

High-Frequency Stock Pairs Trading: Final Research Report

Heyaya Ren

*Department of Statistics, University of Wisconsin–Madison, Madison, USA
renheyaya@gmail.com*

Abstract. This report presents a high-frequency pairs-trading study centered on tickers 600606 and 600604. The project progressed from a literature review and initial minute-level demonstration to a systematic analysis of 3,626 daily MATLAB files covering 2005 through 2020. Returns were constructed at six sampling frequencies to examine how measured comovement changes with the observation interval. The analysis then compared return correlations, normalized the two price series, constructed their relative spread, and adjusted historical prices for corporate actions using the latest available split and dividend dictionary. A rolling 20-trading-day Z-score framework was subsequently applied to generate mean-reversion trading signals. The backtest used fixed \$500 long and \$500 short positions and evaluated cumulative PnL, total return, winning probability, win-to-loss ratio, maximum drawdown, and Sharpe ratios. The results showed that return correlation generally increased as the sampling interval became longer. After corporate-action adjustment, the trading strategy produced net PnL of \$1,665.10 and a net return of 166.51% at a 0.07% round-trip transaction cost. Transaction-cost robustness remained positive through 0.20%, with an estimated break-even round-trip cost of 0.90254%. Overall, the study provides a transparent empirical implementation of high-frequency pairs trading and evaluates its performance under realistic transaction-cost assumptions.

Keywords: pairs trading, high-frequency data, mean reversion, z-score, transaction costs

1. Introduction and literature review

Pairs trading is a relative-value strategy that seeks to profit from temporary divergences between securities whose prices or returns have historically moved together. Gatev et al. [1] provided one of the best-known empirical formulations of the strategy by matching stocks according to the distance between normalized historical price paths and trading subsequent divergences. Elliott et al. [2] developed an analytical mean-reversion framework for the spread, while Lin et al. [3] used cointegration and minimum-profit bounds to integrate loss protection into a pairs-trading procedure.

Later research examined whether simple pairs-trading profits persist and whether they remain economically meaningful after realistic market frictions. Do and Faff [4] found that profitability had declined over time but remained strong during periods of market turbulence. Bowen et al. [5] showed in a high-frequency setting that returns are highly sensitive to transaction costs and execution speed. Do and Faff [6] likewise found that commissions, market impact, and short-selling costs materially reduce profits, although carefully matched pairs can remain profitable.

The literature also developed alternative approaches to pair formation and spread modeling. Avellaneda and Lee [7] generated statistical-arbitrage signals after removing common return components using principal components or sector exchange-traded funds. Caldeira and Moura [8] used cointegration to select a portfolio of pairs and model mean-reverting spreads. Liew and Wu [9] proposed a copula-based approach to model dependence beyond simple correlation or cointegration. Jacobs and Weber [10] showed that pairs-trading profitability varies with the type of news driving pair divergence, investor attention, and limits to arbitrage.

Huck [11] compared distance, stationarity, and cointegration approaches and examined whether volatility timing adds economic value.

Broader comparative evidence reinforces the importance of method selection and market frictions. Rad et al. [12] compared distance, cointegration, and copula strategies across the U.S. equity market using time-varying trading costs. Krauss [13] reviewed the major methodological families in pairs trading, including distance, cointegration, time-series, stochastic-control, and other approaches. These studies provide a useful context for interpreting a transparent Z-score-based implementation as one member of a much larger statistical-arbitrage literature.

High-frequency data introduce an additional issue because measured dependence can change with the sampling interval. Epps [14] documented substantial interval sensitivity in very-short-run stock-price comovement. This observation is directly relevant to the present comparison of return correlations at 1-minute, 5-minute, 10-minute, 1-hour, 2-hour, and daily frequencies. Stübinger and Endres [15] later developed a mean-reverting jump-diffusion pairs-trading model using minute-by-minute data and explicitly incorporated individualized entry and exit thresholds and transaction costs.

Within this literature, the present project is a focused empirical application rather than a comparison of every available pair-selection model. The study first evaluates return correlation across sampling frequencies, then constructs normalized prices and a spread for tickers 600606 and 600604, corrects the historical series for corporate actions, applies a rolling Z-score trading rule, and evaluates PnL and transaction-cost robustness.

