Dynamic Long Short Equity with Active Volatility Hedging

Shixuan An[a,b] **Henry Hai**[a,b] **Zening Wang**[a,b] **Zihao Wei**[a,b]
[a] *Duke University, Durham, NC 27708, USA*
[b] *All authors contributed equally to this work.*

Abstract

Stock pair trading is a form of statistical arbitrage that exploits the relative movements between two closely related stocks. However, this trading strategy faces numerous challenges such as identifying optimal stock pairs, determining precise entry and exit points, and implementing effective risk management practices. In this study, we enhance this strategy by developing an advanced algorithmic approach that forms long-short equity pairs, and propose a Total Least Squares methodology, incorporated with insights from the Constant Elasticity of Variance model, to calculate the hedge ratio. We also implement active risk management for optimization. We conducted experiments on the top 99 stocks with the highest market capitalization in S&P 500 and applied our algorithm across multiple time periods. Our results indicate that our method not only identifies pairs with high potential for profitability but also significantly enhances the returns from these trades. Additionally, our approach facilitates a more dynamic adaptation to market conditions, enhancing both the stability and scalability of pair trading strategies.

1 Introduction

Pair trading is a market-neutral trading strategy that capitalizes on the correlation between two assets that are historically proven to move together. By simultaneously buying one asset and short-selling another, traders aim to profit from the relative movements of the two, regardless of the direction of the market. This approach hinges on the idea that if the prices of the paired assets deviate from their historical relationship, they will eventually revert back to their mean, allowing traders to gain from this adjustment. Pair trading is especially appealing because it can potentially yield profits in both rising and falling markets, providing a hedge against the market. As a sophisticated strategy that evolved from the desks of quantitative analysts, pair trading involves complex statistical methods to identify suitable pairs and to determine the optimal timing for entering and exiting trades.

This paper introduces a dynamic approach to managing long-short equity pairs, which includes actively managing volatility through sophisticated algorithms. This technique aims to enhance the traditional pair trading strategy by improving the selection of stock pairs, optimizing entry and exit timing, and implementing effective risk management practices.

The subsequent sections of this paper are organized as follows. In Section 2 we propose a pair selection method which incorporates finding stock alphas, pairs clustering and filtering. In Section 3 we characterize the optimal trading strategy with a focus on hedge ratio

calculated by total least squares and constant elasticity of variance model. In section 4, risk management techniques are presented. In Section 5, we conduct empirical tests with a large sample of US stocks as a validation for our trading strategy. Section 6 and 7 provides directions for future work and our conclusion to this study.

1.1 Related Work

Pairs trading strategies have evolved significantly with the integration of unsupervised learning techniques to enhance the identification and selection of asset pairs. Clustering algorithms, such as k-means, and agglomerative clustering, are now commonly used to group stocks based on historical price data and firm characteristics, increasing the likelihood of identifying profitable trading opportunities (Han et al., 2023). This marks a significant shift from traditional methods that primarily relied on cointegration and distance metrics (Gatev et al., 2006). While these conventional methods are stable and have been proven effective, the incorporation of machine learning offers a dynamic approach to adapt to changing market conditions and uncover non-obvious pair relationships.

Other methods, such as the time-series approach and cointegration approach, still play a crucial role in pairs trading. The time-series models are particularly useful in specific markets like commodities, where price relationships can be modeled with high precision (Cummins & Bucca, 2012). Similarly, cointegration methods are employed to predict price movements and identify pairs with a high probability of convergence (Krauss et al., 2017).

1.2 Contributions

Specifically, we make the following contributions in our paper:

- We have integrated cointegration tests with clustering methods to enhance pair selection. This innovative approach ensures robust pairing by leveraging both machine learning and statistical analysis techniques, thereby improving the predictive strength and stability of our trading pairs.

- We have pioneered the incorporation of a Total Least Squares methodology to calculate hedge ratio enriched with Constant Elasticity of Variance model insights. This fusion not only enhances beta estimation but also aligns with advanced risk management principles, considering the leverage effect and stochastic volatility in trading.

- We have theorized and implemented multiple novel risk management strategies specifically designed for pairs trading. These strategies have been rigorously tested and documented for their effectiveness, contributing significantly to the robustness and reliability of our overall trading strategy.

2 Pair Selection Methodology

In this paper, we introduced a four-step pair selection methodology designed to enhance trading strategies through a combination of quantitative and qualitative data analysis. Initially, we generated features by analyzing stock prices using returns and incorporating various company and asset-specific characteristics. We then applied dimensionality reduction techniques to eliminate redundant information, followed by clustering to group similar entities. Finally, we formed pairs using momentum sorting and validated them through correlation analysis and the Johansen test to ensure each pair's mean-reverting properties. This structured approach streamlined the selection process and aimed to improve the robustness and profitability of pair trading strategies. The entire methodology is illustrated in Figure 1.

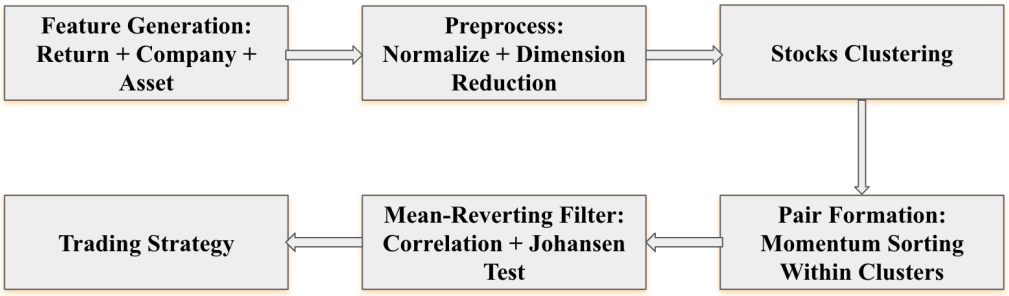

Figure 1: Pair Selection Flow Chart

2.1 Feature Generation

Traditional methods typically relied solely on stock price data for feature generation and dimension reduction. However, stocks possess many significant characteristics beyond their prices, and overlooking these can hinder the identification of robust pair relationships.

Han, He, and Toh (2023) highlighted the importance of incorporating firm characteristics in pairs trading strategies. Lacking access to the database that converts PERMNO to ticker symbols, we instead utilized the Morningstar US Fundamental Dataset available on QuantConnect. Founded in 1984 by Joe Mansueto, Morningstar empowers investors with extensive information and tools. The dataset includes data on over 525,000 investment offerings and real-time global market data covering a wide range of financial instruments (QuantConnect, n.d.).

2.1.1 Return

The stock price is the cornerstone attribute of equity securities and encapsulates the essence of a company's market valuation. It is not just a numerical marker, but a complex amalgam of investor sentiment, market dynamics and underlying company performance. In the field of financial research, the stock price is considered the most telling indicator that distills the intrinsic characteristics and financial health of a company.

In our analytical framework, we adopted the concept of return as the fundamental quantitative measure to assess stock price performance. The return effectively captures the percentage change in the stock price, offering a distilled view of its fluctuation over time. Building upon this approach, we incorporate the measure mom_i, as introduced by Han, He, and Toh (2023), to represent return features. This particular metric allows us to focus on the cumulative return within a defined timespan, which, for the purposes of our study, is restricted to the previous twelve months. Therefore, the formula is modified to accommodate a rolling window approach:

The i-month return feature at the end of month $t-1$ is computed as the cumulative return from month $t-i$ to $t-2$, for $i > 1$, and as the immediate past one-month return for $i = 1$:

$$\text{mom}_i = \begin{cases} r_{t-1}, & \text{if } i = 1, \\ \prod_{j=t-i}^{t-2}(r_j + 1) - 1, & \text{if } i \in \{2, \ldots, 12\}, \end{cases} \tag{1}$$

where r_j represents the return in month j.

2.1.2 Company Information

Company profile information encompasses detailed insights into a company's operations, strategic directions, industry positioning, and financial health, serving as a foundational tool for comparative and predictive financial analysis. These profiles encompass detailed data including industry classification, revenue streams, geographical presence, and management strategies. Such comprehensive data enables analysts to discern subtle yet critical similarities between companies that might not be apparent from financial metrics alone. This deepened insight enhances the robustness of pairing methodologies by ensuring that paired entities share fundamental, strategic, and operational likenesses, thereby reducing the risk of spurious pairings and increasing the likelihood of genuine correlations.

For our analysis, we utilized the Morningstar US Fundamental Dataset available on QuantConnect, specifically extracting 3 key data point attributes from the 'CompanyProfile' objects in the US Fundamentals dataset:

Total Employee Number This parameter quantifies the total number of employees, as stated in the most recent Annual Report, 10-K filing, Form 20-F, or an equivalent report, which specifies the count of employees at the end of the latest fiscal year (QuantConnect, n.d.).

Market Capitalization This parameter is calculated by multiplying the current share price by the total number of outstanding shares. For ADR (American Depository Receipt) share classes, it is determined by the product of the price and the ratio of ordinary shares to ADRs, using the most recent closing price and shares data (QuantConnect, n.d.).

Enterprise Value This parameter is calculated by adding Market Capitalization, Preferred Stock, Long-Term Debt, and Capital Lease Obligations, then adjusting for Cash and Cash Equivalents, Marketable Securities, and other relevant financial instruments (QuantConnect, n.d.)..

