KU LEUVEN

FACULTY OF
ENGINEERING SCIENCE

Unveiling pairs trading opportunities via multi-perspective clustering

Sisheng Liu

Thesis submitted for the degree of
Master of Science in
Electrical Engineering, option
Electronics and Chip Design

Supervisor:
Prof. Johannes De Smedt

Assessors:
Ir. Kn. Owsmuch
K. Nowsrest

Assistant-supervisors:
Ir. An Assistent
A. Friend

Academic year 2023 – 2024

Preface

I would like to thank everybody who kept me busy the last year, especially my promoter and my assistants. I would also like to thank the jury for reading the text. My sincere gratitude also goes to my wive and the rest of my family.

$$1 + 2 = 3$$

Sisheng Liu

Contents

Abstract

The `abstract` environment contains a more extensive overview of the work. But it should be limited to one page.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

LoG	Laplacian-of-Gaussian
MSE	Mean Square error
PSNR	Peak Signal-to-Noise ratio

Symbols

42	"The Answer to the Ultimate Question of Life, the Universe, and Everything" according to
c	Speed of light
E	Energy
m	Mass
π	The number pi

Chapter 1

Introduction

Pairs trading strategy is based on contrarian strategy, where the trader profits by shorting overvalued stocks and longing undervalued stocks, with the bet that the overvalued and undervalued stocks price would revert in the short term. Some studies argue that market microstructure phenomena, such as risk-aversion-related inventory effects or bid-ask bounce, serve as the primary drivers behind these reversals, whereas other stream of studies posit that the predictability of future returns stems from market overreaction and correction or belief reversionSubrahmanyam (2005). For the first driver, Short-term price reversal is driven by lower stock turnover, lower liquidity of stocks and inventory imbalances of market maker who bears inventory riskAVRAMOV et al. (2006). Short-term reversal is related to bid-ask spreads and microstructure phenomena such as risk-aversion-related inventory effectsJegadeesh and Titman (1993). Common strategy of buying historical win stock, short lose stock can temporarily drive stock price from its long-term valueJegadeesh and Titman (1993). For the market overreaction driver. It posits that investors experience waves of optimism and pessimism that would result in temporary price swings Lo and MacKinlay (1989). Reversal effect is more predicable and portfolios of prior "losers", are found to outperform prior "winners" when the market overreacts De BONDT and THALER (1985). In Jacobs (2015) work, the author explored the relationship between market sentiment proxies and 100 kinds of anomalies simulated in the cross-section of expected equity returns. It's found that market sentiment has a substantial impact on the return anomalies. A study on behavior finance suggests that irrationality might have long-term and substantial impact on asset prices in an economy where rational and irrational traders interact Jacobs (2015). When the irrationality persists, it might prevent contrarian arbitrageurs from profiting because the price does not correct itself as expected. This would result in the limit of arbitrage and allows the existence of mispricing. A breif summary from Gatev on pairs trading is that pairs trader wins by being disciplined in trading from undisciplined overreaction of the crowd. All the above explanations supports the proposition that information from the past can be exploited to predict future return. Thus, it contradicts the Efficient Markets Hypothesis (EMH), which posits that the market value of an asset reflects its fundamental value. According to EMH, no investment

strategy can consistently yield positive average returns by relying on past information, as rational agents swiftly adjust their trading strategies in response to new price informationFama (2021). The EMH hypothesis is being constantly challenged by finding new trading opportunities from past information. In the case of pairs trading, traders try to identify temporarily mispriced assets and capitalize on their positions by speculating that the mispricing will be rectified in the future, thereby generating profits. Hence, identifying stocks with price anomalies and having accurate estimation on their future direction and level of reversion becomes crucial in profiting from pairs trading strategy.

Chapter 2

Literature review

Krauss (2017) summarized 7 types of pairs trading approach. The first is distance approach, of which the most cited work is by Gatev et al. (1999). It involves pairs formation stage and trading stage. During formation stage, it exhaustively searches for pairs of which the normalized price series has minimum sum of squared deviations. In the subsequent 6-month trading stage, the top 5 and 20 pairs with the smallest historical distance measure were selected. The trading rule is to open long-short positions when pair prices diverge by more than two historical standard deviations and close positions when prices revert. This strategy aims to profit on price divergences between pairs of stocks and profit from their subsequent reversion. A bootstrap analysis found that selected pairs beats the randomly selected pairs on average. Chen et al. (2019) used correlation of 5-year's historical monthly return for quasi-multivariate pairs and picked top 50 most correlated pairs for trade. The minimization of objective function specified by correlation metrics leads to less strict pairs selection criterion than minimizing objective function specified by distance, thus allowing for more trading opportunities Krauss (2017). Recognizing the declined profitability of Gatev's pairs trading strategy, Do and Faff (2012)Do et al. (2006) improved the profitability by only allowing pairs matched within the same sector and used zero crossing as measurement of potential profitable pair.

The next is cointegration approach in which Vidyamurthy (2004)'s work is the most cited. Vidyamurthy (2004) uses cointegration as the measurement of comovement and apply it in pairs trading strategy by profiting from the spread caused by temporary mispricing. The fundamental difference with distance approach is the test of stationary of spread. Cointegration testing mainly root from the work of Johansen (1988) Engle and Granger (1987) Dickey and Fuller (1979). The cointegration approach finds its theoretical ground on arbitrage pricing theory (APT). In APT framework, an asset's return is the total contribution from different risk factors that the asset is exposed to Ross (1976). It is an improved version of CAPM model developed by Sharpe (1964). In mathematical form, the return of asset i at time t is expressed as chapter 2. μ_i represents risk premium. Risk factor vector $\mathbf{f}_t$ is weighted by vector $\beta_{\mathbf{i}}'$, which stands for how much exposure the asset i is exposed

to each risk factor. ϵ_{it} is the idiosyncratic risk.

$$r_{it} - \mu_i = \beta_{\mathbf{i}}' \mathbf{f}_t + \epsilon_{it} \tag{2.1}$$

The APT states that if two securities have the risk factor, and they have the same risk factor exposure at the same time, the expected return of these stocks during a certain time interval should be the same Vidyamurthy (2004).

The third one is time series approach. Elliott et al. (2005) models spread as Ornstein-Uhlenbeck process that has mean reverting property and the pair is traded when the spread exceeds certain calibrated threshold. It is fully tractable because it uses pure mathematical model and it is quite aligned with the mean reverting property of spread in pairs trading Krauss (2017). Other approaches are stochastic control based pairs trading strategy Liu and Timmermann (2013)Jurek and Yang (2007), coupula based approach Liew and Wu (2013) and PCA based approach Avellaneda and Lee (2008) which are less popular in the academia.

The most relatable approaches to this study are machine learning based pairs trading techniques. With the development of computing power, machine learning and more advanced algorithms played a role in innovating pairs trading strategy. These algorithms can be applied in either stock selection, in which stocks are paired or grouped together by homogeneity from certain perspectives, or forward-looking return prediction in deciding whether to enter long or short position.

Huck (2010) identified overvalued and undervalued stocks using Eleman neural network and ELECTRE III. The Eleman network predicts one-week ahead returns and produces a pairwise spread prediction matrix. ELECTRE III ranks stocks with the predicted spread matrix with undervalued stocks on top and overvalued stocks to bottom. The top and bottom k stocks are bought and sold with equal weights and held for one week. This strategy yields a maximum weekly excess return of 1.1%. It is a popular ML-based pairs trading strategy as it doesn't require explicit model parameters like cointegration coefficients.

In the work of Takeuchi (2013), features from historical individual stock prices were extracted using a stacked restricted Boltzmann machines, and then passed to feedforward neural network (FFNN) classier to predict the probability of next month return being above or below the median return of all the traded stocks. Ranked by the probability of next month return being above median, the top deciles are bought and bottom deciles are sold, creating an equal weighted long-short portfolio. It achieved annualized return of 45.3% over the 1990-2009 test period compared to the baseline momentum strategy of 10.53%.

Flori and Regoli (2021) identified cointegration group for each stocks in S&P 500 as the set of stocks that are cointegrated to a given sock for less than 1% significance. Then for each stocks within its cointegration group, past values of the gap with respect to the mean return of the group, past trading volume and past returns were used as features of multi-variate LSTM features to predict next day's probability of positive return. The predicted probability of positive return is then used to complement the trading signal derived from the pairs trading strategy based on price gap. It is shown to have improved the trading strategy based on price gap.

Han et al. (2023)

[what is the application on unsupervised learning for trading]

[why is TI good] [why is firm good] Han (2018) Green et al. (2017) [what is DTW] [why is DTW good]: because technical indocator might be similar in shape but not in pace [multi-perspective]

Chapter 3

Methodology

This chapter details the research methodology. It begins by introducing the overview of the experiment setup. Then data sources of three perspective feature sets and the preprocessing steps applied to the raw data for the proposed SEFT clustering. We then introduce the baseline single perspective and multi perspective clustering algorithms. Next, the portfolio formation method is discussed. Finally, we outline the metrics used to evaluate trading performance in our study.

3.1 Experiment setup

In the study of Han et al. (2023), historical return characteristics and firm characteristics are simultaneously used as features for clustering and forming portfolio. It is based on the assumption that stocks with similar return features are expected to exhibit similar future trends, and firm characteristics are predictive of future returns from both accounting and asset valuation perspectives. The author showed that including firm characteristics as clustering feature boosted trading performance than clustering using return characteristics along. Inspired by this, we seek to explore cluster-based pairs trading strategy by combining features from three perspectives using multi-perspective clustering. The three perspectives are sector characteristics, firm characteristics, and technical indicator, and thus the three perspectives clustering is also named SEFT clustering. Clustering on sector categorization or firm characteristics is feature-based, using features of each datapoint as input, while clustering on technical indicators is distance-based, using a precalculated distance matrix for input. Data of all three perspectives are directly or indirectly derived from the database maintained by The Center for Research in Security Prices, LLC (CRSP). Since the portfolio is formed based on the clustering result from the previous month and readjusted over again the next month, the feature sets and distance matrix used for clustering are generated at monthly basis for the cluster period from February 2003 to November 2023. The trade period is from March 2003 to December 2023. The scope of trading is all the stocks listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and Nasdaq. On average there are

about 6000 tradable stocks every month within this scope. To perform SEFT clustering, features from all three perspectives are indispensable. However, not all the stocks have features from the the three perspectives and thus make the number of tradable stocks every month less than all the stocks listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and Nasdaq. The selection of candidate tradable stocks are described in Section 3.2. The sector and firm characteristics features span from February 2003 to November 2023 as consistent with the portfolio formation period and can be directly used for feature-based clustering. Data source and preprocessing for sector and firm characteristics are described in Section 3.3 and Section 3.4. For technical indicators clustering, we first compute multidimensional DTW distance matrix from a set of historical technical indicators time series of window size of past 36 months. The computed distance matrix is then used as input for clustering. Therefore, the technical indicator data spans from February 2000 to November 2023 with window size of 36 months, and the computed technical indicator distance matrix spans from February 2003 to November 2023, as consistent with the portfolio formation period and chronologically aligned with features from other perspectives. The data source and computation of technical indicator and multidimensional DTW distance matrix computation is described in Section 3.5 and Section 3.6.

The SEFT clustering and threshold-based trading works as follow. For a given portfolio formation month, stocks are first clustered separately from the three perspectives using K-medoids clustering. For a stock clustered in three perspectives, the distance to all the medoids from each perspectives can be derived. The distance to medoids from three perspectives are then weighted and combined as the new feature. The combined new feature is then used for final clustering using k-medoids. The SEFT clustering and K-medoids clustering are detailed in Section 3.9 and Section 3.8. With the clustered stocks, undervalued and overvalued stocks within each cluster are paired to form an equal weighted long-short portfolio and held for a month. One month after the portfolio formation, the portfolio is liquidized, and the remaining cash is then reinvested for the next round using the same strategy. The threshold trading strategy is introduced in Section 3.10. The same threshold trading strategy is also combined with clustering result over the same set of stocks as SEFT using only features from one of the three perspective, also known as single-perspective clustering, and features from two of the three perspectives, known as dual-perspectives clustering to form single or dual perspective portfolio. Single perspective, dual perspective clustering and their portfolio formation are introduced in Section 3.11 Section 3.12. By comparing the single and dual-perspective with SEFT clustering, we can understand the role of each perspective plays in SEFT clustering. We also compare the SEFT clustering with other two benchmark portfolios. The first benchmark portfolio is formed by S&P500 index. The second is reversal portfolio formed by longing stocks with prior month return below the median of all tradable stocks and shorting stocks with prior month return above the median. The formation of benchmark portfolios is discussed in Section 3.13.

3.2 Tradable stocks preselection

Prior to the introduction of features from three perspectives, this section serves to introduce the criterion of choosing the appropriate set of stocks to trade for each month. K-medoids clustering does not support missing features. It is hence important to ensure the features of stocks to be clustered at each month are not missing. Since SEFT requires feature from three perspectives and the set of stocks with features available in the datasets of three perspectives are different in different portfolio formation period, it is hence important to select stocks with available features in all three perspectives to ensure the SEFT clustering works correctly.

At each portfolio formation date T, the following sets of stocks are defined from datasets of each perspective:

- $\mathcal{S}_T^{TI}$: the set of stocks with no missing values for all 11 technical indicator time series over the past 36 months as introduced in Section 3.5.
- $\mathcal{S}_T^{firm}$: the set of stocks with no missing firm characteristics as introduced in Section 3.4
- $\mathcal{S}_T^{sector}$: the set of stocks with no missing SIC code at T as described in Section 3.3

Tradable candidate stocks for SEFT clustering at each portfolio formation date can be represented as Equation (3.1), in which EndOfMonths is the set of all natural month-end dates.

$$\forall T \in \text{EndOfMonths: } \mathcal{S}_T^{\text{SEFT}} \leftarrow \mathcal{S}_T^{SEctor} \cap \mathcal{S}_T^{Firm} \cap \mathcal{S}_T^{Tech} \tag{3.1}$$

Additionally, we define effective dates of trading $\mathcal{T}$ as the month-end dates which have non-zero number of tradable stocks defined by $\mathcal{S}_T^{\text{SEFT}}$. Mathematically, this means the set of dates with non-zero cardinality on set $\mathcal{S}_T^{\text{SEFT}}$ for all natural month-end date:

$$\forall T \in \text{EndOfMonths: } \mathcal{T} \leftarrow T \quad \text{s.t. } |\mathcal{S}_T^{\text{SEFT}}| > 0 \tag{3.2}$$

To make trading performance comparable among portfolios formed by SEFT, single perspective and dual perspective clustering, we aim to impose same set of candidate stocks for all clustering methods. With SEFT clustering requiring the data availability from all three perspectives, it has the most constrained criterion in selecting tradable stocks, compared to single or dual perspective clustering requiring data availability from only one or two perspective. Therefore, for all clustering methods and the formed portfolio thereof, we impose same set of candidate stocks for each formation period as defined in Equation (3.1). Hence, candidate tradable stocks for all different perspectives at each effective date of trading $T \in \mathcal{T}$ equals to $\mathcal{S}_T^{\text{SEFT}}$, and is simply denoted as $\mathcal{S}_T$[1].