2. Data and initial preparation

The project began with the assigned stock list file `stock_list_20200831.csv`, the example minute-level file `minutemat_20200630.mat`, and a stock split and dividend dictionary. The stock list was reviewed together with GICS industry classification information to support pair selection. The minute-level MATLAB files contain the stock identifier, intraday time, and transaction price. The split and dividend dictionary contains corporate-action information needed later to prevent mechanical ex-rights and ex-dividend jumps from being interpreted as trading signals.

The broader research workflow was designed to move from a small demonstration to a systematic empirical analysis. The early work established the required return calculation, sampling frequencies, and correlation procedure before the analysis was expanded to the larger high-frequency dataset.

3. Initial demonstration analysis

The first empirical task was to select a pair of stocks and work directly with minute-level data. For the selected pair, returns were computed at six sampling frequencies: 1-minute, 5-minute, 10-minute, 1-hour, 2-hour, and daily. Returns were calculated as $P2/P1 - 1$, where $P1$ is the earlier price and $P2$ is the later price. The correlation between the two stocks' return series was then calculated separately at each frequency.

The initial demonstration was expanded to more consecutive trading days so that the analysis could evaluate whether return correlations were persistent or time-varying and whether the relationship changed when more observations were included. This stage established a repeatable procedure for the later full-dataset analysis.

4. Comprehensive multi-file analysis

The analysis was subsequently expanded to the comprehensive price_minute_level dataset. A MATLAB workflow processed 3,626 daily .mat files covering 2005 through 2020. The files were loaded in chronological order, observations for tickers 600606 and 600604 were selected, the stocks were matched on common intraday times, and usable matched prices were retained. Return series were then constructed at the same six sampling frequencies used in the demonstration analysis.

The full-sample results show a clear sampling-frequency effect. Correlations are lowest at the highest sampling frequency and rise as the interval becomes longer. The purpose of this comparison is descriptive: it demonstrates that measured comovement depends materially on the observation interval, consistent with the sampling-interval issue documented by Epps [14]. The return correlations by sampling frequency are summarized in Table 1.

Table 1. Return correlation by sampling frequency

Sampling Frequency	Return Correlation
1-minute	0.083013
5-minute	0.23632
10-minute	0.26042
1-hour	0.26957
2-hour	0.34640
Daily	0.35623

5. Normalized price and spread analysis

After the return-correlation analysis, the two minute-level price series were combined across the sample and normalized so that each series equaled 1 at the beginning of the sample. This placed the two securities on a common scale and made their relative movements directly comparable.

The use of normalized price paths is consistent with the distance-based pairs-trading framework of Gatev et al. [1].

The relative spread was defined as the normalized price of ticker 600606 minus the normalized price of ticker 600604. The normalized series and the spread were examined over time, and the mean and standard deviation of each normalized price series were computed as summary statistics. The summary statistics are reported in Table 2, while the normalized price series and spread are shown in Figures 1 and 2, respectively.

