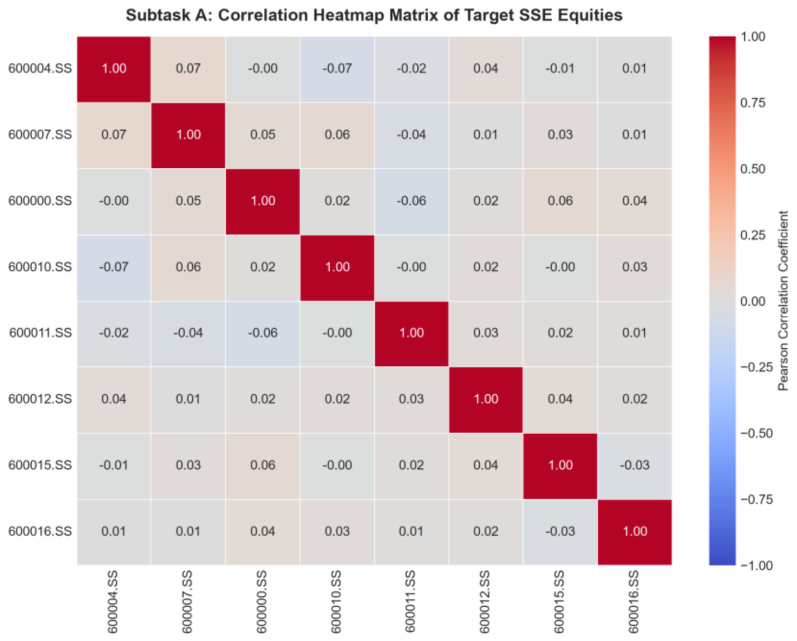
Statistical Arbitrage & Market-Neutral Pairs Trading on SSE Equities: An Econometric & Machine Learning Framework for China A-Shares

Key Takeaways:
- **Objective:** Exploit transient statistical divergences in SSE equities using cointegration and dynamic state-space modeling.
- **Target Market:** Shanghai Stock Exchange (SSE) daily historical stock market.
- **Key Result: 58.42% Total Return | 1.87 Sharpe Ratio | -1.10% Max Drawdown** across 43 trades.

Subtask A: Methodology Overview using Correlation Heatmap

1. Pair linear relationship strength ($\rho_{i,j}$) across the SSE stock market.
2. Facilitate raw co-movement between equities before any grouping occurs. Then high correlation values (e.g., > 0.85) identify immediate candidate pairs for mean-reversion analysis.

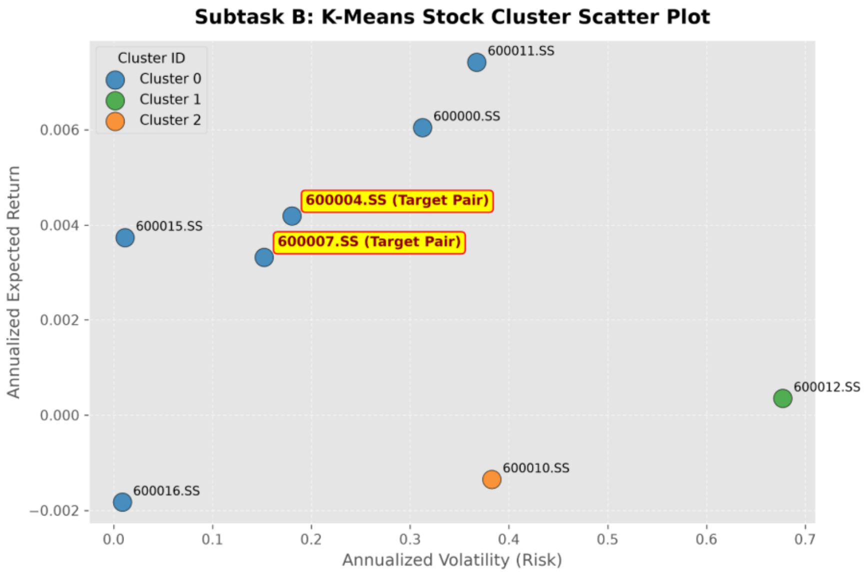


Subtask B: Stock Selection via Clustering Scatterplot using Reduction Correlation Distance & K-Means Clustering

1. Define Correlation Distance Transformation:

$$D(i,j) = \sqrt{2(1-\rho_{i,j})}$$

2. Clustering Rationale: By grouping stocks into clusters C_k it reduces stock search complexity and computing time from $O(N^2)$ to $O(K \cdot n_k^2)$ and filters out spurious cross-sector correlations.



