run stationarity.

Practical Significance – A key contribution of this work for active investment managers is the explicit modeling of transaction costs. The paper demonstrates that a modest 10 basis point cost can significantly reduce the Sharpe ratio of high-turnover strategies, such as in the AUS-CAN pair where the gross Sharpe of 0.51 dropped to a net 0.13. This highlights that execution efficiency can be as important

as statistical modeling. Furthermore, the study identifies "drift risk" in commodity-linked pairs (e.g., OIH-USO), where storage costs and roll yield prevent the formation of stable long-run equilibria, making them less suitable for traditional pairs trading.

Conclusions and Impact – The paper concludes that while cointegration provides an analytical tool for identifying economic linkages, it must be paired with regime-aware risk control and continuous rolling diagnostics. This work provides a reproducible workflow that allows practitioners (investors, portfolio managers, and traders) to identify when mean-reversion adds value and when structural breaks make it a liability. Ultimately, this research characterizes pairs trading not as a riskless arbitrage, but as a methodical relative-value strategy that, when implemented with realistic cost modeling and discipline, regime awareness, robust out-of-sample validation, and adaptive parameters, remains a viable tool for diversification in modern, volatile markets.

1. Introduction

Pairs trading is a market-neutral trading strategy that involves simultaneously taking long and short positions in two (or more) highly correlated assets. The core idea is to profit from the temporary divergence and convergence of the prices of these assets: when the price of one asset deviates significantly from the other, the strategy assumes that they will eventually revert to their historical relationship. This allows

investors, portfolio managers, and traders to go long with the underperforming asset and short on the outperforming asset, anticipating a convergence in their prices. In contrast to directional strategies, a well-constructed long/short spread can achieve mean-reversion (convergence) gains when temporary deviations from an economic or statistical relationship tighten toward equilibrium. Pairs trading, and its multi-asset generalization, provides a transparent, model-light framework for this objective: it links trade construction to measurable properties of the underlying time series (e.g., stationarity, cointegration), supports explicit risk control (entry/exit thresholds, position holding), and scales naturally to liquid instruments such as ETFs and equity indices. Its strength is that a hedged long/short spread can make profits from relative-value price divergences while attenuating market-wide risk.

The study addresses the end-to-end problem of identifying, testing, and trading mean-reverting relationships. In this work, we (i) estimate hedge ratios or cointegration vectors to form candidate spreads; (ii) test for stationarity/cointegration using spread-level unit-root tests and a system-level procedure; and (iii) implement an out-of-sample trading strategy based on ex-ante standardized deviations (rolling z-scores without self-inclusion), with explicit consideration of transaction costs.

The paper presents three main contributions. *First*, an end-to-end, reproducible workflow: from data analysis and hedge estimation to spread construction, ex-ante

normalization, signal/position mapping, execution assumptions, and reporting. *Second*, a systematic comparison of Engle-Granger (residual-based) and Johansen (system-based maximum likelihood) cointegration tests, clarifying when each supports spread formation and how conclusions differ across single-pair and multi-asset settings. *Third*, a robustness and risk analysis that documents parameter choices, stability across subsamples, spread constructions, and practical transaction costs. While other prior studies examined components of this pipeline, this work is comprehensive and applies a single, rigorous, statistics-based methodology, with strict in-sample training and out-of-sample testing, consistently across a broad universe of ETFs, commodities, and index spreads.

We organize the study around four main questions: (i) Do the proposed spreads exhibit in-sample stationarity/cointegration, and does this property persist out-of-sample? (ii) What is the risk-adjusted performance (returns, volatility, Sharpe) of the trading rule relative to buy-and-hold benchmarks, both gross and net of trading costs? (iii) How robust are results to key design choices (thresholds, lookbacks, ranks/lags, intercept specification)? (iv) To what extent do transaction costs affect realizable outcomes?

Section 2 describes the historical data series of the asset classes, and the pre-processing step of dividing the backtesting period into training and testing ranges. Moreover, it analyzes the correlations among the assets. Section 3 presents the

spread construction, and testing methodology details. Furthermore, it describes the trading strategy and execution assumptions. The performance metrics are reviewed in Section 4, while Section 5 analyzes empirical results and presents some significant case studies. Finally, our conclusions are summarized in Section 6. All analyses and the pairs trading framework were implemented in *R statistical software*[7].

2. Data and Pre-processing

The financial assets considered in this work are primarily liquid, index-level ETFs that proxy whole economies (e.g., MSCI country funds), well-defined U.S. sector benchmarks (e.g., energy, financials), complemented by commodity-linked vehicles, and by high-tech, consumer staples, and energy stocks. By using times series aligned on a common trading calendar, we focus on financial instruments that share strong economic linkages (e.g., trade ties, policy regimes, sector composition, and exposure to global risk factors). In this scenario, some linear combinations of their prices show a stable fair-value relationship. Natural asset clusters include Eurozone equities, resource-heavy developed markets, U.S. indices, and intra-sector assets within the same industry or sector. Such clusters can be considered as potential candidates for both single-pair and multi-asset cointegration. The time series considered in this work were downloaded from *Yahoo Finance* and are summarized in Table 1 (Appendix). We analyzed the daily adjusted-close prices from 2007-01-01 to 2025-09-30 (about 19 years of financial data for a total of 4716 trading days), partitioned

into an *in-sample* (training) window from 2007-01-01 to 2017-12-31 (11 years; 2769 trading days), and an *out-of-sample* (testing) window from 2018-01-01 to 2025-09-30 (~7.75 years; 1,947 trading days).

