

Statistical projection of trading profitability in high-frequency pairs trading: empirical evidence from Z-scores and return correlations

Heng Du

Middlebury College, Vermont, USA

hengrdu@outlook.com

Abstract. This paper examines whether historical return correlations and Z-score signals help predict profitable pairs trades using high-frequency equity data. Stocks are matched by industry and market capitalization, and minute-by-minute prices are divided into separate lookback and trading periods of equivalent lengths. Historical correlations are calculated at several frequencies (5-minute, 10-minute, 1-hour, and 2-hour), while trading-period Z-scores are computed from spread parameters estimated during formation. The results provide insufficient evidence that stronger historical correlation predicts higher subsequent gross profit and loss at any of the frequencies examined. The current Z-score has a statistically significant association with gross P&L, but the regression explains almost none of the variation in trading outcomes. Most predicted returns also fall below the lowest transaction-cost threshold, leaving few potentially profitable opportunities under the assumed costs. Overall, these results suggest that historical convergence of stock pairs does not guarantee economic profitability. They also illustrate the reasons for statistically significant signals to be assessed concurrently with its predictability and associated costs. Within the sample studied, historical pair similarity and spread deviations alone offer limited guidance for identifying trades that remain profitable after execution costs.

Keywords: high-frequency, correlation, spread, convergence, P&L, transaction costs

1. Introduction

Pairs trading is a "short-term speculation" strategy commonly used by hedge funds and investment banks. It is based on the selection of two stocks whose prices moved together historically. As stock prices widen temporarily, traders buy the undervalued stock and sell its overvalued counterpart in the hope that history would "repeat itself" in the future to grant profits [1]. This stratagem sheds new light on the scrutiny of market movements with its basis on the price relationships between linked assets, alluring the attention of a plethora of researchers specializing in the discipline. According to Evan Gatev and his research cohort, divergence between prices prompts traders to buy the relatively undervalued stock and selling its overvalued counterpart [1]. Should the two paths converge subsequently, a profit is made. Nevertheless, such returns' dependence on "price dynamics" and "simple contrarian principles" conceals a strong yet important assumption: Observed

historical relationships must remain sufficiently stable for a later divergence. Whether that assumption holds across market conditions remains an important empirical question.

This precondition underlies much of earlier research. The 2006 paper written by Gatev et al. offer foundational insights on the efficacy of pairs trading using U.S. stock data from 1962 to 2002 [1]. Their distance-based methods identify stock pairs with whose normalized prices follow similar historical paths. Even though the authors' results show that pairs trading can produce positive returns under reasonable assumptions of transaction costs and risk factors, extending their methodology to a high-frequency framework raises new obstacles in making inferences about convergence and economic profitability. Specifically, relationships estimated during the lookback period may become unstable when market conditions change. Moreover, effects of microstructure noise on intraday prices may be more pronounced in high-frequency scenarios, resembling convergences without representing a realizable trading opportunity. In his 2017 paper, Christopher Krauss accordingly emphasizes that historical similarity does not establish a stationary relationship, while Do and Faff show that gross pairs-trading returns can decline substantially once transaction costs and other implementation frictions are considered [1]. These concerns become particularly important at high frequencies because predicted returns are often small and repeated trading causes costs to accumulate rapidly.

This study addresses the gap between historical pair similarity, subsequent convergence, and economically implementable profitability. Traditional distance-based research largely evaluates relationships using daily observations, whereas this paper examines whether historical co-movement remains informative when returns and trading signals are measured using intraday data. This manuscript addresses two questions:

1. Does lookback return correlation across multiple intraday frequencies (5-minute, 10-minute, 1-hour, 2-hour) predict out-of-sample trading performance?
2. Do statistically significant Z-score signals yield economically viable profits after accounting for transaction costs?

The remainder of this paper is organized as follows. Section 2 reviews earlier research present and the analytical frameworks used by the authors, followed by statistical computations and interpretations in the next section. The paper takes a deeper dive in analyzing the possible reasons for the results to come into being, and the implications they have in trading. Finally, the last section summarizes this paper and proposes directions for future research.

2. Results

2.1. Literature review

2.1.1. Foundational arbitrage and strategy mechanics

Using U.S. stock market data from 1962 to December 2002 in their 2006 paper "*Pairs Trading: Performance of a Relative-Arbitrage Rule*", Gatev et al. present the sustainability of risk-adjusted returns from pairs trading and demonstrate the systematic opportunities arising from temporary mispricing between closely related securities [1]. Their findings suggest that even in relatively efficient markets, short-term deviations from fundamental value can persist long enough for disciplined, rule-based strategies to exploit them.

To address the potential biases in excess returns, Gatev and his colleagues implement a disciplined trading rule in which positions are liquidated immediately on the day following the crossing of the paired stocks' prices. Yet, the authors adopt a cautious stance by admitting the inherent unviability in distinguishing returns driven by mean reversion from those induced by bid-ask bounce [1]. Extending this concern, the researchers acknowledge the effects of transaction costs, demonstrating that although this variable exerts negative

influences on the magnitude of returns, it does not challenge pairs trading as an economically significant method for investment. This additional adjustment reinforces the argument that the profitability of pairs trading is not merely an artifact of trading frictions, but a mirror of sustained opportunities arising from temporary mispricing between related securities.