Table 2. Summary statistics of the original normalized price series

Ticker	Mean	Standard Deviation
600606	2.4197	1.2297
600604	3.2841	2.2389

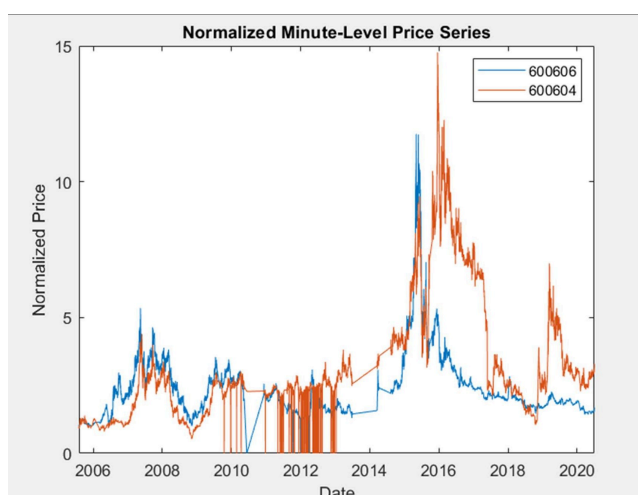


Figure 1. Normalized minute-level price series

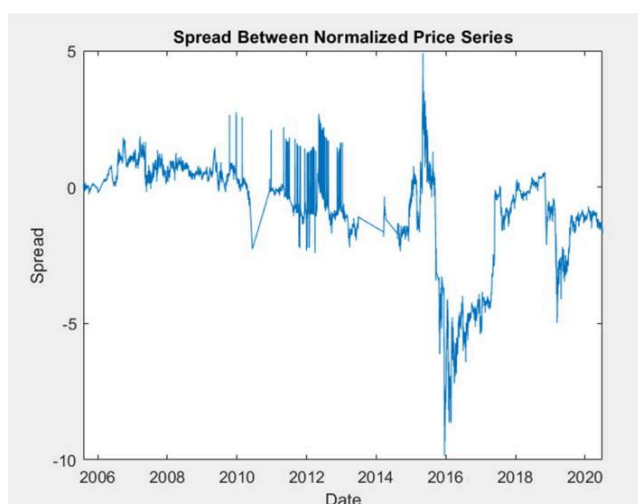


Figure 2. Spread between normalized price series

6. Corporate action adjustments

The original normalized price and spread series contained large discontinuities that could distort mean-reversion signals. To address this issue, the analysis incorporated a stock split and dividend dictionary. The originally referenced 20231231 version was not available in the shared files, so the available alternatives were checked with Professor Chi. He instructed that the latest version be used, and the analysis therefore used 202505_StockSplitDividendDictionary.csv.

For each corporate-action event affecting 600606.SH or 600604.SH, the ex-rights or ex-dividend date, cash dividend, stock dividend proportion, and share-increase proportion were identified. Historical prices before the event date were adjusted using an event-specific factor based on the prior usable close, the cash distribution, and the change in shares. The factors were applied chronologically. In total, 21 relevant corporate-action events were incorporated: 13 for 600606.SH and 8 for 600604.SH.

After these corrections, the adjusted normalized price series and adjusted spread were recomputed. The purpose of this step was to prevent known mechanical corporate-action changes from being treated as mean-reversion signals. The corporate-action adjustment summary is reported

in Table 3, while the adjusted normalized price series and adjusted spread are shown in Figures 3 and 4, respectively.

Table 3. Corporate action adjustment summary

Ticker	Number of Events Applied	Adjustment Input
600606.SH	13	Cash dividends and share changes
600604.SH	8	Cash dividends and share changes
Total	21	202505_StockSplitDividendDictionary.csv