2.1.3 Asset Information

Asset information not only reflects a company's current financial position, but also provides important insights into the underlying stock. By analyzing assets, investors and analysts can gauge a company's operational efficiency, liquidity and overall financial health. This data is critical to assessing a company's ability to handle debt, invest in growth opportunities and generate returns for shareholders. In addition, asset information can reveal trends and patterns that may not be apparent from the income statement alone, such as the accumulation of fixed assets or changes in inventory levels, which can have a significant impact on stock valuations and investor perceptions.

In our research approach, we again utilized 5 data point attributes from the US Fundamentals dataset provided by Morningstar:

Sector Code This code organizes industry groups into 11 distinct sectors, facilitating a structured analysis of diverse market segments. For instance, the code '311' is assigned to companies within the technology sector, as exemplified by Apple Inc. (AAPL). Similarly, the code '206' corresponds to entities in the healthcare domain, such as AbbVie Inc. (ABBV) (QuantConnect, n.d.).

NAICS (North American Industry Classification System) This parameter provides a six-digit numerical classification for individual companies. It is developed by the United States, Canada, and Mexico to standardize business activity statistics across North America, replacing the U.S. SIC system (QuantConnect, n.d.).

Growth Score This score was included to indicate that a stock's earnings per share, book value, revenue, and cash flow are expected to grow rapidly compared to its peers. A lower growth score does not necessarily imply a strong value orientation (QuantConnect, n.d.).

Value Score This score suggests that a stock is priced reasonably relative to its anticipated per-share earnings, book value, revenue, cash flow, and dividends. A high value score denotes a robust value orientation, although it does not automatically imply a growth orientation (QuantConnect, n.d.).

Size Score This score uses a flexible classification system that is robust against overall market movements. The scoring ranges from -100, representing very micro stocks, to 400, representing very large stocks, encompassing various categories from giant-cap to micro-cap stocks. This score helps represents approximately 99% of the U.S. market for actively traded stocks (QuantConnect, n.d.).

2.2 Dimensionality Reduction

In the domain of finance, particularly in the context of stock market analysis, the significance of dimensionality reduction cannot be overstated. Stock markets are influenced by an array of tightly interconnected variables, and deciphering the intricacies of these relationships is pivotal for accurate forecasting (Zhong & Enke, 2017).

Dimensionality reduction serves as a methodological cornerstone in this analytical process, as it enables the distillation of high-dimensional data into a more manageable, lower-dimensional space. Importantly, this reduction is achieved without compromising the essential attributes of the data, thereby preserving the fundamental parameters that accurately encapsulate the data's intrinsic dimensionality (Van Der Maaten, Postma, & Van Den Herik, 2009). Moreover, by focusing on the most consequential aspects that drive market trends, the use of dimensionality reduction simultaneously lowers computational expenses by pruning redundant data dimensions.

In our study, we concentrated on two prominent dimensionality reduction techniques: t-Distributed Stochastic Neighbor Embedding (**t-SNE**) and Density-preserving Stochastic Mapping (**DensMAP**).

t-SNE The t-SNE algorithm, introduced by van der Maaten and Hinton in 2008, is an effective method for reducing the dimensionality of large data sets in machine learning. This technique builds upon the principles of Stochastic Neighbor Embedding (SNE) by converting the Euclidean distances between high-dimensional data points into conditional probabilities that reflect similarities between points. These probabilities, denoted as P_{ij}, indicate the likelihood that one data point is similar to another, based on the formula provided below:

$$P_{j|i} = \frac{\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)}{\sum_{k\neq i}\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2/2\sigma^2)}, \text{ with } P_{ii} = 0. \tag{2}$$

In the high-dimensional space, the probabilities are mathematically defined by the following equation:

$$P_{i|j} = \frac{P_{i|j} + P_{j|i}}{2n} \tag{3}$$

Where n is the total count of data, in the smaller dimensional space we measure the similarities between pairs of points with the Student's t-distribution with one degree of freedom, as shown in the equation below:

$$Q_{i|j} = \frac{(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2)^{-1}} \tag{4}$$

The embedding points, denoted by $\boldsymbol{y}_i'$s,for all object points $\boldsymbol{x}_i'$s, are determined using the Kullback-Leibler divergence, as shown in the following equation:

$$KLD(P_i||Q_i) = \sum_j P_{j|i} \log \frac{P_{j|i}}{Q_{j|i}} \tag{5}$$

The cost function W is given by:

$$W = \sum_i KLD(P_i||Q_i) \tag{6}$$

This minimization is carried out using a gradient-descent optimization method. To optimize the cost function t-SNE used the gradient of the form:

$$\frac{\partial W}{\partial y_i} = 4 \sum_j (P_{j|i} - Q_{j|i})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \tag{7}$$

DensMAP DensMAP introduces a novel approach to embedding by modifying the cost function to include a term that preserves data point density, a feature not directly addressed by methods like t-SNE. This advancement allows DensMAP to maintain local and global structures more effectively. Empirical evidence suggests that integrating density information significantly improves the embeddings' quality (Narayan et al., 2021). The process involves additional computational overhead due to the density consideration, but it results in a more faithful representation of the original data's distribution.

In the DensMAP framework, local density is inferred from the proximity of neighboring points because closeness implies density. The local radius, defined by the mean distance to neighboring points, serves as an indicator of this density. The terms for local densities in the original and the reduced dimensional spaces are represented by:

$$R_P(\boldsymbol{x}_i) := \mathbb{E}_{j\sim P}\left[\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right] = \frac{\sum_{j=1}^n P_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{\sum_{j=1}^n \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}, \tag{8}$$

$$R_Q(\boldsymbol{y}_i) := \mathbb{E}_{j\sim Q}\left[\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2\right] = \frac{\sum_{j=1}^n Q_{ij}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2}{\sum_{j=1}^n \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2}. \tag{9}$$

where P_{ij} and Q_{ij} are symmetric measures of similarity between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, which are given by:

$$P_{ij} = P_{j|i} + P_{i|j} - P_{j|i}P_{i|j} \tag{10}$$

$$Q_{ij} = Q_{j|i} + Q_{i|j} - Q_{j|i}Q_{i|j} \tag{11}$$

The volume occupied by points is related to the radius raised to the power corresponding to the dimensionality of the space—for instance, the volume of a sphere in three dimensions scales with the cube of its radius. Accordingly, the relationship between the local densities in the input and embedding spaces is given by:

$$R_Q(\boldsymbol{y}_i) = \alpha(R_P(\boldsymbol{x}_i))^\beta \quad \Rightarrow \quad r_q^i = \alpha(r_p^i)^\beta + \gamma, \tag{12}$$

where $r_q^i := \ln(R_Q(\boldsymbol{y}_i))$, $r_p^i := \ln(R_Q(\boldsymbol{x}_i))$, and $\gamma := \ln(\alpha)$. Therefore, the relation of logarithms of the local densities should be affine dependence. A measure of linear (or affine) dependence is correlation so we use the correlation of logarithms of local densities:

$$\text{Corr}(r_q, r_p) := \frac{\text{Cov}(r_q, r_p)}{\sqrt{\text{Var}(r_q)\text{Var}(r_p)}}, \tag{13}$$

where the equation of covariance and variance are:

$$\text{Cov}(r_q, r_p) := \frac{1}{n-1} \sum_{i=1}^{n} \left[(r_q^i - \mu_q)(r_p^i - \mu_p) \right], \tag{14}$$

$$\text{Var}(r_q) := \frac{1}{n-1} \sum_{i=1}^{n} \left[(r_q^i - \mu_q)^2 \right], \tag{15}$$

where $\mu_q = \frac{\sum_{i=1}^{n} r_q^j}{n}$, and $\mu_p = \frac{\sum_{i=1}^{n} r_p^j}{n}$.

According to the cost function defined by Narayan et al.,2021:

$$C = -\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (P_{ij} \ln(Q_{ij}) + (1 - P_{ij}) \ln(1 - Q_{ij})) - \lambda \text{Corr}(r_q, r_p) \tag{16}$$

where λ is the regularization parameter which weights the correlation compare to the original cost. This minimization is carried out using a stochastic gradient descent for optimization. Proofs of the derivatives are available in (Narayan et al.,2021, Supplementary Note 2).

2.3 Clustering

Clustering is an unsupervised learning technique that explores inherent structures within data without prior labeling. After reducing dimensionality to facilitate a clearer analysis, we employed clustering to uncover potential relationships among stocks. In the context of stock trading, where ground truth is often absent, these clustering techniques provide a critical framework for hypothesizing and testing relationships among stocks, thus offering valuable insights into market behavior and investment strategies. Specifically, we concentrated on two clustering techniques: **K-means** and **Agglomerative Clustering**.

K-means K-means clustering is a method widely used in unsupervised learning to partition a dataset into a specified number k of clusters. Each observation in the dataset is assigned to the cluster with the nearest centroid, the mean of the points in that cluster. The k-means algorithm aims to find an optimal partitioning of the data by minimizing an objective function, known as the within-cluster sum-of-squares or inertia.

The objective function J that k-means seeks to minimize is given by:

$$J = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2, \tag{17}$$

where C_i represents the set of all data points in cluster i, $\boldsymbol{x}$ is a data point in C_i, and $\boldsymbol{\mu}_i$ is the centroid of C_i. The centroid is calculated as the mean of all points in the cluster, formalized by the equation:

$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \boldsymbol{x}. \tag{18}$$

The process of k-means clustering proceeds through several iterations of two main steps:

1. Assignment of each observation to the closest centroid, based on the Euclidean distance.
2. Updating the centroids to the mean of the observations that have been assigned to them.