Continuing in this vein, the authors use bootstrap analysis to show that simulated pairs yield substantially lower returns, implicating that true pairs trading profits cannot be replicated by simple reversal-based strategies [1]. Moreover, the authors find asymmetries in the contributions of long and short positions within pairs. These findings strengthen the robustness of the strategy's theoretical foundation and highlight the challenges of executing it effectively in practice.

In addition, the researchers prudently acknowledge the limited bounds of conventional risk factors. Specifically, Gatev and his co-authors assert that neither microstructure effects nor asymmetries between stock positions of varied time frames sufficiently account for the observed returns. Rather, the persistence of excess returns points toward a latent and "dormant" risk factor that influences the co-movement of paired stocks. Overall, Gatev et al.'s research establishes pairs trading as a distinct and potentially profitable relative-value strategy [1]. Nevertheless, their findings still have an unresolved question of whether this quantitative edge remains sustainable as fluctuations occur in market conditions and trading frictions.

2.1.2. Friction decay and limits to arbitrage

While acknowledging the foundational contributions done by Gatev's cohort on the profitability of pairs trading in their 2010 paper "*Does Simple Pairs Trading Still Work*", Binh Do and Robert Faff examine whether this "quantitative edge" remains sustainable under changing market conditions and trading frictions. They argue that although pairs trading profitability declines over their sampled period [2], partly attributable to market competition and arbitrage risks. Plus, the increasing occurrence of nonconvergent pairs suggests that historical co-movement does not always guarantee that a temporary divergence will eventually correct itself. Do and Faff's findings therefore qualify Gatev et al.'s distance-based framework by showing that profitability not only depends on historical similarity, but also on future convergence [1].

Launching further from their research in 2010, Do and Faff provide a more refined method in their 2012 paper "*Are Pairs Trading Profits Robust to Trading Costs?*" to account for stock pairs' profitability after transaction costs beyond minimizing the Sum of Squared Deviations of normalized historical price paths (SSD) with the Number of Zero Crossings (NZC), which captures the frequency of spread reversals, thereby accentuating that successful pair selection depends not only on whether two stocks have moved closely together in the past, but also on whether their spread has shown a consistent tendency to reverse and oscillate [3]. Their results show that portfolios filtered from SSD and NZC tend to remain more profitable than those that rely on SSD alone [3]. Yet, improved pair selection does not guarantee profitability. When a pair does converge, the initial spread may be too small to cover the costs of trading. In this case, Do and Faff present a more comprehensive analysis from which they infer that transaction costs directly reduce the available return, market impact further limits the return that can be captured through execution, while short-selling costs place an additional burden on the short side of the strategy.

By emphasizing the cumulative effects of transaction costs, Do and Faff refine Gatev et al.'s framework and show that historical similarity alone is not enough to identify profitable pairs. Their studies point to two important limitations of the "quantitative edge" in pairs trading [2]. First, historical co-movement does not always lead to future convergence, particularly as arbitrage risks increase. Second, even when convergence occurs, the resulting gross return may be insufficient to offset transaction costs, market impact, and short-selling constraints. Together, these findings move the focus beyond identifying predictable price movements to whether those movements are reliable and large enough to remain profitable in practice.

2.1.3. Cointegration breakdown and out-of-sample stability

Despite the refinements in pair selection discussed above, the predictive ability of historical co-movement remains limited when relationships identified during the formation period are transferred to an out-of-sample scenario. In his 2017 paper "*Statistical Arbitrage Pairs Trading Strategies: Review and Outlook*", Christopher Krauss highlights the matter by distinguishing the Gatev et al.'s distance approach from methods that explicitly model long-run equilibrium relationships. He argues that historical similarity in stock price co-movements does not imply persistence, as "spurious relationships based on an assumption of return parity are not mean-reverting" [4]. Cointegration-based approaches place more emphasis on whether there is an equilibrium relationship connecting the price series, rather than only whether they have moved similarly in the past. The distinction provides an important explanation for why distance and correlation measures can perform poorly out-of-sample and therefore demonstrate insufficient predictive ability when applied to a later period. Two stocks can appear closely related during the formation period without their later deviations necessarily returning to their previous level.

In their 2010 paper "*Statistical Arbitrage in US Equities Market*", Marco Avellaneda and Jeong-Hyun Lee try to fill this void through the usage of the Exchange Trade Fund (ETF) as proxies for industrial factors and the application Principal Component Analysis (PCA) to return correlation matrices, along with calculations of eigenvectors and eigenvalues. The eigenvalues indicate the amount of variation explained by the corresponding principal components. The associated eigenvectors determine how individual stocks load onto these components and are used to construct eigenportfolios representing the common factors. Avellaneda and Lee observe that the dominant eigenvector is associated with the market portfolio because its stock coefficients are largely positive, while the corresponding portfolio weights are inversely related to individual stock volatility [5]. By extracting these common components, their approach can distinguish broad market or sector movements from the stock-specific residual component that forms the basis of their statistical-arbitrage signals.