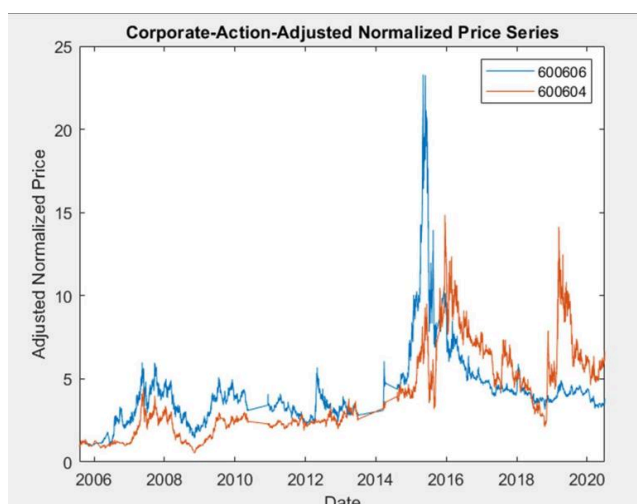


Figure 3. Corporate-action-adjusted normalized price series

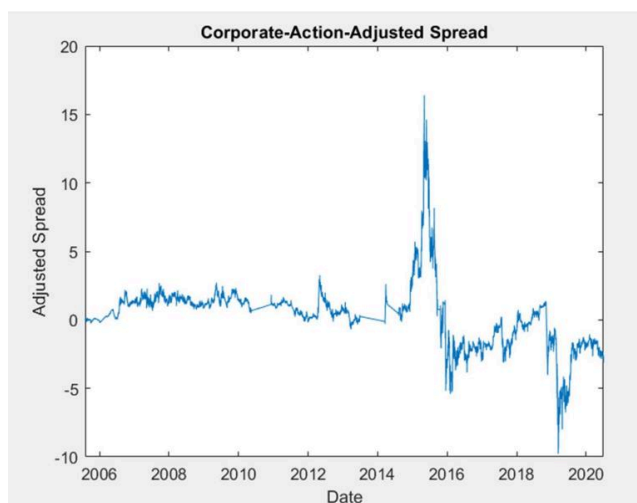


Figure 4. Corporate-action-adjusted spread

7. Rolling Z-score estimation and trading rules

The adjusted spread was converted into a rolling Z-score trading signal. The model used a 20-trading-day lookback window to estimate the spread mean and standard deviation from preceding available trading days. For each current observation, the Z-score was calculated as the current adjusted spread minus the rolling mean, divided by the rolling standard deviation. Standardizing a

mean-reverting spread is consistent with the broader time-series logic used in pairs-trading models such as Elliott et al. [2] and Stübinger and Endres [15].

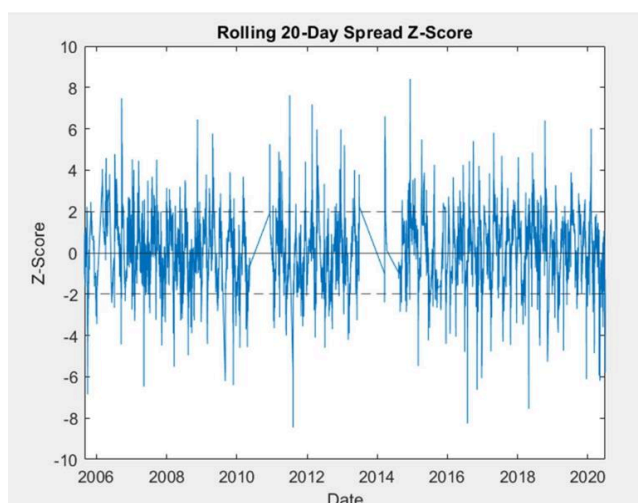


Figure 5. Rolling 20-day spread Z-score

When the Z-score reached or exceeded 2, ticker 600606 was treated as relatively expensive and the strategy entered a short position in 600606 together with a long position in 600604. When the Z-score reached or fell below -2, the strategy entered a long position in 600606 and a short position in 600604. An open trade was closed when the Z-score returned to or crossed zero. The resulting rolling 20-day spread Z-score is shown in Figure 5.

8. PnL simulation and performance measurement

Each trade used a fixed gross pair notional of \$1,000, divided into a \$500 long leg and a \$500 short leg. The quantity of each stock was determined from its entry price so that each leg had \$500 of dollar exposure at entry. Quantities were then held fixed until the exit signal. Minute-by-minute PnL was accumulated across the trading history.

Performance was evaluated using total PnL, total return relative to the \$1,000 gross pair notional, winning probability, win-to-loss ratio, maximum drawdown, and two Sharpe ratios. Winning probability is the percentage of completed trades with positive PnL. The win-to-loss ratio compares the average winning trade with the absolute value of the average losing trade.

Maximum drawdown measures the largest peak-to-trough decline in cumulative PnL. The Sharpe ratio was calculated both with all minute observations, including zero-PnL periods, and with zero-trade minutes excluded. The performance measures are summarized in Table 4.

Table 4. Trading performance summary

Metric	Gross	Net after 0.07% Cost
Total PnL	\$1,805.10	\$1,665.10
Total Return	180.51%	166.51%
Winning Probability	78.50%	78.50%
Win-to-Loss Ratio	0.41508	0.40302
Maximum Drawdown	\$1,479.60	\$1,521.60
Sharpe Ratio with Zero Minutes	0.37584	0.34683

Table 4. (continued)

Sharpe Ratio without Zero Minutes	0.48543	0.44785
-----------------------------------	---------	---------

At the baseline transaction-cost level, the strategy generated net PnL of \$1,665.10 and a net return of 166.51%. The winning probability was 78.50%. The win-to-loss ratio was below 1, indicating that the average losing trade was larger in magnitude than the average winning trade even though winning trades occurred more frequently. The maximum drawdown was \$1,521.60 after costs. The net Sharpe ratio was 0.34683 when zero minutes were included and 0.44785 when zero-trade minutes were excluded. The minute-by-minute cumulative trading PnL is shown in Figure 6.