The algorithm converges when the assignments no longer change. However, the solution may only be a local minimum, and the outcome can be sensitive to the initial selection of centroids.

Despite its inherent limitations, such as the assumption of spherical clusters and sensitivity to the initial centroid placement, k-means clustering is a powerful tool for data analysis. Its simplicity and efficiency make it a popular choice for data partitioning, with applications across various domains. Ongoing research into algorithmic improvements aims to overcome its drawbacks, enhancing its utility in complex real-world applications.

Agglomerative Clustering Agglomerative clustering is a type of hierarchical clustering technique that builds a multilevel hierarchy of clusters by successively merging clusters. This method initiates by considering each data point as a separate cluster and then iteratively combines these atomic clusters into larger and more comprehensive clusters until all points are united into a single global cluster, or a stopping criterion is satisfied.

The process of agglomerative clustering is generally as follows:

1. Initially, treat each observation as a separate cluster, resulting in n clusters for n observations.
2. Calculate the proximity matrix to determine the distances between each pair of clusters.
3. Merge the two closest clusters, reducing the total number of clusters by one.
4. Update the proximity matrix to reflect the distances between the newly formed cluster and the existing clusters.
5. Repeat steps 3 and 4 until all data are grouped into the desired number of clusters or a single cluster remains.

The algorithm employs a linkage criterion that determines the metric used for the distance between sets of observations. While several linkage criteria exist, such as single linkage, complete linkage, average linkage, and Ward's linkage. The choice of which to use may affect the shape and the size of the clusters formed. In the context of this paper, Agglomerative Clustering is utilized with its default linkage setting, namely the Ward linkage, which aims to minimize the within-cluster variance, creating more balanced clusters.

The result of agglomerative clustering can be depicted in a dendrogram, a tree-like diagram that records the sequences of merges or splits. Despite its conceptual simplicity, agglomerative clustering can be computationally intensive, but it is particularly useful for smaller datasets or when the hierarchical structure of clusters is of interest. The method is less scalable to large datasets due to its generally quadratic time complexity.

Hyperparameter Selection The silhouette score is an effective metric for assessing the number of clusters formed by a clustering algorithm, acting as a guide for hyperparameter optimization. It measures how similar an observation is to its own cluster compared to other clusters. The silhouette score takes a value between -1 and 1, where a higher score indicates that objects are well matched to their own cluster and poorly matched to neighboring clusters.

The silhouette score is calculated for each sample within a cluster and is defined as:

$$s = \frac{b - a}{\max(a, b)}$$

where a is the mean distance between a sample and all other points in the same cluster, and b is the mean distance from the sample to all points in the nearest cluster.

To automate the selection of the optimal number of clusters, the following steps can be implemented:

1. For each cluster count k, perform the clustering and calculate the silhouette score for each observation.
2. Compute the average silhouette score for all observations.
3. Select the k that maximizes the average silhouette score, indicating the most appropriate clustering configuration.

This method allows us to determine the most suitable number of clusters objectively by maximizing the average silhouette score, thus facilitating an informed decision on the optimal clustering setup for both k-means and agglomerative clustering algorithms. By utilizing the silhouette score, the choice of k is ensured to enhances both the cohesion within clusters and separation between clusters, making it a crucial hyperparameter in the clustering process.

2.4 Pairs List Selection

In each cluster, inspired by the methodology outlined by Han et al. (2023), we initially select stocks that are in the top and bottom 10% based on their mom_1 performance metric. This step identifies stocks with the highest and lowest momentum.

Johansen Test To examine if pairs of stocks move together over time, we apply the Johansen test for cointegration. This test helps to identify pairs of stocks whose prices share a stable long-term relationship. The Johansen test calculates the trace statistic, a measure used to determine the number of cointegrating relationships among pairs, as follows:

$$\text{Trace Statistic} = -T \cdot \sum_{i=r+1}^{n} \ln(1 - \lambda_i)$$

Here, T is the number of observations, λ_i are the eigenvalues derived from the cointegration test, and r is the number of cointegrating relationships hypothesized under the null hypothesis.

We use a significance level of $p = 0.05$ and require a minimum correlation of 0.8 to consider a pair as cointegrated. Candidate pairs that meet these criteria during the training period are shortlisted. If no pair meets the criteria, the top five pairs with the highest trace statistics (indicating the strongest evidence of cointegration) and a correlation above 0.8 are chosen. Pairs that satisfy the selection criteria across different methods are established as our equity pairs for long-short strategies. These pairs will subsequently undergo performance evaluation in live market conditions to validate the efficacy of our trading model.

2.5 Pair Selection Algorithm Overview

Algorithm 1: Pair Selection Algorithm Based on Clustering and Cointegration Test

Input: tickers T (list), Start Date SD (timestamp), End Date ED (timestamp)
Output: A DataFrame with columns: Month (timestamp), Top Ticker (list), Bottom Ticker (list), and cluster details (list)

Step 1: Calculate and Fetch Features;
$I_{\text{time}} \leftarrow$ Divide the period into months from SD to ED;
$S \leftarrow$ an empty DataFrame;
foreach *month* $i \in I_{time}$ **do**
  $S_i \leftarrow$ an empty DataFrame;
  $P_i \leftarrow$ Fetch daily prices for T from SD to ED;
  $M_i \leftarrow$ Calculate momentum using P_i over the past τ months;
  $F_i \leftarrow$ Retrieve financial metrics (sector, market capitalization, etc.);
  $S_i \leftarrow \text{merge}(S_i, P_i, M_i, F_i)$;
  $S \leftarrow \text{merge}(S, S_i)$;
end
Step 2 & 3: Dimensionality Reduction and Clustering;
$S_{reduced} \leftarrow$ an empty DataFrame, $C \leftarrow$ an empty DataFrame;
foreach *month* $i \in I_{time}$ **do**
  $S_{reduced_i} \leftarrow$ Apply t-SNE/DensMAP to S_i;
  $S_{reduced} \leftarrow \text{merge}(S_{reduced}, S_{reduced_i})$;
  $k_{max_i} \leftarrow$ Find the number of clusters with the highest silhouette scores;
  $C_i \leftarrow$ Perform agglomerative clustering/K-means on $S_{reduced_i}$ with k_{max_i};
  $C \leftarrow \text{merge}(C, C_i)$;
end
Step 4: Pair Formation and Filtering;
$O \leftarrow$ an empty DataFrame;
foreach *month* $i \in I_{time}$ **do**
  $R_i \leftarrow$ an empty DataFrame;
  foreach $c \in C_i$ **do**
    $\mathcal{R}_i \leftarrow$ an empty DataFrame;
    $\mathcal{P}_i \leftarrow$ Generate all possible pair combinations from top 10% and bottom 10% stocks by mom_1;
    foreach *pair* $(t, b) \in \mathcal{P}_i$ **do**
      if *Corr*$(t, b) > 0.8$ **then**
        if *JohansenTest*$(t, b, 0.05)$ **then**
          $\mathcal{R}_i \leftarrow \text{merge}(\mathcal{R}_i, \{(t, b)\})$;
        end
      end
    end
    if $\mathcal{R}_i$ *is empty* **then**
      $\mathcal{R}_i \leftarrow$ Top 5 pairs from $\mathcal{P}_i$ by highest Johansen test trace statistic where each pair (t, b) has *Corr*$(t, b) > 0.8$;
    end
    $R_i \leftarrow \text{merge}(R_i, \mathcal{R}_i, C_i)$;
  end
  $O \leftarrow \text{merge}(O, R_i)$;
end
return O;

3 Trading Signal Generation

Our trading strategy evolves from single mean reversion pair trading to long-short a bunch of candidates selected by our pair selection method. The idea is to match a long position with a short position in two groups of stocks with a high correlation and makes profits from the convergence in price difference (spread). A quantitative framework is established to identify the market regime and capitalize on price discrepancies between two correlated assets through total least squares and Constant elasticity of variance model. The methods to calculate spread and hedge ratio are discussed in the following subsections.

3.1 Single Pair Trading

Single pair trading is a market-neutral trading strategy that involves simultaneously buying and selling two highly correlated financial instruments with the expectation that the relationship between their prices will converge to a historical norm. Traders need to identify trading signals based on the spread, which is the difference in price between two assets. These signals will inform the traders when to enter and exit trades to capitalize on market inefficiencies. The key to a successful pairs trading strategy is to maintain a hedged position, where the trade is market-neutral and exposure to market risk is minimized. This involves calculating the appropriate quantities of each asset to buy and sell, ensuring that the positions are proportionate and that the hedge ratio is maintained. By doing so, traders aim to profit from the convergence of the spread without bearing the brunt of market volatility. For a single pair trading problem setting, we consider a stock pair $\{P, Q\}$, we denote their prices at time t as P_t and Q_t. Based on a given historical look period, we perform linear regression of $log(P_t)$ against $log(Q_t)$, and we assume the regression slope is β (also called hedge ratio), the regression residual ϵ_t, and standard deviation of the residual is σ (Ye, 2023). Let us assume the starting cash or capital is C. The linear regression equation can be expressed as:

$$log(P_t) = \beta log(Q_t) + \alpha + \epsilon_t, zscore_t = \frac{\epsilon_t}{\sigma} \tag{19}$$

We denote portfolio weights $wt_P = \frac{1}{1+\beta}$ for P_t, weight $wt_Q = \frac{\beta}{1+\beta}$ for Q_t and α is a constant.