Undeniably, Avellaneda and Lee's methods therefore provide an important improvement over selecting pairs from raw historical correlation alone. Although PCA allows them to separate broad patterns of market co-movement from stock-specific fluctuations, its reliability depends on estimating many relationships among stocks from a limited number of observations. As the number of estimated relationships grows relative to the available data, some may reflect temporary statistical noise rather than stable patterns of co-movement. Krauss specifically highlights this concern in reviewing Avellaneda and Lee, noting their observation that there are considerably more entries in the correlation matrix than available data points [4]. Thus, even a more developed correlation-based framework faces uncertainty over whether the relationships extracted from historical data will remain informative in a later trading period.

Taken together, Krauss and Avellaneda and Lee demonstrate that historical co-movement alone provides limited evidence of a stable relationship between stocks. Their findings therefore shift the central question from whether stocks moved together historically to whether that relationship remains sufficiently stable to support out-of-sample convergence.

2.2. Methodology and analytical framework

Building on Krauss and Avellaneda and Lee's research, the empirical setup of this paper is whether multi-frequency correlation captures genuine stationary relationships or merely non-executable microstructure noise during out-of-sample Z-score deployment, analyzed through the co-movement of two stocks from the same industry and market cap category. Price data are gathered over April 29 and June 30 of 2020, in which 40 valid trading days are retained after filtering out dates with insufficient trading information. To establish more

comparable sets of stock before investigating their price correlations, candidate pairs are identified by their GICS4 classification and their market cap size buckets, from which stocks of similar size can be matched within the same narrowly defined industry. Market capitalization is divided into five quintiles designated as "micro", "small", "mid", "big", and "large". The criteria employed for creating comparable stock pairs provide a baseline for the subsequent correlation analysis, since pair formation is not determined by historical price co-movement alone. After the candidate pairs are constructed, their one-minute intraday prices are collected and merged into a common dataset covering the full sample period.

To establish a clear separation between pair formation and subsequent trading, the 40-day sample is divided into two halves at the twentieth trading day. The first 20 trading days are designated as the lookback period, while observations after the cutoff date constitute as the trading period. This division is applied consistently across the analysis. During the first twenty days, the historical mean and standard deviation of each pair's normalized price spread are estimated. In the second half of the sample period, the historical spread parameters are held fixed rather than being updated in accordance with the observations. Z-scores and trading signals are then generated only from observations after the cutoff date. In this way, the first half of the sample establishes what is known about the pair historically, and the second half provides a separate period in which the usefulness of that information can be evaluated.

The chronological split also lays the foundation for the theoretical concern brought up by scholastic works discussed in Section 2.1. Krauss specifically argues that strong historical correlation or a small price distance does not guarantee a stationary relationship outside the period in which that relationship is identified [4]. Accordingly, this study uses 5-minute, 10-minute, 1-hour, and 2-hour time buckets exclusively from the first 20-day formation period as baseline historical measures. Together with the initial GICS4 and market-cap matching, this design provides a controlled benchmark for examining whether raw historical co-movement adds useful information once economically similar stocks have already been paired. The separation of the two periods also prevents later observations from entering the historical correlation and spread estimates, alleviating look-ahead bias in the evaluation.

After establishing the empirical benchmark, the next step of this research is centered in quantitative computations of relevant variables like trade spreads. To assess the relative performance of the two stocks on a comparative baseline, prices are normalized by dividing each series by its initial value. This transformation removes differences in absolute price levels and allows the analysis to focus on the evolution of the relationship between the two assets over time. The normalized price series are then aligned and transformed into a wide format, enabling the spread to be calculated as the difference between the normalized prices of the paired stocks.

The spread observed during the first 20 days is used to estimate its historical mean and standard deviation, which provide the reference values for standardizing subsequent deviations. These two parameters are fixed at the 20th-day cutoff and carried forward into the trading period rather than being recalculated using later observations. For every observation after the cutoff date, the current spread is compared with these fixed formation-period parameters to obtain an out-of-sample Z-score. In this way, the Z-score indicates how unusual the current spread is relative to the relationship observed during the formation period, while ensuring that the benchmark itself is not influenced by later price movements.

The trading strategy is then implemented using out-of-sample Z-scores generated from the second 20-day trading period. Trading signals are activated whenever the spread deviates substantially from its historical equilibrium using +2 and -2 as threshold values. Specifically, a short-spread position is initiated when the Z-score exceeds +2, indicating that the spread has widened beyond its normal range and is expected to revert downward. Conversely, a Z-score less than -2 suggests that the spread has become unusually narrow and is

expected to increase toward its historical equilibrium. Minute-by-minute changes in the spread are then used to calculate Profit and Loss (P&L) for the active position. Positive P&L occurs when the spread moves in the anticipated direction of mean reversion, while negative P&L results when the spread continues to diverge from its equilibrium level. The cumulative P&L series aggregates these gains and losses over time, providing a measure of the overall profitability of the pairs trading strategy. Consequently, this variable serves as an indicator of whether the selected pair exhibits sufficient mean-reverting behavior to further generate economically meaningful trading returns.