[1]This notation is used for the rest of the article

3.3 Sector characteristics

In this study, we use Standard Industrial Classification (SIC) code as the identification of sector for each stocks. The SIC code reflects the main type of business of the underlying company U.S. Securities and Exchange Commission. It is possible that the change of the main type of business of some companies or the systematic change of code mapping scheme happens from time to time. For this issue, we obtain the time dependent SIC code mapping of each company so that the clustering of sector characteristics is not affected by neither of these changes.

For portfolio formed by sector-only perspective clustering, we use the SIC code directly as the cluster number assigned to each tradable stocks in the sense that stocks belonging to the same SIC code belongs to the same cluster. This is further detailed in Section 3.11. For dual perspective clustering involving sector characteristics and SEFT clustering, the distances to medoids of each stock from sector perspective is required. In this case, SIC code can't be used directly because it is a categorical variable that does not reflect the position in the feature space. To featurize the sector characteristics at each month-end $t \in \mathcal{T}$, the set of SIC code in $\mathcal{S}_t$ is one-hot coded. Assume stocks in $\mathcal{S}_t$ are categorized into unique sectors contained in set SIC_t in each month t, the one-hot coded SIC feature dataset would be $\mathcal{F}_t^{sic} \in \mathbb{R}^{|\mathcal{S}_t| \times (SIC_t)}$. Dataset $\mathcal{F}_t^{sic}$ is then used to generate distance to medoids by dual-perspective clustering involving sector perspectives or SEFT clustering as introduced in Section 3.9 and Section 3.12.

3.4 Firm characteristics

For firm characteristics, we adopt the methodology of Green et al. (2017) to generate firm characteristics that provides independent information about average US stock returns. The resulting dataset contains more than 120 features. As pointed out in the study of Green et al. (2017), not all these features are predictive in future return. To improve data quality, we follow the practice of Han et al. (2023) to exclude firm characteristics with more than 15% missing values, resulting in total of 78 firm characteristics. The resulting dataset is $\mathcal{F}_t^{firm} \in \mathbb{R}^{|\mathcal{S}_t| \times 78}, t \in \mathcal{T}$.

Same as sector characteristics clustering, for portfolio formed by firm characteristics-only single perspective clustering, the resulting dataset $\mathcal{F}_t^{firm}$ is directly used for clustering and the preceding threshold trading strategy.

[add the table to the appendix]

3.5 Technical indicator

In order to perform clustering on technical indicator perspective at a certain month, various technical indicators time series are first computed. Then the technical indicators of all stocks are used as input to compute pairwise DTW distance matrix between those stocks. Once the distance matrix is computed, it can be used to perform distance-based clustering algorithm. The expectation is that the combination of different technical indicators reflects more information than any single technical indicator. Thus, instead of a single technical indicator, we select multiple types of technical indicators that measures the market dynamics from different perspectives. Because the technical indicators are multiple time series for each stock, we apply multi-dimensional DTW algorithm to derive the pairwise distance. We adopt the technical indicators as summarized in Section 3.5. Computation of technical indicators is executed for all stocks and their available period before they are used for clustering to improve computational efficiency. Past time interval with non-zero window size is required to calculate these technical indicators varies, resulting in NA values for periods without sufficient data. This typically occurs during the initial periods when the stock is first traded on the market. Given that stocks can only be listed once, these NA values are typically expected during the early trading periods. As a result of technical indicator computation, each stock has 11 technical indicator time series with different lengths caused by the different window size chosen for each technical indicator. To ensure identical length of the technical indicators as input for multi-dimensional DTW algorithm, we trim all time series before the first date when all the 11 technical indicators are not missing. Typically the first 10 month's technical indicators are trimmed because 10-month momentum has the longest window size of 10 among other technical indicators.

3.6 multi-dimensional DTW

Dynamic Time Warping (DTW) algorithm is a similarity measure between two time series. It is efficient in detecting two time series with similar shape but different pace by allowing elastic transformation of time series Senin (2008). Consider two univariate time series of equal length N: $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_N)$, $N \in \mathbb{N}$. The algorithm first computes the local cost matrix $C \in \mathbb{R}^{N \times N}$ of all pairs of point from X and Y as described in Equation (3.3):

$$C \in \mathbb{R}^{N \times N} : c_{i,j} = |x_i - y_j|, \quad i, j \in [1 : N] \tag{3.3}$$

With the local cost matrix computed, the DTW algorithm seeks to find an alignment path that goes through the low-cost area in the local cost matrix as the red path shown in Figure 3.1. In formal definition, the alignment path is a sequence of 2D coordinate points $P = (p_1, p_2, \ldots, p_K)$ with $p_l = (x_i, y_j), i, j \in [1 : N], l \in [1 : K]$, where K is the length of the wrapping path. These coordinate points must satisfy the following three conditions:

TIs	Parameters	Description
Exponential Moving Average	length=4	Calculates the Exponential Moving Average over the last 4 months.
Moving Average Convergence Divergence	fast=8 slow=4 signal=9	Calculates the Moving Average Convergence Divergence with fast EMA of 8 periods, slow EMA of 4 months, and signal line EMA of 9 months.
Williams %R	length=4	Measures overbought or oversold conditions of a certain stock by comparing the close of the current period with the high-low range over the last 4 months.
Stochastic Oscillator	k=3 d=3	Calculates the Stochastic Oscillator with %K calculated over 2 months and %D calculated as a 3-month moving average of %K.
Relative Strength Index	length=4	Measures the magnitude of past 4 months price changes to evaluate overbought or oversold conditions.
Rate of Change	length=4	Measures the percentage change in price over the last 4 months.
Momentum	length=2, 4, 6, 8, 10	Calculates momentum indicators with past period ranging from 2 to 10 months.

- **Boundary condition:** the wrapping path must start from the first point of the aligned sequence and end at the last point of the aligned sequence, formally $p_1 = (1, 1), p_K = (N, N)$.
- **Monotonicity condition:** the coordinates of the wrapping path should follow time ordering sequence, formally for $x_1 < x_2 < \cdots < x_K$, $y_1 < y_2 < \cdots < y_K$.
- **Step size condition:** only adjacent jump is allowed, meaning $p_{l+1} - p_l \in \{(1, 1), (1, 0), (0, 1)\}$.

Given an arbitrary wrapping path $p = (p_1, p_2, \ldots, p_K)$ that satisfies the above conditions, $c_p(X, Y) = \sum_{l=1}^{K} c(x_i, y_j)$ is the cost of the wrapping path. The wrapping path P^* with the minimal cost over all possible wrapping paths is the optimal wrapping path. It would be highly computational intensive to perform brutal search over all possible wrapping paths for the optimal path. In order to tackle this optimization problem more efficiently, a dynamic programming methodology is employed to solve the optimal wrapping path by solving the recursive equation define in Equation (3.4), in which $D(i, j)$ is the cumulative distance written as the sum of $c(x_i, y_j)$ in the current cell and the minimum cumulative distance from the adjacent cellsB (1983) Rabiner and Juang (1993) Shokoohi-Yekta et al. (2017).

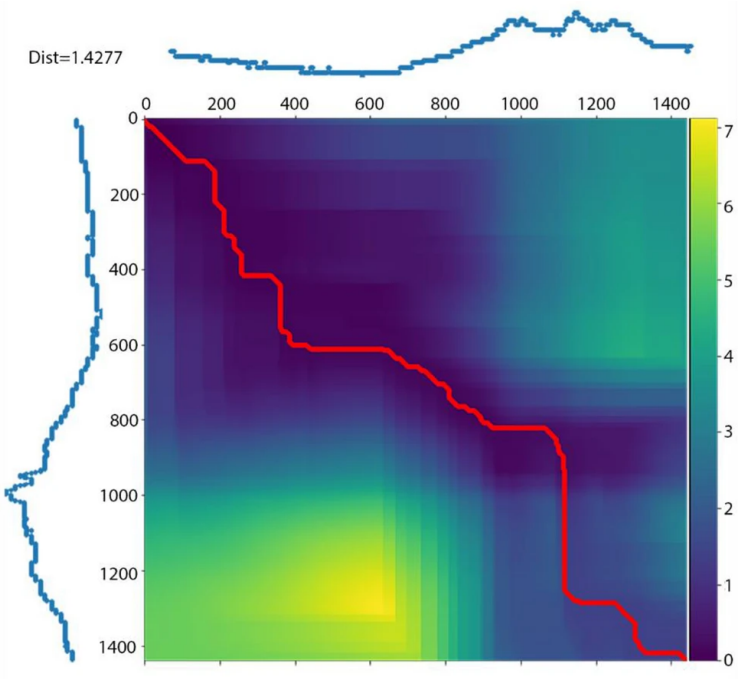

FIGURE 3.1: An example local cost matrix with cost visualized as heatmap. Li et al. (2021)

$$D(i,j) = \min\{D(i-1,j-1), D(i-1,j), D(i,j-1)\} + c(x_i, y_j),\ i,j \in [1,N]. \quad (3.4)$$

When the time series X and Y are M dimension time series where $M > 2$, the problem becomes multidimensional DTW. The multidimensional DTW framework is suitable for the multidimensional technical indicator as introduced in Section 3.5. We can adapt the univariate DTW framework into multi-dimensional DTW just by replacing the scalar difference in Equation (3.3) with multivariate euclidean distance as shown in Equation (3.5) Shokoohi-Yekta et al. (2017).

$$C \in \mathbb{R}^{N \times N} : c_{i,j} = \sum_{m=1}^{M} (x_{im} - y_{jm})^2, \quad i,j \in [1:N] \quad (3.5)$$

Given a portfolio formation month $t \in \mathcal{T}$, $\mathcal{S}_t$ are the set of tradable stocks as defined in Section 3.2. Hence, the pairwise multi-dimenstional DTW distance between any arbitrary stocks in $\mathcal{S}_t$ can be defined as a symmetrical distance matrix $\mathcal{M}_t$ of size $|\mathcal{S}_t| \times |\mathcal{S}_t|$:

$$\mathcal{M}_t \in \mathbb{R}^{|\mathcal{S}_t| \times |\mathcal{S}_t|} : m_{i,j} = D(i,j), \quad i,j \in [1:|\mathcal{S}_t|] \quad (3.6)$$

Since the scale of technical indicators in each dimension differ, we normalize the technical indicators before computing DTW distance to make them invariant to different scales Keogh and Kasetty (2002).

[why dtw is better at financial info]

3.7 PCA preprocessing

In this study, we assume that all variables in the firm characteristics dataset have equal impact on the clustering result. Firm characteristics dataset contains different kinds of financial measurements. From quantitative perspective, these measurements have very different levels of dispersion and scale, and they are not comparable directly. If the firm characteristics features are directly used for clustering, variables with higher scale would have more impact than others when the clustering algorithm tries to form clusters by computing the distance measure Gewers et al. (2021). This would cause biased clustering result, which violates our initial assumption of equal impact of every firm characteristics. One solution to this problem is to normalize every variables in the firm characteristics so that all variables have similar scale and dispersion. The normalization is also applied to the dataset of TI characteristics based clustering with the same reasoning. PCA is then performed after the normalization. The purpose of using PCA in this study is threefold:

- In the firm characteristics dataset, there exists some highly correlated variables because they measure the same financial aspect. The same situation exists in the TI characteristics in that some technical indicators are correlated. PCA decorelates those variables by transforming the original feature space to the direction that explains most of the variance.
- PCA reduces the dimensionality of the original dataset and alleviate the curse of dimension in the proceeding k-meoids clustering.
- Given the reduced dimensionality, the computational resources required for the proceeding clustering is alleviated.

In the PCA, the original dataset X with size $n \times p$ is transformed to the new feature space Y with size $n \times q$ where most of the variance is explained by linearly transformed space. Such linear transformation W is expressed as:

$$Y = XW. \tag{3.7}$$

In order to obtain the transformation matrix W, each column of features in the original data X is first centered by subtracting the mean of each column, resulting in matrix $\tilde{X}$. Then the empirical covariance matrix $Cov(\tilde{X})$ with size $p \times p$ of the centered dataset $\tilde{X}$ is computed:

$$Cov(\tilde{X}) = \frac{1}{n-1}\tilde{X^T}\tilde{X}. \tag{3.8}$$

Then the covariance matrix $Cov(\tilde{X})$ is decomposed by the eigenvalue decomposition:

$$Cov(\tilde{X}) = V\Lambda V^T. \tag{3.9}$$

where V is the eigenvector matrix of size $p \times p$ and Λ is the diagonal eigenvalue matrix of size $p \times p$. The transformation matrix W is then obtained by ordering the columns of eigen-vectors in matrix V decreasingly by their eigen-values in matrix Λ and retain the first q columns of eigenvectors. Thus W is the matrix of size $p \times q$ and the dimensionality of the original data matrix X of size $n \times p$ can be reduced from p to q by multiplying matrix W, resulting in dimensionality reduced matrix Y with size $n \times q$.

3.8 K-medoids clustering

Stock clustering is a task that assigns stocks with similar characteristics into the same group and those with dissimilar characteristics into different groups. For this study, the purpose of using clustering method is to identify stocks with homogenous past information so that stocks with temporary anomaly pricing within a cluster can be used to form long-short portfolio to profit from the reverting property of the anomanous stocks. K-means and K-medoids are two popular clustering methods that segments stocks into groups. They are also explainable cluster algorithms in the sense that the processe of representative stock updating and cluster assigning are relatively simple and intuitive.

In K-means, a user-specified number of centriods are first initialized by randomly sampling from the dataset feature space. Then the algorithm finds clusters by iteratively updating centroids of each clusters and assigning all stocks to the cluster with updated cendroids with the aim of minimizing the within-cluster sum of squares (WCSS) of all clusters until the minimization of WCSS reaches convergence. The cendroid of a cluster in K-means is the mean of all stocks of that cluster. Although the computation of centroid is fast in every iterations, k-means is sensitive to stocks that are far way from the main cluster, also known as outlier, because these outlier stocks distort the cendroid of the cluster. In our study, we use the precalculated multi-dimensional DTW distance matrix as input for technical indicator perspective clustering. k-means does not support pre-calculated distance matrix as input because the computation of cluster mean requires stock in coordinate form. k-meoids clustering can tackle the outlier issue by calculating medoid rather than mean of a cluster. On the other hand, since medoids in k-medoids are actual stocks from the training data, unlike centroids in k-means whose computation requires coordinate-based as input, a distance matrix is sufficient for updating the medoids at each iteration. The fact that medoids are actual representative stocks within a cluster increases financial explainability of the algorithm. Based on the above reasons, we opt for k-medoids clustering as the building block for SEFT clustering.

