Machine Learning for Pairs Trading: a Clustering-based Approach

Francesco Rotondi*[1] and Federico Russo[2]

[1]Department of Finance, Bocconi University, 20136 Milan, Italy
[2]Mathematical and Computing Sciences for Artificial Intelligence, Bocconi University, 20136 Milan, Italy

January 2, 2025

Abstract

In this paper we employ unsupervised learning techniques to identify potential stocks for pairs trading using a clustering algorithm based on three distinct metrics: the Euclidean distance, a PCA-based Euclidean distance and a partial correlation-based distance, the latter representing a novel application in this context. Restricting only to the pairs identified by the clustering algorithm, we implement a straightforward pairs trading strategy that delivers statistically and economically significant excess returns, both in absolute terms and on a risk-adjusted basis, even after accounting for transaction costs. Specifically, focusing on stocks that are or have been constituents of the S&P 500 during the period 2000–2023, we find average monthly excess returns ranging from 36 to 41 basis points, with Sharpe ratios between 0.20 and nearly 0.30 (equivalent to annualized Sharpe ratios of 0.72 to almost 1). The excess returns are uncorrelated with the market or any traditional risk factor. Among the metrics analyzed, the partial correlation-based distance achieves the highest risk-adjusted performance, likely attributable to its superior clustering accuracy, as evidenced by a purity index based on major industry sector classifications. Robustness checks and sensitivity analyses further corroborate these results.

Keywords: trading strategy, pairs trading, machine learning, clustering.

JEL Classification: G10, C38.

1 Introduction and Literature Review

Pairs trading strategies are a well-established example of statistical arbitrage investment opportunities, designed to exploit temporary price deviations between related securities and their

*Corresponding author: rotondi.francesco@unibocconi.it; ORCID: http://orcid.org/0000-0002-5102-6791.

long-term equilibrium (see, e.g., Vidyamurthy (2004) and Elliott et al. (2005) for a detailed overview of pairs trading). The underlying premise is that if the prices of two securities are expected to move in tandem, any temporary misalignment presents an opportunity: one can purchase the undervalued security while short selling the overvalued one, realizing a profit when the prices converge again. However, because the timing of this realignment cannot be determined with certainty, pairs trading is categorized as statistical arbitrage rather than pure arbitrage.

In this paper, we employ unsupervised learning techniques to select tradable pairs using a clustering algorithm based on three distinct metrics: the Euclidean distance, a PCA-based Euclidean distance, and partial correlation, which has been never taken into consideration as a screening tool for pairs. Utilizing only the pairs identified by the clustering algorithm, we implement a straightforward pairs trading strategy that generates statistically and economically significant excess returns in both absolute and risk-adjusted terms. Interestingly, among the metrics employed, the partial correlation-based distance yields the highest risk-adjusted performance, likely attributable to the superior clustering precision it provides, as measured by a purity index based on the major industry sector classifications.

To be effective, pairs trading strategies rely on two key factors: a high degree of similarity between the paired stocks and a strong long-term relationship between them, ensuring that price discrepancies are temporary and likely to correct quickly.

Since the seminal work of Gatev et al. (2006), which employed the Euclidean distance between prices to identify securities with closely aligned price movements, numerous alternative methods for selecting pairs to trade have been proposed in the literature (see Krauss (2017) for an effective literature review on pairs trading techniques). These methods leverage standard statistical and econometric techniques across a range of asset classes and markets. Focusing specifically on the equity market, for instance, Huck (2009) uses multi-criteria decision techniques to detected relatively under-/over-valued securities, Avellaneda and Lee (2010) study the similarities between different securities using principal component analysis, Bogomolov (2013) exploits technical analysis to select pairs of stocks that move together while Huck and Afawubo (2014) and Clegg and Krauss (2018) rely on the stationarity analysis of spreads between prices.

Additionally, Rad et al. (2016) implement pair selection strategies based on cointegration between prices and utilize copulas to estimate the probability of a stock in the pair moving higher or lower than its current price, adjusting trades accordingly. Similarly, Chen et al. (2019) select pairs based on the Pearson correlation of securities' returns.

While the profitability of simple pairs trading strategies appears to be diminishing (see Do and Faff (2010) and Do and Faff (2012)), the emergence of new asset classes and securities[1], as well as advances in quantitative techniques, hold promise for revitalizing these approaches. Indeed, as highlighted by Jacobs and Weber (2015), non-standard pairs trading techniques can still be valuable, particularly when they involve pairs that may not initially appear similar.

In the context of pairs trading, emerging machine learning techniques offer the potential to improve pair selection by identifying similarity patterns that traditional methods may overlook. Building on this idea, Krauss et al. (2017), Sarmento and Horta (2020) and Han et al. (2023) explore unsupervised learning techniques as a preliminary step, subsequently identifying suitable pairs within each cluster. These works pile up on the ones already using unsupervised learning techniques in quantitative finance for portfolio diversification and selection (Dose and Cincotti (2005), Nanda et al. (2010)) and financial time series analysis (Bini and Mathew (2016), Barucci et al. (2021)). Meanwhile, Huck (2010) uses neural network-based forecasting techniques to select pairs while Kim and Kim (2019), Chang et al. (2021) and Kim et al. (2022) propose supervised learning approaches to optimize trading decisions once the pairs have been determined.

Building on these recent advancements, our paper investigates the application of unsupervised learning-based clustering technique, for pair selection among stocks within the S&P 500 index. We specifically examine how different distance metrics, which underpin any clustering algorithm, influence the resulting pair selection and, ultimately, the returns generated by these strategies. Our analysis includes the widely-used benchmark Euclidean distance, first introduced in Gatev et al. (2006), a principal components-based distance considered by Sarmento and Horta

[1] For example, Crépellière et al. (2023) examine such arbitrage opportunities in the cryptocurrency market, while Fanelli et al. (2024) analyze pairs trading strategies in commodity markets, already considered by Cummins and Bucca (2012), following the introduction of a new crude oil futures contract.

(2020), and a novel metric based on partial correlation as described by Kenett et al. (2015). This alternative metric isolates the true correlation between two securities by filtering out spurious correlations caused by shared exposure to a third variable.

The clustering algorithm's output, in terms of the number of optimally formed clusters and the average number of stocks within each cluster, varies significantly depending on the metric employed. However, our analysis reveals that the primary driver of clustering is the stocks' industry sector. This observation is further validated through an analysis of cluster purity with respect to the SIC classification, which supports our hypothesis.

Our pairs trading designed, backtested over the period 2000–2023, generates excess returns that are both statistically and economically significant. Specifically, the average monthly excess returns, net of transaction costs, range between 36 and 41 basis points, aligning with the findings of Do and Faff (2012) and Rad et al. (2016). In risk-adjusted terms, the strategy that delivers the highest monthly Sharpe ratio (0.29, equivalent to an annualized Sharpe ratio of 1.01) is based on partial correlation, which is identified as the most accurate distance metric, as discussed earlier. Measures of downside performance, introduced to address concerns about the appropriateness of the Sharpe ratio for non-normal returns, corroborate both the magnitude and relative performance of the strategies. From an economic perspective, strategies based on partial correlation produce statistically significant alphas when tested against standard risk factors, such as those proposed by Fama and French (1993) and Fama and French (2015), even after incorporating the momentum factor of Carhart (1997) and the liquidity factor of Pástor and Stambaugh (2003).

The remainder of the paper is structured as follows. Section 2 provides a detailed description of the dataset utilized in this study. Section 3 outlines the key components of our trading framework, including the formation period, during which pairs are selected based on the clustering algorithm's output, and the trading period, where the investment strategies are implemented and evaluated. Section 4 presents the empirical findings of our analysis, showcasing various performance metrics in both absolute and risk-adjusted terms. Finally, Section 5 offers concluding remarks.

2 Data

The dataset employed in this study covers the period from January 2000, to December 2023, comprising a total of 6'039 trading days and 288 months. Within this timeframe, the analysis is centered on the common shares of companies listed in the S&P 500 index on each trading day. In total, the study examines 1'098 stocks that were included in the index for at least one month at any point during the specified period. For these stocks, daily bid-ask spread prices are retrieved from the Center for Research in Security Prices (CRSP) database. Furthermore, additional data are collected, including adjustment factors, share codes, sector-level Standard Industrial Classification (SIC) codes, exchange codes, and delisting codes.

Our research design focuses exclusively on stocks classified as common shares, identified by share codes 10 or 11. To account for stock splits and ensure accurate computation of returns, we apply adjustment factors. Regarding industry classification, we simplify the detailed Standard Industrial Classification (SIC) codes provided by CRSP by grouping them into broader categories based on the first two digits of the SIC code, which represent the major industry group[2].

Importantly, our analysis includes stocks that are no longer part of the index as of today, while also tracking those that enter and exit the index during the study period. This approach mitigates survivorship bias and establishes a robust trading design that remains applicable for future analyses involving stocks within the index.

For computing partial correlations (to be introduced in Subsection 3.1.1) and for benchmarking purposes we also retrieve daily levels of the S&P 500 index over the same timeframe.

3 Trading design

In this section, we outline in detail the steps involved in our trading strategies. As highlighted in the introduction, the primary objective of pairs trading is to capitalize on temporary misalignments in the prices of securities that typically exhibit correlated movements. Consequently, the design of such strategies involves two key steps: pair selection and formation over a so-called

[2] Specifically, the classification ranges from codes with first two digits less than or equal to 09 to those between 90 and 99, resulting in ten major groups: agriculture, forestry, and fishing; mining; construction; manufacturing; transportation and public utilities; wholesale trade; retail trade; finance, insurance, and real estate; services; and public administration.

formation period, which is discussed in Subsection 3.1, and pairs trading rules, over a so-called *trading period*, which are addressed in Subsection 3.2. Lastly, Subsection 3.3 focuses on the methods used to evaluate the performance of our trades, as well as the measures adopted to assess their associated risks.