1. At time t, if there is no open position and $zscore_t > entrythreshold$, we short the spread, that is, short $\frac{wt_P * C}{P_t}$ shares of P, and long $\frac{wt_Q * C}{Q_t}$ shares of Q.
2. At time t, if there is no open position and $zscore_t < -entrythreshold$, we long the spread, that is, long $\frac{wt_P * C}{P_t}$ shares of P, and short $\frac{wt_Q * C}{Q_t}$ shares of Q.
3. We hold stock positions (or number of shares) in each stock as constant once positions are opened.
4. If $zscore_t > -exitthreshold$ and long the spread, liquidate positions.
5. If $zscore_t < exitthreshold$ and short the spread, liquidate positions.

Our method will expand from this single pair trading to multiple pairs trading while generally keep the trading rules the same.

3.2 Total Least Squares

Total Least Squares (TLS) estimation determines parameters by minimizing the sum of orthogonal distances, which encompass both the measured distances directly to the regression line and the vertical distances, ensuring a more balanced and accurate fit, particularly when both dependent and independent variables carry measurement errors (Golub & Van-Loan, 1980). Unlike the Ordinary Least Squares approach, which only considers vertical distances and thus errors in the dependent variable, TLS treats the model more holistically.

Because these orthogonal distances remain invariant to shifts in the X and Y coordinates, the calculation of the slope parameter β in the TLS method is consistent. Therefore, it will facilitate the validity of hedge ratio even if we hold the paired position for a long period. In this framework, the observed values of X_i and Y_i (log-price of assets) are each associated with their respective error terms, reflecting the reality that data can be imperfect in both dimensions:

$$Y_i = y_i + e_i \sim \mathcal{N}(0, \sigma_e^2) \tag{20}$$

$$X_i = x_i + u_i \sim \mathcal{N}(0, \sigma_u^2) \tag{21}$$

where x_i and y_i are true values and e_i and u_i are error terms following independent identical distributions. It is assumed that there is linear combination of true values. For convenience,we represent the error variance ratio τ as follows:

$$y_i = \beta_0 + \beta_1 x_i \tag{22}$$

$$Y_i = \beta_0 + \beta_1 x_i + e_i \sim \mathcal{N}(0, \sigma_e^2) \tag{23}$$

$$\tau = \frac{var(Y_i|x_i)}{var(X_i|x_i)} = \frac{\sigma_e^2}{\sigma_u^2} \tag{24}$$

The orthogonal regression estimator is calculated by minimizing the sum of the measured distance and the vertical distance between regression lines by Kim in the following equation (2019):

$$\sum_{i=1}^{n} \{ \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{\tau} + (X_i - x_i)^2 \} \tag{25}$$

$$\beta_1 = \frac{s_{YY}^2 - \tau s_{XX}^2 + \{(s_{YY}^2 - \tau s_{XX}^2)^2 + 4\tau s_{XY}^2\}^{\frac{1}{2}}}{2 s_{XY}} \tag{26}$$

The β_1, which is the hedge ratio obtained from the above equation is used in the same way as that obtained from section 3.1. Similarly, the epsilon value is also used as a trading signal through the Z-score in the state composed of the formation-window size.

3.3 Constant Elasticity of Variance Model

The constant elasticity of variance model (CEV) is a stochastic volatility model that attempts to capture stochastic volatility and the leverage effect (Cox, 1996).

$$dS_t = \mu S_t dt + \sigma S_t^{\theta} dW_t \tag{27}$$

in which S_t is the spot price, t is time, and μ is a parameter characterising the drift, σ is the asset volatility and θ is the elasticity parameter, and W is a Brownian motion.

Unlike traditional models that assume a constant volatility, the CEV model allows volatility to be a function of the underlying asset price, capturing the empirical observation that volatility tends to increase as stock prices decline. This feature aligns with the phenomenon where a company's financial leverage inversely correlates with its equity value, resulting in increased volatility during stock price downturns. Moreover, CEV model offers a more accurate reflection of downside risk and extreme market behavior, specifically heavy-tailed distributions of asset returns, which enhance the effectiveness of risk management strategies.

In our implementation, we use maximum likelihood estimator to estimate the volatility σ of underlying asset to facility the calculation in TLS, precisely, referring as the error terms in equation (20) and (21).

For maximum likelihood estimation, we assume the stock price follows a log-normal distribution with mean and variance given in the CEV model:

$$log(S_t) \sim \mathcal{N}(log(S_0) + (\mu - \frac{1}{2}\sigma^2)t, \sigma^2 t^{2\theta}) \tag{28}$$

Moreover, letting $z_i = \frac{log(S_i)-log(S_{i-1})}{\sqrt{t_i}}$, where t_i is the time interval between ith and $(i-1)$th observations, we have:

$$log(S_t) - log(S_0) = (\mu - \frac{1}{2}\sigma^2)t + \sigma t^\theta Z), where Z \sim \mathcal{N}(0,1) \tag{29}$$

And the log-likelihood would be in the form of:

$$l(\mu, \sigma, \theta) = log(L(\mu, \sigma, \theta)) = log(\prod_i^n f(z_i; \mu, \sigma, \theta)) \tag{30}$$

where $f(z_i; \mu, \sigma, \theta)$ is the probability density function of the standard normal distribution evaluated at z_i.

Writing $f(z_i; \mu, \sigma, \theta)$ in terms of the log-normal distribution, we have:

$$f(z_i; \mu, \sigma, \theta) = \frac{1}{z_i\sqrt{2\pi}} exp(\frac{-(log(\frac{S_i}{S_{i-1}}) - (\mu - \frac{1}{2}\sigma^2)t_i)^{2\theta}}{2\sigma^2 t_i^{2\theta}}) \tag{31}$$

Substituting the above result into log-likelihood, we get:

$$l(\mu, \sigma, \theta) = -\frac{n}{2}log(2\pi) - \sum_i^n log(z_i) - \sum_i^n \frac{(log(\frac{S_i}{S_{i-1}}) - (\mu - \frac{1}{2}\sigma^2)t_i)^{2\theta}}{2\sigma^2 t_i^{2\theta}} \tag{32}$$

3.4 Trading Framework

As depicted in figure 2, our trading strategy generally involves market regime detection, hedge ratio calculation, z-score calculation and equal weighted fund allocation.

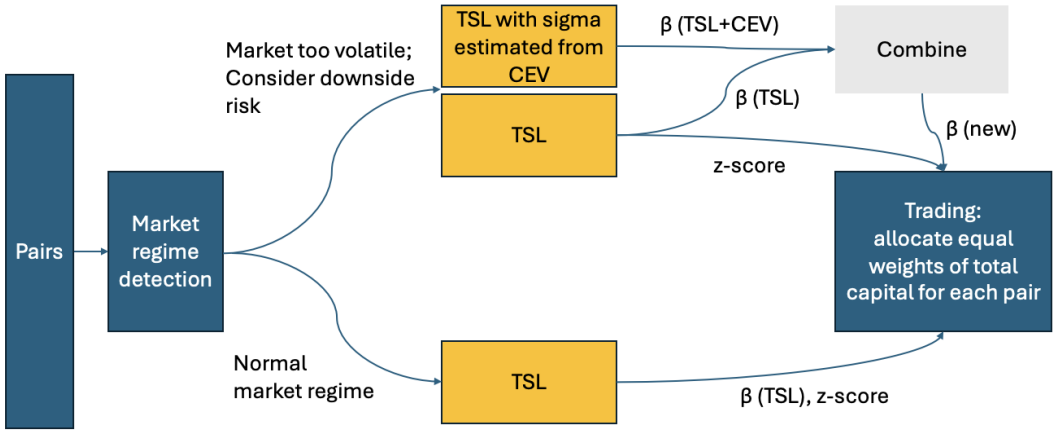

Figure 2: Trading Framework

The strategy begins with selected equity pairs, which are then subjected to market regime detection to determine the current state of market volatility by the 80 days median of VIX value. If the market is deemed overly volatile, indicating elevated risk, the framework incorporates downside risk considerations by employing a Total Least Squares (TSL) approach with volatility estimated from the Constant Elasticity of Variance (CEV) model. This hybrid model aims to enhance the beta estimation by incorporating insights from the CEV model's ability to gauge the changing volatility landscape more accurately. On the other hand, in a normal market regime, the framework relies on the TSL method alone for beta estimation.

The beta values obtained from TSL (and TSL combined with CEV in volatile markets) are then used to calculate the z-scores, which inform the trading signals. The z-score, a

statistical measure that indicates the number of standard deviations from the mean pair price ratio, serves as a critical input for identifying trade entry and exit points.

The final component of the strategy is the trading execution, where equal weights of the total capital are allocated for each pair. This equal-weighting approach simplifies the portfolio construction and ensures a balanced exposure across all trading pairs. The framework combines robust statistical methods with practical trade execution rules, aiming to navigate through different market conditions and exploit opportunities presented by the convergence of prices in paired equities.

3.5 Implementation

This trading strategy is rigorously backtested within the QuantConnect platform, employing a 90-day lookback period for both Beta and z-score calculations. As outlined in Section 3.1, the strategy is initiated with an entry threshold of 2.5 and an exit threshold set at 0.