2.3. Statistical results

Due to the colossal density of the stock price dataset, 75 pairs are randomly sampled from the full set of eligible pair candidates to guarantee the efficient functioning of the programming site itself. The observations that result from this "technical filtering" are assembled into a new database (named "enlarged_pairs_database"). The 20-day formation and trading split is maintained when constructing the prediction dataset. Correlations are calculated at the 5-minute, 10-minute, 1-hour, and 2-hour frequencies using observations from the lookback period.

The resulting models for the relevance between such return correlations and gross P&L provide mixed evidence for the predictive value of the Z-score signals, as shown in Table 1:

Table 1.Regression outputs with trading period gross P&L as the dependent variable

Regressors	Coefficient Estimate β	p -value
Corr_5min	0.562	0.478
Corr_10min	-0.091	0.883
Corr_1hour	0.143	0.742
Corr_2hour	-0.253	0.454
Y-Intercept	-0.019	0.723
Multiple R-squared = 0.0184		

None of the correlation measures demonstrate statistically significant predictive ability. Although the coefficient of the 5-minute return correlation is positive (slope coefficient $\beta = 0.562$), which suggests that stronger short-term synchronization may be associated with higher profitability, the relationship is statistically insignificant (p -value = 0.478). Similarly, the 10-minute (slope coefficient $\beta = -0.091$, p -value = 0.883), 1-hour (slope coefficient $\beta = 0.143$, p -value = 0.742), and 2-hour (slope coefficient $\beta = -0.253$, p -value = 0.454) correlations fail to exhibit significant associations with out-of-sample returns. Here, the 2-hour correlation exhibits the most negative beta value. These outputs suggest that historical correlation does not appear to provide statistically reliable information about subsequent gross P&L. Changing the interval from 5 minutes to 10 minutes, 1 hour, or 2 hours also does not materially improve its predictive value. This finding is consistent with the concern developed from Krauss [4] in Section 2.1: Historical co-movement may describe how closely two securities moved during formation without guaranteeing that the relationship remains stable or useful during the later trading period. Instead, return correlations should be interpreted as supplementary diagnostic characteristics rather than definitive screening criteria, with pair selection potentially requiring additional information such as volatility characteristics and transaction-cost-adjusted profitability.

This lack of predictive consistency can be further observed through the estimated uncertainty of the four correlation coefficients. Figure 1 shows the 95% confidence intervals for the correlation coefficients of the four frequency levels all extend across zero despite differences in their point estimates. The result provides

visual support for the absence of a statistically reliable relationship between formation-period correlation and subsequent gross P&L. More importantly, the alternating signs and overlapping confidence intervals suggest that changing the frequency at which historical co-movement is measured does not materially improve its usefulness as a standalone screening criterion. The limitation therefore appears to concern the information captured by historical correlation itself rather than the selection of a particular return interval.