3.1 Pairs selection

The profitability of pairs trading is intrinsically linked to the degree of comovement between the prices of selected pairs. Insufficient comovement can result in a trader opening a position during a price divergence but being unable to close it if the prices fail to converge again. To reduce the likelihood of such outcomes, great importance is posed on the pair selection phase.

As outlined in the introduction, the earliest formal approach to systematic pairs selection is likely attributable to Gatev et al. (2006). In their work, the Authors compute the pairwise Euclidean distance between all stocks' prices within a given universe and implement trading strategies on the pairs belonging to the first decile of the distance distribution. This methodology, i.e. an initial screening of pairs based on a specific distance metric (most commonly the Euclidean distance), forms the foundation for nearly all subsequent studies cited in the introduction. Many of these studies, however, incorporate additional tests, such as cointegration testing or the evaluation of mean-reversion in relative mispricing, to refine the selection process.

In our trading framework, we adopt an unsupervised machine learning approach to pairs selection, specifically employing a clustering algorithm. Clustering algorithms group data points into clusters based on similarity, with similarity determined by a distance metric, such as the traditional Euclidean distance, which serves as the algorithm's starting point.

In the following, we first consider all stocks that are continuously included in the index during a specified formation period and consider three different distance metrics (Subsection 3.1.1). Then, using this matrix, we apply unsupervised learning to cluster the stocks based on their similarity as measured by these three different metrics (Subsection 3.1.2).

3.1.1 Distance metrics

Efficient unsupervised clustering techniques require the definition of a distance matrix as an essential starting point. Distance matrices, which have been extensively used to analyze relationships among securities in financial markets since the seminal works of Mantegna (1999) and Onnela et al. (2003), quantify pairwise similarities between elements in a dataset. This approach provides a structured and interpretable framework for capturing relationships among stocks.

One of the primary objectives of this study is to assess the impact of different distance metrics on the output of a clustering technique for pairs selection and, ultimately, on the profitability of the resulting strategies. To this end, we employ three distinct distance metrics to construct the distance matrix. The first is the traditional Euclidean distance, defined as the sum of squared deviations between normalized prices (*SSD distance*), as utilized by Gatev et al. (2006). The second is the Euclidean distance calculated on the principal components (*PCA distance*) of securities' returns, as proposed by Sarmento and Horta (2020). Lastly, we use a metric based on partial correlation (*PC distance*), a simple yet effective similarity measure described by Kenett et al. (2015), which isolates the "pure" correlation between two prices by filtering out spurious correlations arising from their shared exposure to a third factor[3].

Let $\{S_t^X\}_{t\in\tau}$ and $\{S_t^Y\}_{t\in\tau}$ be the series of mid prices of securities X and Y that have belonged to the index throughout the entire formation period τ under consideration and let $\{r_t^X\}_{t\in\tau}$ and $\{r_t^Y\}_{t\in\tau}$ be the associated linear returns.

The SSD distance, $SSD_{X,Y}$, between securities X and Y is defined as in Gatev et al. (2006) and reads

$$SSD_{X,Y} = \sum_{t\in\tau} \left(\frac{S_t^X - \bar{S_t^X}}{\hat{\sigma}[S_t^X]} - \frac{S_t^Y - \bar{S_t^Y}}{\hat{\sigma}[S_t^Y]} \right)^2, \tag{1}$$

where $\bar{\cdot}$ and $\hat{\sigma}[\cdot]$ denote, respectively, the sample mean and sample standard deviation of a given series computed over τ.

The PCA distance, $PCA_{X,Y}$, as discussed in Sarmento and Horta (2020), is defined as the

[3] Note that while the SSD distance and the PCA distance are proper metrics, the PC distance is not, as it does not necessarily satisfy the triangle inequality. However, this does not pose a problem for the subsequent clustering step.

Euclidean distance between the synthetic returns series constructed using the first five[4] principal components derived from the variance-covariance matrix of normalized returns. Therefore, the PCA distance is computed as in (1), where the price series S^X and S^Y are replaced by the series of the synthetic returns.

The PC distance, $PC_{X,Y}$, is based on the concept of partial correlation. Partial correlation quantifies the extent to which a variable influences the correlation between a pair of other variables, isolating the direct relationship from confounding effects. This measure is particularly valuable in situations where a strong Pearson correlation between two variables might arise not from a direct relationship but from shared exposure to a third variable. In the context of pairs trading, partial correlation can be employed to assess the influence of the stock market index on the correlation between a pair of stocks. We define here the partial correlation between the returns r_X and r_Y of securities X and Y, controlling for the returns r_M of the market index (i.e. of the S&P500) as

$$\rho_{par}[r_X, r_Y | r_M] = \frac{\rho[r_X, r_Y] - \rho[r_X, r_M]\rho[r_Y, r_M]}{\sqrt{(1-\rho^2[r_X, r_M])\,(1-\rho^2[r_Y, r_M])}},$$

where $\rho[\cdot,\cdot]$ denotes the Pearson correlation coefficient. Because partial correlation values range between -1 and 1, after computing its sample estimator $\hat{\rho}_{par}[r_X, r_Y | r_M]$ over the formation period τ, the PC distance is derived by applying the following transformation to rescale sample partial correlation:

$$PC_{X,Y} = 1 - |\hat{\rho}_{par}[r_X, r_Y | r_M]|\,.$$

This transformation ensures that when two stocks exhibit perfect partial correlation (either positive or negative), their distance is zero.

The left panels of Figure 1 display heatmaps of the three distance matrices derived from the previously described distance metrics, calculated as of the final trading day in the dataset over a three-year formation period. This formation period spans 753 trading days, during which 441 stocks consistently remained constituents of the index. To enhance clarity and readability, the

[4] As discussed in Section 6 of Sarmento and Horta (2020), the resulting PCA distance does not depend significantly on the precise number of principal components used. Therefore, we stick to their approach and set this number equal to five.

heatmaps are restricted to the first 100 stocks, starting from those with the earliest inclusion in the index. As observed, despite their differing scales, the three distance metrics exhibit noticeably distinct patterns. In particular, the SSD distance heatmap shows relatively high and evenly distributed dissimilarities, as indicated by the broader color range (spanning from low to high values) and its sensitivity to absolute price differences. The PCA distance heatmap reflects lower dissimilarities overall, likely due to the dimensional reduction effect of principal components, which emphasize shared variance across stocks. In contrast, the PC distance heatmap exhibits a distinct scaling, with values concentrated in a narrower range, reflecting its focus on filtering spurious correlations and isolating the “pure” relationships between stock pairs. This variation in scaling and dissimilarity levels underscores the differing nature of the metrics in capturing relationships within the data.

3.1.2 Clustering

After the computation of any distance matrix over the formation period, we perform our pairs selection applying a clustering algorithm. Clustering algorithms, a subset of unsupervised machine learning methods[5], group data points (in our case the stocks within our formation period) into clusters such that points within a cluster are more similar to each other than to those in other clusters.

Among the diverse range of clustering algorithms available, two of the most widely used are the k-means algorithm and DBSCAN. The k-means algorithm, initially formalized by MacQueen (1967), partitions data into a predetermined number of clusters, where each point is assigned to the cluster with the nearest mean (often referred to as the cluster center or centroid), which serves as a representing point for the cluster. A notable limitation of this approach is that it requires the user to specify the number of clusters a priori, which can be challenging when the optimal number of clusters is unknown.

In contrast, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, introduced by Ester et al. (1996), identifies clusters based on regions of high density, grouping points that are closely packed while designating sparsely distributed points as noise.

[5] See Chapter 2 of Hull (2019) for a simple yet effective introduction to unsupervised learning algorithms.

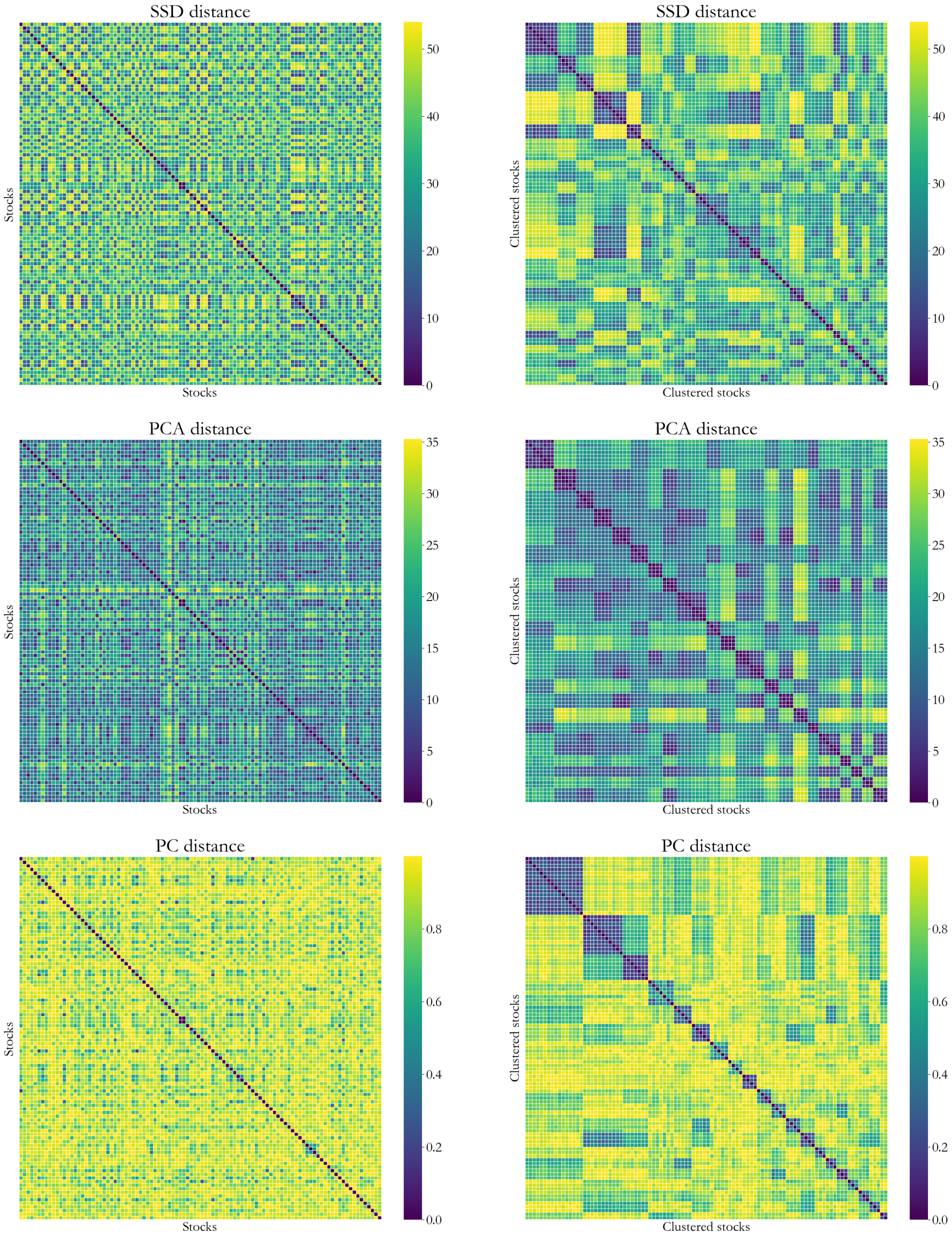

Figure 1: Distance matrices before (left panels) and after (right panels) the clustering, computed on the last trading day of December 2023, over the previous three years of data.