All model identifications, i.e., unit-root diagnostics on individual series, like Engle-Granger residual tests for pairs and Johansen rank determination for multi-asset spreads and hedge-ratio estimation, are performed exclusively in the training window; these specifications are then frozen and carried forward unchanged into the testing window, where trading rules (z-score bands, time/price exits, and trading commission cost deductions) are evaluated using only information available at each decision time. This approach prevents look-ahead bias and data leakage, and also in-sample overfitting/optimism bias, thus avoiding that tuning to noise in the training set would inflate the reported performance in the testing range. A clear separation between model selection and performance measurement provides a robust confirmation that the mean-reverting property (stationary spread) and the estimated hedge ratios are persistent beyond the estimation period.

Moreover, by spanning multiple regimes (post-GFC expansion, commodity cycles, COVID-19 shock, inflation/repricing period), the out-of-sample window serves as a stress test against structural breaks, enabling assessment of rank stability under changing volatility and correlations, as was reported in [3]. The train/test separation follows best practices in quantitative finance, ensuring robust out-of-sample

validation of trading strategies, i.e., estimates on history, validates on unseen data, and yields performance metrics that are more indicative of out-of-sample efficacy rather than in-sample overfitting.

Correlations are estimated in-sample on arithmetic daily returns to avoid spurious level effects (log returns were evaluated and found not to differ significantly from arithmetic returns). These estimates serve only for initial screening, not pairs selection. High positive correlations help identify economically coherent pairs (e.g., EWA–EWC, KO–PEP) where cointegration is plausible but not granted. Moreover, correlation captures short-run co-movement in returns, not long-run equilibrium in price levels. Pairs trading requires cointegration: a stationary linear combination of non-stationary prices.

In the training window, equity correlations are systematically higher than commodity correlations. Equity returns are dominated by shared market factors (GFC, QE/taper), while commodities reflect idiosyncratic drivers (inventory, weather, term-structure carry). Rolling correlations (Figures 1-2, Appendix, the vertical dashed line marking the boundary between training and testing windows) reveal temporal dynamics but are not sufficient for pairs selection.

3. Methodology

For two-asset spreads, we use the standard OLS-hedged specification:

$$s_t = p_{1,t} - \beta p_{2,t} - \alpha,$$

where $p_{1,t}$ and $p_{2,t}$ are price levels, β is the hedge ratio, and s_t is the spread at period t. We set α = 0, justified theoretically and empirically. A non-zero intercept (α ≠ 0) would imply persistent level differences; these are absorbed by the rolling mean during z-score normalization, avoiding overfitting and improving out-of-sample robustness. For multi-asset spreads, we use the Johansen test. After selecting lag length (using Akaike Information Criterion AIC[4]), we obtain cointegration vectors. The first eigenvector (largest eigenvalue) provides hedge ratios:

$$s_t = \beta_1 p_{1,t} + \beta_2 p_{2,t} + \cdots + \beta_n p_{n,t}.$$

The spread is tested for stationarity, and if passes the Augmented Dickey–Fuller test ADF[5] and the Phillips–Perron test PP[6], then it can be treated as mean reverting for trading.

Correlation measures short-run co-movement but does not imply cointegration. Two trending series can display high correlation without a stable pricing link (spurious correlation). Cointegration-based pairs trading relies on a stationary, mean-reverting spread and it establishes the long-run equilibrium, while stationarity implies that deviations from that equilibrium are temporary.

Cointegration Testing

Engle–Granger (EG) Test: A two-step procedure. First, test individual series for unit roots (ADF/PP). Then, regress one price on the other and test the residuals for stationarity; stationary residuals indicate cointegration.

Johansen Test: A multivariate cointegration test used to determine whether multiple asset prices share a stable long-run equilibrium relationship.

Normalization

The raw spread is converted to a rolling z-score:

$$z(s)_{t,L} = \frac{s_t - \mu_{t|t-1,L}}{\sigma_{t|t-1,L}},$$

where:

$$\mu_{t|t-1,L} = \text{mean}(s_{t-L:t-1})$$

$$\sigma_{t|t-1,L} = \text{sd}(s_{t-L:t-1}),$$

and L = 21 (trading days per month, which is a practical look-back period). We use no-self-inclusion (μ and σ computed up to t-1) for conservative, causally clean signals. This avoids self-dampening and ensures thresholds are independent of the observation triggering the trade.

Investing and Trading Strategy

Trading/investing signals are triggered when z-score crosses predetermined bands:

- *Enter long*: z-score drops below -2 (previous day above -2);
- *Exit long*: z-score rises above -1;
- *Enter short*: z-score rises above +2 (previous day below +2);
- *Exit short*: z-score falls below +1.

Positions persist until an exit signal occurs. Signals depend on data up to day t but are applied from t+1 onward, eliminating look-ahead bias.

The strategy is simple, disciplined, and market-neutral, thus allowing an effective assessment of pairs trading. However, performance depends on parameter choices and the stability of the cointegrated relationship. Transaction costs can erode returns, especially with high-turnover trading and frequent portfolio rebalancing.

4. Performance Metrics

Strategies are evaluated against buy-and-hold benchmarks using:

- Annualized return: $r_a = (\prod_{t=1}^{n}(1 + r_t))^{\frac{252}{n}} - 1$.
- Annualized standard deviation: $\sigma_a = \sigma_{daily}\sqrt{252}$.
- Annualized Sharpe ratio: $SR_a = \frac{r_a - r_{risk-free,a}}{\sigma_a}$.