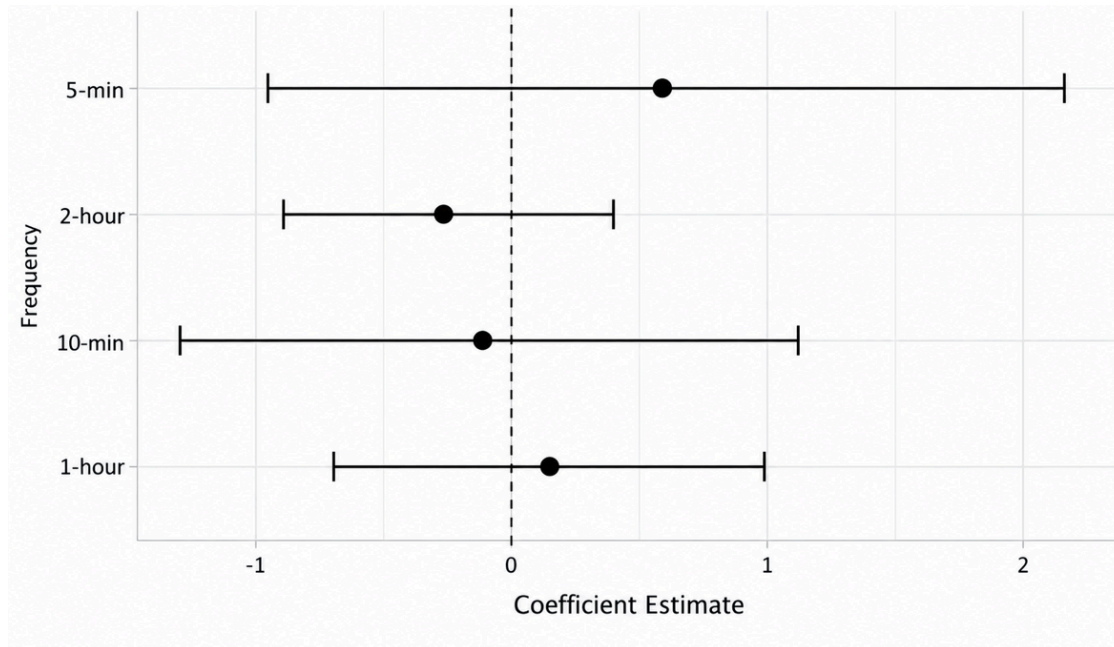


Figure 1. Correlation coefficients by frequency

Continuing in this vein, a separate set of predictive regressions considers whether the current Z-score, rather than formation-period correlation, contains information about subsequent trading outcomes. Maintaining the separation between lookback and trading period, historical spread parameters like the mean and standard deviation (denoted as "historical_mean" and "historical_sd", respectively) are estimated using only observations from the first twenty days and are subsequently matched with observations occurring after the cutoff date. These parameters remain fixed throughout the trading period. In addition, "profitable" is a binomial variable that equals one when positive gross P&L is generated over the next one-minute spread movement. The current z-score is calculated by standardizing the current spread using the frozen historical parameters. Apart from future spread, profitability and gross P&L were incorporated as prediction parameters. Estimating the deviation of these variables helps assess whether the predictive model translates into successful and financially rewarding trades. The outputs of profitability and gross P&L predictions are listed as follows in Table 2 and Table 3:

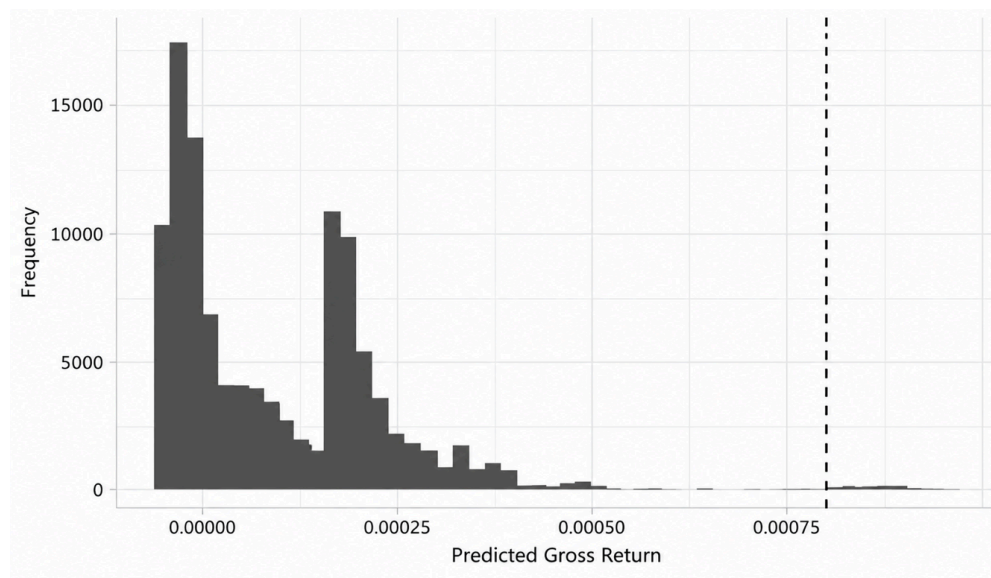
Table 2. Predictive regression outputs with profitability as the dependent variable

Regressors	Coefficient Estimate β	p -value
Current_z	-0.029	0.000
position	0.133	0.000
Historical_sd	1.864	0.000
Y-Intercept	-1.224	0.000

Table 3. Regression outputs with gross P&L as the dependent variable

Regressors	Coefficient Estimate β	p -value
Current_z	8.008 e-06	0.014
position	1.912 e-05	0.183
Historical_sd	-6.244 e-04	0.000
Y-Intercept	4.106 e-05	0.002
R-squared = 6.254 e-05, Adjusted R-squared = 5.185 e-05		

The profitability regression model indicates that the current z-score is negatively associated with the probability of generating a profitable trade (slope coefficient $\beta = -0.029$, p -value < 0.001). The gross P&L regression reports a positive but small slope coefficient on the current Z-score (slope coefficient $\beta = 8.008 \text{ e-}06$, p -value = 0.014), showing a negligible unit of increase per Z-score deviation. At the same time, the model demonstrates exceptionally low explanatory power (R-squared = 6.254 e-05), which indicates that the independent variable is hardly associated with any deviations in gross P&L. Yet, given the large sample size of over 280,000 observations, even minute relationships like this can attain statistical significance with its small standard error and large t statistic. In this case, the relationship between gross P&L and Z-score deviations with a slope coefficient of 8.008 e-06 should be viewed as an economically detectable, rather than a meaningful trading advantage by itself. The "fragile" correlation does not establish that the total predicted return from a Z-score signal is sufficiently large to compensate the transaction costs incurred in executing the trade. The main issue is therefore not simply whether the model detects a statistical relationship, but whether the resulting gross edge is large enough to remain profitable after the costs required to capture it.

