Leveraging Statistical Arbitrage: A Pairs Trading Strategy with Market Data

Aru Mehta, Seung Hwan Lee, John Andersson, Jaesung Kim

mehta.aru1@northeastern.edu, andersson.j@northeastern.edu, kim.jaes@northeastern.edu, lee.seunghwa@northeastern.edu

Northeastern University

Intermediate Programming with Data (DS 2500)

Professor Laney Strange

April 15th 2025

Problem Statement and Background

The problem our team aims to address is the challenge of identifying reliable, data-driven trading opportunities in financial markets that in theory are always efficiently priced. Specifically, we focus on implementing a pairs trading strategy which is a type of statistical arbitrage that exploits inefficiencies between two historically correlated assets, such as ETFs, FX (Foreign Exchange) pairs, or commodities. Usually, this strategy is used by high-frequency trading firms to exploit milli-second inefficiencies in the equities market but given that we don't have the infrastructure to compete with them we aim to look at traditionally overlooked asset pairs. Furthermore, the lack of accessible and robust tools for retail investors to systematically identify cointegrated asset pairs, analyze their spread behavior, and execute profitable trades represents a key gap in the current landscape of algorithmic trading.

We were all motivated to pursue this project due to our backgrounds in finance and the growing role of machine learning and statistical modeling in the world of trading. Unlike other trading strategies, pairs trading offers a market-neutral approach, meaning that even if the market swings with volatility the strategy will still be profitable, which is still extremely attractive in the current market environment of high volatility. This approach is helpful in markets where certain assets exhibit strong long-term correlations but may temporarily diverge due to short-term market inefficiencies. Anybody in the financial world can benefit from more efficient trading systems which allow them to capture Alpha (return over a given benchmark) while simultaneously limiting risk in their portfolios.

Introduction to Data

Our analysis relies on financial data obtained from two sources which include the ETF Screener tool on ETFdb and Alpha Vantage API. These sources provide extensive market data which includes ETF, FX, and commodities performance data and historical price trends and financial indicators required for quantitative analysis.

The financial data used in this project presents no ethical issues regarding data privacy because of its public nature. The aggregate or institutional financial data lacks personally identifiable information (PII) so it does not contain potential biases such as names, ages, gender or ethnicity. The analysis operates under the premise that no extra privacy or ethical protection measures are necessary. Our method uses public data sources through APIs to maintain transparent results that can be replicated while following the data policies of the involved platforms. Additionally, we utilized the BeautifulSoup library to web scrape data from ETFdb and no privacy or ethicality concerns were uncovered given that the website has a documentation site on how to extract data for various ETF symbols across exchanges.

Data Science Approaches

Our project uses several data science techniques to identify and evaluate profitable pairs trading opportunities across various markets and asset classes. The first step involved the Engle-Granger cointegration test, a two-step method that assesses whether two non-stationary time series form a stationary linear combination. In simpler terms, this means finding two assets that have changing mean, variance, and other statistical properties changing over time then combining the assets into one time-series on a common interval. We then can fit an Ordinary Least Squares (OLS) regression model between two asset price series:

$$Yt = \alpha + \beta Xt + \varepsilon t$$

where Yt and Xt are the price series of two assets, and εt represents the residuals or "spread." We then applied the Augmented Dickey-Fuller (ADF) test εt to test for stationarity. The ADF test uses a combination of T-statistics and F-statistics to detect the presence of a unit root in time series. ADF test in pairs trading is done to check the co-integration between two assets (presence of unit root). If there is a unit root present in the time series, it implies that the time series is non-stationary and the assets are not co-integrated. Hence, assets cannot be traded together. Alternatively, if the null hypothesis gets rejected and the assets show co-integration; it implies that the time series is stationary and the assets can be traded [2] . In our case, a low p-value (< 0.1 in our filter) indicates cointegration, suggesting that the spread between the two chosen assets in our project will revert to a long-term mean. To monitor these conditions we calculated the Z-score of the spread which will be used for entry/exit points:

$$Zt = \frac{\varepsilon t - \mu}{\sigma}$$

where μ is the mean of the spread and σ is the standard deviation. This standardization allows us to define thresholds for trade entry/exits. Typically, a Z-score below -2 prompts a long position (expecting convergence) and a score above 2 triggers a short position. These thresholds ensure we only act on statistically significant divergences and without human biases or emotions.