To accurately reflect real-market conditions, the strategy incorporates the volume share model alongside the Interactive Brokers fee model to simulate slippage and transaction costs, respectively. The volume share model is adeptly chosen for its relevance to pair trading, as it dynamically adjusts slippage based on the order's proportion to the historical volume of trade. This ensures that the market impact is mitigated, preserving the intrinsic price relationship of the traded pairs and allowing for the exploitation of price discrepancies without distorting the market. Additionally, the inclusion of a price impact variable within the model aids in fine-tuning the trade size, which is vital for sustaining the pairs' equilibrium and ensuring the strategy's market neutrality.

Regarding transaction costs, the strategy adopts the Interactive Brokers model to closely emulate actual trading expenses. This cost structure is particularly precise, charging US equity trades at $\$0.005$ per share with a minimum fee of $\$1$ and capping the maximum fee at 0.5% of the trade value. Such detailed cost modeling is critical for developing a realistic assessment of the strategy's net profitability and for ensuring that the strategy can withstand the associated costs of active trading.

4 Active Risk Management

The long-short nature of our pairs trading strategy ensures relative market neutrality. However, the mean-reverting tendencies of the pairs become less predictable when the market is in extreme conditions. This section will focus on the active risk management strategies we have explored to mitigate these risks and enhance the resilience of our trading approach. We will delve into our exploration of several novel methods, including employing option strategies for volatility hedging and using large language models for volatility forecasting, along with standard risk management techniques such as stress testing, that ensure robust performance across varying market scenarios.

4.1 Option Strategies for Volatility Hedging

As a part of our initial proposal, we proposed to hedge the risk of divergence that stems from the change in fundamentals or other idiosyncratic shocks, by longing the individual stock's volatility using straddles or strangles. We observed that, materialized idiosyncratic risks would incur a spike in the implied volatility of the stock and, in most cases, simultaneously diverge its share price from its comparable peers. Figure 3 is an example of this behavior–NextEra Energy's (NEE) stock price sharply declined when it announced its sale agreement for Florida City Gas (source). And its share price diverged from Duke Energy (DUK). Therefore, employing straddles or strangles allows us to capitalize on increased volatility while protecting against downside risks associated with unanticipated divergences, effectively stabilizing our portfolio against idiosyncratic shocks.

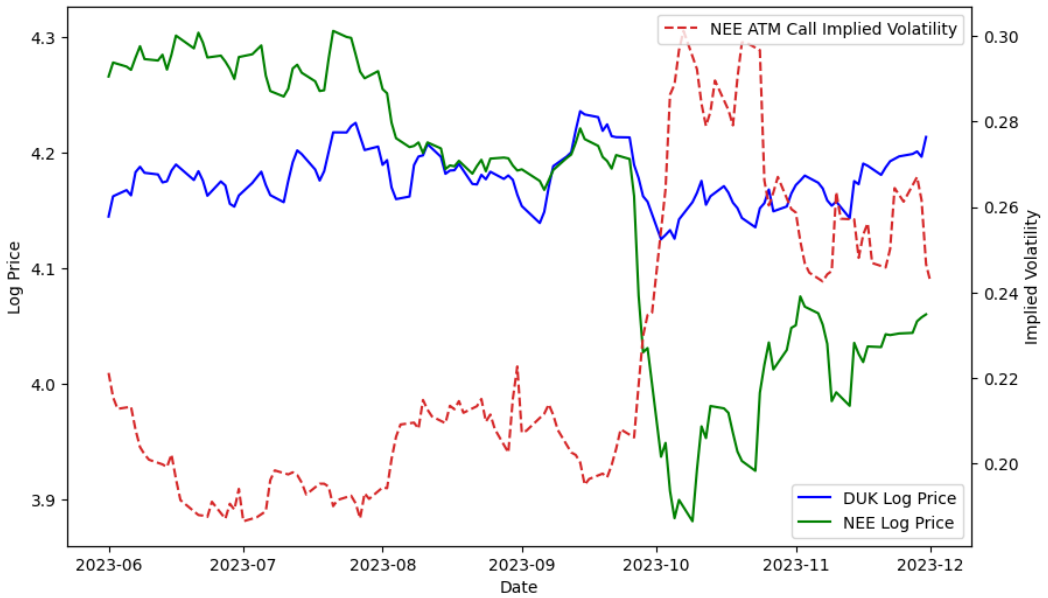

Figure 3: NEE ATM Call IV vs Spread Between NEE and DUK

The downside of this strategy is the cost associated with purchasing options. These costs primarily arise from two sources: the price of the options themselves, and the impact of theta decay. Theta, represented by θ, quantifies the rate at which the value of an option decreases as the expiration date approaches. This decay can significantly affect the profitability of the strategy, particularly if the expected price movements do not materialize within the expected timeframe.

Our pairs trading strategy has evolved from a static buy (sell) and hold approach to a dynamic strategy that may potentially trade multiple pairs each month. This adaptation allows us to respond more effectively to changing market conditions and optimize our hedging strategies. Yet, it also increased the cost of options, depriving us of some potential profits and making the strategy less economical in practice. Ultimately, this led to the exclusion of this particular dynamic option-based hedging from our final strategy. However, the concept of utilizing volatility for risk management was valuable and inspired us to employ other methods to effectively reduce the variance and drawdown of our strategy's returns. These alternative approaches will be discussed in more detail later in this section.

4.2 Large Language Model

In our exploration of risk management techniques, we investigated the potential of leveraging large language models (LLMs) for forecasting idiosyncratic risks before their manifestation. Our reasoning behind this is that LLMs can process a large amount of natural language information in a relatively short time, and may be able to parse and combine different information and yield unexpected insights. While Dang et al. (2020) demonstrated the application of the BERT model (Bidirectional Encoder Representations from Transformers) for detecting volatility, the predictive capabilities of LLMs in this context remain largely unexplored. Therefore, we initiated an independent investigation using the Tiingo stock news dataset, which provides headlines for individual stocks when supplied with a specific ticker.

Our study focused on three distinct stocks: Meta Platforms, Inc. (META), NextEra Energy, Inc. (NEE), and Blackstone Inc. (BX), chosen for their diverse market behaviors during the study period. META was selected due to its significant price increase of approximately 20% from February 1 to February 2, 2024, triggered by better-than-expected earnings reports and the announcement of its inaugural dividend. Conversely, NEE experienced a notable price drop of 10 from September 26 to September 27, 2023, attributed to the impacts of the aforementioned spin-off. BX was included as a contrasting case, showing no significant price movements from January 3 to February 2, 2024, thereby serving as a baseline for normal volatility within the study period.

The experimental design of our forecasting study involved the use of the Tiingo API to retrieve news titles for a 30-day period preceding significant price movements in the stocks of META and NEE. With the objective to determine if the GPT-4 model could predict upcoming price spikes based on the news data, we queried to classify the likelihood of a price spike as 'true', 'false', or 'uncertain'. For the control group BX, which exhibited no significant price movements during the period under study, the model was similarly tasked with analyzing 30 days of news titles prior to February 1, 2024, to predict the possibility of a price spike.

Ticker	Price Shock (Ground Truth)	GPT-4 Prediction
BX	False	Uncertain
META	True	Uncertain
NEE	True	Uncertain

Table 1: Ground Truth versus GPT-4 Predictions for BX, META, and NEE.

The results of our experiment (Table 1), which consistently yielded an 'uncertain' prediction across both test cases and the control group. This outcome may reflect the limitations of general LLMs, which are not specifically trained for financial predictions and thus may lack the nuanced understanding required to make accurate forecasts from news data alone. To enhance the predictive capability of LLMs in this domain, future research could explore the

integration of specialized financial training datasets, more sophisticated prompt engineering, or the development of hybrid models that combine LLM outputs with traditional quantitative financial analysis tools.

4.3 Volatility in Pairs Trading

Our pairs trading strategy benefits from the tendency of two stocks selected via the methodology mentioned in the above sections to converge. Even though our strategy is largely market-neutral, we observed that during periods of abnormally high volatility pair prices are less likely to cross the closing threshold, resulting in larger variance and depressed returns during the periods. We believe that the reduced returns during period of excessive volatility are caused by the high correlation in the market, a pheromone supported by both probability theory and empirical evidence (Loretan and English, 2000). Because of the high correlation, the selected pairs would fluctuate more closely with each other which makes them less likely to approach our take-profit thresholds. Thus, eventually resulting in the liquidation of the positions at loss.

Trading under conditions of extreme volatility is generally undesirable for our strategies, as it renders more false trading signals and the convergence of asset prices less predictable. During our backtesting period from 2017 to 2021, we found that setting a VIX threshold at 50 could effectively mitigate drawdowns. This threshold was chosen based on the VIX level on March 9, 2020, which was the first of four instances when circuit breakers were triggered during the market turmoil of March 2020. While the VIX threshold may seem like an arbitrary level, this mechanism is essential to prevent our strategy from taking unnecessary risks in extreme market conditions. To address the issue of volatility clustering, first documented by Mandelbrot in 1963, where high volatility can persist and fluctuate around our threshold, we implemented a cool-down period of 10 trading days during which no trading activity would occur, once the threshold is hit. This measure is crucial because, without it, our algorithm might re-enter the market prematurely after a slight dip below our VIX threshold, only to face a subsequent jump in VIX. Such a scenario would force another round of liquidations, incurring unnecessary costs and potentially leading to significant losses. By instituting this cool-down period, we aim to ensure that our strategy avoids re-entry into the market during unstable periods and thus protects against the repeated financial impacts of high volatility.