Unlike k-means, DBSCAN does not require the number of clusters to be specified beforehand. Instead, it relies on parameters such as the maximum distance between points to qualify as neighbours and the minimum number of points needed to form a cluster.

Building on DBSCAN, OPTICS (Ordering Points To Identify the Clustering Structure), introduced by Ankerst et al. (1999), is a more advanced clustering algorithm that excels at detecting clusters of varying densities, which is a limitation of DBSCAN. This capability makes OPTICS particularly advantageous for identifying robust stock groupings in financial datasets.

In this work, we use OPTICS to cluster stocks, setting to two the minimum cardinality of each cluster (to encourage a large number of possibly small clusters) and feeding the algorithm with the distance matrices detailed in the previous subsection.

The right panels of Figure 1 present the same distance matrices as in the left panels, but with rows and columns rearranged according to the optimally formed clusters, starting from the largest ones. This reordering highlights intra-cluster similarities, as indicated by the dense blocks of low-distance values along the diagonal, while inter-cluster dissimilarities appear as higher-distance values outside these blocks. This pattern is particularly visible when the PC distance is considered. Among the 441 stocks included in each distance matrices, the clustering algorithm optimally identifies 48 clusters when using the SSD distance, 109 clusters with the PCA distance, and 78 clusters with the PC distance. The sizes of these optimally formed clusters vary significantly. Figure 2 illustrates the distribution of their cardinalities. Notably, aside from a small number of large clusters primarily derived from the PC distance (visible in the upper left corner of the clustered distance matrix in Figure 1), most clusters contain only two or three securities. This substantially reduces the number of pairs to be considered for investment strategies, providing a promising filtering mechanism.

Consistent with this observation, the OPTICS algorithm identifies a significant number of outliers (stocks left unclustered): 314 stocks when starting from the SSD distance, 151 stocks with the PCA distance, and 210 stocks with the PC distance. Figure 3 provides a network representation of the optimally formed clusters derived using the SSD distance, chosen for its relatively small number of clusters, which enhances readability. In this representation, points correspond to stocks and are color-coded according to their major industry group. The clustering

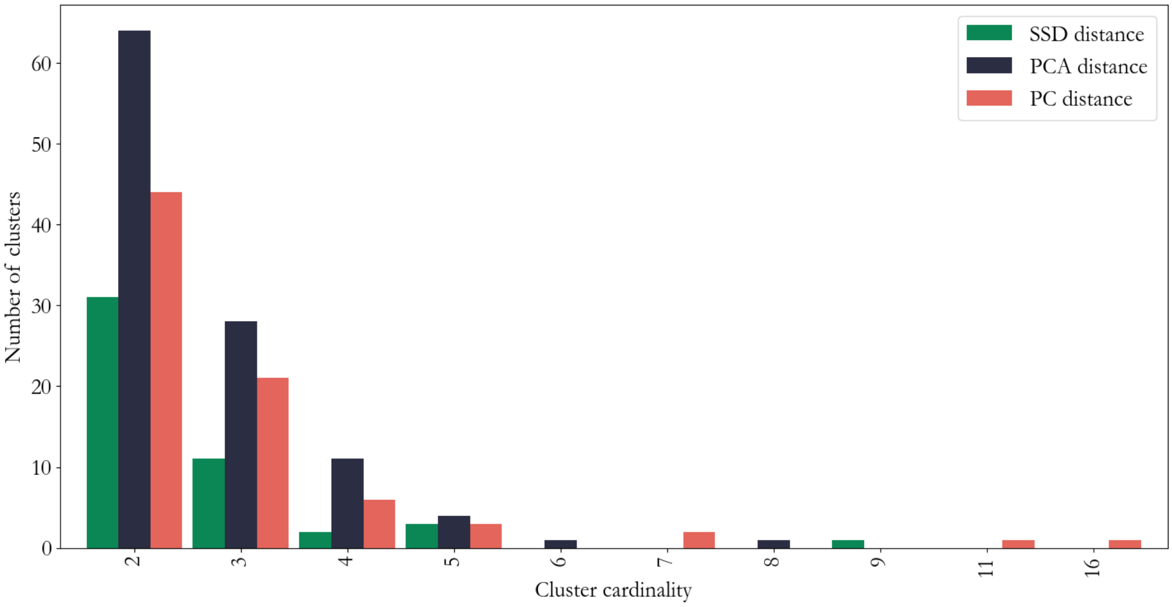

Figure 2: Distribution of the cardinality of the optimally formed clusters as of the last trading day of December 2023.

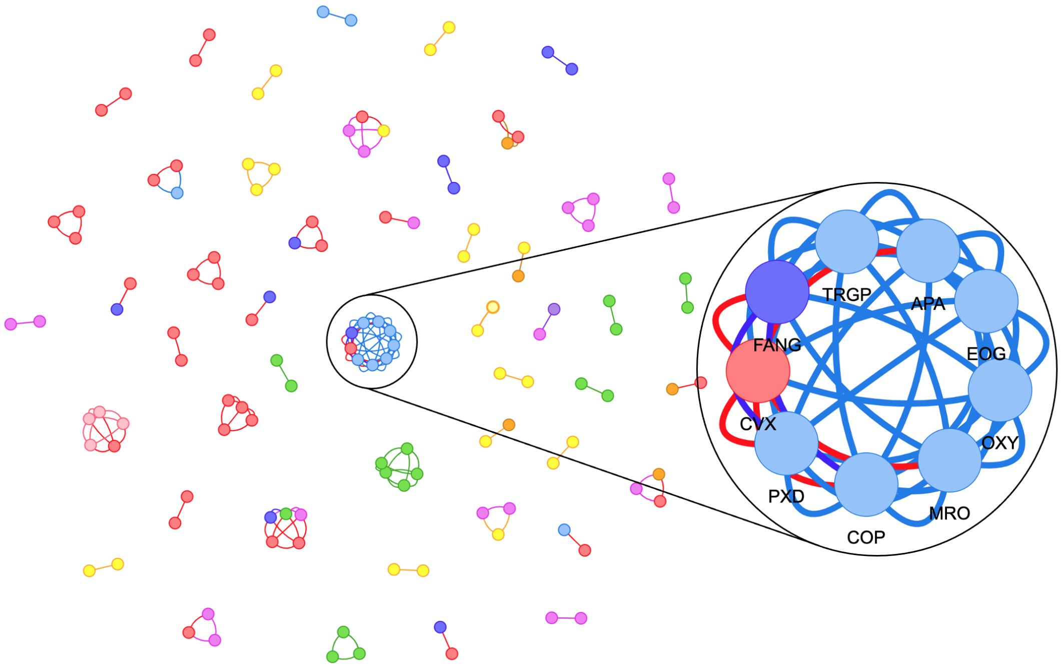

Figure 3: Network representation of the optimally formed clusters according to the SSD distance as of the last trading day of December 2023.

is observed to be predominantly influenced by industry group affiliation. For example, the largest cluster, which, as shown in Figure 2, comprises nine stocks, includes seven companies (depicted in light blue) from the mining industry group[6]. This group encompasses industries involved in metal and coal mining, the extraction of oil and gas, and the quarrying of nonmetallic minerals. Additionally, the cluster includes one company (Chevron Corporation, CVX, shown in red) from the retail industry group, specifically the "Automotive Dealers and Gasoline Service Stations" subgroup, and one company (Diamondback Energy, FANG, shown in purple) from the manufacturing industry group, within the "Petroleum Refining and Related Industries" subgroup. Thus, even when stocks from different industry groups are grouped together, the underlying businesses of the clustered companies are closely related.

To formalize the observation that industry group affiliation is a primary factor driving the clustering, we compute the purity index of the optimally formed clusters with respect to the major industry group of each stock. The purity index is defined as:

$$\text{Purity Index} = \frac{1}{N}\sum_{i=1}^{K}\max_{j}|C_i \cap T_j|, \tag{2}$$

where N is the number of overall stocks within the formation period, K is the number of optimally formed clusters, C_i is the set of stocks in cluster i and T_j is the set of stocks in sector j. A purity value close to one means that clusters tend to be formed by stocks in the same sector.

As of the last trading day of December 2023, the purity index of the optimally formed clusters was calculated to be 0.81 for clustering based on the SSD distance, 0.77 for the PCA distance, and 0.94 for the PC distance. Although the purity index is high across all distance metrics, the partial correlation metric demonstrates a notable advantage. Its higher purity value suggests it is particularly effective in filtering out spurious correlations and clustering stocks that exhibit genuine similarity.