To further improve our understanding of the market environment, we used a Gaussian Hidden Markov Model (HMM) to detect latent "regimes" in the market based on Z-score patterns. In its simplest terms, A stochastic process is a discrete time Markov chain if the probability of our current state only depends on our previous state (so, P (Xn = an|X0 = a0, X1 = a1, . . . Xn−1 = an−1) = P (Xn = an|Xn−1 = an−1)) [6]. This unsupervised learning model clusters observations into different states, such as trending or mean-reverting, which will help us adjust our risk parameters given the asset.

Lastly, we evaluated our strategy using common financial performance metrics. The Sharpe Ratio

$$\text{Sharpe Ratio} = \frac{E[Rp - Rf]}{\sigma p}$$

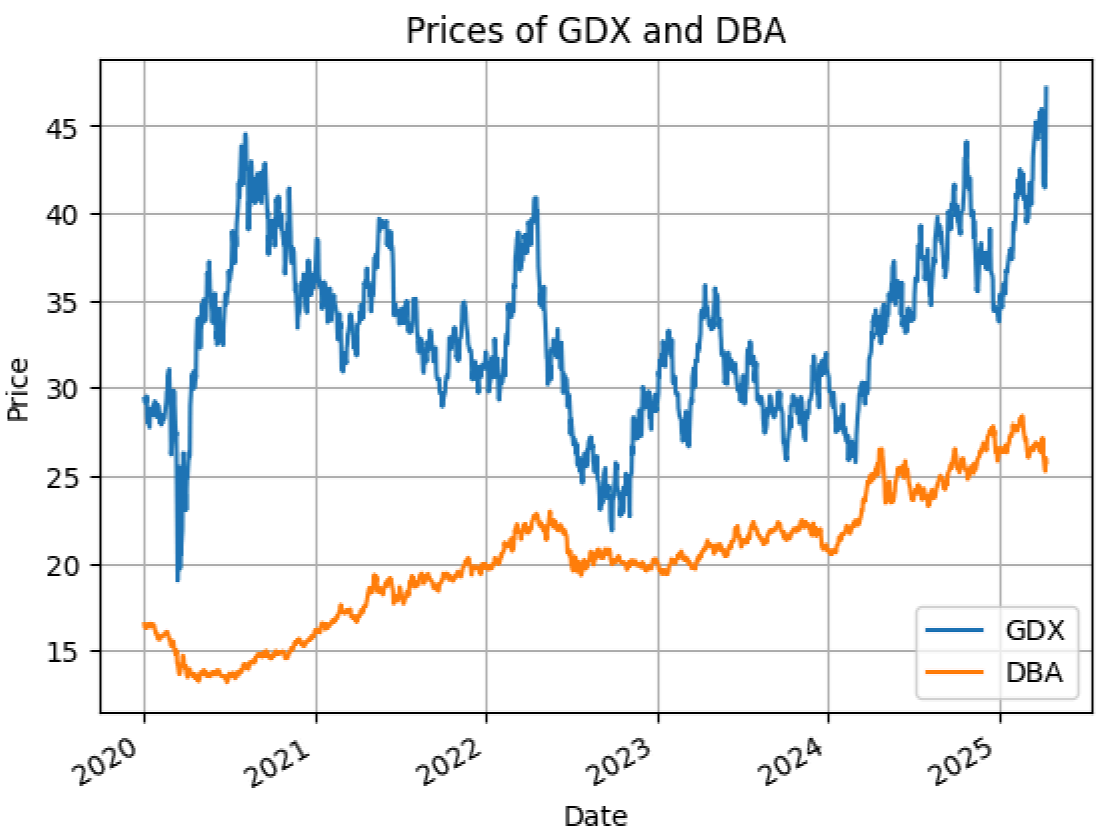
measures risk-adjusted return which is a common metric in portfolio management. Rp is portfolio return, Rf is the risk-free rate (assumed zero in our case since it's market neutral), and σp is portfolio volatility. We also tracked maximum drawdown, in our case is the loss from a peak to a trough in the portfolio value. These metrics are important in real-time as they allow us to measure profitability.