The k-medois clustering works as follow. Given a dataset with n stocks of p features $\mathcal{F} \in \mathbb{R}^{n \times p}$, and the specified number of clusters to be partitioned into k, the pair-wise distance between every two stocks x_i and x_j are first calculated as a distance matrix $\mathcal{M} \in \mathbb{R}^{n \times n} : d_{ij} = ||x_i, x_j||, i, j \in [1 \dots p]$, where $||x_i, x_j||$ is the euclidean distance between stock x_i and x_j. Note that in technical indicator perspective clustering, the distance matrix is replaced with the multi-dimensional DTW distance matrix $\mathcal{M}_t$ as described in Section 3.6. As initialization step, the sum of weighed distance $\mathbf{v}_j$ is calculated as in Equation (3.10) for each stock j. Stocks are then sorted by $\mathbf{v}_j$ in ascending order and the first k stocks with the smallest $\mathbf{v}_j$ value are selected as the initial medoids. Then all stocks are assigned to the cluster with the nearest medoid initialized. This method tend to prioritizes stocks in the middle of the cluster. Then the sum of euclidean distances of every stocks to their assigned medoids (WCSS) is calculated as Equation (3.11).

New medoids are then computed for each cluster to minimize the sum of distance to the rest of the stocks in the cluster as Equation (3.12). The previous medoids of all clusters are replaced by the new medoids.

Each stocks are reassigned to the cluster with the nearest medoid as in Equation (3.13). The WCSS score for all points are updated. If the decrease of WCSS score is 0, then the algorithm stops.

Despite k-medoids is robust to outliers when updating representative stocks for each cluster, it does not remove outliers. Hence, it might increase the portfolio performance by removing outlier stocks in clustering. The outlier removal works as follows. For each stock i, sort the neighborhood stocks by distance in ascending order and obtain the distance at γ percentile as d_i^γ. If the distance to the medoid that the cluster is assigned to dtm_i is larger than the d_i^γ, it will then be considered as outlier and removed. For the main analysis, the outliers are not removed to ensure the same set of candidate stocks are traded every month for better performance comparison. The impact of anomaly removal is analysed in robustness analysis.

3.9 SEFT clustering

Inspired by the TROPICBertrand et al. (2023-05-24) (TRace attributes, cOntrolow Plus Iot Clustering), we propose SEFT (SEctor, Firm characteristics, Technical indicator) clustering algorithm with the goal to enhance pairs trading performance over each single perspective clustering. The SEFT clustering combines information from sector, firm characteristics, and technical indicator, in the expectation that multiperspective information provides better definition of stock homogeneity than single perspective clustering does in price anomaly exploitation. SEFT clustering consists of two stages. For all stocks in $\mathcal{S}_t$ in a given portfolio formation month $t \in \mathcal{T}$, the stocks are first clustered seperatly from sector, firm characteristics, and

Algorithm 1 K-Medoids Clustering Algorithm Park and Jun (2009)

Require: K: number of clusters; $\mathcal{F}$: a data set of n stocks or $\mathcal{M}$: a precalculated distance matrix of $n \times n$.

Ensure: A set of K clusters

Intialize medoids

Calculate euclidean distance matrix of all pair-wise points: $\mathcal{M} \in \mathbb{R}^{n \times n} : d_{ij} = ||x_i, x_j||, i, j \in [1 \dots p]$. If the input is precalculated distance matrix $\mathcal{M}$, it will be directly used for the proceeding computation.

Calculate v_j for stock j as follows:

for $j = 1$ to n **do**

$$v_j = \frac{\sum_{i=1}^{n} d_{ij}}{\sum_{l=1}^{n} d_{il}}, \quad j = 1, \dots, n \tag{3.10}$$

end for

Select k stocks with the smallest v_j values as initial medoids.

Obtain the initial cluster result by assigning each object to the nearest medoid.

Calculate the sum of distances from all stocks to their medoids (WSCC):

$$\text{WSCC} = \sum_{m=1}^{k} \sum_{i \in C_m} d_{im} \tag{3.11}$$

repeat

    Update medoids:

    for each cluster C_m, where $m = 1, \dots, k$ **do**

        Find a new medoid m^* for cluster C_m:

$$m^* = \arg \min_{i \in C_m} \sum_{j \in C_m} d_{ij} \tag{3.12}$$

        Update the current medoid in cluster C_m by replacing it with the new medoid m^*.

    end for

    Assign stocks to medoids

    for each object $i = 1, \dots, n$ **do**

        Assign stock i to cluster C_m where

$$m = \arg \min_{m \in \{1, \dots, k\}} d_{im} \tag{3.13}$$

    end for

    Update WCSS

until WSCC change from previous iteration is zero.

return clustering result

Algorithm 2 Outlier removal for K-meoids Han et al. (2023)

Require: γ: lower γ percentile of neighborhood stocks distance of a stock to be compared to its distance to medoids; $\mathcal{F}$: a data set of n stocks; $\mathcal{I}$: medoids of all clusters

Ensure: Updated dataset with outliers removed $\mathcal{F}^*$

1: **for** each stock i **do**
2: Calculate distance of stock i to its cluster centroid as dtm_i.
3: Calculate lower γ quantile distance of all its neighborhood stocks as d_i^γ.
4: **if** $\text{dtm}_i > d_i^\gamma$ **then**
5: remove stock i
6: **end if**
7: **end for**
8: **return** Updated dataset with outliers removed $\mathcal{F}^*$

technical indicator perspectives using k-medoids. With the clustered stocks from a single perspective, their distance to medoids (DTM) to each clusters from that perspective is derived. Note that the number of medoids for each perspectives is equivelent to the respective K used for the K-medoids clustering. For the second stage, the DTM calculated from each perspectives are combined as the new feature for final clustering. Figure 3.2 demonstrates the workflow of SEFT clustering

For sector characteristics clustering, stocks with the same SIC code are considered as in the same cluster. The DTM from sector perspective is simply the one-hot coded dataset as described in Section 3.3 with the 0/1 flipped in the sense that the distance between the SIC of a given stock and the SIC of another stock is zero only when the latter stock share the same SIC code and one when they have differnet SIC code. It is denoted as $DTM_t^S \in \mathbb{R}^{|\mathcal{S}_t| \times |SIC_t|}$, where SIC_t is the set of unique SIC code assigned to all stocks in $\mathcal{S}_t$.

For firm characteristics clustering, some features such as market capitalization would have higher scale than other features such as employee growth rate. Hence, each firm characteristics feature is first preprocessed with MiniMax scalling in order to align to the same scale. Then the rescaled features are proceeded for PCA to extract principle components. We choose the same number of principle components p^F for every month to make results from different experiment scenarios more comparable. It turns out that $p^F = 20$ gives the optimal performance for SEFT clustering. The extracted principle components are then proceeded for K-medoids clustering to extract DTM of each stock to every medoids, denoted as $DTM_t^F \in \mathbb{R}^{|\mathcal{S}_t| \times k^F}$. For the same reason as choosing p^f for PCA, we choose the same k^F for k-meoids in every momnth. We choose $k^F = 75$ for every months as it gives the optimal performance for SEFT clustering. Consequently, there are 75 DTMs for each stocks in $\mathcal{S}_t$ in each portfolio formation month t.

For technical indicator clustering, we use the DTW distance matrix derived as

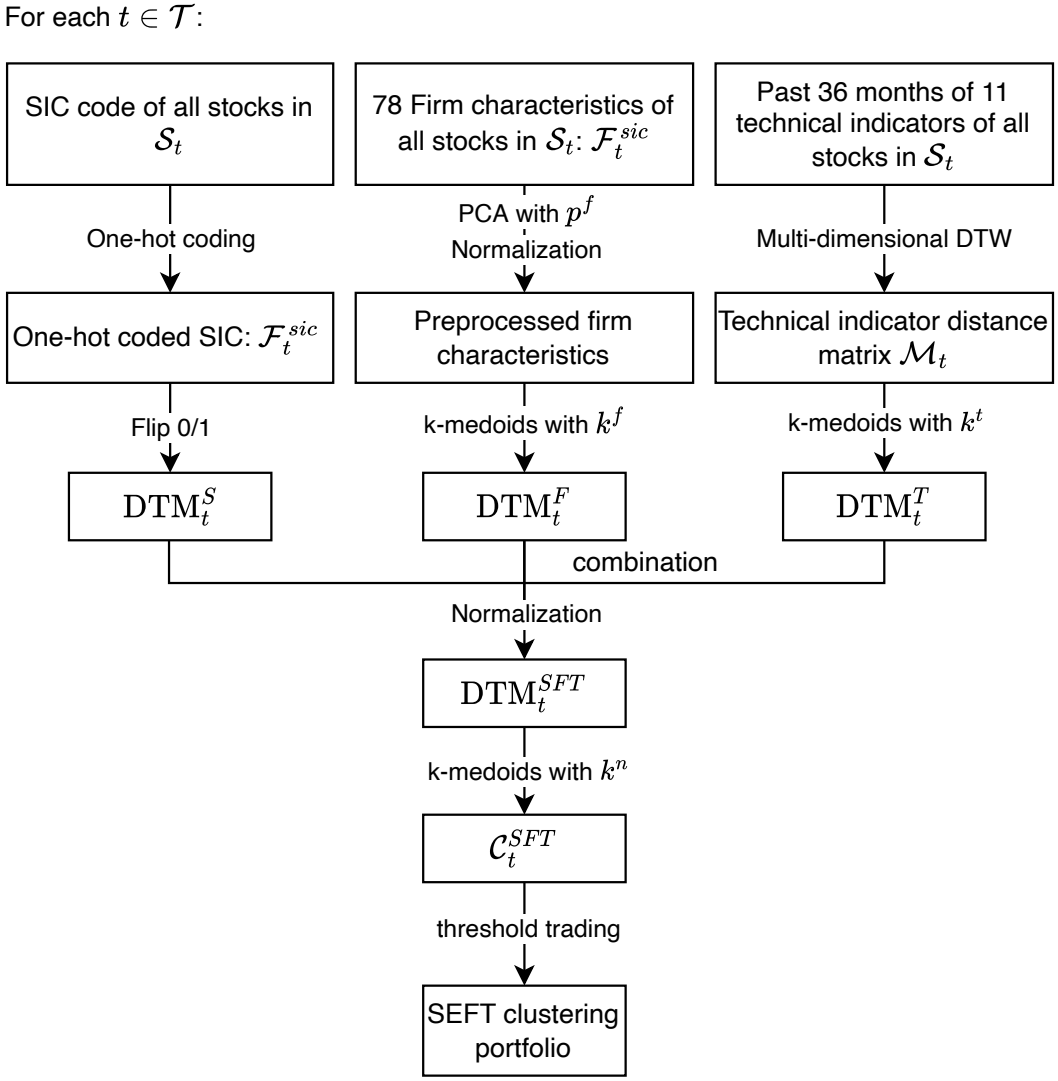

FIGURE 3.2: Workflow of SEFT clustering.

described in Section 3.6 as input for k-medoids clustering using the same k^T every month. We choose $k^T = 30$ as it gives the most promising trading performance. Hence for each stock in $\mathcal{S}_t$, there are 30 DTMs in each portfolio formation month t, denoted as $DTM_t^T \in \mathbb{R}^{|\mathcal{S}_t| \times k^T}$

In the second stage, the DTM features derived from three perspectives from each month t are combined and normalized to form a new dataset $DTM_t^{SFT} \in \mathbb{R}^{|\mathcal{S}_t| \times (|SIC_t| + k^F + k^T)}$. The aim of normalization is to enforce equal importance on every DTM features. The resulting dataset is then proceeded for the final k-meoids clustering using the same k^n for every month. We choose $k^n = 200$ as its resulting portfolio has the optimal trading performance. The resulting clustering result is then used to form SEFT clutsering long-short portfolio using threshold strategy as described in Section 3.10.

3.10 Trade with threshold method

Given a set of stocks with clustering label in a given portfolio formation month, we use the threshold trading strategy to form long-short portfolio Han et al. (2023). Algorithm 3 describes the procedure of forming long-short portfolio with threshold trading strategy. In a portfolio formation month t, stocks belonging to the same cluster, as defined in a given clustering result $\mathcal{C}_t$, are sorted based on their monthly return prior to the formation month (mom1) as described in Line 8. The stock with the highest mom1 is paired with the stock having the lowest mom1. Similarly, the stock with the second highest mom1 is paired with the stock having the second lowest mom1, and so on as in Line 9. The mom1 return gap of every pairs are calculated as in Line 10. The standard deviation of mom1 return gap of all pairs are calculated as in Line 11, and only pairs with return gap exceeding a percentage α of standard deviation are traded. When a pair is traded, the lower mom1 stock is longed and a higher mom1 stock is sold as the former is deemed undervalued, and the latter is deemed overvalued. Such trade is done for all clusters with all traded stocks equally weighted in terms of invested value. The whole portfolio is cashed out one month after the portfolio formation, and the remaining cash balance is reinvested for the next round trading using the same strategy. In the main analysis, based on the model specifications in Section 3.9, we choose the threshold $\alpha = 0.6$, which results in the optimal portfolio performance. we also ignore the trading cost incurred. In the robust analysis, the effect of trading cost is analysed.

3.11 Single perspective clustering portfolio

The purpose of forming portfolio using single perspective is to have a better understanding on the role and impact of clustering from seperate perspective in SEFT clustering. When the portfolio is formed by sector only clustering, stocks with the identical SIC code is considered as in the same cluster, and the threshold trading

Algorithm 3 Threshold trading strategy

Require: $\mathcal{T}$: tradable formation months; $\mathcal{S}$: set of tradable candidate stocks in all formation months; $\mathcal{C}$: cluster mapping of stocks in $\mathcal{S}$ in all formation months; fee: trading fee; α: trading threshold.