[6] These companies are: Targa Resources Corp. (TRGP), APA Corporation (APA), EOG Resources Inc. (EOG), Occidental Petroleum Corporation (OXY), Marathon Oil Corporation (MRO), ConocoPhillips (COP) and Pioneer Natural Resources Company (PXD).

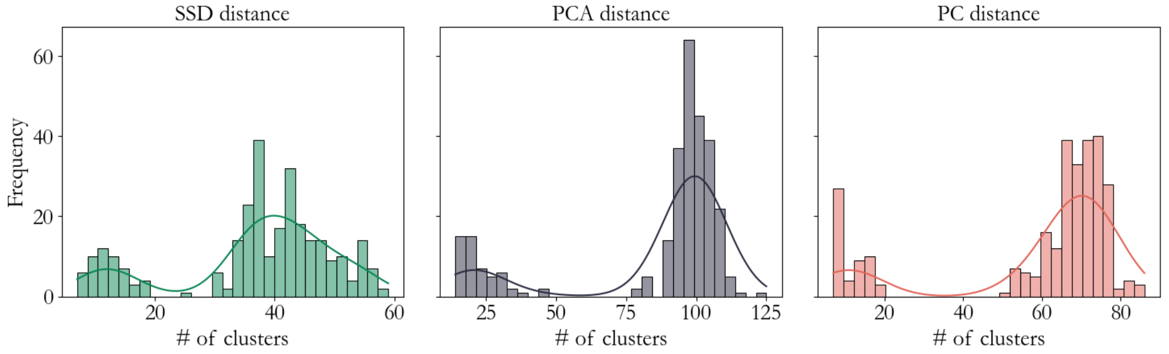

Figure 4: Distribution of the monthly number of optimally formed clusters over a three-year rolling formation period over the entire timeframe.

We extend the previous analysis by considering a rolling three-year formation period across the entire sample. Figure 4 presents the empirical distributions of the number of optimally formed clusters based on the three distance metrics on the first trading day of each month within our timeframe. These distributions exhibit a bimodal pattern, a phenomenon attributable to the earlier years of the sample, particularly the early 2000s, when frequent changes in the index constituents resulted in a relatively smaller pool of stocks available for clustering. Consistent with the findings for the last trading day of December 2023, the PCA distance consistently produces the largest number of clusters, followed by the PC distance, with the SSD distance yielding the fewest clusters.

Figure 5 illustrates the purity index of the optimally formed clusters for each distance metric over time. As observed, the clusters derived using the PC distance consistently achieve the highest purity index, which remains close to one. This result is particularly promising in the context of the trading strategies discussed in the subsequent section.

Before introducing the trading strategies implemented with the clustered stocks, it is important to address the possibility of further reducing the number of pairs considered for investment. Although, as illustrated in Figure 2, most of the optimally formed clusters are relatively small, and the majority of available stocks are classified as outliers and left unclustered, the resulting number of pairs restricted to those formed within the same cluster can still be substantial. This

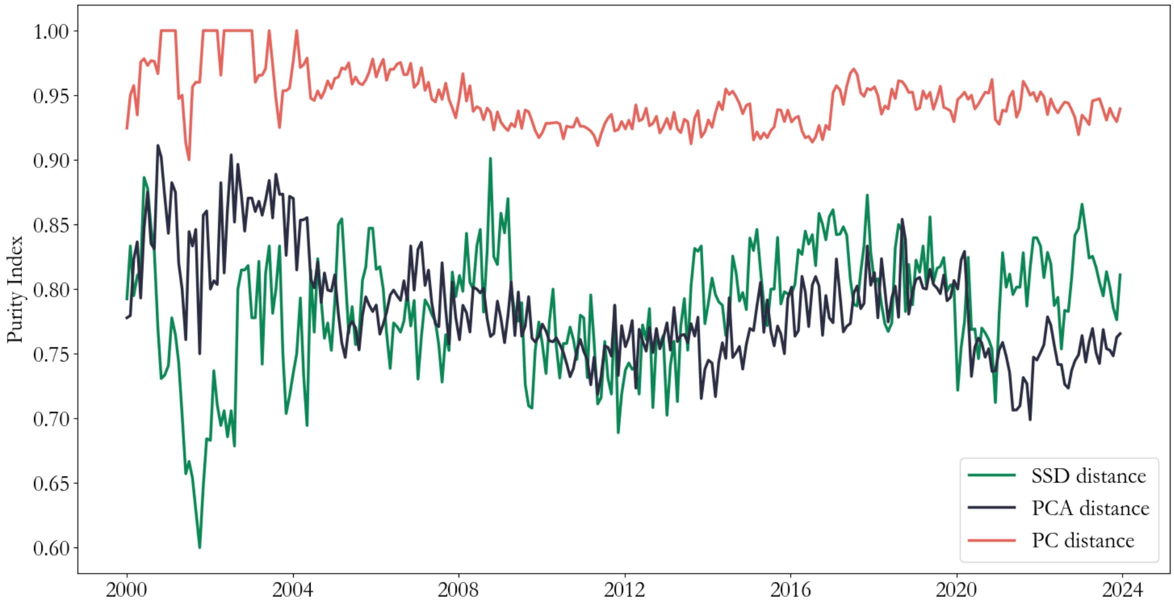

Figure 5: Purity index computed by (2) out of the optimally formed clusters over a three-year rolling formation period over the entire timeframe.

is primarily due to the presence of a few outlier clusters containing a large number of stocks, which could pose significant challenges for investors attempting to monitor the behavior of all tradable pairs.

To mitigate this concern, previous studies (e.g., Sarmento and Horta (2020)) have proposed additional filters to narrow the set of tradable pairs further. These filters include testing for cointegration between stock prices in a pair or assessing the mean-reverting behavior of the relative mispricing, often quantified by the Hurst exponent. However, as the focus of our analysis is on evaluating the impact of the distance metric used in the clustering algorithm, our benchmark analysis does not incorporate these additional controls.

For completeness, we conduct a side analysis in which we re-run the entire backtest while testing pairs for cointegration, employing the univariate approach of Engle and Granger. As we will discuss in subsequent sections, the number of tradable pairs decreases significantly, as many securities' prices are not cointegrated. Nevertheless, the overall performance of the trading strategy remains qualitatively unchanged.

3.2 Trading rule

Assume that stocks X and Y belong to the same cluster, making them eligible for pairs trading.

Following the seminal methodology of Gatev et al. (2006), we generate trading signals when the so-called *spread* between X and Y, in normalized terms, gets statistically different from zero. The spread, traditionally denoted by $\{Z_t\}_{t\in\tau}$, is defined[7] as the OLS residuals of the regression:

$$S_t^Y = \gamma S_t^X + \varepsilon. \tag{3}$$

This regression is estimated within the formation period[8]. Accordingly, we define the spread as:

$$Z_t = S_t^Y - \hat{\gamma}_{OLS} S_t^X \tag{4}$$

for any $t \in \tau$. This serves as the starting point for standard univariate cointegration tests. Following the Engle and Granger procedure, the next step would involve testing the stationarity of the series Z. However, as previously mentioned, we do not include this additional control in the main analysis, as our focus is solely on the non-parametric pairs selection derived from the clustering algorithm. We investigate the impact of this cointegration check in Subsection 4.2.

For each day t during the trading period, we compute the realized spread Z_t using (4) and normalize it with the sample mean[9] and standard deviation calculated over a subsample[10] of the formation period τ. Specifically, we consider the so-called *z-score*, denoted as $\tilde{Z}_t$, defined as:

$$\tilde{Z}_t = \frac{Z_t - \bar{Z}t}{\hat{\sigma}[Z_t]} \tag{5}$$

[7] There is no strict consensus in the pairs trading literature regarding whether the spread should be defined using a regression based on prices or log prices. For our benchmark analysis, empirical tests indicate that the final results are almost identical. Therefore, we adopt the regression based on prices to simplify the interpretation of position sizing in (6).

[8] We acknowledge that the residuals of the regression of Y on X differ from those of the regression of X on Y. If an additional check for cointegration between X and Y is included and the residuals are tested for stationarity, the conventional approach suggests selecting the configuration with greater statistical evidence of stationarity. Here, we select X and Y almost arbitrarily, based on their addition date to the index. However, both configurations were tested, and the numerical results were nearly identical.

[9] Note that the regression used to generate the residuals in (4) does not include an intercept term. Consequently, the sample mean of the residuals is generally non-zero.

[10] In our benchmark analysis, we define this subsample as the preceding six months. In Subsection 4.3, we conduct a sensitivity analysis concerning the length of this subsample, which reveals that its impact is relatively minor.

where $\bar{\cdot}$ and $\hat{\sigma}[\cdot]$ represent the sample mean and sample standard deviation, respectively, computed over a subsample of τ. As is common in many pairs trading algorithms, the regression coefficient estimate $\hat{\gamma}_{OLS}$ and the sample moments used to compute the z-score are determined at the start of the trading period and remain fixed until its conclusion.

Once the z-score is calculated, a position is initiated when the z-score deviates significantly from zero, indicating it is “statistically different” from zero. This occurs when the z-score crosses predetermined thresholds, denoted as $\pm z$. Common thresholds are based on assumptions of standard normality: $z = \pm 2$ (or ± 1 and ± 3) correspond to statistical significance at the conventional 95% confidence level (or 70% and 99% confidence levels, respectively). Specifically, when the z-score drops below $-z$, it suggests that Y is becoming relatively undervalued compared to X, prompting a long position in Y and a short position in X. Conversely, if the z-score exceeds z, it signals that Y is relatively overvalued compared to X, leading to a short position in Y and a long position in X. Afterwards, a position is closed either if the normalized spread reverts and crosses zero or if one of the stocks in the pair is delisted. It is important to note that if the position is closed due to the spread reverting to zero, the trade results in a gain. However, in case of delisting, the closure is suboptimal, and the trade may result in a loss.

The position sizing is determined to align with the spread in (4), ensuring that the gross exposure of the two positions remains fixed at one USD. Denoting by N_j the number of shares of stock $j \in X, Y$ that are either bought or sold short, we impose:

$$N_Y = \frac{1}{\hat{\gamma}_{OLS} S^X + S^Y} \text{ and } N_X = \hat{\gamma}_{OLS} N_Y, \tag{6}$$

where $\hat{\gamma}_{OLS}$ was the OLS estimate of the original regression in (3).