Together, the combination of these techniques forms an automatic trading system free of human biases. From selecting cointegrated pairs and generating trading signals, to evaluating their effectiveness in real-time backtests. Every method serves a unique purpose in ensuring our pair-trading strategy is able to replicated across a variety of assets and risk parameters

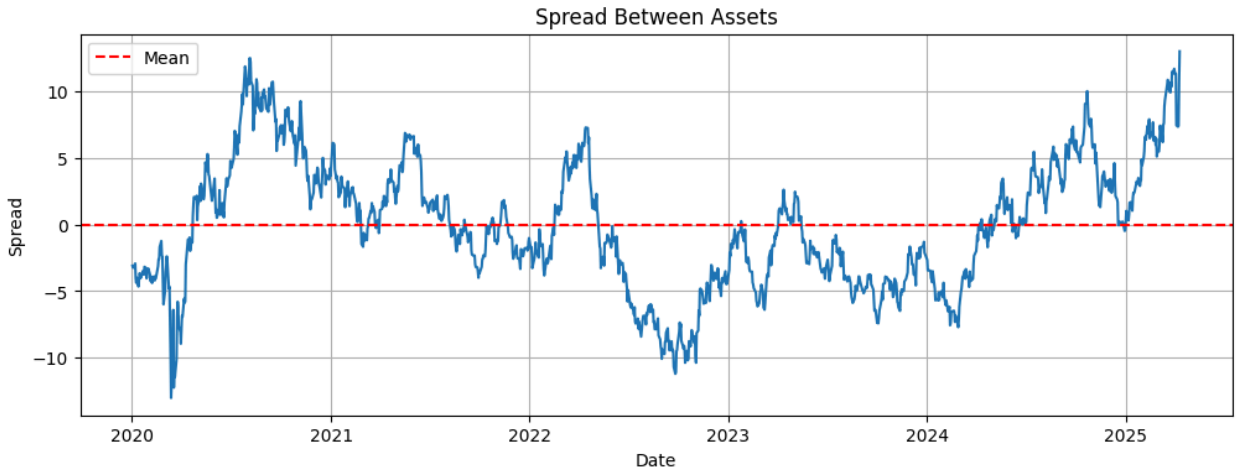
Results and Conclusions

Our final model identified GDX (Gold Miners ETF) and DBA (Agriculture Futures ETF) as the strongest pair for statistical arbitrage, delivering a cumulative return of 127.61% since 2020 with a Sharpe ratio of 1.13 (>1.0 is deemed profitable in the long-term) and a maximum drawdown of just 9.58%, well below traditional risk parameters of 20%. Given the real-time data pull of our functions, we assume that the most profitable pair can change given the day or even the hour if the time interval or risk parameters are adjusted.

Plot 1 - Prices of the Two Highest Sharpe Ratio Pair (GDX & DBA)