Figure 4: Strategy PnL Without the Stop-trading Mechanism Based on VIX

Figure 5: Strategy PnL With the Stop-trading Mechanism Based on VIX

Figure 4 and Figure 5 showcases the PnL between our strategy without the VIX-based stop-trading mechanism. The results are clear–with the stop-trading mechanism, the strategy yields a higher net profit and is less volatile during the high volatility period of early 2020.

In refining our strategies to mitigate risks and enhance returns, we have adopted the practice of shorting the VIX ETF, specifically UVXY, during periods of elevated VIX. This decision leverages the mean-reverting nature of the VIX and the consistent downward trend of UVXY, a leveraged ETF designed to deliver 1.5 times the daily performance of the S&P 500 VIX Short-Term Futures Index. The inherent structure of UVXY, characterized by daily rebalancing and the decay in the value of futures contracts, causes it to depreciate over time, making it an attractive target for short-selling during volatility spikes.

However, we approach this strategy with significant caution. Our position sizes are deliberately small to minimize potential losses, and we have implemented a stringent stop-loss mechanism, as seen in the Drawdown Control section below. This cautious approach allows us to capitalize on the potential benefits of shorting UVXY while substantially reducing the risk of large drawdowns during unexpected market movements.

4.4 Other Risk Management Techniques

Z-score-based Stop-loss

To effectively mitigate the risks associated with pair trading, we have instituted a Z-score-based stop-loss mechanism that liquidates positions when the z-score exceeds a certain level. While the ideal scenario would assume a normal distribution of Z-scores—implying a minimal probability of extreme deviations—financial markets frequently defy such assumptions. In reality, there can be instances where the price differential between two correlated stocks widens excessively. More critically, there are circumstances where the diverging prices of the paired stocks may not revert to the historical mean, posing a significant risk of persistent losses.

To address these challenges, this stop-loss mechanism is calibrated based on the Z-score. A higher Z-score suggests a widening spread between the stocks thus a greater risk of loss. Hence, Z-score-based stop-loss mechanism necessitates a dynamic stop-loss strategy that

can adapt to changing market conditions and effectively control the risk by terminating the position before losses escalate beyond an acceptable limit.

Drawdown Control

To further enhance our risk management framework, we have implemented a drawdown control mechanism. This mechanism is centered around a 7% trailing stop-loss from the most recent high water mark. Once this threshold is triggered, all trading positions are promptly liquidated, and trading activities are suspended for a period of time. This immediate action serves to prevent further erosion of capital in the face of declining market values.

The decision to halt trading is particularly crucial during times when specific stocks or entire industries experience periods of elevated uncertainty. Such conditions often result in unpredictable market movements and heightened volatility, which may not be reflected by the VIX index. Following the activation of the stop-loss, a cooldown period is instituted before reentry into the market. The pause in trading allows the market conditions to simmer down. After this cooling-off period, the algorithms reassess the market data to determine the optimal timing for re-entry. This mechanism, with the implementation of a trailing stop-loss and a lagged re-entry, effectively helps reduce drawdowns and limits excessive risk-taking, ensuring a controlled exposure to market uncertainties.

5 Experiments and Results

In this study, we crafted an experimental design to rigorously evaluate the effectiveness of our trading strategy. The in-sample training period was established from January 1, 2017, to January 1, 2021, a span that was selected to provide a robust dataset for model training and validation.

To assess the predictive power and resilience of our trading approach, we implemented three distinct out-of-sample testing intervals: the first from January 1, 2016, to January 1, 2017, the second from January 1, 2022, to November 1, 2022, and the third from October 1, 2023, to April 1, 2024. We strategically chose these time periods to test the performance of the strategy under different market conditions to get a comprehensive picture of its time robustness and generalizability.

5.1 Stock Selection Result

In the domain of candidate selection for pair trading, two predominant methodologies were observed. The first involved an exhaustive search across the entire universe of assets (Krauss, Do, & Huck, 2017). Although this methodology was computationally intensive, it was capable of uncovering exceptionally intriguing pairs. The second methodology categorized assets into sectors prior to pairing (Do & Faff, 2010), hoping to significantly reducing the likelihood of spurious correlations (Sarmento & Horta, 2020). However, it may also inadvertently overlook potential intersectoral relationships.

Recognizing the advantages and limitations of both methodologies, we aimed to integrate these approaches to not only discover interesting pairs but also to incorporate the sector assumption. Specifically, we initially narrowed our focus to stocks within the S&P 500. Given the computational constraints of the QuantConnect platform, we further refined our selection to the top 99 stocks by market capitalization as of January 1, 2017, the commencement of our in-sample training period. Additionally, since we already incorporated sector information as a feature in the feature generation process, we performed an exhaustive search within these 99 tickers. The final selection results are detailed in Table 2.

Tickers								
AAPL	ABBV	ABT	ADBE	ADP	AIG	AMGN	AMT	AMZN
AVGO	AXP	BA	BAC	BIIB	BK	BLK	BMY	CAT
CHTR	CL	CMCSA	COP	COST	CRM	CSCO	CVS	CVX
D	DHR	DIS	DOW	DUK	EOG	EPD	F	FB
FDX	GD	GE	GILD	GM	GOOG	GS	HAL	HD
HON	IBM	INTC	JNJ	JPM	KHC	KMI	KO	LLY
LMT	LOW	MA	MCD	MDLZ	MET	MMM	MO	MRK
MS	MSFT	NEE	NFLX	NKE	NVDA	ORCL	OXY	PCLN
PEP	PFE	PG	PM	PNC	PYPL	QCOM	SBUX	SCHW
SO	SPG	T	TJX	TMO	TMUS	TXN	UNH	UNP
UPS	USB	UTX	V	VZ	WBA	WFC	WMT	XOM

Table 2: Curated Investment Tickers

5.2 Feature Generation Result

Throughout the training and testing phases of the study, we carefully calculated returns for each month from mom_1 through mom_{12} using Equation 1. Concurrently, we sourced additional company profiles and asset information for each month from the Morningstar

dataset. To further enhance our dataset, we factorized categorical features such as sector codes and NAICS codes, assigning them integer labels starting from 1.

For categorical feature sector code and NAICS code in int form, we first factorize them and reassign label from 1 to them. Through this methodical approach, we generated 19 comprehensive feature sets during the feature generation process. These features encompass not only stock price dynamics but also a variety of other relevant financial indicators. This multifaceted dataset provides a solid foundation for analyzing the interactions between stock performance and broader company metrics, thereby improving the predictive accuracy of our trading strategies. The feature set example for the first month of the training period is detailed in Table 3.

Ticker	m1m	...	m12m	Sector	MC	People	GS	VS	SS	NAICS
AAPL	0.047	...	0.06	311	6.08E+11	11600	74.08	74.79	423.68	1
ABBV	0.029	...	0.14	206	1.01E+11	30000	77.28	63.10	307.86	2
ABT	0.008	...	-0.10	206	5.65E+10	75000	47.83	46.38	270.26	3
...	...	...	...	...	...	...	...	...	...	...

Table 3: Feature Set

5.3 Dimension Reduction and Clustering Result

In our study, we evaluated two dimensionality reduction techniques—t-SNE and DensMAP—and two clustering methodologies—K-means and Agglomerative Clustering—on the stock dataset.

We assessed the clustering outcomes by calculating the Silhouette Score for four distinct combinations: 1. t-SNE + Agglomerative Clustering, 2. t-SNE + K-means, 3. DensMAP + Agglomerative Clustering, and 4. DensMAP + K-means. The Silhouette Score serves as a metric for cluster quality, with higher scores indicating more well-defined cluster separation. According to the results presented in Table 4, while K-means and Agglomerative Clustering performed comparably, DensMAP achieved a superior quality of clusters compared to t-SNE, demonstrating its effectiveness in discerning more coherent groupings.

Further, inspired by the findings of Han, He, and Toh (2023), who noted that Agglomerative Clustering outperformed other methods when applied to datasets enriched with firm-specific features, we opted to finalize our pair selection strategy by integrating DensMAP for dimensionality reduction and Agglomerative Clustering for cluster formation.