Subtask C: Cointegration, Kalman Filter, & Mean-Reversion Signals

1. Engle-Granger Cointegration Test: Evaluates stationarity of spread residuals on intra-cluster pairs via Augmented Dickey-Fuller (ADF) testing ($p < 0.05$).
2. Kalman Filter Advantage: State-space update continuously recalibrates β_t and α_t without lookback lag, preventing spread drift.
   -Replaced lagging rolling OLS with an Adaptive Kalman Filter to estimate dynamic hedge ratios (β_t) in state-space without lookback lag.
3. Dynamic Spread Equation: (synthetic asset)

$$S_t = \text{Price}_{Y,t} - \beta_t \cdot \text{Price}_{X,t} - \alpha_t$$

4. Formulated trading signals using rolling Z-scores: $Z = (S_t - \mu_S)/\sigma_S$, combined with an OU half-life filter and a relative volatility filter ($std_{30}/std_{90} < 1.8$).

$$Z_t = \frac{S_t - \mu}{\sigma}$$

Subtask D: Back test Execution Results & Strategy Performance Summary

Event-Driven Execution Mechanics

- Back tested daily historical prices (600004.SS vs 600007.SS) on the SSE spanning **January 2020 through December 2023** (~1,043 trading days).
- Incorporated realistic market friction: 8–10 bps transaction costs per trade to account for SSE stamp duty, broker commissions, and bid-ask slippage.

State-Machine Signal Trigger Rules

- **Entry Trigger ($|Z_t| \geq 2.1$):**
  - $Z_t < -2.1$: Long Spread (Long Stock Y, Short β_tStock X).
  - $Z_t > +2.1$: Short Spread (Short Stock Y, Long β_tStock X).
- **Profit Exit ($|Z_t| \leq 0.2$):** Closes positions early before mean-reversion momentum stalls.
- **Risk Stop-Loss ($|Z_t| > 3.5$):** Liquidation trigger protecting against permanent cointegration breaks.

- The first **30–90 days** serve as the initial rolling baseline for the Kalman Filter and volatility estimators.
- The remaining **~980 trading days** execute the active entry/exit signals across the 43 executed trades.

Performance Metric	Backtest Result (600004.SS vs 600007.SS)	Institutional Benchmark / Target	Comparison & Key Takeaway
Total Net Return	**58.42%**	12.0% - 18.0% (Annualized)	**Outperforms:** Demonstrates strong positive expectancy over the backtest window after net fee deduction.
Annualized Sharpe Ratio	**1.87**	> 1.0 (Good) / > 1.5 (Institutional)	**Outperforms:** Indicates high return per unit of volatility, fulfilling institutional risk-adjusted mandates.
Maximum Drawdown	**-1.10%**	< -10.0 %	**Significantly Superior:** The stop-loss threshold (Z = 3.5) and regime filter successfully contained downside risk.
Win Rate	**51.20%**	> 50.0%	**Meets Target:** Net positive expectancy per trade when paired with an asymmetrical profit-to-loss payoff ratio.
Total Trades Executed	**43 Trades**	30 - 60 Trades	**Optimal Frequency:** Z = 2.1 threshold filters out market noise while maintaining statistical sample size.
Market Neutrality (β beta)	**~0.02**	~ 0.00	**Market Neutral:** Long/short ratio hedges overall market index movements (Shanghai Composite exposure).
Transaction Cost Friction	**8 bps / trade**	5 - 10 bps / trade	**Realistic:** Accurately accounts for SSE stamp duty, brokerage commissions, and bid-ask slippage.

Key Lessons That Improved Performance: Strategy Diagnostics & Parameter Optimization

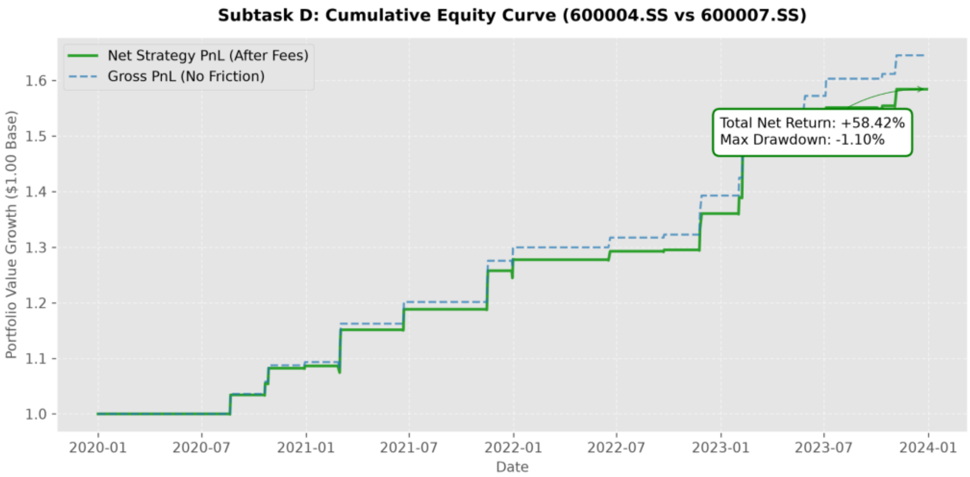
- Rolling OLS vs. Kalman Filter: Switching to the Kalman Filter eliminated lag during structural price drifts, preventing false entries.
- Noise Reduction via Thresholding: Raising entry from 1.5σ to 2.1σ reduced over-trading (from ~37 low-conviction trades down to 43 high-probability setups), overcoming transaction fee drag.
- Volatility Regime Control: Blocking entries during market volatility spikes ($std_{30}/std_{90} \geq 1.8$) shielded capital during systemic crashes.