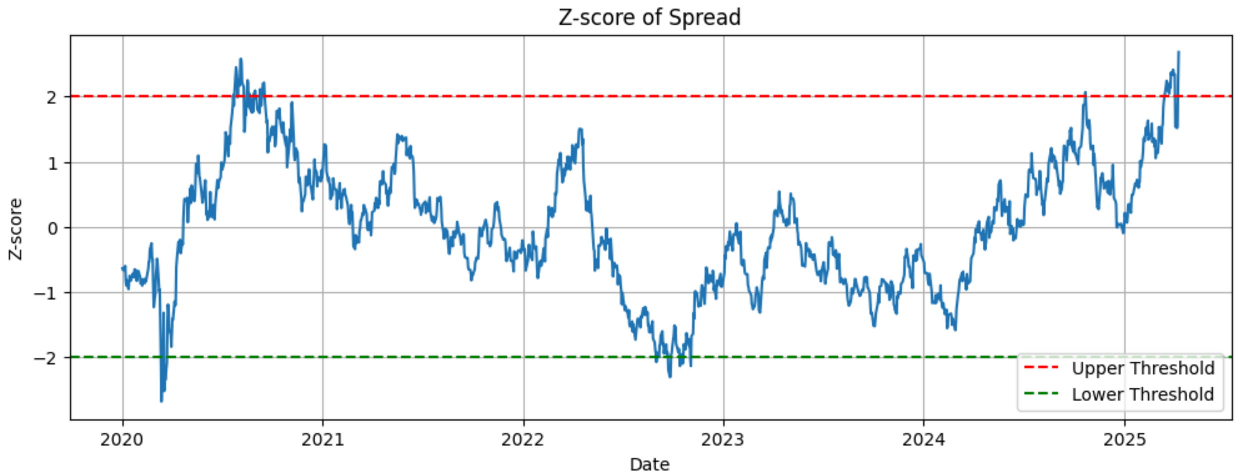


Plot 2 - % Spread Between the Two Given Asset Pairs



The second plot shows the percentage spread between two selected asset pairs over time from 2020 to early 2025. The spread calculation shows the difference between two asset returns or prices to reveal their relationship changes across time. The spread shows cyclic patterns through its peaks and troughs which indicate times when assets move apart from each other and move closer together. The red dashed line shows the historical mean spread which serves as a reference point to determine if assets trade above or below their usual relative value. When the spread moves far from its mean value it indicates possible statistical arbitrage or mean reversion trading opportunities. The large positive spread during early 2025 indicates that one asset has developed a higher value than the other asset based on fundamental market conditions. The spread analysis works to detect relative value trends between ETFs and other asset pairs.

Plot 3 - Z-Score of Spread (If > 2 or <-2, possible mean reversion)

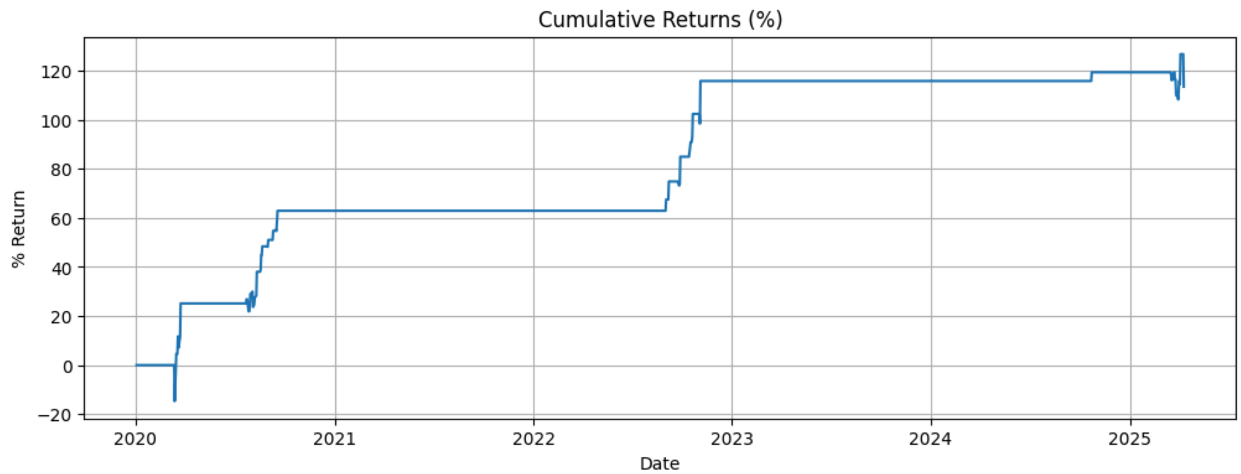


The chart displays the Z-score of the spread between the asset pair to identify times when the spread moves away from its average. A Z-score in excess of +2 or less than -2 is often seen as a sign of mean reversion. The chart reveals multiple such thresholds have been crossed—the most notable ones being in late 2020, early 2023, and early 2025—indicating potential trading opportunities when the spread was statistically overextended. Such signals can help in identifying the periods in which the price relationship between the assets may reverse to the mean, which can be useful in pair trading strategies.