These metrics answer: how much the strategy earns, how volatile returns are, and whether returns compensate for risk.

5. Experimental Results and Case Studies

AUS–CAN Case Study (EWA–EWC) – EWA (Australia) and EWC (Canada) are commodity-driven economies with similar sector profiles (financials, natural resources). The spread passes EG cointegration tests. The performance metrics are reported in Table 2, where rfint and rfintc summarize the trading-strategy results without and with trading commissions, respectively, while rfaus and rfcan are the buy-and-hold results for AUS and CAN, respectively. The strategy outperforms on a risk-adjusted basis before costs but underperforms after costs.

Table2. AUS (EWA)-CAN (EWC) out-of-sample performance metrics

Statistic	rfint	rfintc	rfaus	rfcan
Annualized Return	0.038	0.010	0.061	0.093
Annualized Std Dev	0.076	0.075	0.242	0.201
Annualized Sharpe (Rf=0%)	0.506	0.131	0.251	0.464

Transaction costs erode gains in this high-turnover strategy. During trending markets (post-2020), buy-and-hold outperforms as the spread stabilizes.

CAN–RSA Case Study (EWC–EZA)

Canada and South Africa share resource dependence (energy, mining, materials). The EWC–EZA spread shows a significant mean-reversion.

Table 3. CAN (EWC)-RSA (EZA) out-of-sample performance metrics

Statistic	rfint	rfintc	rfcan	rfrsa
Annualized Return	0.086	0.056	0.093	0.039
Annualized Std Dev	0.087	0.087	0.201	0.316
Annualized Sharpe (Rf=0%)	0.992	0.650	0.464	0.123

The strong performance of the CAN–RSA spread is underpinned not only by statistical mean-reversion, but also by macroeconomic parallels between the Canadian and South African economies. Both countries are highly resource-dependent, with significant exposure to commodities such as oil, gold, and base metals. The ETFs representing these markets, EWC (Canada) and EZA (South Africa), reflect this structural similarity in their sector compositions. These shared economic drivers often lead to correlated responses to global commodity price movements, forming the basis for a potentially cointegrated relationship in their equity indices, and supporting the trading strategy's foundation: deviations in the

relative pricing between EWC and EZA are likely temporary and driven by short-term noise rather than fundamental divergence, as shown in Figure 3 (Appendix).

AUS–CAN–RSA Multi-Asset Case Study

The three-asset spread (EWA, EWC, EZA) captures common commodity-driven trends.

Table 4. AUS-CAN-RSA out-of-sample performance metrics

Statistic	r_aus_can_rsa	r_aus_can_rsa_c	rfaus	rfcan	rfrsa
Annualized Return	0.075	0.043	0.061	0.093	0.039
Annualized Std Dev	0.093	0.093	0.242	0.201	0.316
Annualized Sharpe (Rf=0%)	0.805	0.460	0.251	0.464	0.123

Even after accounting for transaction costs (r_aus_can_rsa_c), the strategy maintains positive performance (Sharpe ratio of 0.46), comparable to the standalone Sharpe of the best-performing ETF (EWC at 0.46). Importantly, both strategy variants exhibit superior stability relative to the individual ETFs, especially EZA (South Africa), which suffers deep and extended drawdowns as shown in Figure 8 (Appendix). This highlights the benefit of capturing a stationary, mean-reverting spread that arises from the relative dynamics among Australia, Canada, and South Africa. All these countries are major commodity exporters with strong exposure to natural resources, and their equity markets, represented by EWA (Australia), EWC (Canada), and EZA (South Africa), share common drivers like global demand for metals, oil, and energy. However, these economies respond differently to local shocks and policy decisions, leading to short-term deviations among their respective ETFs. These deviations are

often temporary, allowing the strategy to exploit the reversion to equilibrium. The relatively low volatility of the constructed spread and the consistent performance despite fluctuating ETF returns confirm that the three-asset spread effectively filters out idiosyncratic risk while preserving exposure to a stable common stochastic trend.

GLD–SLV Case Study

GLD (gold) and SLV (silver) failed in-sample EG cointegration tests. Despite this, the strategy generated profits from 2018 to January 2020, with annualized return ~16% after transaction costs (Table 5). Post-2020, the relationship broke down due to COVID-19 and shifting market dynamics (gold as safe-haven, silver as industrial commodity). The spread became non-stationary, and the strategy underperformed. This case demonstrates that cointegration is not guaranteed to persist; structural breaks can render previously profitable pairs ineffective (Figure 5, Appendix).

Table 5. GLD-SLV performance metrics (2018 – Jan 2020)

Statistic	rfint	rfintc	rfgld	rfslv
Annualized Return	0.190	0.161	0.089	0.018
Annualized Std Dev	0.556	0.555	0.107	0.178
Annualized Sharpe (Rf=0%)	0.341	0.289	0.830	0.101

The market changed dramatically around March 2020. The onset of the COVID-19 pandemic drove an unprecedented flight to safe-haven assets and inflation hedges, causing gold prices to rally, while silver initially lagged and as an industrial commodity introduced volatility and retail speculation. With the spread no longer mean-reverting after January 2020 as displayed in Figure 4 (Appendix), the pair

strategy stopped generating profitable signals and what used to be short-term mispricing turned into prolonged trend deviations. By the time convergence occurred, it was often after long stretches of divergence, thus leading to the strategy's underperformance, as reported in Table 6.