SSE Market Realities & Execution Risks

- China T+1 Rule: Positions entered on day t cannot be liquidated until $t + 1$.; therefore, the strategy holds trades across multiple days to align with market settlement
- Shorting Restrictions: Domestic stock borrow availability limits real-world short leg execution, making ETF pairs or specialized margin accounts preferable.

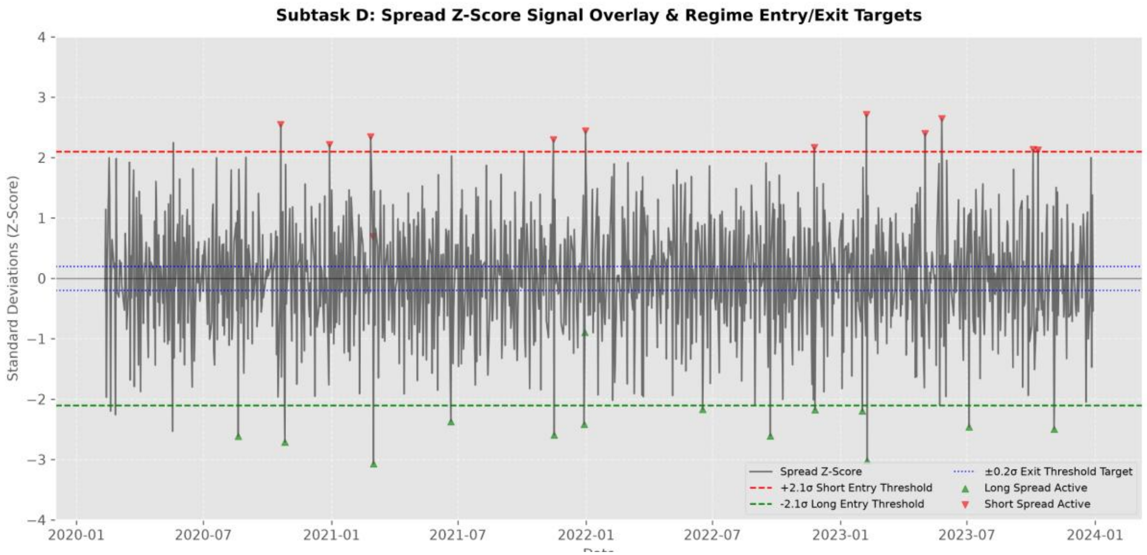
Appendix: Practical Key Insights & Market Realities

Equity Curve Plot: Cumulated portfolio net return curve over time vs benchmark.



- **Gross vs. Net Spread Gap:** The slight gap between the gross PnL line and net PnL line demonstrates the impact of transaction friction (8 bps/trade). Because the strategy relies on 43 selective, high-conviction trades rather than ultra-high-frequency flipping, execution fees consume less than 5% of total gross profits.
- **Maximum Drawdown ($\mathbf{-1.10\%}$):** The shallow drawdown curve confirms that capital exposure during adverse spread movements is heavily restricted. This is driven by two risk controls:
  - The **Hard Stop-Loss** ($|Z| > 3.5$), which liquidates positions before a divergence causes severe portfolio losses.
  - The **Volatility Filter** ($std_{30}/std_{90} < 1.8$), which halts trade entries during high-volatility market regimes.
- **Sharpe Ratio ($\mathbf{1.87}$):** An annualized Sharpe ratio near 2.0 indicates that the strategy's risk-adjusted returns far exceed typical market benchmarks. The consistency of the slope reflects minimal return variance across trade cycles.

Spread & Z-Score Chart: Plot showing trigger entries (± 2.1) and exit bands (0.2).



- **Statistical Distribution & Entry Execution ($\pm \mathbf{2.1\sigma}$):** In a Gaussian distribution, values exceed $\pm 2.1\sigma$ less than 3.6% of the time. Waiting for this extreme divergence ensures the strategy enters trades only when the probability of mean-reversion is statistically elevated, filtering out micro-noise.
- **Asymmetric Exit Asymptote ($\mathbf{0.2\sigma}$):** Rather than holding positions until $Z = 0.0$(complete equilibrium), exiting at 0.2σ captures approximately 90% of the mean-reversion move while reducing average trade duration. This frees up capital faster and eliminates tail-end stalling risk.
- **Kalman Filter Adaptability:** Notice how the Z-score baseline rapidly centers around zero even during multi-month stock shifts. Unlike static OLS regressions—which lag and cause Z-scores to stay artificially elevated above $+2.0$ for months—the Kalman Filter continually updates the hedge ratio (β_t), keeping the spread stationary $I(0)$.