**Figure 2.** Expected returns with transaction threshold

The predictability of profits from the above regressions should be interpreted jointly with transaction cost thresholds of 0.07%, 0.10%, and 0.20%. Only when the predicted gains exceed these thresholds can the statistical relationships be regarded as economically meaningful. In an economic significance analysis, transaction cost thresholds of 0.07%, 0.10%, and 0.20% were integrated into the prediction function for gross

P&L for trading periods after the cutoff date. Among the 105,644 predicted trading strategies, only 689 generated expected returns greater than 0.07%, with a maximum of approximately 0.09%. The histogram below visually reinforces these results through an overwhelming concentration of predicted returns below the 0.07% reference line indicates that most statistically identified opportunities are economically too small to withstand even the least restrictive cost assumption.

The economic effect of transaction costs becomes more apparent when their impact is considered cumulatively over the trading period. Whereas Figure 2 evaluates whether individual predicted opportunities are sufficiently large to clear the transaction-cost hurdle, Figure 3 compares cumulative gross P&L with the corresponding net performance under the 0.07%, 0.10%, and 0.20% cost assumptions. The performance paths generally move together because they originate from the same underlying trading signals, but the net P&L paths remain progressively below gross P&L as increasingly restrictive cost assumptions are imposed. This separation illustrates friction decay: repeated transaction costs gradually absorb part of the returns generated by the strategy and widen the difference between gross and realizable performance. The effect is particularly important given the relatively small predicted returns shown in Figure 2, since even modest execution costs can materially reduce an already limited gross trading edge. Consequently, the existence of statistically detectable spread movements does not necessarily translate into economically meaningful profitability once the costs of repeatedly entering and exiting positions are incorporated.

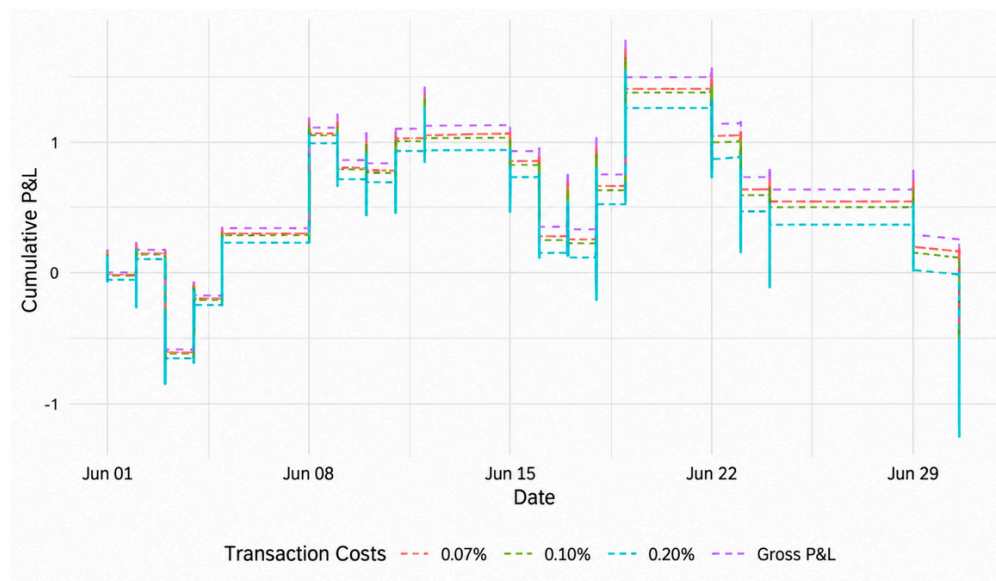


Figure 3. Cumulative P&L across transaction cost tiers

These findings suggest that the expected profits determined from future spread are insufficient to offset the transaction cost thresholds for the robustness checks used earlier. In this case, the regression model used in this analysis demonstrate insufficient economic significance, suggesting that observed statistical relationships do not readily translate into more arbitrage opportunities and chances for profits under realistic market conditions, complicating the earlier assertion on the necessity to obtain positive differences between predicted gains and transaction cost thresholds to evaluate economic significance, and highlighting the existence of other factors.

2.4. Discussions

From the empirical results, the first issue is whether historical correlation at any frequency helps identify pairs that subsequently perform better. As shown in Section 2.3, the regression results do not provide evidence of

such a pattern. None of the four lookback-period correlations is significantly related to subsequent gross P&L. The 5-minute correlation has a p -value of 0.478, and the 2-hour correlation has a p -value = 0.454, while the 10-minute and 1-hour measures are similarly insignificant. The estimated coefficients also change signs across frequencies, making it difficult to assert that the lack of significance is simply caused by choosing an unsuitable return time bucket. Besides, the absence of significance across all four intervals over the lookback period suggests a broader paradox in which stocks can appear closely related during the 20-day formation period without stronger historical correlation translating into better performance during the subsequent trading period. This shifts the discussion away from the choice of a particular correlation frequency and toward a more fundamental question of what historical correlation captures about the relationship between the paired stocks.