Table 6. GLD-SLV out-of-sample performance metrics

Statistic	rfint	rfintc	rfgld	rfslv
Annualized Return	-0.593	-0.603	0.145	0.132
Annualized Std Dev	1.004	1.004	0.147	0.282
Annualized Sharpe (Rf=0%)	-0.591	-0.601	0.983	0.469

In practical terms, a trade that would "buy silver & sell gold" on an overshoot (or vice-versa) started to either not revert or take too long to revert, making profits elusive.

In summary, the GLD–SLV pairs trade lost its edge because the statistical foundation (a stationary spread) was undermined by extraordinary macroeconomic events and shifting market dynamics. What had been considered as a stable precious metals arbitrage turned into a much more unpredictable relationship after 2020, causing the strategy's early gains to plateau and rendering it largely ineffective in the subsequent years.

NED–FRA Case Study (EWN–EWQ)

EWN (Netherlands) and EWQ (France) passed in-sample cointegration tests but underperformed out-of-sample. EWN outperformed EWQ consistently (annualized

return 10.0% vs. 7.4%), creating a trending (not mean-reverting) spread. The mean-reverting strategy bet against the trend, generating losses, after transaction costs.

Table 7. NED-FRA out-of-sample performance metrics

Statistic	rfint	rfintc	rfned	rffra
Annualized Return	0.0274	-0.0003	0.1002	0.0743
Annualized Std Dev	0.0682	0.0682	0.2249	0.2180
Annualized Sharpe (Rf=0%)	0.4025	-0.0037	0.4454	0.3407

This highlights that cointegration is not permanent. Structural shifts (e.g., technology weight in EWN vs. luxury/energy in EWQ) can break long-run equilibrium. The Dutch and French economies have faced different structural conditions since 2018, leading to economic and macro divergence. For example, the Netherlands (EWN) is heavily weighted toward global-tech and trade-sensitive stocks, whereas France (EWQ) leans more on luxury brands, energy, and domestic industries. ASML (semiconductors) alone is ~29% of EWN, while LVMH (luxury) is ~8% of EWQ. In other words, just because two series cointegrated in one period does not guarantee it holds later, and a temporary equilibrium broke down as markets evolved.

Pairs-trading research emphasizes that profitability is based on the stability of the cointegrating relationship. In our case, any slow drift or new trend (e.g., EWN consistently outpacing EWQ) would violate the stationarity assumption. Once the spread no longer reverts to its mean, mean-reversion signals generate losses.

In summary, the EWN–EWQ strategy underperformed because the out-of-sample environment violated its core assumptions. Structural or sectoral shifts and volatility

regimes likely broke the equilibrium that existed in the training range, turning the spread into a trending or unstable series. This analysis highlights a key lesson: cointegration-based strategies can be fragile. They require stable economic linkages and must withstand macro changes.

Cointegration Assessment

EG test results (Tables 8, Appendix) show that most commodity-related pairs (like GLD–SLV, OIH–USO, USO–GLD) fail cointegration tests. Commodity prices are driven by idiosyncratic fundamentals (storage costs, seasonality, geopolitical shocks) with no arbitrage mechanism to enforce a stable price ratio. Equity and index pairs exhibit more stable relationships due to shared economic drivers.

Outperformance without Cointegration: SPY–AAPL–MSFT

The SPY–AAPL–MSFT spread failed Johansen in-sample cointegration tests (Table 9, Appendix) and visual inspection shows multi-year trends and level shifts, not mean-reversion (Figure 6, Appendix). Hence, these results reveal a significant departure from the strategy's primary statistical requirements:

- *Stationarity rejection*: Both visual inspection and formal Johansen testing confirm no cointegration during the 2007–2017 training period.
- *Unit-root characteristics*: The spread exhibits multi-year trends and level shifts, behaving like a unit-root process with drift rather than a stationary process.

- *Persistent deviations*: The spread stays on one side of its mean for years, indicating that the horizontal reference mean is not a stable attractor.
- *Heteroskedasticity*: The series shows time-varying variance, with volatility clustering and scaling alongside price levels, which is inconsistent with stable mean-reversion.

The "Cointegration Paradox"

The novelty of this case study lies in the empirical finding that this non-cointegrated spread outperformed buy-and-hold benchmarks for all three individual assets during the 2018–2025 testing window.

Performance metrics: Despite the lack of a long-run equilibrium, the strategy delivered higher cumulative returns than SPY, AAPL, or MSFT on both a gross and net-of-costs basis (Table 10).

Table 10. SPY-AAPL-MSFT out-of-sample performance metrics

Statistic	r_spy_aapl_msft	r_spy_aapl_msft_c	rfspy	rfaapl	rfmsft
Annualized Return	0.4257	0.3879	0.1426	0.2691	0.2751
Annualized Std Dev	0.5854	0.5850	0.1964	0.3113	0.2860
Annualized Sharpe (Rf=0%)	0.7272	0.6631	0.7258	0.8646	0.9621

Regime-dependent gains: The strategy capitalized on transient, regime-dependent corrections in relative prices. While the relative trend often ran against the mean-reversion rule (notably in 2021–2023), the strategy recovered sharply as repeated "divergence–convergence" episodes emerged (Figure 7, Appendix).

Practical Significance for Active Managers

For practitioners of active investing, this analysis provides three vital insights:

- *Capturing transient alpha*: The spread framework allows active managers to capture short-horizon relative-value premiums that static allocations cannot realize. By using state-dependent entry/exit thresholds, the strategy adapts to market shifts through continuous rebalancing.
- *Active vs. passive exposure*: While a buy-and-hold position simply transmits market fluctuations, the spread framework converts mispricing episodes into realized returns.
- *Risk management of drift risk*: The paper warns that trading non-cointegrated spreads introduces drift risk. Therefore, strict risk controls and an awareness of structural outperformance (e.g., mega-cap tech vs. the broader market) are essential for successful implementation.