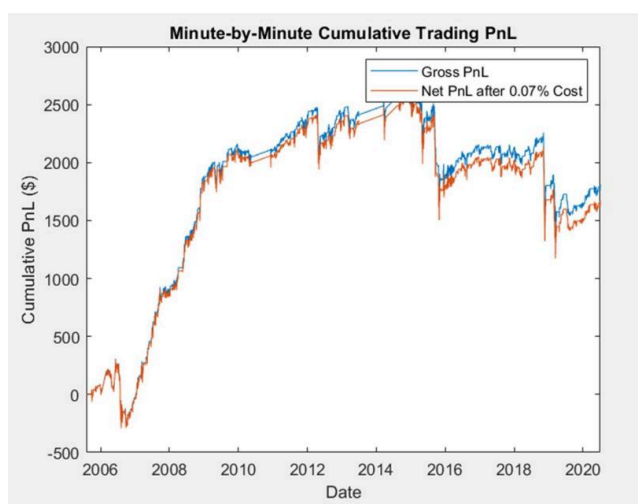


Figure 6. Minute-by-minute cumulative trading PnL

9. Transaction cost robustness and break-even analysis

The final empirical step incorporated China A-share transaction costs. The baseline round-trip cost was set to 0.07% and applied to the \$1,000 gross pair notional for each completed trade. Robustness was evaluated by increasing the round-trip cost from 0.07% to 0.20% in 0.01 percentage-point increments. This robustness exercise is motivated by the pairs-trading literature's finding that market frictions and execution assumptions can materially affect profitability [5, 6, 12]. The transaction-cost robustness results are summarized in Table 5.

Table 5. Transaction cost robustness

Round-Trip Cost	Net PnL	Net Return	Profitable
0.07%	\$1,665.10	166.51%	Yes
0.08%	\$1,645.10	164.51%	Yes
0.09%	\$1,625.10	162.51%	Yes
0.10%	\$1,605.10	160.51%	Yes
0.11%	\$1,585.10	158.51%	Yes
0.12%	\$1,565.10	156.51%	Yes
0.13%	\$1,545.10	154.51%	Yes
0.14%	\$1,525.10	152.51%	Yes
0.15%	\$1,505.10	150.51%	Yes

Table 5. (continued)

0.16%	\$1,485.10	148.51%	Yes
0.17%	\$1,465.10	146.51%	Yes
0.18%	\$1,445.10	144.51%	Yes
0.19%	\$1,425.10	142.51%	Yes
0.20%	\$1,405.10	140.51%	Yes

The strategy remained profitable at every tested transaction-cost level through 0.20%. The estimated break-even round-trip transaction cost was 0.90254%, approximately 0.90%. This value is substantially above the 0.07% baseline used in the primary result. The break-even transaction cost is reported in Table 6.

Table 6. Break-even transaction cost

Measure	Result
Break-Even Round-Trip Cost	0.90254%

10. Limitations

This study has several limitations. First, the final systematic backtest focuses on one fixed pair, 600606 and 600604, so the reported performance should not be interpreted as evidence that the same result would hold across a broad universe of China A-share pairs. Second, the sample covers 2005 through 2020, and market structure, liquidity, regulation, short-sale conditions, and trading technology may differ outside this period.

Third, the corporate-action adjustment removes known dividend and share-change discontinuities but does not model every possible structural change affecting either company. Long suspensions, reorganizations, changing fundamentals, or regime shifts may alter the economic relationship between the securities. Fourth, the transaction-cost model uses a fixed percentage round-trip cost and sensitivity range rather than separately modeling bid-ask spreads, slippage, market impact, short-borrow availability, or time-varying execution costs. These frictions are important in high-frequency pairs trading [5, 6].

Finally, the strategy uses a transparent rolling Z-score rule rather than optimizing across alternative pair-selection methods or stochastic models. The reported total return is defined relative to the fixed \$1,000 gross pair notional used in the backtest rather than as a compounded portfolio return.

11. Future work

Although the active empirical phase of this project is complete, future research could apply the same workflow to many pairs and compare same-industry and cross-industry relationships. Pair formation could also be compared across distance, cointegration, and copula methods, as examined by Caldeira and Moura [8], Liew and Wu [9], Rad et al. [12], and reviewed by Krauss [13].