Ensure: R: monthly return of all traded stocks in the whole trading period.

for each formation month $t \in \mathcal{T}$ **do**
  if t=$\mathcal{T}$`[0]` **then** ▷ if on the first trade date
    Set initial cash \$1
  end if
  for each cluster $i \in [1, 2, \ldots, k]$ **do**
    $\mathcal{S}_t^i \leftarrow$ `getclu_i`$(\mathcal{S}_t, \mathcal{C}_t, i)$ ▷ identify the set of stocks in cluster i
    $\mathcal{R}_t \leftarrow$ `getmom1`$(\mathcal{S}_t^i)$ ▷ obtain prior one month-return of all stocks in $\mathcal{S}_t^i$
    $\mathcal{R}_t^* \leftarrow [r_1, r_2, \ldots, r_m], r_1 < r_2 < \cdots < r_m$ ▷ arrange stocks in $\mathcal{S}_t^i$ with prior month return $\mathcal{R}_t$ in ascending order.
    $\mathcal{P} \leftarrow \left\{(r_{(i)}, r_{(m-i+1)}) \mid i = 1, 2, \ldots, \lfloor \frac{m}{2} \rfloor\right\}$ ▷ pair stock with the highest return with the one having the least return, second highest return with second least return etc.
    $SP \leftarrow \left\{r_{(i)} - r_{(m-i+1)} \mid i = 1, 2, \ldots, \lfloor \frac{m}{2} \rfloor\right\}$ ▷ calculate the return spread of each pair.
    $\sigma \leftarrow$ `std_dev`(SP) ▷ calculate the standard deviation of prior one month return spreads for all pairs.
    for each pair $(r_{(i)}, r_{(m-i+1)}) \in \mathcal{P}$ **do**
      if $r_{(i)} - r_{(m-i+1)} > \alpha\sigma$ **then**
        Long stock i and short stock $m - i + 1$ with equal weights and trading fee deducted;
      end if
    end for
  end for
  Hold all positions for 1 month
  Cash out all held positions and calculate the aggregated return R_t from all traded stocks.
end for
return R: monthly return of all traded stocks in the all trading periods.

strategy is directly applied to the resulting cluster. When the portfolio is formed by firm characteristics-only clustering, the firm characteristics is first processed with MiniMax scaling, and then proceeded to PCA using the same number of components $p^f = 20$ as in Section 3.9. Next, the preprocessed dataset is used as input for k-meoids clustering. The resulting clustering result can then be used for threshold trading strategy. When the portfolio is formed by technical indicator-only clustering, the precalculated multidimensional distance matrix is used as input for k-meoids clustering and the resulting clustering can be proceeded for threshold trading strategy. Figure 3.3 (i., ii., iv.) shows the workflow of forming single perspective S, F, and T clustering portfolios.

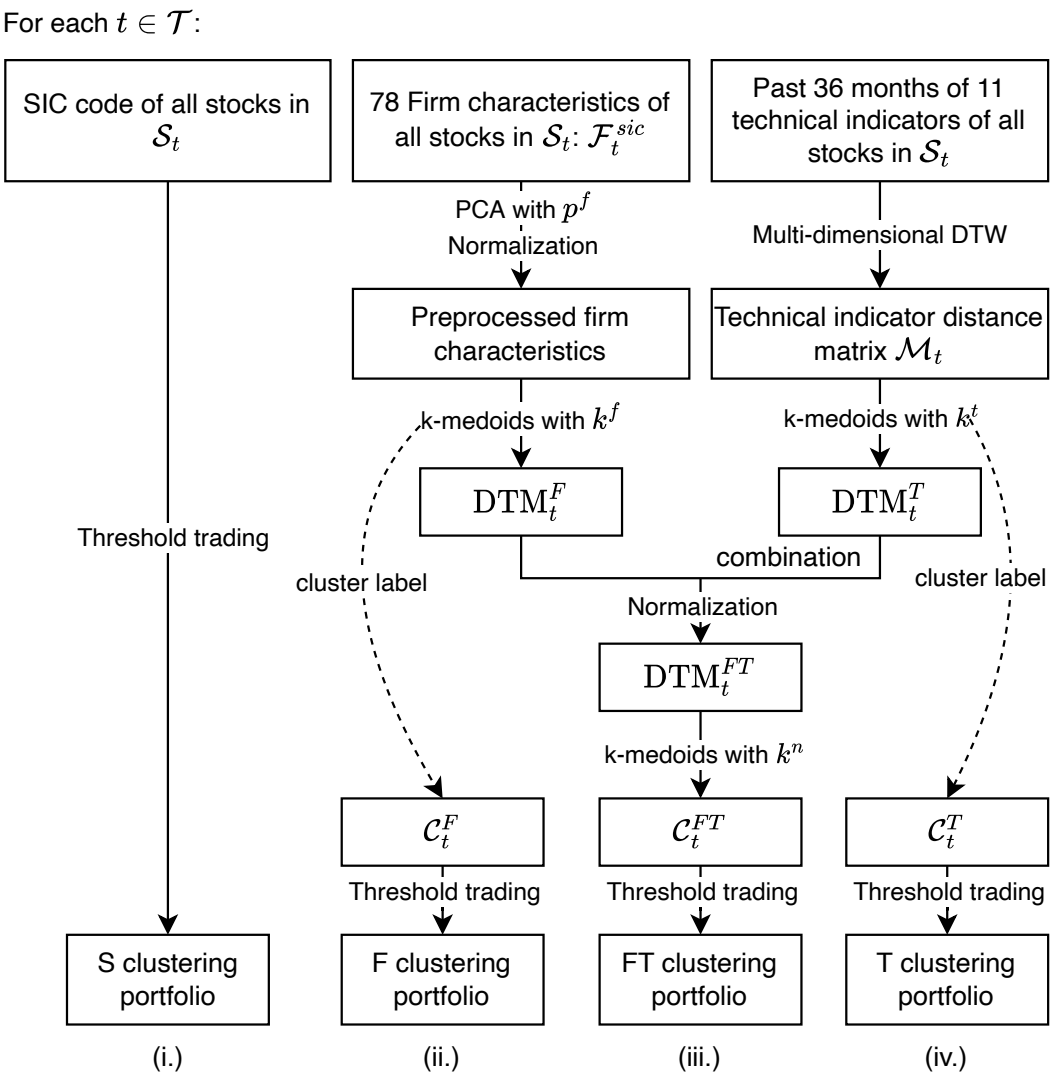

Figure 3.3: Workflow of single perspective clustering (i., ii., iv.) and dual perspective FT clustering (iii.).

3.12 Dual perspective clustering portfolio

It is also interesting to dissect the tri-perspective SEFT clustering into single perspective and dual perspective clustering. Hence, we also form portfolio using dual-perspective clustering and compare it with the missing single-perspective clustering to see the impact on the SEFT clustering from both sides. Before two perspectives are combined for clustering, the distance to meoids (DTM) from both perspectives are first calculated the same way as SEFT clustering does as described in Section 3.9. Then the DTM from two perspectives are combined, normalized and clustered the same way as described in Section 3.9. The resulting dual-perspective clusterings are sector-firm characteristics perspective (SF), sector-technical indicator perspective (ST) and firm characteristics-technical indicator perspective (FT) clusterings. They are then proceeded to threshold trading strategy. Figure 3.3 (iii.) shows the workflow of dual-perspective FT clustering.

3.13 Benchmark portfolios

The benchmark portfolio is used to compare our trading strategies against the main market evolution. We use the portfolio formed by S&P500 index and reversal strategy in which candidate tradable stocks in $\mathcal{S}_t$ are longed if the prior mom1 is below the median mom1 and shorted if prior mom1 is above the median mom1.

Chapter 4

Empirical analysis

In this chapter, we first analyze the clustering characteristics of SEFT methodology versus single perspective clustering for each month using the parameters calibrated as described in Section 3.9. Then we compare the trading performance using SEFT clustering versus other clustering and benchmark performance. A deep-dive in sub-period performance is then provided. Next we do a factor regression to see the impact of market dynamics on the trading performance. Note that the parameters used for single, dual perspectives, and SEFT clustering in the aforementioned analysis are identical and based on the optimal parameters for SEFT clustering as specified in Section 3.8 and Section 3.10. To determine whether the relatively better performance of SEFT clustering is sensitive to parameter variations, we also conduct a robustness analysis across different choices of parameters. Note that in this chapter, typically in tables and illustrations, S stands for sector perspective clustering, F for firm characteristics' perspective clustering and T being technical indicator perspective clustering. When two letters are combined, it stands for dual perspectives clustering using perspectives that the two letters respectively stands for. SFT clustering is our proposed SEFT clustering which perform clustering using all three perspectives.

4.1 Cluster characteristics

For each month $t \in \mathcal{T}$, stocks in $\mathcal{S}_t$ are consistently segmented into 200 clusters with SEFT clustering as specified in Section 3.8. Sector and firm characteristics based clustering seem to form larger cluster than technical indicator and SEFT clustering as seen from Figure 4.1. The largest cluster has twice number of stocks than the second-largest cluster from 2003 to 2011 for sector based clustering. Sector based clustering forms larger cluster because sector has more uneven segmentation over the candidate stocks than clusters defined from other perspectives. Since PCA is applied before firm characteristics is used for clustering, dimensionality is reduced and thus within cluster distance reduced. Thus, larger cluster is formed when data points become closer.

Delving deeper into the sector composition of the largest four clusters from

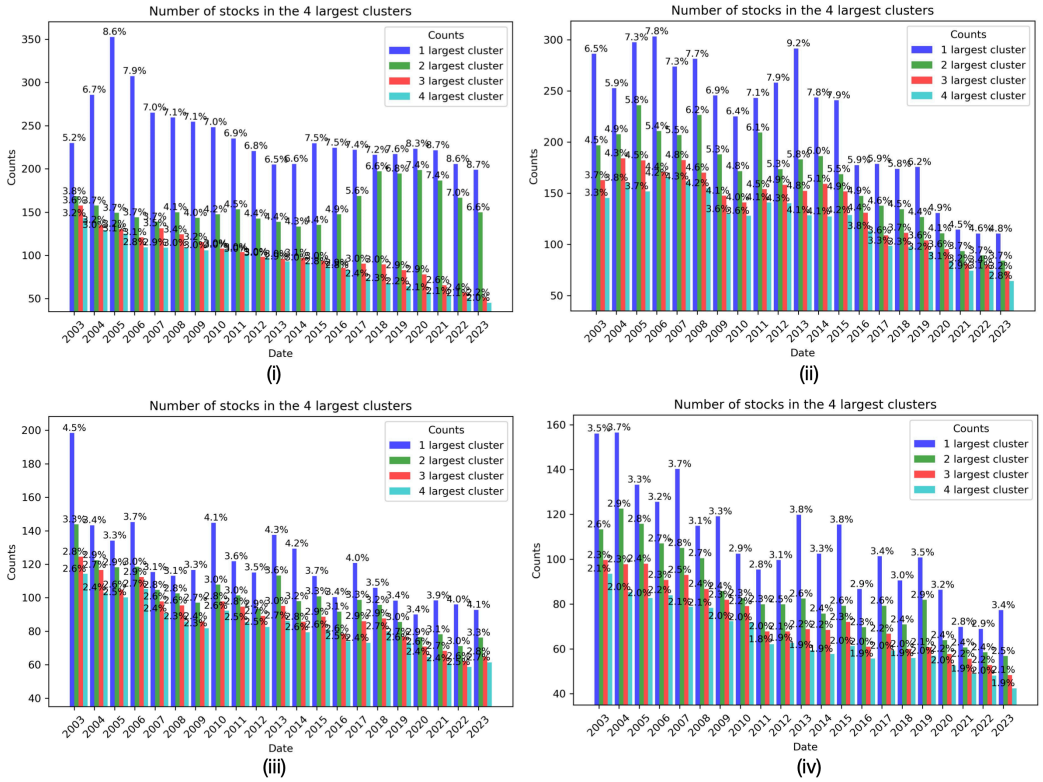

Figure 4.1: Clusters with the top 4 average number of stocks from (i.) sector based clustering, (ii.) firm characteristics based clustering (iii.) technical indicator based clustering and (iv.) SEFT clustering and their percentages to the whole candidate stocks pool.

sector-based clustering, as described in Figure 4.1, we observe that the commercial banking sector consistently has the highest number of stocks across all trading periods. This is followed by the pharmaceutical preparation manufacturing, software publisher, and semiconductor and related device manufacturing sectors, which shuffle back and forth in rank. This pattern closely aligned with the evolution of sector composition of the US stock market. The COVID pandemic seemed to boost the growth of surgical and medical instrument manufacturing sector as it appeared in the top 4 ranking as of 2020.

To deepen the understanding on the interaction among the three singular perspectives, we explore pairwise similarity between two of the three perspectives for each portfolio formation month. Adjusted Mutual Information (AMI) provides measurement for the similarity between two clustering results. It is an ideal similarity metric for measuring pairwise similarity among three perspectives because it does not depend on the absolute label, and it is adjusted for chance. Figure 4.2 shows

Year	1st	2nd	3rd	4th
2003	CB	SI	SP	PPM
2004	CB	SI	SP	PPM
2005	CB	SI	SP	PPM
2006	CB	SP	SI	PPM
2007	CB	SP	PPM	SRDM
2008	CB	PPM	SP	SRDM
2009	CB	PPM	SP	SRDM
2010	CB	PPM	SRDM	SP
2011	CB	PPM	SRDM	SI
2012	CB	PPM	SRDM	SI
2013	CB	PPM	SRDM	SI
2014	CB	PPM	SRDM	SP
2015	CB	PPM	SRDM	SP
2016	CB	PPM	SP	SRDM
2017	CB	PPM	SP	SRDM
2018	CB	PPM	SP	SRDM
2019	CB	PPM	SP	SRDM
2020	CB	PPM	SP	SMIM
2021	CB	PPM	SP	SMIM
2022	CB	PPM	SP	SMIM
2023	CB	PPM	SMIM	SRDM

TABLE 4.1: The sectors of the top 4 clusters in sector based clustering. The abbreviations of sector names are PPM:Pharmaceutical Preparation Manufacturing CB:Commercial Banking SI:Savings Institutions SRDM:Semiconductor and Related Device Manufacturing SMIM:Surgical and Medical Instrument Manufacturing SP:Software Publishers CB:Commercial Banking SI:Savings Institutions

that clustering result between sector and firm characteristics share the highest and the most volatile similarity than two other pairs over time. Clustering between firm characteristics and technical indicator has relatively smaller similarity and the smallest between sector and technical indicator. Possible explanation is that companies within the same sector tend to exhibit more similar firm characteristics and stock market evolution, thus technical indicator, and there's less explicit link between firm characteristics and technical indicator.

With the fixed threshold optimized for SEFT clustering, around 30% to 40% of the stocks are chosen to form long-short portfolio in each cluster. This can be reflected in Figure 4.3, where the average number of stocks traded comprises approximately 30% to 40% of the average total candidate stocks per year.