6. Conclusions

In-sample cointegration is not a strict prerequisite for out-of-sample profitability. Disciplined mean-reversion rules can exploit temporary mispricings among fundamentally related securities even when they lack a permanent stationary equilibrium.

Short-horizon mean-reversion rules can exploit transient, regime-dependent corrections in relative prices without requiring a perfectly stationary spread. However, the lack of cointegration implies drift risk (non-stationarity); hence,

rigorous risk control is essential. The experimental results presented in this work are consistent with markets periodically correcting relative mispricing among fundamentally related securities; the spread framework converts these divergence–convergence episodes into returns that buy-and-hold positions in single assets do not capture.

The out-of-sample (OOS) performance of all single-pair and multi-asset spreads was evaluated over a strict testing window from January 1, 2018, to September 30, 2025. This period is particularly significant for active investment managers as it encompasses diverse market regimes, including the post-GFC expansion, the COVID-19 volatility shock, and the subsequent high-inflation and interest-rate repricing environment. The strategy's efficacy was measured using the most significant performance metrics, and benchmarked against a passive buy-and-hold approach for each individual constituent asset.

The empirical results across the OOS window present a mixed but consistent picture of active vs. passive management. Several spreads delivered superior risk-adjusted returns compared to their best-performing constituent legs, demonstrating how multi-asset combinations can effectively filter idiosyncratic risk to preserve a stable common stochastic trend.

A critical advantage of the mean-reverting spread structure is its defensive nature during market stress. In the AUS-CAN case study, the unadjusted strategy

experienced shallower and less frequent drawdowns than the individual ETFs; specifically, during the March 2020 crash, the ETFs suffered drawdowns exceeding –40%, while the strategy's losses remained much more contained (Figure 9, Appendix).

The impact of trending markets: The strategy generally lagged behind passive benchmarks during strong trending markets. For instance, after the mid-2020 recovery, the directional rally in global equities allowed buy-and-hold positions in assets like EWC (Canada) to outperform the spread strategy, which inherently bets on relative value rather than outright momentum.

The critical role of transaction costs for practitioners: The study highlights that execution efficiency is as vital as the statistical model. The transition from gross to net performance revealed substantial erosion of the statistical edge. In high-turnover pairs like AUS-CAN, the gross Sharpe ratio plummeted after costs. This suggests that active managers must trade off the frequency of rebalancing signals against the reality of slippage and fees to maintain a viable net alpha.

A central finding of this OOS summary is that in-sample cointegration is informative but not decisive for future profitability. Some pairs that exhibited robust in-sample cointegration, such as EWN-EWQ (Netherlands/France), underperformed in the testing window due to structural or sectoral shifts, such as the divergent performance

of Dutch semiconductors vs. French luxury goods, which transformed a stationary spread into a trending one.

Conversely, the SPY-AAPL-MSFT multi-asset spread managed to outperform buy-and-hold benchmarks despite a lack of formal in-sample cointegration. By dynamically capturing transient relative-value premiums through state-dependent entry/exit thresholds, the strategy realized gains that static allocations could not.

The OOS summary characterizes pairs trading not as a "riskless arbitrage," but as a disciplined relative-value strategy. While stationarity remains the standard for risk management, the results suggest that active managers can extract value from non-stationary spreads provided they utilize regime-aware risk controls and continuous rolling diagnostics. Ultimately, the structured approach developed and evaluated in this research offers attractive risk-return profiles and meaningful diversification compared to directional benchmarks in modern, volatile markets.

References

[1] Engle, R. F., & Granger, C. W. J. (1987). Cointegration and Error Correction: Representation, Estimation, and Testing. Econometrica, 55(2), 251-276.

[2] Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica, 59(6), 1551-1580.

[3] Zakamulin, V. (2014). The real-life performance of market timing with moving average and time-series momentum rules. Journal of Asset Management, 15(4), 261–278.

[4] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

[5] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association, 74(366), 427-431.

[6] Phillips, P. C. B., & Perron, P. (1988). Testing for a Unit Root in Time Series Regression. Biometrika.