One explanation lies in the difference between co-movement and the stability of the relative price relationship. Return correlation captures whether the two stocks tend to move in the same direction over a given interval, but it says little about whether small differences in those movements are temporary. For example, two stocks may respond similarly to market or industry conditions and therefore maintain a high return correlation, while one stock consistently earns slightly higher returns than the other. These small differences may have little effect on their measured correlation, but they accumulate over time and can cause the normalized price spread to drift away from its earlier level. A pair can therefore display strong return co-movement without having a spread that repeatedly returns toward its historical mean. This distinction helps explain the apparent contradiction in the results: stronger formation-period correlation does not necessarily predict higher subsequent P&L because the strategy ultimately profits from the reversal of relative-price differences, rather than from the continuation of common price movements.

This distinction also resonates with Gatev et al.'s characterization of a "common factor" in their 2006 paper [1]. Literally, such a variable can explain why two stocks move together but cannot always guarantee that their relative price spread is mean-reverting or implicates profits. Two stocks may share substantial exposure to the same market or industry factor while retaining idiosyncratic components that drift apart during the trading period. Consequently, pairs with stronger exposure to common historical influences do not always exhibit more reliable subsequent convergence or generate higher gross P&L. The insignificant coefficients across all four correlation frequencies are consistent with the interpretation that correlation identifies broad co-movement but does not measure the specific spread dynamics from which the strategy earns its returns.

The results across frequencies provide a further reason to treat this as a problem with the content of correlation rather than its timing. If the 5-minute measure were insignificant only because very short-horizon returns contain excessive noise, the estimates might become more stable when the data are aggregated to one or two hours. That pattern does not appear. The longer-interval coefficients remain insignificant and do not move toward a common sign. In his 1986 paper "*Noise*", Fischer Black observes that "the price of a stock reflects both the information that information traders trade on and the noise that noise traders trade on" [6]. Aggregating returns may smooth some temporary fluctuations, but it cannot identify whether the co-movement that remains reflects a stable relative-price relationship, a shared reaction to news, or market noise. The repeated insignificance is therefore not resolved by using a slower frequency.

The construction of the trading-period Z-score introduces a related issue. Its mean and standard deviation are estimated during the first 20 days and then held constant. A later Z-score consequently measures the distance between the current spread and its formation-period benchmark. It does not test whether that benchmark remains valid after the cutoff date. Krauss warns that "spurious relationships based on an assumption of return parity are not mean-reverting" [4]. In this case, a large Z-score may indicate a temporary divergence, but it may also reflect a lasting change in relative valuation caused by new firm-specific

information. If the relationship has changed, a greater distance from the old mean does not imply stronger pressure to return to it. The Z-score should therefore be interpreted as a measure of deviation from the historical benchmark rather than a direct probability of convergence. This is an important difference. A signal can become more extreme because the earlier spread relationship has broken down, even though pairs trading customarily views a larger deviation as a stronger contrarian opportunity. Here, increasing the position in response to the Z-score will increase exposure to continued divergence. The signal identifies the location of the spread relative to an earlier distribution, but it does not identify the economic process that placed the spread there.

Abrupt spread movements make this limitation more visible. The standalone Z-score used here does not distinguish a gradual intraday deviation from an overnight jump. Yet an intraday liquidity imbalance may be temporary, whereas an overnight announcement may move one stock toward a new persistent valuation. Both observations can produce a similar standardized deviation even though their expected paths are different. Pooling them under the same signal can weaken the statistical relationship between the Z-score and subsequent P&L.

This helps explain the apparently conflicting evidence in the gross-P&L regression with "current_z". The corresponding coefficient is positive and statistically significant (slope coefficient $\beta = 8.008 \text{ e-}06$, $p\text{-value} = 0.014$), but the entire model's R-squared is only $6.254 \text{ e-}05$. These statistics answer different questions. The p -value concerns whether the average slope can be distinguished from zero under the regression assumptions. R-squared measures how much of the variation in gross P&L is explained by the model. With more than 280,000 observations, a very small average effect can be estimated precisely enough to attain significance. Even so, the Z-score explains only about 0.0063% of gross-P&L variation. The result is statistically detectable, but it provides little forecasting power for an individual observation.

The near-zero R-squared implies that observations with similar Z-scores often lead to very different outcomes. One spread may converge, another may continue to widen, and a third may reverse by too little to be economically useful. This variation suggests that a large Z-score can arise under different market conditions and does not necessarily imply the same likelihood or magnitude of convergence. This is consistent with the possible explanation by Joel Hasbrouck, who believes that information asymmetry conveyed through trades can have persistent effects on prices as fudge factors [7]. Although his analysis does not directly examine spreads in pairs trading, it illustrates that observed price movements need not represent purely temporary disturbances, thereby providing additional reason for caution in interpreting the magnitude of a Z-score alone as an indication of subsequent convergence.