4.2 Trading performance analysis

Table 4.2 summarizes the monthly return statistics and annually adjusted trading performance metrics of the long-short portfolio using benchmark clustering, SEFT clustering and other perspective clustering over the whole trading period. Table A.1 and Table A.2 are respectively the return statistics and trading performance metrics of the long and short leg of the portfolio. Figure 4.4 is the box plot describing the

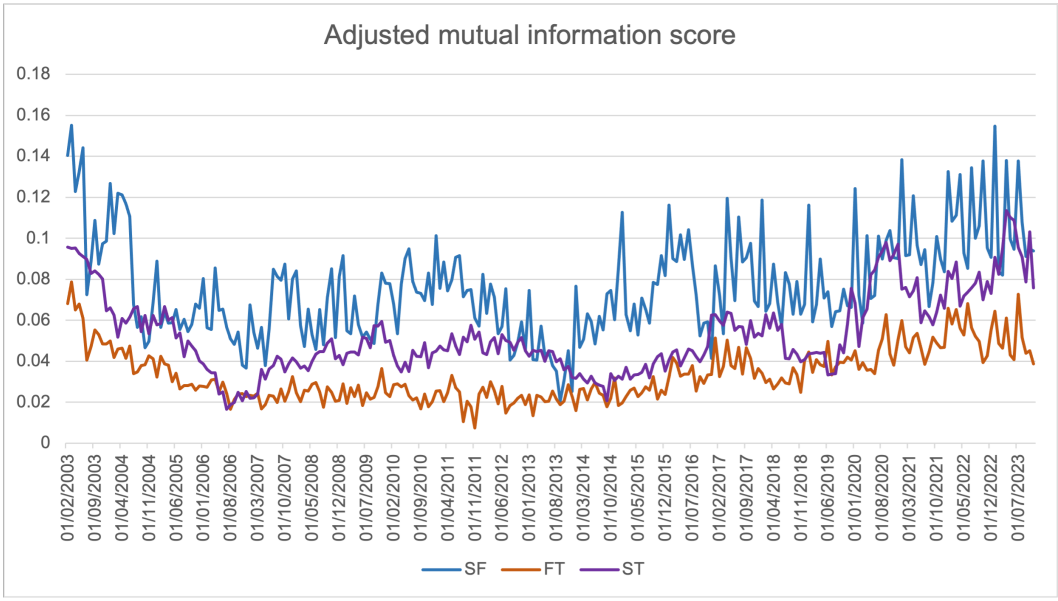

Figure 4.2: Evolution of adjusted mutual information score between sector and firm characteristics(SF), firm characteristics and technical indicator(FT), sector and technical indicator (ST) based clustering.

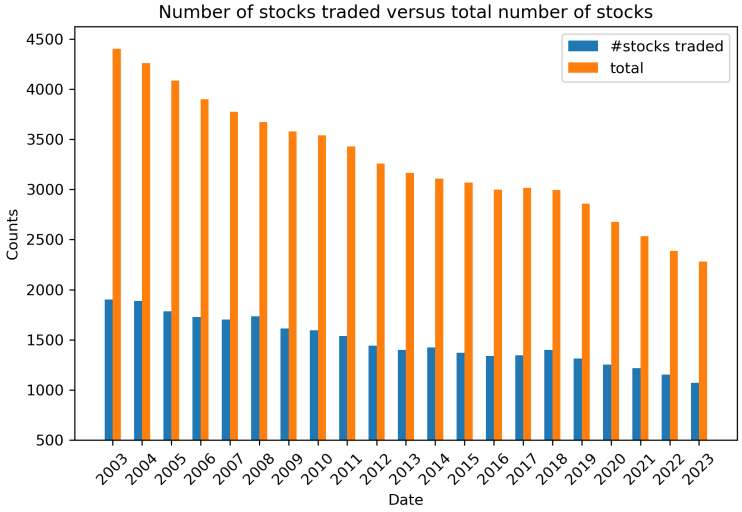

Figure 4.3: Average number of stocks traded per year

return distribution of portfolios formed by different clustering methods. Figure 4.5 is the logarithm of cumulative return of portfolios formed by different clustering methods. We use the following metrics as the main measurement for strategy performance comparison.

- Monthly return $= r_p$
- Monthly risk-free rate $= r_i$ [1]
- Monthly return standard deviation $\sigma_p = \sigma(r_p)$
- Shape ratio $= \frac{r_p - r_i}{\sigma_p} \times \sqrt{12}$
- Downside Deviation$\sigma_d = \sigma(r_p), r_p < 0$
- Sortino Ratio $= \frac{r_p - r_i}{\sigma_d}$
- Gross Profit $= \sum r_p, r_p > 0$
- Gross Loss $= \sum r_p, r_p < 0$
- Profit Factor $= \frac{\text{Gross Profit}}{|\text{Gross Loss}|}$
- Max Drawdown(MDD): Maximum loss occurred since the portfolio value reached a peak during the whole trading period.
- Calmar Ratio $= \frac{\text{Annualized Return}}{|\text{MDD}|}$
- Turnover $= \sum |r_p|$

In the long-short portfolio, the ST clustering has the highest annualized mean return of 0.23, but also the highest annualized volatility of 0.222 among other clustering methods, thus making it less preferred taking the risk characteristics into account. From Figure 4.4 and Figure 4.5, it can be seen that the portfolio formed by ST clutsering experience several acute fluctuations and has wider range of fluctuations compared to other perspective clusterings. The reverse portfolio has the least annualized volatility of 0.073 among all strategies, but the second lowest annualized mean return of 0.108, making it a safer portfolio to invest, but unattractive in profitability because of its low mean return.

The SEFT clustering seems to be the most optimal and promising strategy in balancing both the portfolio return stability and growth as it has the highest Sharpe ratio of 1.46 and highest t-statistics of 8.4 among all the strategies. The high Sharpe ratio seen in SEFT clustering is mainly because of much reduced volatility compared to the volatility seen in its single and dual perspective counterparts.

[1] We use 3.5% as anual risk-free rate, approximately equivalent to monthly adjusted rate of 0.28%.

For single perspective portfolio, all three perspectives has similar annualized return of 0.15, with sector perspective having the lowest volatility and technical indicator perspective having the highest volatility. By combining two single perspectives into dual perspective, the effect on return and volatility differ depending on different combination. When sector and firm characteristics perspectives(SF) are combined, the mean return falls in between the mean returns of two single perspectives and the volatility is considerably lower than both perspectives. Since the volatility reduction is larger than the mean return change, the Sharpe ratio with SF perspectives combined is seen considerably increased than both of the single perspective Sharpe ratios. When firm characteristics and technical indicators perspectives(FT) are combined, both the mean return and volatility are reduced. Since the decrease in return is trivial compared to the decrease in volatility, the Sharpe ratio is slightly larger than both perspectives. A considerable increase in both mean return and volatility is seen with sector perspective and technical indicator(ST) perspective combined. The resulting Sharpe ratio of ST perspective clustering falls in between the Sharpe ratio of S and T perspective.

From the above evidence, we found that adding firm characteristics to both sector perspective and technical indicator perspective clustering can reduce the volatility of the single perspective clustering. By incorporating firm characteristics into ST clustering, the Sharpe ratio increases substantially from 0.884 to 1.461. The volatility decreases from 0.22 to 0.09, which is more than a 50% reduction, with only a minor sacrifice in the mean return. The downside risk and maximum drawdown of ST clustering are also reduced substantially by incorporating firm characteristics. This is consistent with the finding of Han et al. (2023) that excluding firm characteristics deteriorates trading performance and firm characteristics improves the identification of homogenous stocks in long-short portfolio pairs trading. Other than incorporating firm characteristics to ST perspective clustering, the combination of any dual perspective clustering with the single perspective not contained in the dual perspectives reduces the volatility as seen in Table 4.2. The mean return of SF clustering is increased by incorporating technical indicator perspective and so does the mean return of FT clustering by incorporating sector perspective. By comparing SEFT clustering with three single perspective clusterings, the mean return of SEFT is larger than all three single perspectives, and the volatility of SEFT is lower than all three single perspectives. All evidence shows that SEFT clustering combines three perspectives effectively and comprehensively in better idenfitying similar stocks for long-short portfolio pairs trading.

4.3 Sub-period analysis

We divide the trading period of 21 years into 4 consecutive sub-periods to compare the trading performance of portfolios formed in each sub-periods. The trading performance statistics of each sub-period is summarized in Table 4.3. The cumulative

	Single perspective			Dual perspective			SEFT	Benchmarks	
Monthly return statistics	**S**	**F**	**T**	**SF**	**FT**	**ST**	**SFT**	**reverse**	**sp500**
Mean Return	0.012	0.013	0.012	0.013	0.012	0.019	0.014	0.009	0.008
Standard Deviation	0.034	0.035	0.041	0.031	0.029	0.064	0.026	0.021	0.043
Standard Error	0.002	0.002	0.003	0.002	0.002	0.004	0.002	0.001	0.003
t-statistic	5.714	5.872	4.854	6.594	6.371	4.742	8.406	6.765	2.970
Min	-0.039	-0.129	-0.080	-0.054	-0.037	-0.107	-0.044	-0.034	-0.169
0.25	-0.003	-0.001	-0.003	-0.004	-0.003	-0.003	-0.002	-0.002	-0.016
0.5	0.004	0.006	0.005	0.007	0.006	0.007	0.009	0.005	0.012
0.75	0.017	0.018	0.016	0.021	0.019	0.025	0.022	0.014	0.034
Max	0.376	0.313	0.328	0.220	0.263	0.753	0.140	0.114	0.127
Skewness	5.964	4.650	4.652	2.651	3.894	7.100	1.920	2.076	-0.602
Kurtosis	54.706	36.084	28.539	12.389	25.672	71.777	5.451	6.873	1.340
VaR(95%)	-0.014	-0.013	-0.018	-0.017	-0.016	-0.022	-0.014	-0.015	-0.073
ES(95%)	-0.019	-0.029	-0.034	-0.032	-0.023	-0.049	-0.022	-0.022	-0.097
Annualized risk-return metrics	**S**	**F**	**T**	**SF**	**FT**	**ST**	**SFT**	**reverse**	**sp500**
Annualized Return	0.147	0.157	0.149	0.155	0.140	0.230	0.167	0.108	0.096
Annualized Vol	0.117	0.122	0.141	0.107	0.100	0.222	0.091	0.073	0.148
Sharpe Ratio	0.958	1.005	0.818	1.124	1.052	0.884	1.461	1.010	0.418
Downside Deviation	0.007	0.016	0.013	0.012	0.008	0.020	0.008	0.008	0.033
Sortino Ratio	4.936	2.276	2.598	3.025	3.934	2.848	4.784	2.747	0.548
Gross Profit	3.701	3.949	4.020	4.142	3.649	5.932	4.116	2.914	5.175
Gross Loss	-0.645	-0.671	-0.907	-0.910	-0.738	-1.136	-0.639	-0.663	-3.168
Profit Factor	5.734	5.885	4.434	4.551	4.942	5.222	6.438	4.396	1.633
Profitable Years	21	21	20	20	20	19	20	20	17
Unprofitable Years	0	0	1	1	1	2	1	1	4
Max Drawdown	0.053	0.164	0.145	0.091	0.045	0.175	0.056	0.079	0.526
Calmar Ratio	2.760	0.958	1.027	1.710	3.088	1.317	2.976	1.374	0.183
Turnover	4.346	4.620	4.926	5.052	4.388	7.068	4.755	3.577	8.343

TABLE 4.2: Long-short portfolio monthly return statistics and annualized trading performance metrics using different clustering methods.

value of portfolios in each sub-period is shown in Figure 4.7.

The first sub-period spans from 2003 to 2007. From Table 4.3, portfolio formed by SEFT clustering has the highest Sharpe ratio of 1.137 among all the strategies. This shows that SEFT clustering has the best balance between both return and risk compared to other clustering and benchmark portfolio in this period. We also notice a negative Sharpe ratio in reverse portfolio. This indicates that the mean return of the long-short portfolio formed by reversal strategy is lower than the interest free rate during the first sub-period. We see that the S&P 500 portfolio has the highest annualized mean return, but also the highest volatility among all portfolios. This makes the S&P 500 portfolio a less prudent strategy. From Figure 4.7, we can visually observe that the cumulative return of S&P 500 portfolio is the most turbulent among all strategies, despite of its highest mean return. Considering S&P 500 reflects the overall market condition, SEFT clutsering portfolio can effectively absorb and smooth out several market turbulences in the sub-period. During the year of 2007 when it was the onset of financial crisis, two severe oscillations in cumulative return are observed on S&P 500 portfolio and the portfolio ended up ticking down

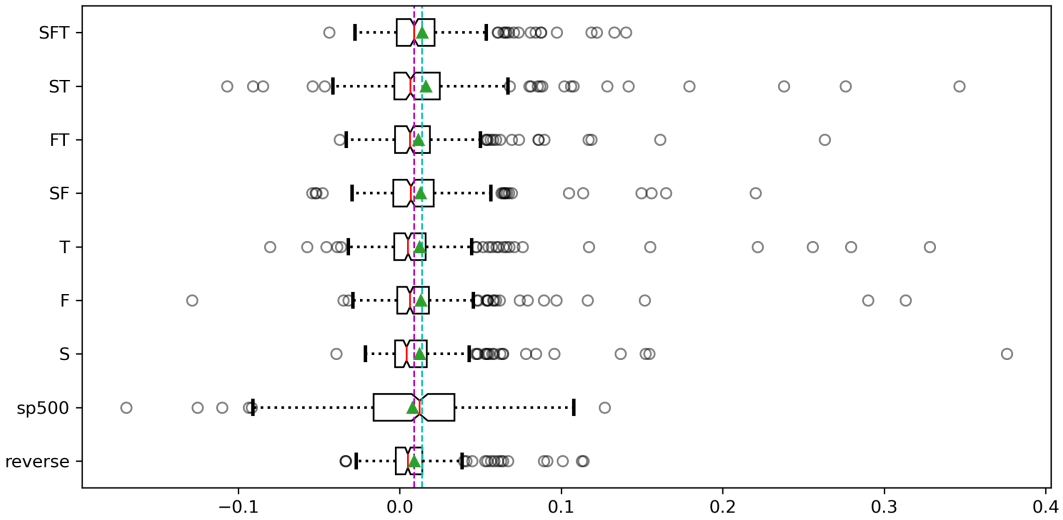

Figure 4.4: The return distribution of portfolios formed by different clustering methods. The magenta and cyan dashed vertical lines mark the horizontal position of the median and mean return of portfolio formed by SEFT clustering respectively for better visual comparison to other return distributions.

for 16 % within only two months from October 2007 to the year-end. During the severe downturn of US stock market, the perspective clustering portfolios made it to absorb the shock and even made positive return. This is primarily because the short portfolio profited more from the declining market trend than the losses incurred by the long portfolio, as illustrated in Figure 4.6.