[7] R: The R Project for Statistical Computing

Appendix

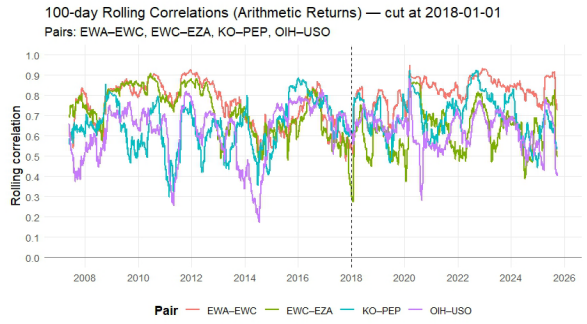

Figure 1. 100-day rolling correlations

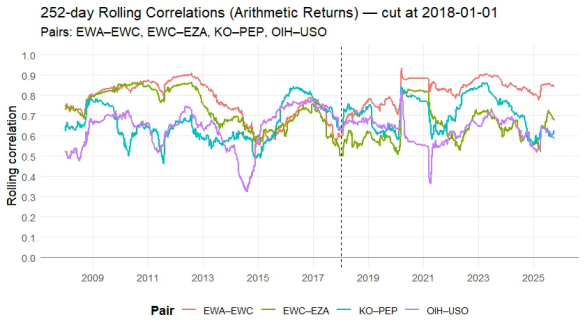

Figure 2. 252-day rolling correlations

Table 1. Assets

Ticker	Issuer	Underlying	Ticker	Issuer	Underlying
EWA	iShares	MSCI Australia Index	SLV	iShares	LBMA Silver Price
EWC	iShares	MSCI Canada Index	XLE	SPDR	Energy Select Sector Index
EZA	Shares	MSCI South Africa Index	IEO	Shares	Select Oil Exploration & Production Index
EWG	iShares	MSCI Germany Index	OH	VanEck	MVIS US Listed Oil Services 25 Index
EWQ	iShares	MSCI France Index	USO	USCF Investments	Front-month WTI crude oil futures
EWN	Shares	MSCI Netherlands IMI 25/50 Index	SU	Common stock	Canada Suncor Energy Inc.
EWZ	iShares	MSCI Brazil 25/50 Index	WTI	Common stock	W&T Offshore, Inc.
LQD	iShares	USD Liquid Investment Grade Index	VALE	Common stock	Brazil vate S.A. - Iron Ore & Nickel
WI	Vanguard	CRSP US Total Market Index	BHP	Common stock	BHP Group Limited — Metals & Mini
SPY	SPDR	S&P 500 Index	XOM	Common stock	Exxon Mobil Corporation
DA	SPDR	Dow Jones Industrial Average	cvx	Common stock	Chevron Corporation
XLF	SPDR	Financial Select Sector Index	KO	Common stock	The Coca-Cola Company
KRE	SPDR	S&P Regional Banks Select Industry Index	PEP	Common stock	PepsiCo, Inc.
GLD	SPDR	LBMA Gold Price PM	WMT	Common stock	Walrnart Inc.
GDX	VanEck	NYSE Arca Gold Miners Index	COST	Common stock	Costco Wholesale Corporation
IAU	Shares	LBMA Gold Price	AAPL	Common stock	Apple Inc.
			MSFT	Common stock	Microsoft Corporation

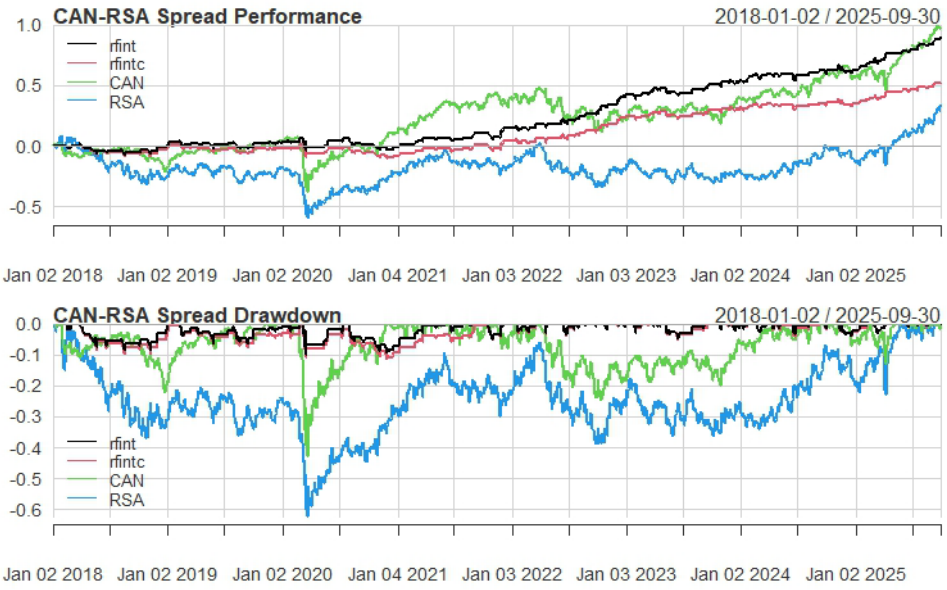

Figure 3. EWC-EZA pairs strategy comparison (testing range)

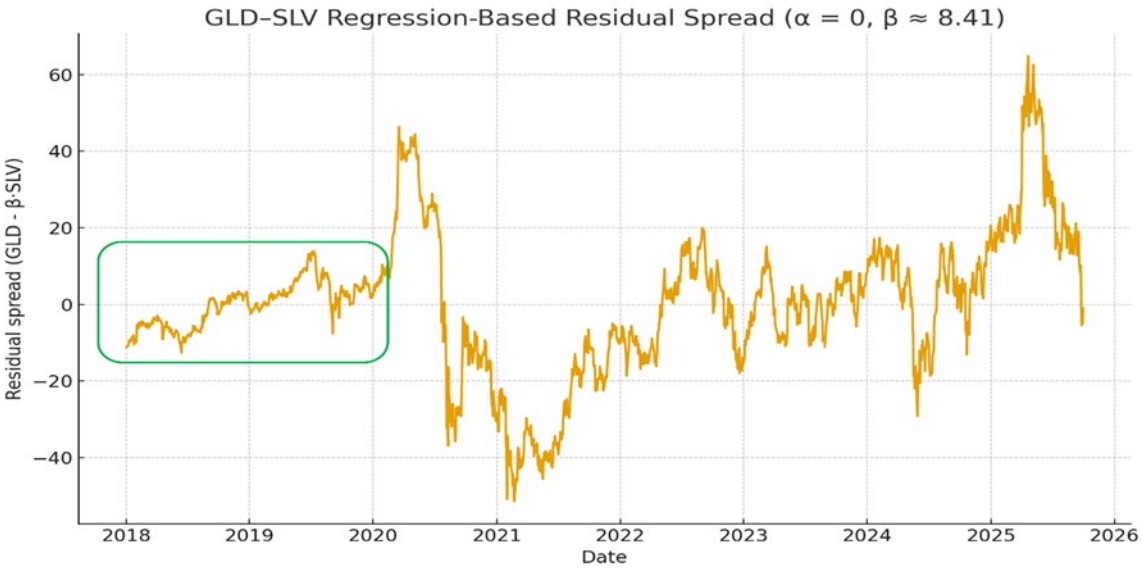

Figure 4. GLD-SLV residual spread (testing range)