Plot 4 - Potential Trade Signals Given Z-Score (1 = Long Entry, -1 = Short Entry)

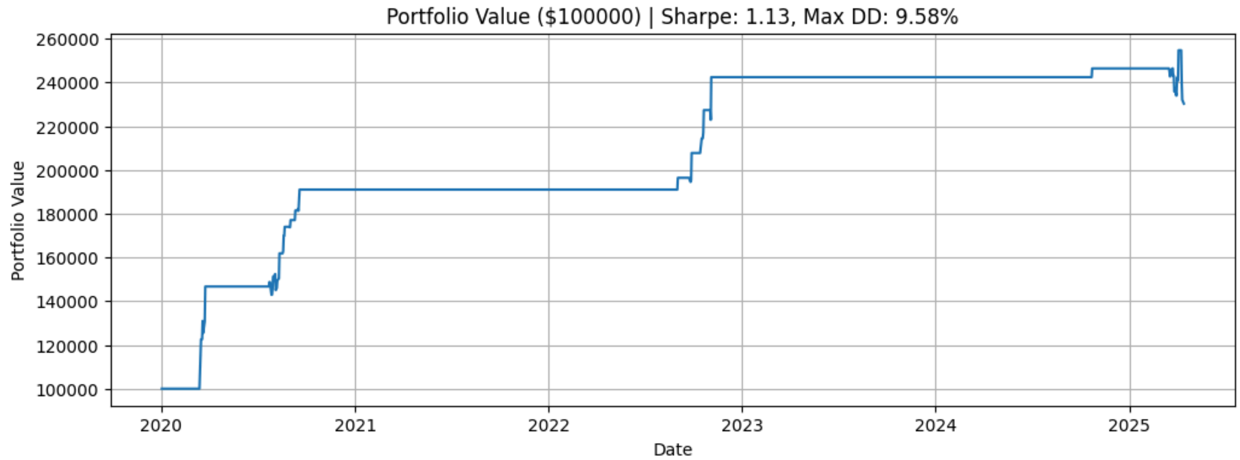
This plot converts the Z-score thresholds into trade signals where +1 is a long entry signal and -1 is a short entry signal. The visual shows discrete, infrequent positions taken during moments of statistical divergence. Trade entries were found to cluster during early 2020, late 2022, and 2025, which corresponds with the extreme Z-score readings and extreme market volatility. The infrequent nature of these entries highlights the conservative trigger criteria of the strategy that seeks high-confidence opportunities with significant mean reversion signals.

Plot 5 - Cumulative Returns Between Pair Since (2020 - Present)



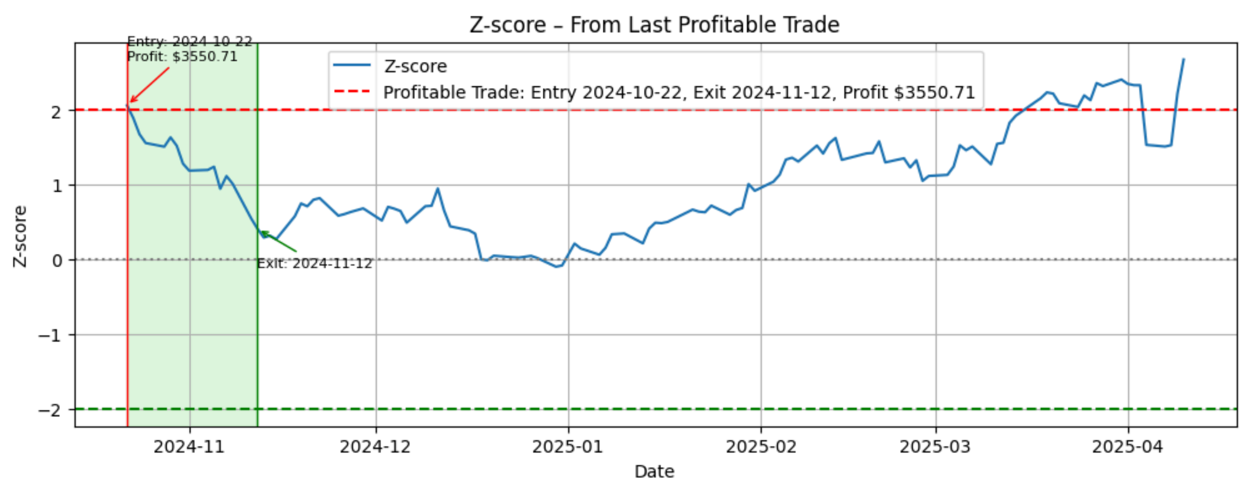
The cumulative return graph presents the outcome of the pair trading strategy from 2020 onward. In general, the strategy has produced a positive cumulative return, reaching more than 120% in the early part of 2023. The step-like growth pattern indicates that profits are made in concentrated bursts after successful trade execution during significant spread divergence periods. The consistent positive returns with minimal drawdown periods indicate the effectiveness of z-score thresholds as entry and exit points.