The second sub-period spans from 2008 to 2012, during which severe market downturn happened at the beginning of the period. We see a continued substantial drop of the overall market as indicated by S&P 500 portfolio in the whole year of 2008 from Figure 4.7. Perspective clustering portfolios greatly absorbed the market downside impact as their drops compared to the drop seen in S&P 500 portfolio are trivial. From Figure 4.6, we can see that the loss from the long portfolio in the year 2008 is greatly compensated by the profit from the short portfolio, resulting in limited loss from the market downturn. After the great downturn, The S&P 500 experienced several fluctuations onwards from the year of 2009 with upwards trend in general. All clustering portfolios exhibit steady growth and seem to alleviate downtrend impact from the several market fluctuations. From the long and short portfolios trend in Figure 4.6, we see that the long portfolio roughly moves in tandem with S&P 500 portfolio, whereas short portfolio has the opposite trace. This shows that the short portfolio can absorb the downside market shocks in the long-short portfolio. Looking into the performance statistics in Table 4.3, SEFT clustering maintained its superiority in balancing between growth and prudence as we can see from its highest Sharpe ratio of 1.58 among all portfolios. It has the

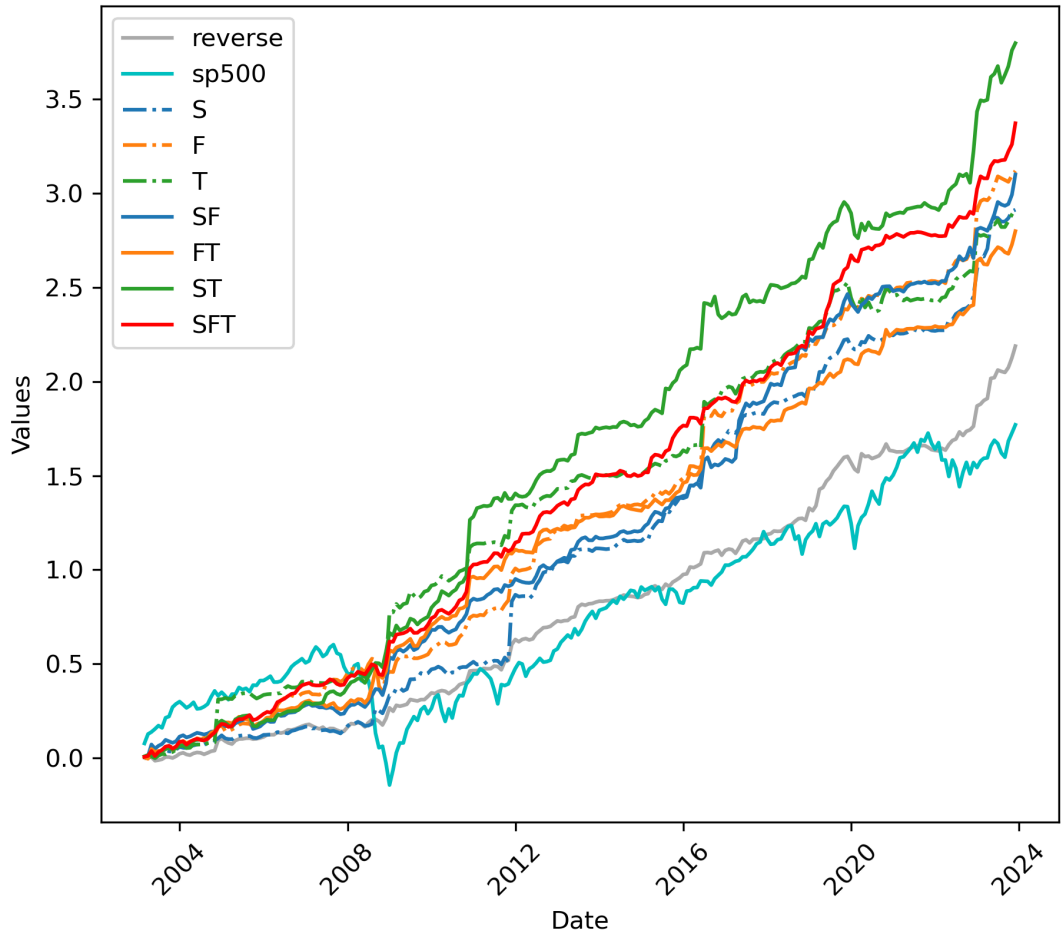

FIGURE 4.5: The logarithm of cumulative returns of portfolios formed by different clustering methods over the whole trading period, starting from $1 in Feburary 2003.

highest profit factor of 7.17, indicating the highest ratio of total gross profit to total gross loss. S&P 500 portfolio had only Sharpe ratio of 0.05 as it's totally exposed to the market downturn condition caused by the financial crisis.

The third sub-period spans from 2013 to 2017, during which the market exhibits more rapid growth than the previous sub-period. The SEFT clustering portfolio performed well with a Sharpe ratio of 1.36, only surpassed by the sector clustering portfolio, which had a Sharpe ratio of 1.45. The superiority of sector clustering in this sub-period indicates that short-term price anomaly corrections are more effectively and consistently identified within the same sector than within same characteristics from other perspectives. From economic perspective, stocks within the same sector are more impacted by the same economic driver, such as industrial regulations or homogenous market behavior within the same sector, than other clusterings. When

sector characteristic is combined with characteristics from other perspectives, the performance might be comprised as other characteristics might distort the identification of homogenous stocks and short term price anomaly.

The fourth sub-period spans from 2018 to 2023. A major market event was COVID-19 pandemic starting from the year 2020 and overshadowed the market for the next 2 years onwards. The pandemic had a great impact second to the financial crisis in 2008 as reflected from the S&P 500 portfolios droping by 30% from February 2020 to the valley bottom in March. Clustering portfolios were hit by the pandemic at the same time, but less severe than the S&P 500 portfolio. For S, F, FT, and SF clustering portfolios, the impact from pandemic seems relatively trivial compared to other perspectives. This indicates that firm characteristics and sector characteristics helped identify short-term price anomalies and their short-leg portfolios hedged the market downturn. This is probably because the impact from the pandemic to a company can be instantly reflected from its firm characteristics such as financial conditions or accounting information, whereas technical indicators had a slower response to the impact from the pandemic. The change in firm characteristics can primarily originate from sudden drop in market behavior from demand side and strict government regulations from production side. The impact from pandemic can also be sector dependent, meaning that the pandemic regulations were imposed on certain sectors and should have homogenous impact on companies belonging to the same sector. This explained the reason for sector characteristics being able to identify price anomalies as well. With three perspectives combined, SEFT clustering outperformed other clustering portfolios with the highest Sharpe ratio of 1.74. By combining three individual perspectives, the mean return of SEFT clustering is significantly improved compared to single-perspective clustering portfolios without too much increase on volatility.

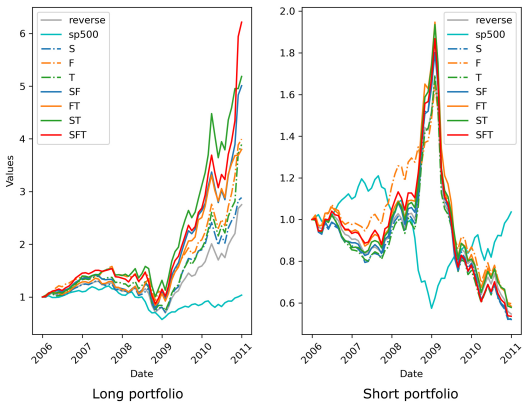

Figure 4.6: Close-up look between the cumulative return of long and short portfolio during the years from 2006 to 2010 starting at $1.

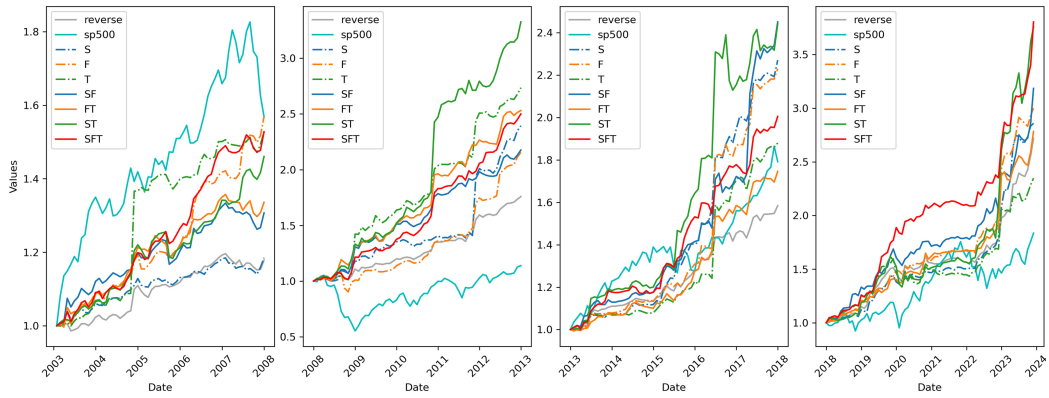

FIGURE 4.7: The cumulative value of different portfolios in 4 sub-periods with each period and strategy starting at $1.

Sub-period	Risk metrics	reverse	sp500	S	F	T	SF	FT	ST	SFT
2003-2007	Annualized Return	0.034	0.096	0.035	0.093	0.092	0.056	0.060	0.079	0.088
	Annualized Vol	0.038	0.092	0.031	0.052	0.118	0.056	0.048	0.054	0.047
	Sharpe Ratio	-0.008	0.669	0.019	1.127	0.490	0.387	0.539	0.816	1.137
	Sortino Ratio	-0.018	1.202	0.031	2.207	3.021	0.556	0.995	0.971	1.822
	Profit Factor	2.187	2.111	2.306	4.709	4.965	2.347	2.648	3.643	3.752
	Max Drawdown	0.037	0.141	0.038	0.032	0.028	0.054	0.045	0.057	0.031
	Calmar Ratio	0.921	0.681	0.935	2.927	3.292	1.037	1.329	1.374	2.821
2008-2012	Annualized Return	0.116	0.044	0.189	0.162	0.216	0.161	0.193	0.255	0.189
	Annualized Vol	0.068	0.189	0.180	0.123	0.175	0.094	0.113	0.173	0.098
	Sharpe Ratio	1.200	0.050	0.859	1.035	1.038	1.341	1.405	1.276	1.581
	Sortino Ratio	3.833	0.065	6.970	1.155	6.735	4.611	4.831	6.718	3.940
	Profit Factor	4.932	1.191	5.479	4.010	6.303	4.967	6.028	7.114	7.178
	Max Drawdown	0.030	0.475	0.044	0.164	0.039	0.031	0.040	0.031	0.056
	Calmar Ratio	3.857	0.093	4.347	0.986	5.543	5.123	4.819	8.131	3.374
2013-2017	Annualized Return	0.094	0.122	0.169	0.171	0.134	0.188	0.115	0.192	0.143
	Annualized Vol	0.055	0.098	0.093	0.150	0.133	0.132	0.074	0.163	0.079
	Sharpe Ratio	1.086	0.894	1.455	0.908	0.753	1.169	1.084	0.963	1.367
	Sortino Ratio	2.290	1.473	9.445	6.306	4.108	3.546	5.013	2.261	4.522
	Profit Factor	4.725	2.491	8.857	9.434	5.175	6.246	4.612	4.497	5.621
	Max Drawdown	0.027	0.089	0.017	0.030	0.037	0.052	0.027	0.109	0.032
	Calmar Ratio	3.490	1.371	10.134	5.676	3.614	3.611	4.176	1.757	4.501
2018-2023	Annualized Return	0.175	0.119	0.185	0.195	0.153	0.205	0.182	0.240	0.234
	Annualized Vol	0.103	0.180	0.111	0.137	0.132	0.124	0.133	0.174	0.115
	Sharpe Ratio	1.368	0.470	1.354	1.175	0.900	1.372	1.113	1.180	1.744
	Sortino Ratio	3.898	0.753	5.140	7.459	1.715	3.567	4.865	2.230	9.853
	Profit Factor	5.169	1.597	6.605	7.522	3.122	4.746	6.166	3.804	8.915
	Max Drawdown	0.079	0.248	0.053	0.029	0.145	0.091	0.034	0.175	0.032
	Calmar Ratio	2.227	0.481	3.473	6.715	1.053	2.255	5.283	1.372	7.224

TABLE 4.3: The return metrics of each portfolio at each sub-period

4.4 Factor analysis

To determine if the performance of the portfolio formed by perspective clustering is affected by systematic market risk, we perform risk factor regressions on four factor models as Han et al. (2023) did. We regress on the portfolio formed by SEFT clustering since it's the most promising portfolio among all perspective clustering methods. The factor models we use are Fama-French three-factor model (FF3) FAMA and FRENCH (1996), Fama-French three-factor model extended with momentum and reversal factor(FF3+2), Fama-French five-factor model (FF5)Fama and French (2015) and q^5 model Hou et al. (2021). It is interesting to see if there's effect of momentum and short-term reversal since momentum is incorporated in the technical indicator perspective clustering and the long-short portfolio is in itself a contrarian portfolio profiting by exploiting the return reversal. Table 4.4 shows the regression results of the four factor models.

When the return of long-short portfolio formed by SEFT is regressed against three-factor model (FF3), we get an alpha of 0.0041(t=2.44) in moderate significance with $p < 0.05$ and a significant positive market beta of 0.0015(t=3.90). This suggests that the portfolio formed by SEFT is not market neutral, and the positive beta is a sign that the portfolio favors more on high beta stocks on the long leg than the short leg, despite its dollar neutrality given that the investments in all portfolio formation month in both long and short legs are equally weighted in value. This is consistent with the evidence that the positive return of SEFT portfolio mainly come from the long leg portfolio. Similar market beta can be observed also in FF5 and q^5 model.

The extended Fama-French three-factor model has the highest explanatory power with the highest R^2 of 0.23 among all factor models and a significant monthly alpha of $0.0049(t = 3.1)$. It has highly significant ($p < 0.01$) exposure on momentum and reversal factor with coefficients of $-0.0010(t = -2.9)$ and $-0.0011(t = 5.5)$, and moderately significant ($p < 0.05$) exposure on high minus low factor (HML) with coefficient of $0.0010(t = 2.036)$. The negative and significant risk loading on momentum indicates that the portfolio has more tendency to long historically inferior stocks, expecting its reversion in the future, than to long good performance stocks. This, by no surprise, is consistent with the contrarian nature of price anomaly exploitation in the long-short portfolio. The positive and significant loading on reversal factor is also expected for the same reason. The negative and moderately significant HML factor suggests that the portfolio favors more on stocks with higher book-to-market ratio than value stocks. We see negative and significant factor loading on RMW factor in FF5 model, which indicates the existence of stocks whose firm is unprofitable but with rapid growthMosoeu and Kodongo (2022). In q^5 model, we see ROE factor being negative and significant. This would suggest that the market favors firms with weaker profitability than stronger profitability.