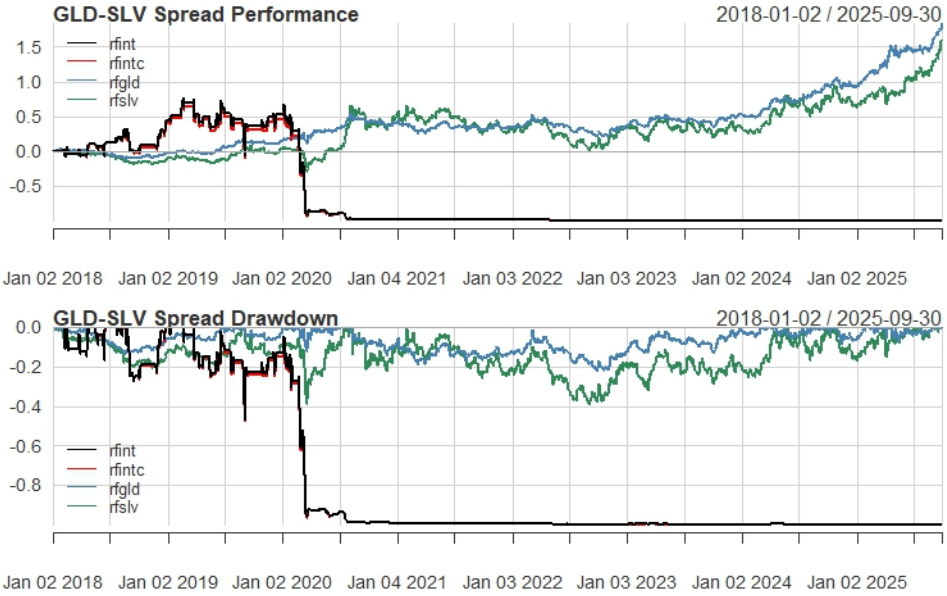

Figure 5. GLD-SLV pairs strategy comparison (testing range)

Table 8. Engle-Granger test results (training range) – single-pair spreads

Test	AUS_CAN	CAN_RSA	AUS_RSA	GER_FRA	NED_FRA	GER_NED	AUS_BHP
ADF p-value	0.01	0.02	0.14	0.09	0.03	0.26	0.27
PP p-value	0.01	0.01	0.03	0.06	0.01	0.26	0.66

Test	CAN_SU	AUS_GLD	CAN_WTI	EWZ_VALE	RSA_GLD	SPY_DIA	AAPL_MSFT
ADF p-value	0.23	0.62	0.36	0.06	0.47	0.99	0.61
PP p-value	0.05	0.71	0.40	0.02	0.46	0.98	0.64

Test	GLD_GDX	GLD_IEO	XLE_IEO	LQD_SPY	GDX_IAU	GLD_SLV	XLE_OIH
ADF p-value	0.61	0.55	0.17	0.89	0.56	0.38	0.09
PP p-value	0.52	0.63	0.07	0.91	0.48	0.42	0.08

Test	KO_PEP	XLF_KRE	XOM_CVX	WMT_COST	VTI_SPY	OIH_USO
ADF p-value	0.80	0.41	0.66	0.47	0.49	0.37
PP p-value	0.84	0.53	0.55	0.58	0.41	0.45

Table 9. Johansen test results (training range) – multi-asset spreads

Spread	K	test_r2	crit10_r2	crit5_r2	crit1_r2	test_r1	crit10_r1	crit5_r1	crit1_r1	test_r0	crit10_r0	crit5_r0	crit1_r0
AUS_CAN_RSA	4	4.48	6.5	8.18	11.65	18.87	15.66	17.95	23.52	34.08	28.71	31.52	37.22
GER_NED_FRA	2	0.00	6.5	8.18	11.65	10.11	15.66	17.95	23.52	22.12	28.71	31.52	37.22
CAN_SU_WTI	9	2.49	6.5	8.18	11.65	10.47	15.66	17.95	23.52	25.34	28.71	31.52	37.22
AUS_RSA_GLD	3	3.66	6.5	8.18	11.65	9.48	15.66	17.95	23.52	33.83	28.71	31.52	37.22
SPY_AAPL_MSFT	2	1.71	6.5	8.18	11.65	8.36	15.66	17.95	23.52	15.15	28.71	31.52	37.22