Figures 6 through 9 illustrate the outcomes of the four distinct combinations with dimensions reduced to both 2D and 3D, and with a fixed cluster count of three, solely for demonstration purposes. In our formal study, we consistently reduced the dimensionality to 3D while the number of clusters varied according to different months based on the Silhouette Score to adapt to temporal variations in the data.

	t-SNE	DensMAP
K-means	0.24	0.38
Agglomerative	0.26	0.43

Table 4: Silhouette Score Result

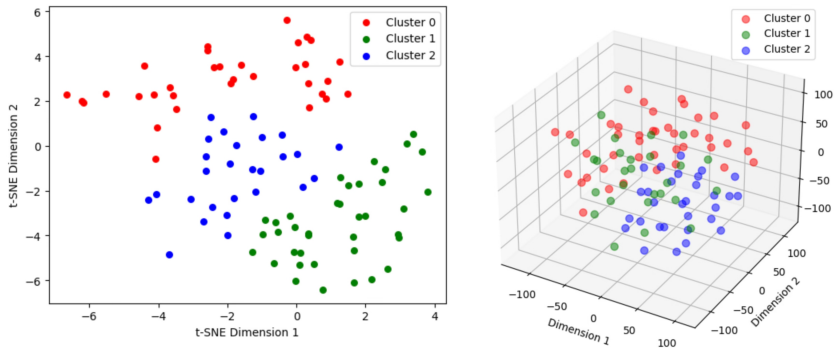

Figure 6: Visualization of t-SNE and Agglomerative Clustering

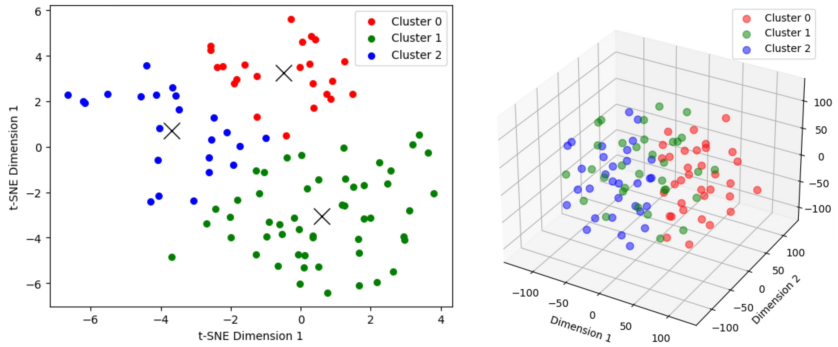

Figure 7: Visualization of t-SNE and K-means

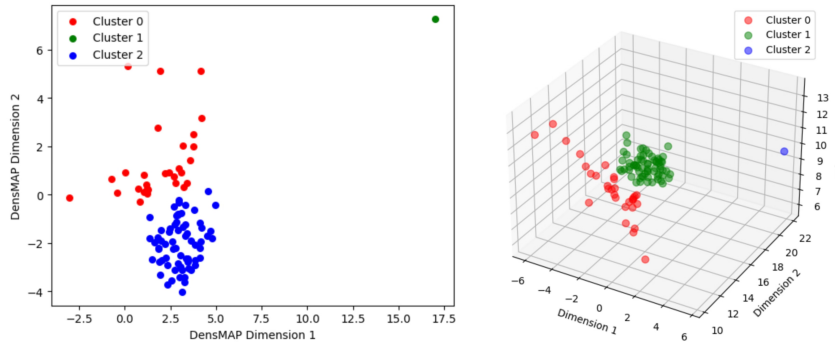

Figure 8: Visualization of DensMAP and Agglomerative Clustering

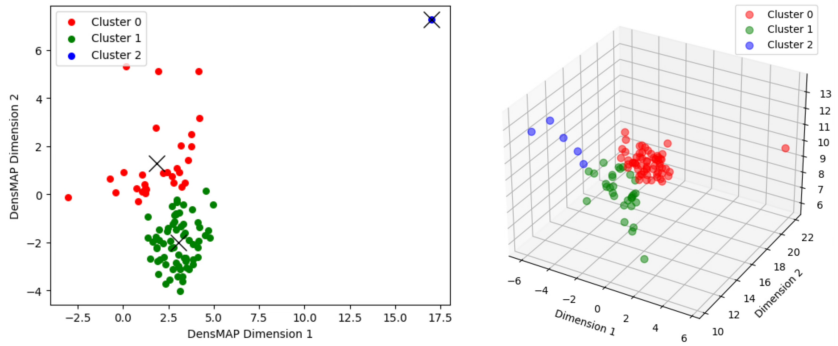

Figure 9: Visualization of DensMAP and K-means

5.4 Pair Selection Result

The final output was comprised of four principal components. The initial segment represents the temporal aspect, indicating the period during which the selected pair was utilized. For instance, the data in the first row pertains to the pair information employed from January 1, 2017, to January 31, 2017. The subsequent two components are the top ticker and bottom ticker, respectively, which collectively constitute the trading pair. The remaining columns furnish detailed information about the clusters, serving as the foundational data essential for formulating subsequent trading signal strategies. A partial view of the pair selection data frame from the training period is presented in Table 10.

	Time	Top	Bottom	cluster0	cluster1	cluster2	cluster3	cluster4
0	2017-01-31	['CRM SZQUJUA9SVOL', 'TJX R735QTJ8XC9X', 'ORCL...	['AAPL R735QTJ8XC9X', 'MO R735QTJ8XC9X', 'SBC ...	['AIG R735QTJ8XC9X', 'ALD R735QTJ8XC9X', 'CHV ...	['AAPL R735QTJ8XC9X', 'ADBE R735QTJ8XC9X', 'AM...	['AVGO UEW4IOBWVPT1', 'BA R735QTJ8XC9X', 'CHTR...	['AMGN R735QTJ8XC9X', 'BK R735QTJ8XC9X', 'F R7...	['BEL R735QTJ8XC9X', 'BMY R735QTJ8XC9X', 'GILD...
1	2017-02-28	['IDPH R735QTJ8XC9X', 'CHV R735QTJ8XC9X', 'PFE...	['ADBE R735QTJ8XC9X', 'AVGO UEW4IOBWVPT1', 'AB...	['BLK ROIDIDJRNXID', 'CAT R735QTJ8XC9X', 'CHTR...	['AMZN R735QTJ8XC9X', 'CL R735QTJ8XC9X', 'COST...	['AAPL R735QTJ8XC9X', 'ABBV VCY032R250MD', 'AD...	['ABT R735QTJ8XC9X', 'AMGN R735QTJ8XC9X', 'CRM...	['BEL R735QTJ8XC9X', 'BMY R735QTJ8XC9X', 'F R7...
2	2017-03-31	['NVDA RHM8UTD8DT2D', 'KFT S5HU4FPL6G6D', 'OXY...	['AAPL R735QTJ8XC9X', 'FPL R735QTJ8XC9X', 'BMY...	['CAT R735QTJ8XC9X', 'CHV R735QTJ8XC9X', 'EOG ...	['AUD R735QTJ8XC9X', 'COST R735QTJ8XC9X', 'FB ...	['BK R735QTJ8XC9X', 'CVS R735QTJ8XC9X', 'DIS R...	['ALD R735QTJ8XC9X', 'CMCSA SJTSDFE19S9X', 'IB...	['AAPL R735QTJ8XC9X', 'AMZN R735QTJ8XC9X', 'BE...
3	2017-04-30	['F R735QTJ8XC9X', 'IDPH R735QTJ8XC9X', 'COST ...	['TXN R735QTJ8XC9X', 'ABBV VCY032R250MD', 'LOW...	['AMGN R735QTJ8XC9X', 'BK R735QTJ8XC9X', 'BMY ...	['CHV R735QTJ8XC9X', 'EOG R735QTJ8XC9X', 'GE R...	['ABBV VCY032R250MD', 'AUD R735QTJ8XC9X', 'BLK...	['ABT R735QTJ8XC9X', 'CL R735QTJ8XC9X', 'COST ...	['BA R735QTJ8XC9X', 'CHTR UPXX4G43SIN9', 'CMB ...
4	2017-05-31	['SCH R735QTJ8XC9X', 'FDX R735QTJ8XC9X', 'LMT ...	['MWD R735QTJ8XC9X', 'BA R735QTJ8XC9X', 'CHTR ...	['ABT R735QTJ8XC9X', 'CL R735QTJ8XC9X', 'CVS R...	['AMGN R735QTJ8XC9X', 'BLK ROIDIDJRNXID', 'CAT...	['ADBE R735QTJ8XC9X', 'AVGO UEW4IOBWVPT1', 'EO...	['ALD R735QTJ8XC9X', 'CSCO R735QTJ8XC9X', 'DUK...	['AIG R735QTJ8XC9X', 'BK R735QTJ8XC9X', 'CMCSA...

Figure 10: Partial View of Pair Selection Data Frame

5.5 Backtesting Result

Figure 11: Portfolio Return 2017-2020

The in-sample backtesting, conducted from January 1, 2017, to January 1, 2021, yielded a total return of 154.64% over the four-year period, as depicted in Figure 11. The strategy

not only demonstrated a winning rate of 59% but also achieved a profit-loss ratio of 2.77, culminating in a Sharpe ratio of 1.446. Additionally, the maximum drawdown was contained to 9.00%, reflecting the effectiveness of our risk management measures within the strategy's design.

Additionally, out-of-sample backtesting was conducted to validate the strategy's robustness beyond the in-sample timeframe. The results of this backtesting, which further demonstrate the strategy's resilience across different market conditions, are presented in table 5.

Period	Returns	Sharpe Ratio	Maximum Drawdown
Out-of-Sample: 1/1/2016 - 1/1/2017	-1.45%	-0.237	5.3%
Out-of-Sample: 1/1/2020 - 11/1/2020	10.47%	0.601	7.9%
Blind Out-of-Sample: 10/1/2023 - 4/1/2024	4.80%	0.177	6.2%

Table 5: Out of Sample Backtesting Results

5.6 Risk Management Stress-testing

Figure 12: 2008 Financial Crisis Stress-test Result

To ensure the robustness of our trading parameters and validate the effectiveness of our risk management strategies, we conducted stress tests using historical data from the tumultuous 2008 financial crisis. This period, marked by extreme market volatility and significant economic downturns, served as a rigorous testing ground for our algorithms. Figure 12 illustrates the performance of our strategy from Sept 1, 2008, to January 1, 2009, which despite challenging conditions, managed to achieve a positive return of 5.12% and the drawdown is approximately 3.9%. it compares favorably against the S&P 500, which experienced a drawdown of roughly 40% and a negative return of about 33% over the same period. This result further demonstrates the resilience and relative stability of our strategy.