Future work could also investigate structural-break or regime-aware estimation, use explicit out-of-sample or walk-forward pair selection, and compare the current rolling Z-score rule with richer mean-reversion models such as the jump-diffusion framework of Stübinger and Endres [15]. A more detailed execution model could incorporate time-varying bid-ask spreads, slippage, market impact,

and short-borrow constraints. These possibilities are presented only as future research directions and are not additional requirements for the completed project.

12. Conclusion and project closure

The completed research progressed from literature review and data familiarization to a systematic high-frequency pairs-trading backtest. The early analysis established return construction at multiple frequencies and showed that the measured return correlation between the two stocks increases as the sampling interval becomes longer. The comprehensive MATLAB analysis then extended the work across 3,626 daily files, normalized the two price series, constructed their relative spread, and provided the data foundation for the trading model.

The final stage adjusted historical prices for dividends and share changes, estimated a rolling 20-day Z-score signal, implemented systematic entry and exit rules, simulated \$500 long and \$500 short positions, and evaluated performance with and without transaction costs. At the required 0.07% round-trip cost, the strategy produced net PnL of \$1,665.10 and a net return of 166.51%. The cost analysis remained profitable through the required 0.20% robustness range and produced an estimated break-even round-trip cost of approximately 0.90%.

These results complete the streamlined empirical requirements established for the project. No additional backtests, sample expansions, regression-based pair-selection exercises, or further model variations are performed in this report because the active research and empirical modeling phase has been formally concluded.

References

- [1] Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3), 797-827. <https://doi.org/10.1093/rfs/hhj020>
- [2] Elliott, R. J., Van Der Hoek, J., & Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance*, 5(3), 271-276. <https://doi.org/10.1080/14697680500149370>
- [3] Lin, Y.-X., McCrae, M., & Gulati, C. (2006). Loss protection in pairs trading through minimum profit bounds: A cointegration approach. *Journal of Applied Mathematics and Decision Sciences*, 2006, Article 73803. <https://doi.org/10.1155/JAMDS/2006/73803>
- [4] Do, B., & Faff, R. (2010). Does simple pairs trading still work? *Financial Analysts Journal*, 66(4), 83-95. <https://doi.org/10.2469/faj.v66.n4.1>
- [5] Bowen, D. A., Hutchinson, M. C., & O'Sullivan, N. (2010). High-frequency equity pairs trading: Transaction costs, speed of execution, and patterns in returns. *The Journal of Trading*, 5(3), 31-38. <https://doi.org/10.3905/jot.2010.5.3.031>
- [6] Do, B., & Faff, R. (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2), 261-287. <https://doi.org/10.1111/j.1475-6803.2012.01317.x>
- [7] Avellaneda, M., & Lee, J.-H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7), 761-782. <https://doi.org/10.1080/14697680903124632>
- [8] Caldeira, J. F., & Moura, G. V. (2013). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *Brazilian Review of Finance*, 11(1), 49-80.
- [9] Liew, R. Q., & Wu, Y. (2013). Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19(1), 12-30. <https://doi.org/10.1057/jdhf.2013.1>
- [10] Jacobs, H., & Weber, M. (2015). On the determinants of pairs trading profitability. *Journal of Financial Markets*, 23, 75-97. <https://doi.org/10.1016/j.finmar.2014.12.001>
- [11] Huck, N. (2015). Pairs trading: Does volatility timing matter? *Applied Economics*, 47(57), 6239-6256. <https://doi.org/10.1080/00036846.2015.1068923>
- [12] Rad, H., Low, R. K. Y., & Faff, R. (2016). The profitability of pairs trading strategies: Distance, cointegration and copula methods. *Quantitative Finance*, 16(10), 1541-1558. <https://doi.org/10.1080/14697688.2016.1164337>
- [13] Krauss, C. (2017). Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2), 513-545. <https://doi.org/10.1111/joes.12153>

- [14] Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a), 291-298. <https://doi.org/10.1080/01621459.1979.10482508>
- [15] Stübinger, J., & Endres, S. (2018). Pairs trading with a mean-reverting jump-diffusion model on high-frequency data. *Quantitative Finance*, 18(10), 1735-1751. <https://doi.org/10.1080/14697688.2017.1417624>