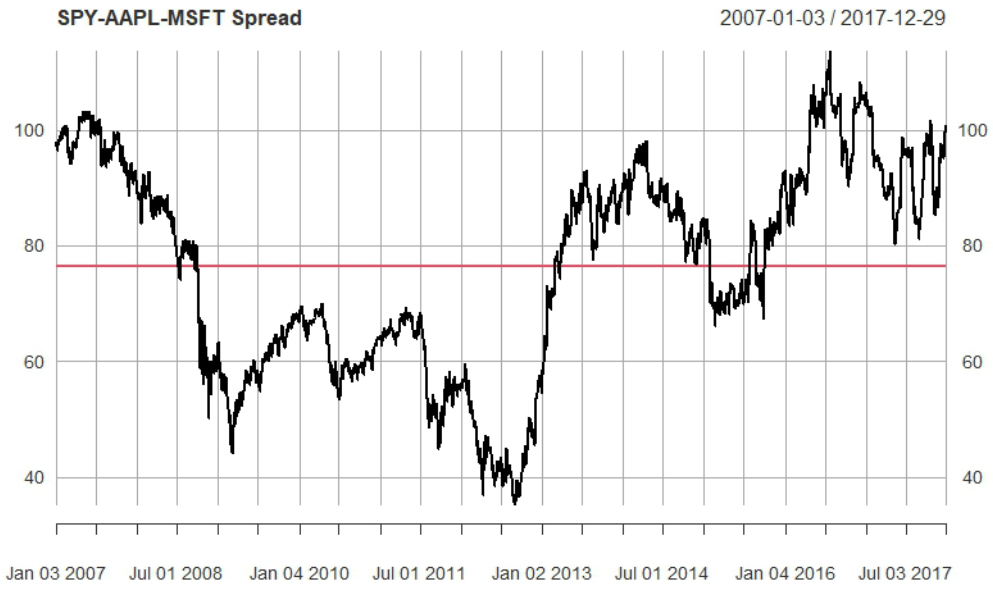

Figure 6. SPY-AAPL-MSFT spread (training range)

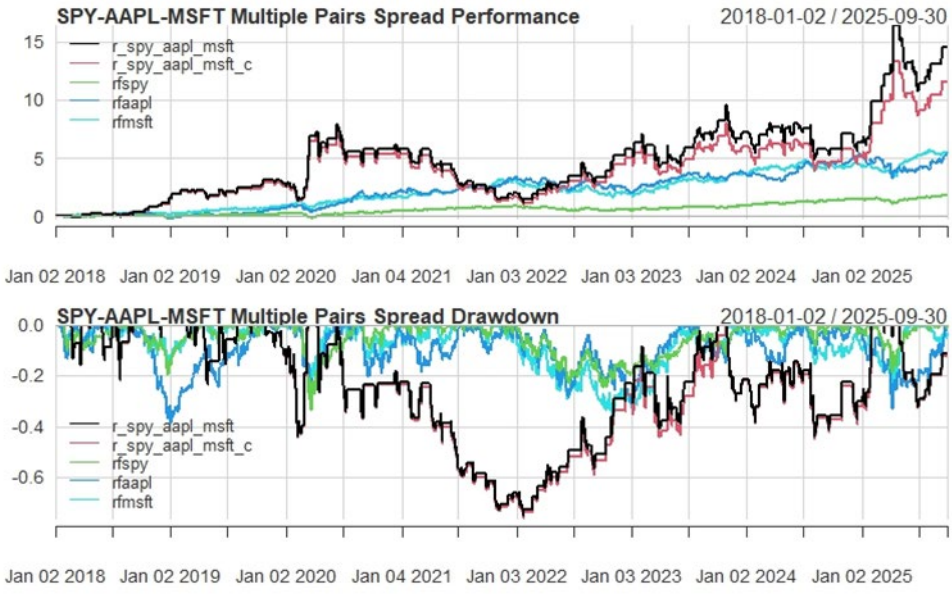

Figure 7. SPY-AAPL-MSFT multi-asset strategy comparison in the testing range

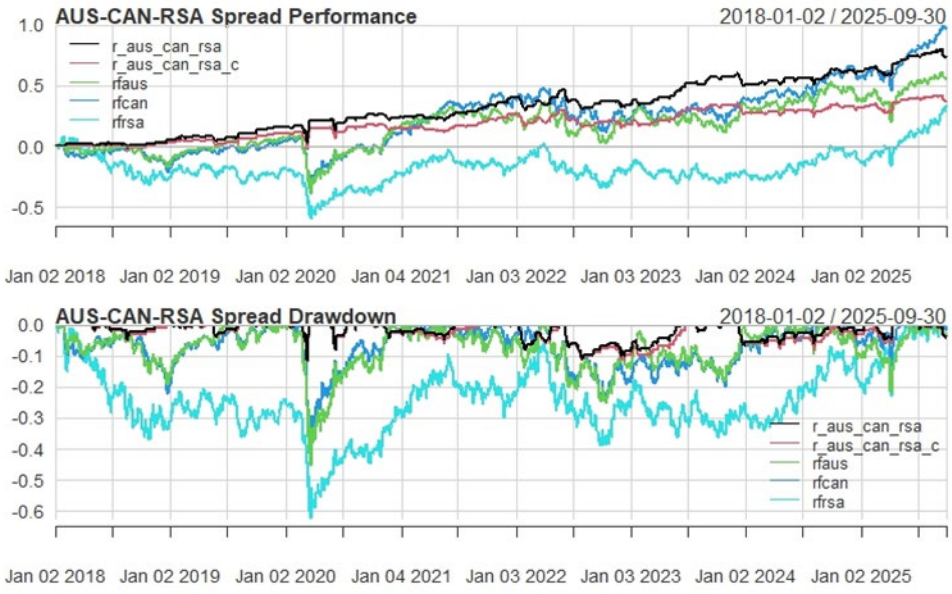

Figure 8. EWA-EWC-EZA multi-pairs strategy comparison (testing range)

Figure 9. EWA-EWC pairs strategy comparison (testing range)