The strategy outlined above does not impose an upper limit on the number of positions that may remain open simultaneously. In principle, all pairs identified during a given formation period, as described in Subsection 3.1, could result in open positions. Figure 6 illustrates the distribution of the monthly number of pairs selected by the clustering algorithm. On average, the SSD distance identifies 89.31 pairs, the PCA distance selects 227.86 pairs, and the PC distance yields 195.06 pairs. While these figures may appear substantial, it is important to note that the total number of potential pairs, assuming 400 stocks are included in the formation period, is approximately 80,000. This highlights the significant reduction in the number of potentially tradable pairs achieved through our trading design.

Despite the relatively large number of pairs to monitor, the number of pairs actually traded remains relatively low. Figure 7 illustrates the distribution of daily traded pairs, considering a rolling three-year formation period and a one-month trading period over the entire dataset. On average, adopting the SSD distance for clustering results in 22.41 pairs being traded. This number increases to 57.86 when the PCA distance is used, and it is 43.93 for the PC distance. As a reference, for each trading period, both Gatev et al. (2006) and Rad et al. (2016) prescribes monitoring exactly twenty pairs, of which, on average, half are actually traded. If a trading desk considers this number of pairs excessive, we propose two approaches to reduce the number of actively traded pairs, as discussed in Subsections 4.2 and 4.3. The first approach involves narrowing the selection to pairs formed by cointegrated stocks only, while the second entails increasing the z-score threshold required to trigger the opening of a position. Among these methods, the first tends to reduce overall returns, whereas the second appears to enhance them.

Finally, we incorporate transaction costs into our analysis, accounting for them both when a trade is opened and when it is closed. To this end, we use the bid and ask prices provided by CRSP for all stocks. We disregard proportional transaction costs, which, as noted by Do and Faff (2012), have been substantially decreasing over time for large investors (such as a hypothetical hedge fund that might implement our trading design).

3.3 Performance evaluation

Since the size of pairs trading strategies varies over time depending on the number of pairs/stocks traded each day, calculating returns is non-trivial. To address this issue, we first compute the mark-to-market profit and loss (PnL, henceforth) generated by each selected pair throughout the trading period. At the conclusion of the period, we aggregate the PnLs from all selected pairs and scale the total by the gross capital actually deployed during the trading period. This approach ensures a consistent and accurate representation of returns relative to the capital employed.

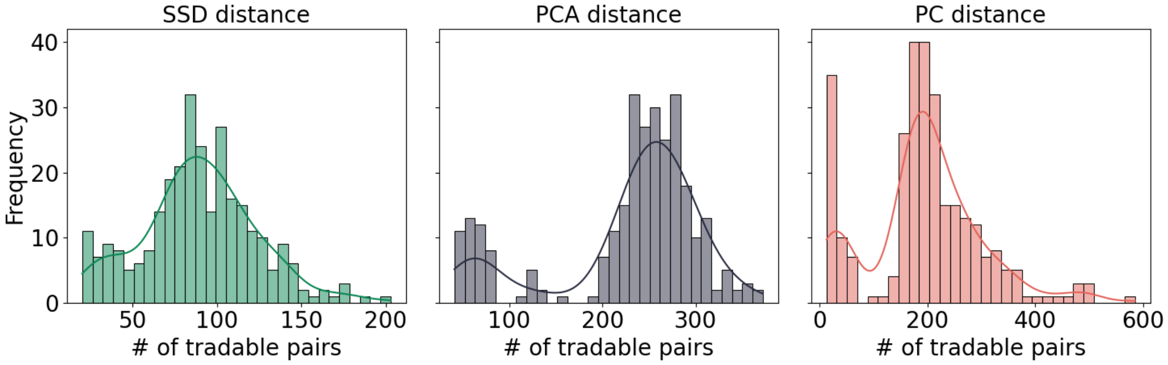

Figure 6: Distribution of the monthly number of tradable pairs, over a rolling three-year formation period and a one-month trading period.

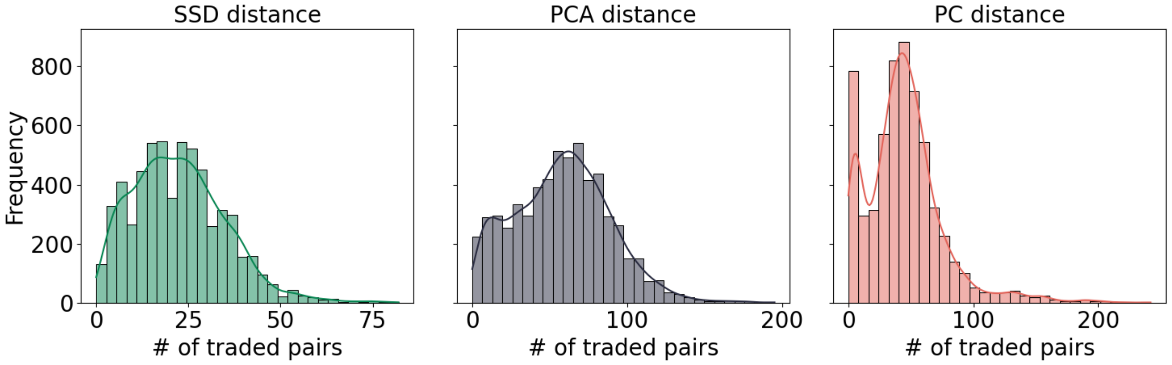

Figure 7: Distribution of the number of actually traded pairs, on a daily basis, over a rolling three-year formation period and a one-month trading period.

Focusing on a single pair involving stocks X and Y, the daily Profit and Loss (PnL) on day t, provided the position is open, can be expressed as:

$$PnL_t^{X,Y} = \begin{cases} (N_Y S_t^Y - N_X S_t^X) - (N_Y S_{t-1}^Y - N_X S_{t-1}^X) & \text{if } Y \text{ is bought and } X \text{ is sold} \\ (N_X S_t^X - N_Y S_t^Y) - (N_X S_{t-1}^X - N_Y S_{t-1}^Y) & \text{if } X \text{ is bought and } Y \text{ is sold.} \end{cases} \tag{7}$$

Since the sizing of the pair is one USD, the return of a single pair is equal to its PnL. If, on the contrary, there is no open position at day t, the PnL for the selected couple is equal to zero. Notice that long and short positions in (7) are marked-to-market as actual bid/ask prices are taken into consideration. Daily PnLs from couple X-Y are then aggregated into monthly PnLs.

To evaluate the overall performance of our pairs trading strategy, we aggregate the returns from all individual pairs. Gatev et al. (2006) propose two approaches for measuring excess returns: excess return on committed capital and fully-invested excess return. The excess return on committed capital scales PnLs by the (constant) number of pairs selected for trading at the beginning of each trading period, while the fully-invested excess return scales PnLs by the number of pairs actively traded during the trading period on a day-by-day basis.

In this work, we adopt a similar methodology by scaling the sum of the monthly PnLs from all actively traded pairs, collected in a set denoted by $\mathcal{TD}_m$ for month m, by the maximum gross exposure observed over the month. This exposure corresponds to 1 USD times the number of pairs that have been traded at least once during the month, represented by the cardinality of the set $\mathcal{TC}$. We argue that this approach provides a more accurate depiction of the capital effectively employed. Formally, our benchmark monthly excess return[11] r_m is defined as:

$$r_m = \frac{\sum_{(X,Y)\in\mathcal{TC}_m}\left(\sum_{t\in m} PnL_t^{X,Y}\right)}{|\mathcal{TC}_m|}. \tag{8}$$

[11] In line with the pairs trading literature (e.g., Gatev et al. (2006), Rad et al. (2016), and Sarmento and Horta (2020)), we define excess returns as the returns on invested capital without subtracting the monthly risk-free interest rate from the accrued monthly PnLs. This definition reflects the implicit assumption in the trading design that when a position in a pair is not open, the capital allocated to that pair remains uninvested and does not earn the risk-free rate. Furthermore, pairs trading strategies are typically implemented in a dollar-neutral manner, which eliminates the opportunity cost of capital and thus obviates the need to adjust for the risk-free rate.

Here, the slight abuse of notation $t \in m$ indicates the summation over all days t belonging to the same month m.

4 Results

In this section, we evaluate the performance of the pairs trading framework introduced in the previous section, considering the three distance metrics used as inputs for clustering separately. Specifically, in Subsection 4.1, we provide a comprehensive analysis of the return series, examining their statistical and financial characteristics, including descriptive statistics, distribution, absolute and risk-adjusted performance, and economic significance. In Subsection 4.2, we present a couple of supplementary analyses, while in Subsection 4.3, we explore the sensitivity of strategy performance to key parameters of the trading design.

4.1 Benchmark results

In our benchmark analysis, we employ a rolling three-year formation period combined with a one-month trading period. Within our timeframe, this approach yields a series of 288 monthly excess returns, calculated as described in (8). Additionally, we standardize the spread in (4) using sample moments estimated on the previous six months and we adopt the standard opening threshold for the strategy, $z = \pm 2$, which corresponds to the spread between the prices of the pairs being statistically distinct from its mean at the 95% confidence level.

In the subsequent analysis, we first examine the empirical characteristics of the three series of monthly excess returns generated by the different distance metrics used as inputs for the clustering algorithm (Subsection 4.1.1). Next, we assess the economic significance of these excess returns produced by our trading strategy by benchmarking them against a large set of effective risk factors (Subsection 4.1.2).