Plot 6 - Portfolio Value of $100,000 with Sharpe Ratio and Max Drawdown



The plot shows the growth of a $100,000 portfolio that employs the pair trading strategy with a Sharpe ratio of 1.13 and a maximum drawdown of 9.58%. The portfolio value more than doubled over the observed period, with major upward moves occurring in 2020, mid-2022, and early 2023. The strategy has a relatively modest drawdown and a Sharpe ratio greater than 1, which means that it provides favorable risk-adjusted returns, which makes it an attractive model for disciplined statistical arbitrage.

Plot 7 - Last Profitable Trade with Entry/Exit Points and Profit



The final graph focuses on the most recent profitable trade with clear short-entry on October 22, 2024, exit on November 12, 2024, and profit of $3,550.71 from a $10,000 capital allocation. The z-score at the time was above the upper threshold indicating overextension and a short position was taken which proved to be successful as the spread reverted to the mean. This example illustrates how z-score-based entries can be used for exact and lucrative trading execution.

Future Work

The methods that we used to compare two different assets and understand the relationship between them show an evident and strong relationship. Yet, there can be more future work used to evaluate individual assets and to compare them through finding a covariance in two analyses. Most of our current research produces a visualization of comparison. However, if we take a different approach of assessing individual assets in a unique manner and map their relationship, it may produce a meaningful result. One of the first things that we can do is add a stationary check process where we ask why we need a stationary check (e.g., ARMA - autoregressive–moving-average). This is specifically important given that we are dealing with time-series data.

Another factor that we can place strong emphasis on is the noise term of asset returns. While the amount and definition of noise will differ depending on the number of variable terms we set for the asset return, it would be a good trade-off point (accuracy vs. limitation of data) that we can study further. Some of the models that we can utilize for noise aspects are the ARCH (Autoregressive Conditional Heteroskedasticity) or GARCH models, which are basically ARMA models used for analyzing the noise term.

Lastly, in terms of the cointegration test, it would be interesting to enable the comparison of assets beyond just two assets through using the Johansen Test, which is a multivariate cointegration test that will open up this research to a higher level compared to the Engle-Granger model, which is limited to two variables. For example, this will allow the comparison of trading gold, silver, and platinum together, which means that further interpretation of certain assets leading to multiple changes in different assets can be visualized and studied at the same time. While it is important to continuously expand on the potentials of this research, it is also important to keep in mind that any discrepancies or arbitrage opportunities found in this process can be effectively used to generate a meaningful trading/projection model.

References

1) https://databento.com/blog/build-a-pairs-trading-strategy-in-python

2) https://blog.quantinsti.com/augmented-dickey-fuller-adf-test-for-a-pairs-trading-strategy/

3) Alexander, C. (2008). Market Risk Analysis: Practical Financial Econometrics (Vol. 2). Wiley.

4) Tsay, R. S. (2010). Analysis of Financial Time Series (3rd edition). Wiley.

5) Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. Econometrica, 55(2), 251–276.

6) file:///C:/Users/mehta/Downloads/Fullpacket.pdf

7) https://etfdb.com/

8) https://www.alphavantage.co/