	FF3	FF3+2	FF5	q^5
Intercept	0.00414 **	0.00488 ***	0.00482 ***	0.00551 ***
	(2.442)	(3.100)	(2.778)	(3.142)
Market	0.00159 ***	0.00031	0.00153 ***	0.00120 ***
	(3.898)	(0.735)	(3.630)	(2.679)
SMB	0.00118	0.00071		
	(1.636)	(1.056)		
HML	-0.00043	-0.00106 **		
	(-0.796)	(-2.036)		
Momentum		-0.00115 ***		
		(-2.913)		
Reversal		0.00318 ***		
		(5.547)		
SMB5			0.00038	
			(0.490)	
HML5			-0.00036	
			(-0.513)	
RMW5			-0.00185 **	
			(-1.980)	
CMA5			-0.00021	
			(-0.188)	
R_ME				-0.00057
				(-0.735)
R_IA				-0.00041
				(-0.480)
R_ROE				-0.00191 **
				(-2.278)
R_EG				-0.00097
				(-0.886)
R^2	0.09409	0.23024	0.10546	0.12025
Adj. R^2	0.08305	0.21447	0.08713	0.10222
Num. obs.	250	250	250	250
*p<0.10	**p<0.05	***p<0.01		

TABLE 4.4: Factor regression results of portfolio formed by SEFT clustering

Chapter 5

Robustness analysis

To check if the optimal performance of SEFT clustering is due to data snooping, we perform robustness analysis for all clustering portfolios by comparing portfolio performance with each parameter changed independently. Table 5.1 Table 5.2 Table 5.3 are Sharpe ratio, annualized mean return and annualized volatility of clustering portfolios under different scenarios. The baseline scenario uses the set of parameters optimized for SEFT clustering portfolio as specified in the main analysis in Section 3.9. For all other stressed scenarios, only the parameter mentioned in the scenario is shifted to a certain value with all other parameters remain the same as the baseline scenario. In these stressed scenarios, p^f is the number of principle components that the firm characteristics are reduced to. Downstream of p^f, k^f is the number of clusters specified for kmeoids in computing DTM^F. k^t is the number of clusters specified for kmeoids in computing DTM^T. k^n is the number of clusters specified for kmeoids in the final stage clustering. γ scenarios remove stocks with DTM exceeding γ quantile neighborhood distance before being proceeded for threshold trading strategy. The above parameters are model parameters that would determine the final clustering result. The trade-time scenarios specify downstream trade-time parameters. To determine the contribution of small-cap stocks in the portfolio, we use mktcap scenarios in which stocks with certain lower quantile market capitalization in all candidate stocks are removed before trading. We are also interested in different α scenarios in which α of standard deviation is used as the threshold in forming long-short portfolio. Finally, fee scenarios specify the basis point of trading cost deducted from each trade.

5.1 Model parameters

As a general observation from Table 5.1, we found that SEFT clustering portfolio maintains its good performance in most scenarios with parameters shifted within a certain range. From Table 5.2 and Table 5.3, ST clustering portfolio in general has the highest mean return and also highest in volatility. SEFT clustering portfolio has the least volatility in most scenarios. For the first stage multi-perspective clustering, when p^f, k^f and k^t are shifted within a certain range, SEFT clustering

Scenario	value	S	F	T	SF	FT	ST	SFT
baseline		0.958	1.005	0.818	1.124	1.052	0.884	**1.461**
pf(20)	5	0.958	1.050	0.818	1.128	1.054	0.884	**1.236**
	10	0.958	1.078	0.818	1.290	1.063	0.884	**1.275**
	30	0.958	0.824	0.818	1.283	1.016	0.884	**1.306**
kf(75)	15	0.958	0.998	0.818	1.014	0.831	0.884	**1.256**
	35	0.958	0.865	0.818	1.296	1.041	0.884	**1.300**
	95	0.958	1.143	0.818	1.186	1.052	0.884	**1.248**
	115	0.958	1.168	0.818	1.160	0.804	0.884	**1.324**
kt(30)	10	0.958	1.005	0.937	1.124	0.970	**1.088**	1.079
	50	0.958	1.005	0.865	1.124	0.923	1.033	**1.245**
	70	0.958	1.005	0.917	1.124	1.077	1.105	**1.129**
	90	0.958	1.005	0.813	1.124	0.922	**1.207**	1.129
kn(200)	25	0.958	1.017	0.925	0.862	**1.171**	0.921	0.792
	50	0.958	0.744	0.974	0.885	**1.309**	0.939	1.139
	100	0.958	0.962	0.785	0.978	1.068	**1.076**	0.962
	500	0.958	0.974	0.872	1.094	0.776	1.160	**1.176**
	700	0.958	1.029	0.787	0.841	0.893	**1.089**	1.074
	1000	0.958	**1.096**	0.709	0.873	0.925	0.980	1.001
γ(0)	0.009	0.958	0.861	0.497	1.094	0.944	0.847	**1.491**
	0.004	0.958	0.804	0.357	1.082	0.757	0.920	**1.161**
mktcap(0)	0.025	0.746	0.997	0.697	0.953	0.766	0.689	**1.150**
	0.05	0.537	0.905	0.600	0.838	0.674	0.577	**0.915**
	0.1	0.368	**0.808**	0.472	0.649	0.404	0.512	0.736
α(0.6)	0.2	0.916	0.816	0.662	0.969	0.919	0.893	**1.214**
	0.4	0.864	0.942	0.742	1.107	0.976	0.927	**1.352**
	0.8	0.968	1.139	0.899	1.177	1.028	0.975	**1.417**
	1	0.970	**1.277**	0.898	1.127	1.066	0.968	1.272
	1.2	0.955	**1.284**	0.867	1.093	0.998	1.025	0.974
fee(0)	25 bps.	0.830	0.882	0.711	0.985	0.902	0.816	**1.297**
	50 bps.	0.703	0.758	0.604	0.846	0.753	0.748	**1.131**

Table 5.1: The Sharpe ratio of all stressed scenarios. The value in the bracket of each scenario is the value of the parameter used in the non-stressed baseline scenario. The boldface value is the highest Sharpe ratio among all clustering portfolios under the same scenario.

Scenario	value	S	F	T	SF	FT	ST	SFT
baseline		0.147	0.157	0.149	0.155	0.140	**0.230**	0.167
pf(20)	5	0.147	0.177	0.149	0.183	0.152	**0.230**	0.167
	10	0.147	0.156	0.149	0.199	0.166	**0.230**	0.173
	30	0.147	0.146	0.149	0.151	0.134	**0.230**	0.182
kf(75)	15	0.147	0.141	0.149	0.194	0.137	**0.230**	0.186
	35	0.147	0.142	0.149	0.182	0.143	**0.230**	0.157
	95	0.147	0.164	0.149	0.156	0.150	**0.230**	0.160
	115	0.147	0.145	0.149	0.151	0.129	**0.230**	0.142
kt(30)	10	0.147	0.157	0.142	0.155	0.120	**0.247**	0.163
	50	0.147	0.157	0.157	0.155	0.138	**0.192**	0.158
	70	0.147	0.157	0.162	0.155	0.134	**0.229**	0.161
	90	0.147	0.157	0.161	0.155	0.136	**0.199**	0.175
kn(200)	25	0.147	0.310	0.325	0.367	0.217	**0.482**	0.309
	50	0.147	0.251	0.226	**0.317**	0.206	0.305	0.257
	100	0.147	0.195	0.155	0.210	0.177	**0.231**	0.204
	500	0.147	0.116	0.110	0.123	0.101	**0.155**	0.128
	700	**0.147**	0.136	0.109	0.103	0.115	0.139	0.122
	1000	**0.147**	0.131	0.105	0.106	0.119	0.119	0.110
γ(0)	0.009	0.147	0.131	0.088	0.174	0.127	**0.221**	0.162
	0.004	0.147	0.111	0.069	0.170	0.097	**0.185**	0.127
mktcap(0)	0.025	0.116	0.129	0.095	0.123	0.134	**0.178**	0.133
	0.05	0.089	0.108	0.088	0.099	0.120	**0.152**	0.114
	0.1	0.071	0.094	0.072	0.109	0.119	**0.147**	0.095
α(0.6)	0.2	0.106	0.106	0.101	0.114	0.098	**0.136**	0.120
	0.4	0.131	0.132	0.121	0.141	0.120	**0.170**	0.135
	0.8	0.187	0.213	0.178	0.190	0.180	**0.269**	0.190
	1	0.232	0.276	0.197	0.212	0.207	**0.286**	0.200
	1.2	0.258	**0.331**	0.232	0.235	0.219	0.314	0.212
fee(0)	25 bps.	0.132	0.142	0.134	0.140	0.125	**0.215**	0.152
	50 bps.	0.117	0.127	0.119	0.125	0.110	**0.200**	0.137

TABLE 5.2: The mean return of all stressed scenarios. The boldface value is the highest return among all clustering portfolios under the same scenario.

portfolio maintains the highest Sharpe ratio across other portfolios. When k^t equals 10 and 90, the SEFT clustering is only exceeded by ST clustering.

For the second stage clustering, the SEFT portfolio maintains the highest Sharpe ratio when k^n equals 200 and 500 and becomes suboptimal for other selections of k^n. When k^n equals 20 and 50, FT clustering has the optimal Sharpe ratio across all portfolios. This may be because the combined DTM features from firm characteristics and technical indicator inherently have larger group of spatially close stocks. Thus, clustering with lower kn would result in less clusters and more homogenous stocks within a larger cluster. From Table 5.3, we see that the volatility of FT clustering is reduced compared to F and T clustering with k^n equals 20 and 50. This suggests that the combination of DTM from F and T perspective would stabilize the return from time to time under smaller number of k^n. When sector perspective DTM features are incorporated, the stocks are more scattered in the feature space

Scenario	value	S	F	T	SF	FT	ST	SFT
baseline		0.117	0.122	0.141	0.107	0.100	0.222	**0.091**
pf(20)	5	0.117	0.135	0.141	0.131	0.112	0.222	**0.108**
	10	0.117	0.113	0.141	0.128	0.124	0.222	**0.108**
	30	0.117	0.135	0.141	**0.091**	0.098	0.222	0.113
kf(75)	15	0.117	**0.107**	0.141	0.158	0.123	0.222	0.121
	35	0.117	0.124	0.141	0.114	0.104	0.222	**0.094**
	95	0.117	0.113	0.141	0.102	0.110	0.222	**0.100**
	115	0.117	0.095	0.141	0.100	0.118	0.222	**0.081**
kt(30)	10	0.117	0.122	0.115	0.107	**0.088**	0.196	0.119
	50	0.117	0.122	0.141	0.107	0.112	0.153	**0.099**
	70	0.117	0.122	0.139	0.107	**0.092**	0.176	0.112
	90	0.117	0.122	0.156	**0.107**	0.110	0.136	0.125
kn(200)	25	**0.117**	0.271	0.314	0.386	0.156	0.486	0.346
	50	**0.117**	0.291	0.197	0.319	0.131	0.289	0.195
	100	**0.117**	0.167	0.154	0.180	0.134	0.182	0.176
	500	0.117	0.084	0.087	0.081	0.086	0.104	**0.080**
	700	0.117	0.099	0.095	**0.081**	0.090	0.096	0.082
	1000	0.117	0.088	0.100	0.082	0.092	0.086	**0.076**
γ(0)	0.009	0.117	0.112	0.107	0.127	0.098	0.220	**0.086**
	0.004	0.117	0.096	0.096	0.125	0.083	0.163	**0.080**
mktcap(0)	0.025	0.110	0.095	0.087	0.093	0.131	0.208	**0.086**
	0.05	0.101	0.081	0.089	**0.077**	0.127	0.204	0.087
	0.1	0.099	**0.074**	0.080	0.114	0.209	0.220	0.082
α(0.6)	0.2	0.078	0.087	0.100	0.082	**0.069**	0.113	0.070
	0.4	0.112	0.103	0.117	0.096	0.088	0.146	**0.074**
	0.8	0.158	0.157	0.159	0.132	0.142	0.240	**0.110**
	1	0.203	0.189	0.181	0.158	0.162	0.260	**0.130**
	1.2	0.234	0.231	0.228	0.183	0.185	0.273	**0.183**
fee(0)	25 bps.	0.117	0.122	0.140	0.107	0.100	0.222	**0.091**
	50 bps.	0.117	0.122	0.140	0.107	0.100	0.221	**0.090**

Table 5.3: The volatility of all stressed scenarios. The boldface value is the lowest volatility among all clustering portfolios under the same scenario.

and thus need higher number of k^n for clustering in order to achieve better performance.

In Table 5.2 and Table 5.3, we found that both the volatility and mean return decreases as k^n increases. We also found that the number of stocks traded increases as k^n increases in Table 5.4. As k^n increases, more clusters are formed despite of fewer stocks incorporated in each cluster on average. This indicates that more proportion of stocks are traded in each cluster on average in larger k^n than in smaller k^n. Given the same proportion parameter α and higher proportion of stocks traded in each cluster in larger k^n scenario, it can be speculated that the standard deviation of spread in each cluster should be lower in larger k^n scenario than smaller k^n scenario to allow more stocks to be traded. One explanation for lowered return spread standard deviation in higher k^n is that the past one-month return of stocks

in one cluster tend to be more similar when fewer stocks are grouped into the cluster. Thus, there are fewer opportunities to exploit return anomalies in larger k^n scenarios. With fewer total number of stocks traded in smaller k^n scenarios, both the return and the volatility are higher in smaller k^n scenarios than in higher k^n scenarios.

The effect of anomaly stock removal γ is also considered. A stock is considered as outlier and removed when the DTM to its closest cluster is larger than the γ quantile distance of all its neighborhood stocks. This means a higher γ imposes less strict criterion in outlier removal, and perhaps no outliers are removed when γ is large enough and neighborhoods are close enough to each other. In the empirical study, the algorithm only starts to remove outlier stocks when γ is less than 2%. This implies that stocks are quite evenly distributed around their medoids rarely with extremely far-away outliers, mainly attributed to the advantage of kmeoids clustering in dealing with outliers or high data dimensionality. When γ decreases to from 2% to 0.9%, the Sharpe ratio of SEFT portfolio is increased to 1.491 from 1.461 in the baseline scenario with no outlier removal. The slight increase on Sharpe ratio is mainly driven by reduced volatility as seen from Table 5.3. This implies that there exists a small number of outlier stocks that, when removed, can improve the portfolio performance. When γ continues dropping to 0.4%, the Sharpe ratio of SEFT portfolio dropped sharpely to 1.161. The drop is mainly caused by the sharp drop in the mean return accompanied by relatively moderate drop in volatility. This indicates that some high potential stocks are regarded as outlier stocks and removed from the portfolio, causing great drop in mean return. In both γ scenarios, SEFT exceeded other portfolios in terms of Sharpe ratio.

In conclusion, the performance superiority of SEFT clustering portfolio is robust in most model parameter scenarios.

kn(200)	S	F	T	SF	FT	ST	SFT
25	1483	1195	1151	1205	1234	1187	1215
50	1483	1274	1247	1297	1319	1271	1299
100	1483	1361	1343	1386	1397	1359	1386
500	1483	1556	1608	1644	1577	1650	1667
700	1483	1578	1648	1708	1572	1722	1734
1000	1483	1559	1635	1735	1497	1773	1768

TABLE 5.4: The average number of stocks traded when different k^n is applied

5.2 Trade-time parameters

To see if small-cap stocks contribute to the profit of long-short portfolio, we remove stocks with specified lower quantile market capitalization of all tradable stocks before the threshold trading process. We found that removing small-cap stocks causes reduction in Sharpe ratio, which is mainly driven by the decline of mean return.

This implies that small-cap stocks have substantial contribution to the long-short portfolio profitability. Despite the drop of Sharpe ratio, SEFT still maintains the highest Sharpe ratio when lower quantile market capitalization removal is less than 5%.