4.1.1 Descriptive statistics and performances

Table 1 summarizes the key descriptive statistics for the series of monthly excess returns, computed as detailed in Subsection 3.3, for the three distance metrics used in the clustering al-

	Mean	T-ratio	St. Dev.	Min.	Median	Max.	Pos.	Neg.
SSD dist.	0.0041	4.27***	0.0161	-0.0727	0.0023	0.0922	62.6%	37.8%
PCA dist.	0.0036	3.48***	0.0175	-0.0693	0.0019	0.0828	58.7%	41.3%
PC dist.	0.0038	4.90***	0.0132	-0.0649	0.0026	0.0797	70.1%	29.9%

	Skewness	Kurtosis	$\mathrm{VaR}_{95\%}$	$\mathrm{CVaR}_{95\%}$	Sharpe ratio	Ann. Sharpe ratio
SSD dist.	0.88	6.67	-0.0141	-0.0289	0.2526	0.88
PCA dist.	0.26	3.67	-0.0164	-0.0370	0.2085	0.72
PC dist.	0.25	8.48	-0.0111	-0.0263	0.2904	1.01

Table 1: Summary statistics of the monthly excess returns series within the benchmark analysis, computed as described in Subsection 3.3 for the three distance metrics, each featuring 288 observations. ***/**/* denote significance at 1%/5%/10% level.

gorithm. The average monthly returns across the three strategies range from 36 basis points (bps) for the PCA distance to 41 bps for the SSD distance. These values align with the average monthly excess returns reported by Rad et al. (2016) and Do and Faff (2010). When compounded to annualized figures, these returns translate to a range of 4.40% to 5.03%, which is economically significant. Furthermore, the average returns are statistically significant at the 1% confidence level. Notably, the proportion of positive monthly excess returns exceeds 50% across all strategies, peaking at 70.1% for the PC distance.

While the SSD distance outperforms the PC distance in terms of average excess return, it exhibits higher volatility. Consequently, the PC distance achieves the highest risk-adjusted performance, with a monthly Sharpe ratio of nearly 0.3, corresponding to an annualized Sharpe ratio of approximately 1. Figure 8 presents the five-year rolling monthly Sharpe ratios for the investment strategies based on the three different distance metrics. Consistent with the results reported in Table 1, the strategy utilizing pairs selected according to the PC distance demonstrates superior performance across the majority of the periods.

Figure 9 illustrates the empirical distribution of the three return series alongside their kernel density estimates. Consistent with the skewness and kurtosis metrics reported in Table 1, the returns are positively skewed and leptokurtic, suggesting appealing upside potential for investors. However, formal tests reject the hypothesis of normality.

As noted by Eling and Schuhmacher (2007) and Eling (2008), the non-normality of returns

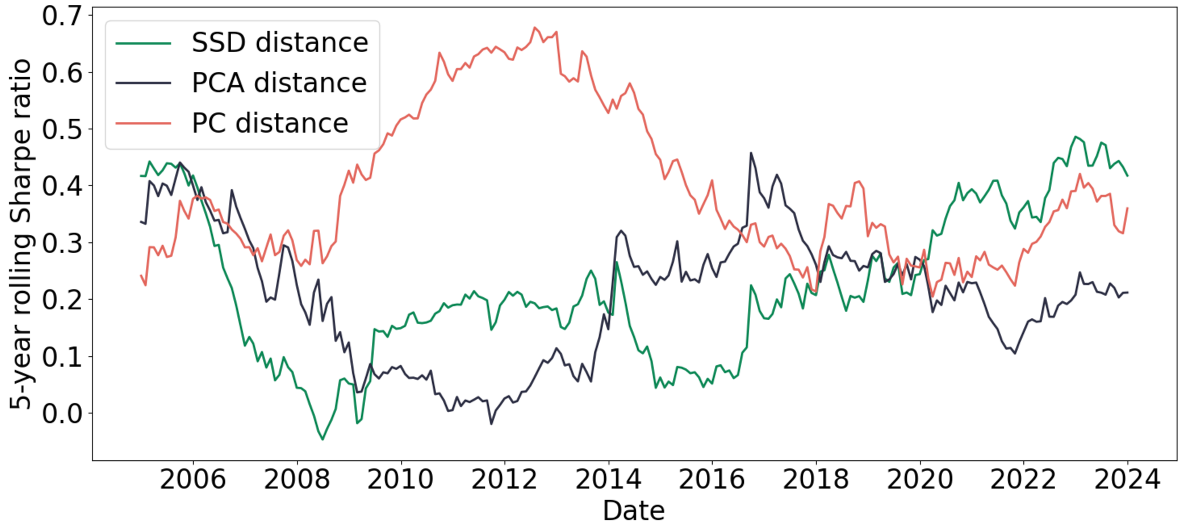

Figure 8: Five-year rolling monthly Sharpe ratio of the three different strategies within the benchmark analysis.

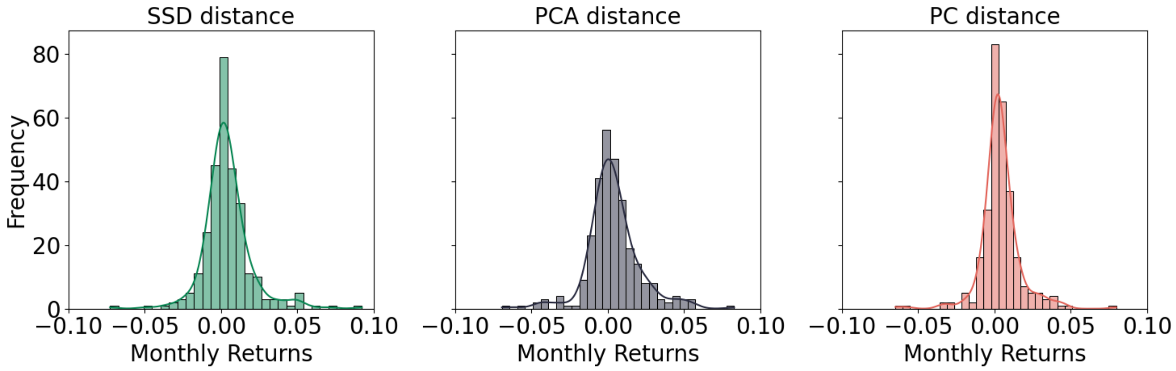

Figure 9: Empirical distribution and kernel density estimator of the monthly excess returns series within the benchmark analysis.

	Omega ratio	Sortino ratio	Max drawdown	Burke ratio
SSD distance	2.2607	0.2824	-0.0845	0.1524
PCA distance	1.8852	0.2329	-0.1468	0.1284
PC distance	2.6259	0.3011	-0.1159	0.2112

Table 2: Downside performance measures of the monthly excess returns within the benchmark analysis.

suggests that the Sharpe ratio may overestimate risk-adjusted performance. To address this concern, Table 2 provides a range of downside performance measures, as detailed in Schuhmacher and Eling (2011). Specifically, we compute the following: the Omega ratio, which quantifies excess return per unit of downside risk by dividing the sum of positive excess returns by the absolute value of the sum of negative deviations; the Sortino ratio, calculated as the average positive excess return divided by the standard deviation of negative returns; the maximum drawdown, which measures the largest peak-to-trough decline in portfolio value, representing the worst potential loss from a peak; and the Burke ratio, which compares the average return to downside risk, represented by the standard deviation of drawdowns. As shown in the table, the results, which are consistent with those reported by Rad et al. (2016), indicate that the PC distance outperforms the other metrics on nearly all downside performance measures, with the exception of the maximum drawdown.

Figure 10 illustrates the cumulative excess returns of a 1 USD investment in the three strategies over the entire timeframe. While the final values of the three investments are comparable, the strategies based on the SSD distance and PCA distance experience a brief period of exceptionally high performance during the early 2000s. In contrast, the strategy based on the PC distance exhibits more stable and consistent growth in investment value over time.

4.1.2 Economic significance

We now analyze the economic significance of the excess returns generated by our three strategies within the benchmark example and assess whether these excess returns are compensations for specific risk exposures. To do so, we regress the monthly excess returns on a variety of standard risk factors. Specifically, we first consider the five risk factors proposed by Fama and French (2015). Next, we restrict to the first three (namely the original ones in Fama and French (1993)),

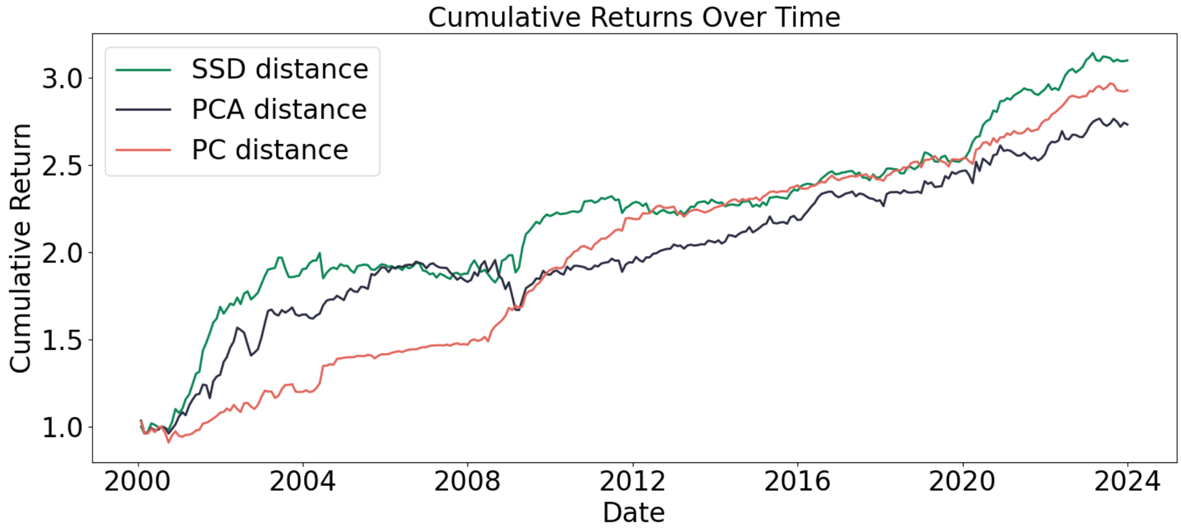

Figure 10: Cumulative excess return of a 1 USD investment in the three different strategies within the benchmark analysis.

and we add the momentum factor from Carhart (1997) and the liquidity factor introduced by Pástor and Stambaugh (2003). Finally, we incorporate all these factors into a combined model.