To further evaluate the effectiveness of our risk management measures, we conducted an additional experiment by disabling all risk management logic and rerunning the same backtest over the specified period. The results of this test are presented in Figure 14. Without the risk management mechanisms, the strategy suffered a drawdown exceeding 15% and yielded a return of approximately 2%.

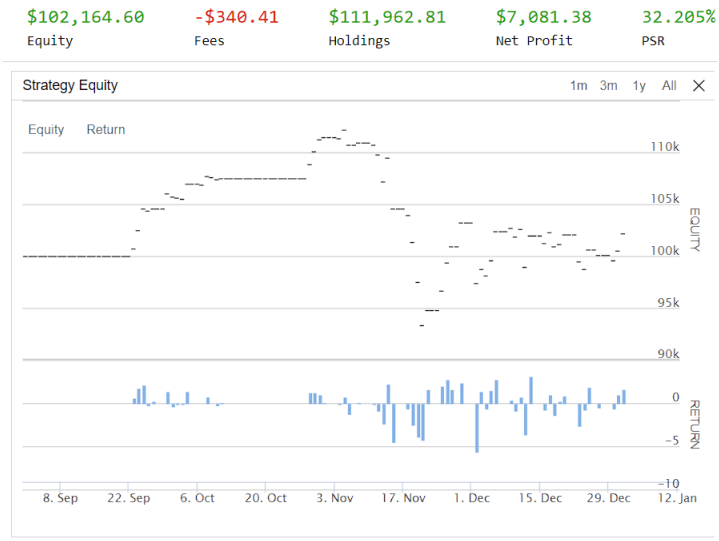

Figure 13: 2008 Financial Crisis Stress-test Result Without Risk Management

Figure 14: COVID Stress-test Result With Risk Management

The stress test over the COVID market turmoil (Mar 1, 2020 - Apr 1, 2020) yielded a negative monthly return of -2.22%.

6 Live Trading Result

Our algorithm has been deployed in a live trading environment since April 23rd, and as of now, no trades have been executed as the specified conditions for making transactions have not been met (see Figure 15). We will continue to observe and analyze market conditions closely, maintaining our algorithm in readiness to execute trades when the relevant criteria are satisfied.

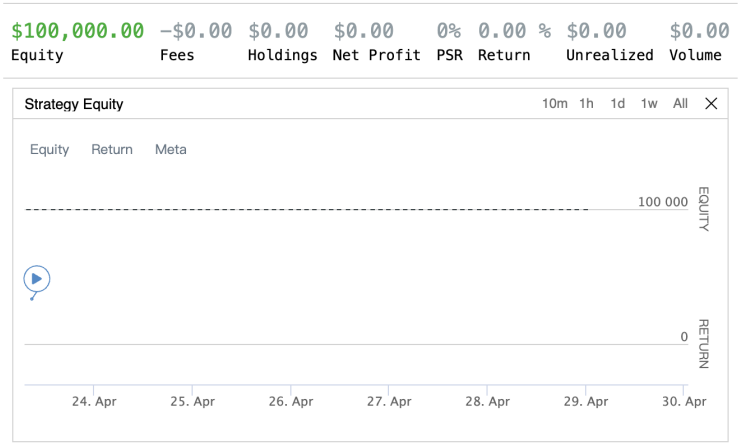

Figure 15: Live Trading Result

7 Future Work

In our ongoing pursuit of excellence, we plan to extend the scope of our strategy to encompass a broad spectrum of asset classes, including cryptocurrencies, futures, and foreign exchange markets. This endeavor will not only diversify the applicability of our approach but also present an opportunity to harness the distinctive characteristics inherent to each asset class.

Moreover, we aim to apply our refined strategy across various global stock markets, with particular attention to the European, Chinese, and Indian financial landscapes. Given the unique nature of each market's dynamics, we anticipate that further nuanced adjustments will be required. Consequently, an element of bespoke fine-tuning will be an integral part of this expansion process.

From a technical standpoint, since the dimension reduction and clustering techniques we have employed thus far are grounded in traditional methodologies, we will continue our exploration of more sophisticated approaches in future work. To this end, we propose the integration of cutting-edge dimension reduction methods. Specifically, we will delve into the potential of deep neural networks, such as Autoencoder Networks.

Additionally, recognizing that stocks are multi-faceted entities with myriad time series features, we are considering the application of tensor decomposition methods for dimensionality reduction. Tensor decomposition stands to offer significant advantages over traditional methods, notably in its ability to maintain the multi-dimensional nature of the data, thereby preserving inherent interrelationships and providing a richer, more holistic understanding of the underlying structures.

In the realm of clustering, we are set to explore the frontiers of biclustering and triclustering in high-dimensional spaces. These sophisticated clustering techniques have the potential to uncover correlations and patterns that may be overlooked by traditional methods, leading to a more nuanced and in-depth analysis of market data.

When it comes to trading, market regime detection should be given considerable attention, because it allows strategy to be adapted to the prevailing market conditions, enhancing the potential for profitability and risk management. In our strategy, regime detection is based on a simple rule. By identifying the current regime more accurately, whether it be a bull or bear market, a high-volatility environment, or a period of market stability, optimal strategies can be adjusted with position sizing, risk tolerance, and even the selection of assets to trade in a way that aligns with the inherent characteristics of that regime.

Regarding volatility analysis, we could train, and fine-tune a hybrid transformer-based model tailored for predicting market volatility. This model will leverage the powerful capabilities of transformers to analyze time-series data and extract complex patterns that are indicative of volatility trends. By accurately forecasting periods of high volatility, it will enable us to actively trade volatility as a hedging mechanism alongside our current pairs trading strategy. The ultimate goal is to create a robust system that not only capitalizes on volatility as an asset class but also uses these predictions to mitigate risks inherent in pairs trading.

8 Conclusion

In this study, we introduced an advanced algorithmic approach that not only enhances the selection and management of stock pairs but also optimizes risk assessment, and strategic adaptability to fluctuating market conditions. Our research has demonstrated the multiple benefits of our approach, including improved profitability and the ability to adapt to market changes. Recognizing that there is room for improvement, we could refine our strategy by employing advanced dimensionality reduction techniques, extending our analysis to high-dimensional biclustering and triclustering methods, and integrating a hybrid transformer-based volatility analysis model. These enhancements are intended to deepen the analytical capabilities of our algorithm and extend its application to a wider range of assets and markets, strengthening our trading strategy and making it more robust and comprehensive for future endeavors in the area of dynamic pair trading.

Acknowledgments and Disclosure of Funding

The authors would like to express their sincere gratitude to Professor David Ye for his invaluable guidance and support in the course Math 585: Algorithmic Trading. The insights and expertise shared by Prof. Ye were instrumental in shaping the research in this paper. The authors are also grateful for the participatory discussions and educational environment provided by the course, which greatly enhanced their understanding of the principles and practices of multiple algorithmic trading.

The authors would like to disclose that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Cox, J. C. (1996). The constant elasticity of variance option pricing model. *The Journal of Portfolio Management*, pp. 15-17.
2. Cummins, M., Bucca, A., (2012). Quantitative spread trading on crude oil and refined products markets. *Quantitative Finance 12*, 1857–1875.
3. Dang, W., Zhou, B., Zhang, W., & Hu, S. (2020, June). Time Series Anomaly Detection Based on Language Model. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (pp. 544-547).
4. Do, B., & Faff, R. (2010). Does simple pairs trading still work?. *Financial Analysts Journal, 66*(4), 83-95.
5. Gatev, E., Goetzmann, W.N., Rouwenhorst, K.G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies 19*, 797–827.
6. Golub, G. H., & VanLoan, C. F. (1980). An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, vol. 17, no. 6, pp. 883–893.
7. Han, C., He, Z., & Toh, A. J. W. (2023). Pairs Trading via Unsupervised Learning. *European Journal of Operational Research, 307*(2), pp. 929-947.

8. Kim, T., & Kim, H. Y. (2019). Optimizing the pairs-trading strategy using deep reinforcement learning with trading and stop-loss boundaries. *Complexity*, Hindawi, pp. 1-20.
9. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689-702.
10. Loretan, M., & English, W. B., (2000, June). Evaluating changes in correlations during periods of high market volatility. *Bank for International Settlements*.
11. Mandelbrot, B. (1967, October). The Variation of Some Other Speculative Prices. *The Journal of Business*, vol. 40, no. 4, pp. 393–413.
12. Narayan, Ashwin, Berger, Bonnie, & Cho, Hyunghoon. (2021). Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *Nature Biotechnology, 39*:765–774.
13. QuantConnect. (n.d.). Morningstar US Fundamental Data. Retrieved April 27, 2024, from `https://www.quantconnect.com/docs/v2/writing-algorithms/datasets/morningstar/us-fundamental-data`.
14. Sarmento, S. M., & Horta, N. (2020). Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158, 113490.
15. Van Der Maaten, L., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research, 10*(66-71), 13.
16. Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2625.
17. Ye, D., (2023) Lecture 8: Algorithmic Trading – Financial Data and Modeling. *Duke University*.
18. Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert systems with applications, 67*, 126-139.