Next we test on the robustness of trading threshold parameter. We found that SEFT clustering maintains the highest Sharpe ratio when α is less than 1, and when α equals 1, the Sharpe ratio of SEFT clustering portfolio is only exceeded by F clustering portfolio by 0.5%. When α equals to 1.2, F clustering portfolio has a lot higher Sharpe ratio than other portfolios. From Table 5.2 and Table 5.3, we can see a clear monotonic positive change of both mean return and volatility with respect to the threshold parameter. The explanation is twofold. First, higher threshold selects pairs with spread diverged more in the past. When the threshold is set higher, the pairs selected tend to converge more in magnitude since they had a wider initial divergence in the past one month. On the other hand, a higher threshold reduces the number of stocks traded, resulting in a less diversified portfolio and higher portfolio volatility. For SEFT portfolio, the α that gives the best balance between mean return and volatility is when α equals 0.6, as reflected from the Sharpe ratio in the baseline scenario.

Finally, we test on the robustness of trading fee. When a 25 bps trading fee is imposed on each trade, the mean return is reduced from 0.23 to 0.215 with volatility unchanged. It's straight forward that the fixed trading fee reduces mean return and does not impact the volatility. The SEFT portfolio maintains the highest Sharpe ratio in two trading fee scenarios, and it still has a promising Sharpe ratio of 1.131 when 50 bps trading fee is imposed.

To conclude, despite some extent of compromised performance when certain trade-time parameters are shifted, SEFT portfolio maintains its performance superiority and robustness in most trade-time scenarios.

Chapter 6

Conclusion

The final chapter contains the overall conclusion. It also contains suggestions for future work and industrial applications.

Appendices

Appendix A

The First Appendix

	Single perspective			Dual perspective			SEFT	Benchmarks	
Monthly return statistics	**S**	**F**	**T**	**SF**	**FT**	**ST**	**SFT**	**reversion**	**sp500**
Mean Return	0.040	0.039	0.043	0.045	0.038	0.059	0.043	0.035	0.008
Standard Deviation	0.095	0.094	0.105	0.093	0.091	0.146	0.086	0.076	0.043
Standard Error	0.006	0.006	0.007	0.006	0.006	0.009	0.005	0.005	0.003
t-statistic	6.602	6.627	6.434	7.661	6.559	6.363	7.950	7.347	2.970
Min	-0.250	-0.226	-0.242	-0.221	-0.242	-0.244	-0.215	-0.236	-0.169
0.25	-0.003	-0.006	-0.003	-0.006	-0.005	0.003	0.000	-0.001	-0.016
0.5	0.029	0.029	0.031	0.035	0.030	0.040	0.035	0.034	0.012
0.75	0.078	0.073	0.072	0.088	0.073	0.090	0.082	0.068	0.034
Max	0.753	0.683	0.641	0.535	0.621	1.535	0.441	0.344	0.127
Skewness	2.067	2.397	2.413	1.140	1.592	5.262	0.794	0.508	-0.602
Kurtosis	13.894	14.695	10.984	4.549	8.285	45.016	3.003	2.925	1.340
VaR(95%)	-0.089	-0.080	-0.081	-0.088	-0.089	-0.080	-0.083	-0.081	-0.073
ES(95%)	-0.135	-0.126	-0.128	-0.131	-0.134	-0.134	-0.128	-0.126	-0.097
Annualized risk-return metrics	**S**	**F**	**T**	**SF**	**FT**	**ST**	**SFT**	**reversion**	**sp500**
Annualized Return	0.474	0.473	0.514	0.540	0.452	0.704	0.520	0.424	0.096
Annualized Vol	0.328	0.326	0.364	0.322	0.315	0.505	0.298	0.263	0.148
Sharpe Ratio	1.341	1.346	1.315	1.571	1.328	1.326	1.626	1.479	0.418
Downside Deviation	0.052	0.045	0.048	0.048	0.050	0.053	0.048	0.048	0.033
Sortino Ratio	2.421	2.809	2.896	3.052	2.435	3.644	2.938	2.329	0.548
Gross Profit	13.431	13.325	14.153	14.734	12.973	17.805	14.129	12.019	5.175
Gross Loss	-3.552	-3.463	-3.449	-3.483	-3.553	-3.143	-3.302	-3.194	-3.168
Profit Factor	3.781	3.848	4.103	4.230	3.652	5.666	4.279	3.763	1.633
Profitable Years	19	20	20	19	19	20	19	19	17
Unprofitable Years	2	1	1	2	2	1	2	2	4
Max Drawdown	0.454	0.407	0.453	0.436	0.379	0.364	0.434	0.462	0.526
Calmar Ratio	1.045	1.162	1.135	1.238	1.192	1.932	1.198	0.916	0.183
Turnover	16.984	16.787	17.602	18.217	16.526	20.947	17.431	15.212	8.343
	11.807	10.785	12.149	12.551	11.638	12.698	11.791	13.378	11.072

TABLE A.1: Long portfolio monthly return statistics and annualized trading performance metrics using different clustering methods.

	Single perspective			Dual perspective			SEFT	Benchmarks	
Monthly return statistics	**S**	**F**	**T**	**SF**	**FT**	**ST**	**SFT**	**reversion**	**sp500**
Mean Return	-0.015	-0.013	-0.018	-0.019	-0.014	-0.020	-0.015	-0.017	0.008
Standard Deviation	0.062	0.055	0.064	0.065	0.059	0.066	0.060	0.056	0.043
Standard Error	0.004	0.003	0.004	0.004	0.004	0.004	0.004	0.004	0.003
t-statistic	-3.861	-3.810	-4.415	-4.637	-3.841	-4.847	-4.097	-4.895	2.970
Min	-0.274	-0.204	-0.301	-0.259	-0.230	-0.317	-0.260	-0.223	-0.169
0.25	-0.051	-0.049	-0.052	-0.055	-0.050	-0.055	-0.051	-0.047	-0.016
0.5	-0.017	-0.013	-0.017	-0.016	-0.016	-0.019	-0.017	-0.017	0.012
0.75	0.014	0.019	0.017	0.017	0.018	0.012	0.019	0.015	0.034
Max	0.251	0.191	0.210	0.217	0.198	0.212	0.203	0.209	0.127
Skewness	0.241	-0.049	-0.359	-0.303	0.196	-0.328	-0.043	0.173	-0.602
Kurtosis	2.795	0.993	2.728	1.550	1.349	2.715	1.315	1.846	1.340
VaR(95%)	-0.101	-0.094	-0.120	-0.134	-0.103	-0.124	-0.108	-0.104	-0.073
ES(95%)	-0.140	-0.134	-0.173	-0.177	-0.138	-0.174	-0.142	-0.135	-0.097
Annualized risk-return metrics	**S**	**F**	**T**	**SF**	**FT**	**ST**	**SFT**	**reversion**	**sp500**
Annualized Return	-0.181	-0.159	-0.215	-0.230	-0.173	-0.243	-0.186	-0.207	0.096
Annualized Vol	0.214	0.190	0.222	0.226	0.205	0.229	0.207	0.193	0.148
Sharpe Ratio	-1.007	-1.016	-1.122	-1.168	-1.009	-1.212	-1.064	-1.251	0.418
Downside Deviation	0.041	0.038	0.048	0.049	0.039	0.049	0.041	0.038	0.033
Sortino Ratio	-1.510	-1.480	-1.509	-1.560	-1.542	-1.634	-1.558	-1.844	0.548
Gross Profit	4.019	3.739	3.837	3.882	4.019	3.811	3.958	3.397	5.175
Gross Loss	-7.787	-7.045	-8.314	-8.669	-7.618	-8.881	-7.832	-7.720	-3.168
Profit Factor	0.516	0.531	0.462	0.448	0.528	0.429	0.505	0.440	1.633
Profitable Years	5	4	4	5	4	3	4	3	17
Unprofitable Years	16	17	17	16	17	18	17	18	4
Max Drawdown	0.986	0.975	0.994	0.995	0.983	0.997	0.987	0.991	0.526
Calmar Ratio	-0.183	-0.163	-0.216	-0.231	-0.176	-0.244	-0.188	-0.209	0.183
Turnover	11.807	10.784	12.151	12.551	11.637	12.692	11.790	11.117	8.343

Table A.2: Short portfolio monthly return statistics and annualized trading performance metrics using different clustering methods.

Bibliography

K. Li, K. Sward, H. Deng, J. Morrison, R. Habre, M. Franklin, Y.-Y. Chiang, J. L. Ambite, J. P. Wilson, and S. P. Eckel, "Using dynamic time warping self-organizing maps to characterize diurnal patterns in environmental exposures," *Scientific reports*, vol. 11, no. 1, p. 24052, 2021.

A. Subrahmanyam, "Distinguishing between rationales for short-horizon predictability of stock returns," *Financial Review*, vol. 40, no. 1, pp. 11–35, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0732-8516.2005.00091.x

D. AVRAMOV, T. CHORDIA, and A. GOYAL, "Liquidity and autocorrelations in individual stock returns," *The Journal of Finance*, vol. 61, no. 5, pp. 2365–2394, 2006.

N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of Finance*, vol. 48, no. 1, pp. 65–91, 1993. [Online]. Available: http://www.jstor.org/stable/2328882

A. W. Lo and A. C. MacKinlay, *When are Contrarian Profits Due to Stock Market Overreaction?*, ser. NBER working paper series no. w2977. Cambridge, Mass: National Bureau of Economic Research, 1989.

W. F. M. De BONDT and R. THALER, "Does the stock market overreact?" *The Journal of Finance*, vol. 40, no. 3, pp. 793–805, 1985. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1985.tb05004.x

H. Jacobs, "What explains the dynamics of 100 anomalies?" *Journal of banking & finance*, vol. 57, pp. 65–85, 2015.

E. F. Fama, "Efficient capital markets a review of theory and empirical work," in *The Fama Portfolio*. Chicago: University of Chicago Press, 2021, pp. 76–121.

C. Krauss, "Statistical arbitrage pairs trading strategies: Review and outlook," *Journal of economic surveys*, vol. 31, no. 2, pp. 513–545, 2017.

E. G. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, *Pairs Trading: Performance of a Relative Value Arbitrage Rule*, ser. NBER working paper series no. w7032. Cambridge, Mass: National Bureau of Economic Research, 1999.

H. Chen, S. Chen, Z. Chen, and F. Li, "Empirical investigation of an equity pairs trading strategy," *Management Science*, vol. 65, no. 1, pp. 370–389, 2019.

B. Do and R. Faff, "Are pairs trading profits robust to trading costs?" *The Journal of financial research*, vol. 35, no. 2, pp. 261–287, 2012.

B. Do, R. Faff, and K. Hamza, "A new approach to modeling and estimation for pairs trading," in *Proceedings of 2006 financial management association European conference*, vol. 1. Citeseer, 2006, pp. 87–99.

G. Vidyamurthy, *Pairs Trading: quantitative methods and analysis*. John Wiley & Sons, 2004, vol. 217.

S. Johansen, "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control*, vol. 12, no. 2, pp. 231–254, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165188988900413

R. F. Engle and C. W. J. Granger, "Co-integration and error correction: Representation, estimation, and testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987. [Online]. Available: http://www.jstor.org/stable/1913236

D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

S. A. Ross, "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, vol. 13, no. 3, pp. 341–360, 1976. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022053176900466

W. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964. [Online]. Available: https://EconPapers.repec.org/RePEc:bla:jfinan:v:19:y:1964:i:3:p:425-442

R. J. Elliott, J. Van Der Hoek, and W. P. Malcolm, "Pairs trading," *Quantitative finance*, vol. 5, no. 3, pp. 271–276, 2005.

J. Liu and A. Timmermann, "Optimal Convergence Trade Strategies," *The Review of Financial Studies*, vol. 26, no. 4, pp. 1048–1086, 01 2013. [Online]. Available: https://doi.org/10.1093/rfs/hhs130

J. W. Jurek and H. Yang, "Dynamic portfolio selection in arbitrage," in *EFA 2006 meetings paper*, 2007.

R. Q. Liew and Y. Wu, "Pairs trading: A copula approach," *Journal of Derivatives & Hedge Funds*, vol. 19, pp. 12–30, 2013.

M. Avellaneda and J.-H. Lee, "Statistical arbitrage in the u.s. equities market," *Political Methods: Quantitative Methods eJournal*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:15863073

N. Huck, "Pairs trading and outranking: The multi-step-ahead forecasting case," *European journal of operational research*, vol. 207, no. 3, pp. 1702–1716, 2010.

L. Takeuchi, "Applying deep learning to enhance momentum trading strategies in stocks," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:17043131

A. Flori and D. Regoli, "Revealing pairs-trading opportunities with long short-term memory networks," *European Journal of Operational Research*, vol. 295, no. 2, pp. 772–791, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221721001995

C. Han, Z. He, and A. J. W. Toh, "Pairs trading via unsupervised learning," *European Journal of Operational Research*, vol. 307, no. 2, pp. 929–947, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037722172200769X

Y. Han, "What firm characteristics drive us stock returns," *Journal of Economic Dynamics and Control*, vol. 3185335, 2018.

J. Green, J. R. M. Hand, and X. F. Zhang, "The characteristics that provide independent information about average u.s. monthly stock returns," *The Review of financial studies*, vol. 30, no. 12, pp. 4389–4436, 2017.

U.S. Securities and Exchange Commission, "Standard industrial classification (sic) code list," https://www.sec.gov/corpfin/division-of-corporation-finance-standard-industrial-classification-sic-code-list, accessed: May 1, 2024.

P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, no. 1-23, p. 40, 2008.

K. J. B, "The symmetric time-warping problem : From continuous to discrete," *Time Warps, String Edits, and Macromolecules : The Theory and Practice of Sequence Comparison*, pp. 125–161, 1983.

L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, ser. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993. [Online]. Available: https://books.google.be/books?id=XEVqQgAACAAJ

M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dtw to the multi-dimensional case requires an adaptive approach," *Data mining and knowledge discovery*, vol. 31, pp. 1–31, 2017.

E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 102–111.

F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. D. F. Costa, "Principal component analysis: A natural approach to data exploration," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–34, May 2021. [Online]. Available: http://dx.doi.org/10.1145/3447755

H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3336–3341, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741740800081X

Y. Bertrand, J. De Weerdt, and E. Serral Asensio, "A novel multi-perspective trace clustering technique for iot-enhanced processes: A case study in smart manufacturing," 2023-05-24.

E. F. FAMA and K. R. FRENCH, "Multifactor explanations of asset pricing anomalies," *The Journal of Finance*, vol. 51, no. 1, pp. 55–84, 1996. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb05202.x

E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1–22, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304405X14002323

K. Hou, H. Mo, C. Xue, and L. Zhang, "An augmented q-factor model with expected growth," *Review of Finance*, vol. 25, no. 1, pp. 1–41, 2021.

S. Mosoeu and O. Kodongo, "The fama-french five-factor model and emerging market equity returns," *The Quarterly Review of Economics and Finance*, vol. 85, pp. 55–76, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1062976920301460