Table 3 presents the results from these three regressions. Notably, the alphas are consistently statistically significant and closely align with the average excess returns reported in Table 1. Furthermore, the Market factor (MKT) is generally not statistically significant, except for the PCA distance under the five-factor model. This result aligns with the market-neutral nature of pairs trading strategies[12] and indicates that beta-hedging, a common practice in algorithmic strategies involving equities, is not required in this context. The Small Minus Big (SMB), High Minus Low (HML), Conservative Minus Aggressive (CMA), and Liquidity (LIQ) factors are also largely insignificant across all model specifications.

Interestingly, the Robust Minus Weak (RMW) factor, which measures profitability, exhibits a positive and significant relationship with excess returns when included. This suggests that robust firms, characterized by strong profitability, are less likely to underperform due to their sound fundamentals, while weaker firms may face continued challenges, aligning the strategy with the profitability tilt captured by the RMW factor. Additionally, the Momentum (MO)

[12] This finding is further supported by Table 5, which demonstrates that there is essentially no sample correlation between the excess returns of our strategies and those of the equity index from which the stocks are selected.

	Alpha	MKT	SMB	HML	RMW	CMA	MO	LIQ
SSD dist.	0.0033***	0.0193	-0.0081	0.0287	0.1160**	0.043		
	(0.001)	(0.0259)	(0.0447)	(0.037)	(0.0574)	(0.0526)		
PCA dist.	0.0027***	0.0692**	0.0344	0.0778*	0.0808*	-0.0167		
	(0.001)	(0.0322)	(0.0464)	(0.0453)	(0.0471)	(0.0748)		
PC dist.	0.0031***	0.0172	0.0372	-0.0154	0.1309***	-0.0002		
	(0.0008)	(0.0165)	(0.0328)	(0.0271)	(0.0393)	(0.0369)		
SSD dist.	0.0044***	-0.0261	-0.0455	0.056			-0.0781***	-0.0341*
	(0.0012)	(0.0274)	(0.0282)	(0.038)			(0.0225)	(0.0179)
PCA dist.	0.0034***	0.0168	0.0114	0.0755**			-0.0777***	0.0234
	(0.0011)	(0.0296)	(0.0375)	(0.0372)			(0.0243)	(0.0213)
PC dist.	0.0040***	-0.0221	-0.0122	0.0124			-0.0439**	0.0044
	(0.0007)	(0.0204)	(0.0344)	(0.0222)			(0.0179)	(0.0152)
SSD dist.	0.0035***	-0.0034	0.0112	-0.0183	0.1371**	0.0746	-0.0881***	-0.0345**
	(0.001)	(0.0229)	(0.0409)	(0.0345)	(0.0537)	(0.048)	(0.0197)	(0.0169)
PCA dist.	0.0028***	0.0312	0.0509	0.0311	0.0940**	0.0351	-0.0835***	0.0226
	(0.001)	(0.0311)	(0.0428)	(0.0535)	(0.0477)	(0.0712)	(0.0254)	(0.0209)
PC dist.	0.0032***	-0.0026	0.0476	-0.0434	0.1403***	0.027	-0.0509***	0.0025
	(0.0007)	(0.0187)	(0.0319)	(0.0269)	(0.0368)	(0.0348)	(0.0179)	(0.0153)

Table 3: OLS regressions of the monthly excess returns of the three different strategies within the benchmark analysis against the Fama-French 5 factor model (Fama and French (2015)), the Fama-French 3 factor model (Fama and French (1993)) augmented by the momentum factor (Carhart (1997)) and the liquidity factor (Pástor and Stambaugh (2003)) and a factor model including all the aforementioned factors. Newly-West standard errors with five lags in parentheses. ***/**/* denote significance at 1%/5%/10% level.

factor demonstrates a consistently negative and significant relationship with excess returns. This finding is expected, as pairs trading is inherently a mean-reversion strategy, which bets against trends by shorting stocks that have outperformed (anticipating a reversal) and longing stocks that have underperformed (expecting recovery).

Finally, the alphas persist for all strategies, even under the full-factor model that incorporates all the factors mentioned above. This robustness highlights the distinctiveness and significance of the excess returns generated by the strategies.

4.2 Further analysis

We now conduct a few additional analyses to refine our understanding of the performance and characteristics of the trading strategies. First, we modify the trading design by incorporating

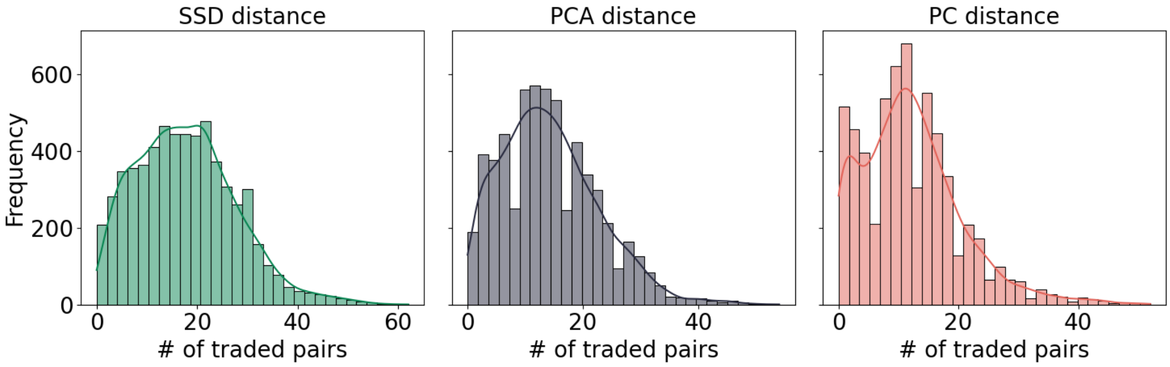

Figure 11: Distribution of the number of actually traded cointegrated pairs, on a daily basis.

a stationarity check on the spread defined in (4). Second, although pairs trading strategies are predominantly employed by hedge funds rather than individual investors, we compare the results of our strategies against a benchmark passive investment approach, namely a buy-and-hold strategy on the market index.

As briefly mentioned in Subsection 3.2, where we introduced the spread Z_t as the residuals of the regression in (3), testing for cointegration has been a standard practice in pairs trading since the seminal works of Vidyamurthy (2004) and Gatev et al. (2006). Cointegration testing is reasonable, as two cointegrated price series produce a spread that is stationary and potentially mean-reverting, which is a key assumption underpinning pairs trading strategies.

To investigate how the excess returns of our strategies are affected when this additional control is introduced, we test for cointegration using the standard univariate Engle-Granger approach over the formation period. Introducing this requirement excludes several tradable pairs that would otherwise qualify under our initial trading design. Figure 11, which parallels Figure 7, shows the number of stock pairs that pass the cointegration test and are subsequently traded on a daily basis. Imposing this criterion approximately halves the number of traded pairs. Specifically, when using the SSD distance the number of traded pairs is 17.78, which becomes 14.18 when the PCA distance is applied, and 12.12 for the PC distance.

However, as shown in Table 4, this reduction in the number of tradable pairs significantly affects the performance of the strategies. In particular, the Sharpe ratios of the PCA- and

	Mean	T-ratio	Std. Dev.	Sharpe ratio	Ann. Sharpe ratio	Pos.	Neg.
SSD dist.	0.0047	4.49***	0.0176	0.2646	0.9166	62.8%	37.2%
PCA dist.	0.0027	2.36**	0.0194	0.1391	0.4819	53.8%	46.2%
PC dist.	0.0032	3.92***	0.0138	0.232	0.8037	63.6%	36.4%

Table 4: Summary statistics of the monthly excess returns series within the benchmark analysis, computed as described in Subsection 3.3 for the three distance metrics, each featuring 288 observations. ***/**/* denote significance at 1%/5%/10% level.

	SSD dist.	PCA dist.	PC dist.	S&P500
SSD dist.	1			
PCA dist.	0.5447	1		
PC dist.	0.2636	0.3814	1	
S&P500	-0.0284	-0.0485	-0.0669	1

Table 5: Sample correlations of the monthly excess returns of the three pairs trading strategies within the benchmark sample and the monthly returns of the S&P 500.

PC-based strategies decrease substantially, while the Sharpe ratio of the SSD-based strategy remains nearly unchanged.

Pairs trading strategies are a form of algorithmic trading, typically employed by hedge funds rather than individual investors, due to their reliance on sophisticated statistical models and frequent rebalancing. However, comparing the excess returns of these strategies to a passive buy-and-hold investment in a market index remains relevant, as it provides a baseline for evaluating their performance relative to a widely accessible and low-cost benchmark.

As a benchmark, we evaluate a buy-and-hold strategy applied to the S&P 500 index, from which the stocks in our study are selected. Over the entire time frame, a 1 USD investment in the index yields a terminal value of 3.24. By comparison, Figure 10 shows that the terminal values for the SSD distance, PCA distance, and PC distance strategies are 3.65, 2.06, and 2.42, respectively. The average return for the index is 51 basis points (bps); however, this sample mean has a t-ratio of 1.94 and is not statistically significant, reflecting a high standard deviation of 4.46%, which is nearly three times greater than the standard deviations observed in the pairs trading strategies. Consistent with this, the monthly Sharpe ratio for the index is only 0.1143 (0.3960 on an annualized basis).

Interestingly, the correlation matrix of excess returns, presented in Table 5, indicates that the excess returns from the pairs trading strategies are largely uncorrelated with the market.

4.3 Sensitivity analysis

We now conduct a sensitivity analysis on the key parameters of our trading design: the length of the subsample used to compute the z-score (i.e., the sample size for estimating the mean and standard deviation in (5)), which is set to six months in the benchmark analysis, and the z-score threshold z that triggers the opening of a position, set to 2 in the benchmark case.