Hence, the regression captures only a small average relationship between the Z-score and gross P&L, while leaving most differences in individual trading outcomes unexplained. This may also shed light on the "ambiguity" that underlies the contrasting directions of gross P&L and profitability regressions. Plus, the rightward skewness of the high-frequency return distribution implicates that some larger mean-reverting gains can raise average gross P&L even when most individual signals do not produce positive outcomes. Immediate execution frictions can further reduce the likelihood that these small predicted movements translate into profitable trades. Thus, a positive average relationship between the Z-score and gross P&L does not necessarily imply that a larger Z-score increases the probability of a successful trade.

The binary profitability model highlights another point for caution. The current Z-score is negatively associated with the probability of a profitable observation, even though its coefficient in the continuous gross-P&L regression is positive. The two results do not necessarily contradict each other because the models describe different aspects of the outcome distribution. A limited number of larger gains could raise average P&L without increasing the proportion of observations that are profitable. Positive and negative Z-scores may

also have asymmetric effects that are obscured when they enter the model as one signed variable. The evidence does not support a simple rule in which a larger Z-score uniformly produces a better trade.

Observed-price mechanics may account for part of the reversal. Richard Roll observes that "trading costs induce negative serial dependence in successive observed market price changes" [8]. If one stock trades near the ask while the other trades near the bid, the recorded spread can widen even when their underlying relative values have changed little. Later transactions on opposite sides of the quotes can make the spread narrow again. The data then display divergence followed by convergence, but a trader must pay the bid-ask spreads to capture that movement. Some of the statistical reversal may therefore be produced by the same microstructure costs that prevent it from becoming realizable P&L.

The issue is also reflected in the transaction cost analyses. Most predicted returns are already below the least restrictive assumed cost threshold of 0.07%, suggesting that the problem is not simply that large gross profits are reduced by high trading costs. Rather, many of the identified opportunities are too small to be economically viable in the first place. This is consistent with Do and Faff's 2012 paper, in which the authors similarly argue that transaction costs substantially reduce pairs-trading profitability and note that some divergence thresholds may be too small to cover trading costs even when convergence occurs [4]. Overall, the transaction cost results suggest that detecting mean reversion and being able to profit from it are separate issues. In this sample, the Z-score may identify a statistical relationship, but the expected magnitude of that relationship is generally insufficient to make it economically meaningful after costs.

In general, the evidence suggests that pair selection, convergence testing, and trade filtering should be treated as separate tasks. Industry, size, and correlation can establish that two firms are broadly comparable and share common influences. Additional tests, such as cointegration, stationarity, zero crossings, residual half-life, or jump-adjusted dynamics, are needed to assess whether the relative spread is likely to reverse. The expected size of that correction must then be compared with trading costs before a position is opened. The results do not show that all pairs-trading opportunities are absent. Rather, they show that historical similarity and an extreme Z-score are not sufficient evidence of stable convergence or executable profit. This fills the main gap raised in Section 2.1 by locating the weakness in the distance between common co-movement, residual mean reversion, and the return that can be retained after costs.

3. Conclusion

This paper presents an empirical analysis of the return prospects of pairs trading using high-frequency data from April to June 2020. The use of high-frequency observations makes it possible to examine short-lived spread movements and evaluate whether relationships identified during the formation period remain useful at the frequencies at which trading decisions are made. The statistical results show that strong historical co-movement alone does not guarantee that relative-price deviations will converge in a later trading period. This conclusion remains consistent across multiple intraday return intervals, suggesting that it is not simply an outcome of selecting an unsuitable sampling frequency. The quantitative models also illustrate that while the current Z-score is statistically associated with subsequent gross P&L, the near-zero explanatory power for the corresponding regression model indicates that the magnitude of a spread deviation is not a reliable indicator of economic profitability by itself. Moreover, by integrating the effects of transaction cost thresholds into comparisons of predicted returns, the paper also demonstrates that only a small proportion of trading opportunities may remain profitable after costs. Taken together, these findings contribute a practical distinction between pair similarity, convergence potential, and implementable profitability, showing that these should be treated as separate stages of a high-frequency pairs-trading strategy.

Future studies can be directed in several trajectories. For one thing, researchers could use samples that elapse longer trading periods and cover diverse conditions of volatility and liquidity conditions. Furthermore, the criteria employed for pair selection could combine industry and size matching with additional tests like cointegration, spread stationarity, and factor-adjusted returns, in which the corresponding models could consider external "variables" like overnight jumps and firm-specific updates, while allowing nonlinear or asymmetric effects of positive and negative Z-scores. Finally, future work could replace fixed cost thresholds with realized, time-varying execution costs and conduct a fully out-of-sample comparison of gross and net performance. These extensions can help identify not merely which pairs look similar historically, but which deviations are likely to converge and remain profitable after the full cost of implementation.

References

- [1] Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3), 797–827. <https://doi.org/10.1093/rfs/hhj020>.
- [2] 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>.
- [3] 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>.
- [4] 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>.
- [5] 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>.
- [6] Black, F. (1986). Noise. *The Journal of Quantitative Finance*, 41(3), 15.
- [7] Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance*, 46(1), 179–207. <https://doi.org/10.1111/j.1540-6261.1991.tb03749.x>
- [8] Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4), 1127–1139. <https://doi.org/10.1111/j.1540-6261.1984.tb03897.x>.