Table 6 presents three key performance metrics for the strategies as a function of these parameters. The top panel reports the average monthly excess returns in basis points, the middle panel provides the associated t-ratios, and the bottom panel shows the monthly Sharpe ratios. The average excess returns are consistently above 20 basis points (with one exception for the PCA distance) and tend to increase as the opening threshold z rises. This increase in z reduces trading activity, likely concentrating trades on the most profitable pairs and further decreasing the number of traded pairs and the employed capital. Interestingly, if a trading desk finds the number of traded pairs (analyzed in conjunction with Figure 7) to be excessive, increasing the opening threshold to $z = 2$ or $z = 3$, which corresponds to a z-score statistically different from zero at nearly the 99% confidence level, may offer a more effective alternative than introducing the cointegration test. As discussed in Subsection 4.2, the latter appears to adversely impact performance.

The t-ratios are almost always above the critical values for the 1% confidence level, indicating that the average excess returns are statistically significant across nearly all parameter configurations.

Finally, the Sharpe ratios also show an upward trend with increasing z, reinforcing that, when evaluated on a risk-adjusted basis, the PC distance remains the clustering metric that delivers the highest performance.

Average monthly excess return (bps)

	SSD distance					PCA distance					PC distance				
	1	1.5	2	2.5	3	1	1.5	2	2.5	3	1	1.5	2	2.5	3
1 m	25	28	30	28	29	17	22	28	29	29	23	27	24	24	25
3 m	29	33	33	36	34	19	25	30	38	43	29	32	35	37	35
6 m	26	31	38	43	39	15	17	36	46	48	24	31	41	42	38
12 m	25	28	33	43	54	16	30	35	55	49	22	33	35	48	32

T-ratio

	SSD distance					PCA distance					PC distance				
	1	1.5	2	2.5	3	1	1.5	2	2.5	3	1	1.5	2	2.5	3
1 m	3.18	3.27	2.78	2.68	2.63	2.28	2.85	3.22	3.78	3.54	4.18	4.33	4.48	4.48	4.25
3 m	3.79	3.8	3.75	3.51	3.33	2.47	2.86	3.27	3.98	4.2	4.02	4.39	4.38	4.83	4.35
6 m	2.93	3.54	4.27	3.75	2.93	2.12	2.14	3.48	4.67	4.28	3.59	3.82	4.9	5.61	4.65
12 m	2.49	3.48	3.22	4.08	2.59	2.06	3.16	3.18	4.35	3.83	3.24	3.4	3.74	4.4	3.92

Sharpe ratio

	SSD distance					PCA distance					PC distance				
	1	1.5	2	2.5	3	1	1.5	2	2.5	3	1	1.5	2	2.5	3
1 m	0.19	0.19	0.16	0.16	0.15	0.13	0.17	0.2	0.22	0.21	0.25	0.25	0.28	0.26	0.25
3 m	0.22	0.22	0.22	0.21	0.2	0.15	0.17	0.19	0.23	0.25	0.24	0.26	0.26	0.28	0.26
6 m	0.17	0.21	0.25	0.22	0.17	0.12	0.13	0.21	0.27	0.25	0.21	0.23	0.29	0.33	0.27
12 m	0.15	0.21	0.19	0.24	0.15	0.12	0.19	0.19	0.26	0.23	0.19	0.2	0.22	0.26	0.23

Table 6: Average monthly excess returns (in basis points), associated t-ratios, and monthly Sharpe ratios for the three strategies under the benchmark example, presented as a function of the subsample length used to standardize the spread Z_t in (4) (rows) and the z-score threshold for position opening (columns). Critical t-ratio values for statistical significance are 1.65, 1.97, and 2.59 at the 10%, 5%, and 1% levels, respectively.

5 Conclusions

In this paper, we analyzed the application of an unsupervised learning-based clustering technique for selecting pairs of stocks within the S&P 500 index. We focused on understanding how various distance metrics, which form the basis of clustering algorithms, influence the pair selection process and, in turn, the returns generated by these strategies. Our study considered several distance metrics, including the commonly used Euclidean distance, a principal components-based distance, and a novel metric derived from partial correlation. The partial correlation metric aims to isolate the true relationship between two securities by eliminating spurious correlations that arise from shared exposure to a third variable.

Our findings demonstrated that the output of the clustering algorithm, specifically, the number of clusters formed and the average size of each cluster, varied significantly depending on the chosen distance metric. However, we identified that the primary factor driving clustering was the industry sector of the stocks. This conclusion was further supported by an analysis of cluster purity based on the SIC classification, which validated our hypothesis.

When applying pairs trading strategies and backtesting them over the period from 2000 to 2023, we observed excess returns that were both statistically and economically significant. In terms of risk-adjusted performance, the strategy based on partial correlation produced the highest monthly Sharpe indicating that this distance metric was the most effective. Measures of downside risk further supported the robustness of the strategy, highlighting its favorable performance relative to others. From an economic standpoint, strategies based on partial correlation generated statistically significant alphas when tested against standard risk factors, such as those developed by Fama and French, even after incorporating additional factors like momentum and liquidity.

Additional analysis revealed that, in risk-adjusted terms, these strategies are competitive with a simple buy-and-hold approach on the index itself. Furthermore, to reduce the number of traded pairs, one can either focus on pairs formed by cointegrated stocks or raise the threshold for opening a position. Among these options, focusing on cointegrated pairs generally leads to lower overall returns, while increasing the threshold tends to improve them.

Disclosure statement

The Authors report there are no competing interests to declare.

Funding statement

No funding was received.

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, page 49–60. Association for Computing Machinery.

Avellaneda, M. and Lee, J.-H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7):761–782.

Barucci, E., Bonollo, M., Poli, F., and Rroji, E. (2021). A machine learning algorithm for stock picking built on information based outliers. *Expert Systems with Applications*, 184:115497.

Bini, B. and Mathew, T. (2016). Clustering and regression techniques for stock prediction. *Procedia Technology*, 24:1248–1255. International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015).

Bogomolov, T. (2013). Pairs trading based on statistical variability of the spread process. *Quantitative Finance*, 13(9):1411–1430.

Carhart, M. M. (1997). On persistence in mutual fund perdormances. *Journal of Finance*, 52(1):57–82.

Chang, V., Man, X., Xu, Q., and Hsu, C. (2021). Pairs trading on different portfolios based on machine learning. *Expert System*, 38(3).

Chen, H., Chen, S., Chen, Z., and Li, F. (2019). Empirical investigation of an equity pairs trading strategy. *Management Science*, 65(1):370–389.

Clegg, M. and Krauss, C. (2018). Pairs trading with partial cointegration. *Quantitative Finance*, 18(1):121–138.

Crépellière, T., Pelster, M., and Zeisberger, S. (2023). Arbitrage in the market for cryptocurrencies. *Journal of Financial Markets*, 64:100817.

Cummins, M. and Bucca, A. (2012). Quantitative spread trading on crude oil and refined products markets. *Quantitative Finance*, 12(12):1857–1875.

Do, B. and Faff, R. (2010). Does simple pairs trading still work? *Financial Analysts Journal*, 66(4):83–95.

Do, B. and Faff, R. (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2):261–287.

Dose, C. and Cincotti, S. (2005). Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*, 355(1):145–151.

Eling, M. (2008). Does the measure matter in the mutual fund industry? *Financial Analysts Journal*, 64(3):54–66.

Eling, M. and Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking Finance*, 31(9):2632–2647.

Elliott, R. J., Van der Hoek, J., and Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance*, 5(3):271–276.

Ester, M., Kriegel, H., Sander, J., and Xiaowei, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fanelli, V., Fontana, C., and Rotondi, F. (2024). A hidden Markov model for statistical arbitrage in international crude oil futures markets. *Working paper, available at SSRN.*

Gatev, E., Goetzmann, W. N., and Rouwenhorst, K. G. (2006). Pairs trading: performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827.

Han, C., He, Z., and Toh, A. J. W. (2023). Pairs trading via unsupervised learning. *European Journal of Operational Research*, 307(2):929–947.

Huck, N. (2009). Pairs selection and outranking: An application to the sp 100 index. *European Journal of Operational Research*, 196(2):819–825.

Huck, N. (2010). Pairs trading and outranking: The multi-step-ahead forecasting case. *European Journal of Operational Research*, 207(3):1702–1716.

Huck, N. and Afawubo, K. (2014). Pairs trading and selection methods: is cointegration superior? *Applied Economics*, 47(6):599–613.

Hull, J. (2019). *Machine Learning in Business: an introduction to the world of data science.* GFS Press.

Jacobs, H. and Weber, M. (2015). On the determinants of pairs trading profitability. *Journal of Financial Markets*, 23:75–97.

Kenett, D., Huang, X., Vodenska, I., Havlin, S., and Stanley, H. (2015). Partial correlation analysis: applications for financial markets. *Quantitative Finance*, 15(4):569–578.

Kim, S.-H., Park, D.-Y., and Lee, K.-H. (2022). Hybrid deep reinforcement learning for pairs trading. *Applied Sciences*, 12(3).

Kim, T. and Kim, H. (2019). Optimizing the pairs-trading strategy using deep reinforcement learning with trading and stop-loss boundaries. *Complexity*, 2019(3):1–20.

Krauss, C. (2017). Statistical arbitrage pairs trading strategies: review and outlook. *Journal of Economic Surveys*, 31(2):513–545.

Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the SP 500. *European Journal on Operations Research*, 259(2):689–702.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 5:281–297.

Mantegna, R. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11:193–197.

Nanda, S., Mahanty, B., and Tiwari, M. (2010). Clustering indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12):8793–8798.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertész, J., and Kanto, A. (2003). Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E*, 68:056110.

Pástor, L. and Stambaugh, R. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685.

Rad, H., Low, R., and Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 16(10):1541–1558.

Sarmento, S. M. and Horta, N. (2020). Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158:113490.

Schuhmacher, F. and Eling, M. (2011). Sufficient conditions for expected utility to imply drawdown-based performance rankings. *Journal of Banking Finance*, 35(9):2311–2318.

Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis.* Wiley, Hoboken (NJ).