VSB – TECHNICAL UNIVERSITY OF OSTRAVA
FACULTY OF ECONOMICS

FINANCE

Výběr a optimalizace portfolia na základě strategie faktorového investování
Portfolio Selection and Optimization Based on Factor Investing Strategy

Student: Bc. Qian Gao
Supervisor of diploma thesis: doc. Ing. Aleš Kresta, Ph.D.

Ostrava 2023

Content

1 Introduction

Portfolio construction and optimization is an important topic in finance, aiming to achieve excess returns over the benchmark. These operations are carried out based on active portfolio management, compared to passive returns, by selecting the right market timing to adjust the portfolio thus minimize risk and maximize return. With the increase of public information disclosure, the available information to determine investment potential has been enhanced in several dimensions. Traditional portfolio construction methods are mainly based on the market value and risk level of assets, but this approach is difficult to cope with the escalating diversified and complex investment market. The relationship between the complexity of multidimensional information and the achievement of excess returns is becoming increasingly elusive. This deviates from the original purpose of portfolio construction and requires screening and judging the potential profitability of the underlying investments to further optimize the portfolio. Factor investment strategy, as an emerging investment concept, can help investors achieve portfolio optimization and risk control more effectively by mining and utilizing the factor information behind asset prices.

As an alternative systematic approach to portfolio construction, factor investing aims to capture the performance of underlying fundamental, technical and systematic risk factors, such as size, value, momentum and quality. From there, individual stocks associated with these factors are screened, thereby narrowing the scope of portfolio optimization. Enabling these factors to provide more accurate predictions by taking into account various market factors when constructing a portfolio. In different market environments, factor investment strategies can reduce the volatility of a portfolio, improve its performance, and mitigate the impact by diversifying the portfolio's risk during dramatic market changes.

In the 1970s, as the U.S. gradually shifted to institutional markets, i.e., investing through professional investors or institutions, scholars such as Barra and Fama attempted to find major categories of factors to explain the returns, and the models and methods converged as standardization emerged returns and the explanation of regressors gradually diminished. However, financial markets in emerging countries such as China started late and thus may logically have a marginal decline in the explanation of factors as well, while the Chinese stock market still has a strong appeal to investors with its uniqueness and importance. In addition, factor investment strategies will likely become a more important

and promising area of investment as artificial intelligence and big data technologies continue to develop. The application of factor investment strategies can also help investors achieve more intelligent and systematic investment decisions.

Based on the above, the objective of this thesis is to explore the application of factor investment strategies in portfolio construction and to discuss its practical application. Since individual investors are at a disadvantageous position in the market compared with institutional investors, the main discussion in this thesis is from the standpoint of individual investors, trying to construct a set of investment analysis processes suitable for individual investors, based on which we try to explain stock market returns by various factors and discriminate factor characteristics through machine learning, so as to construct a portfolio suitable for investors themselves and optimize it, and improve the efficiency and risk control ability.

In addition, this thesis draws on the latest research reports from investment banks on multi-factor model testing, and with the help of quantitative platforms such as Ricequant and Joinquant, the universality and usability of the research environment are maintained to the greatest extent, and the data can be accessed directly. By using alphalens and sklearn modules to compute and analyze the factor pool and get the results, we can perform payoff analysis, calculate the information coefficients and correlations of the factors, and determine whether the factors pass the single-factor test. An example of single-factor analysis is provided as a reference in order to provide investors with more accurate and comprehensive information for investment decisions. Subsequently, this thesis selects stocks based on the availability of factors to obtain the possibility of future returns through portfolio optimization. Therefore, this thesis is organized as follows.

Chapter 1, the introduction provides an overview of the thesis and its objectives. Chapter 2 introduces the theoretical foundations and methods of factor investing. The theoretical foundations of factor investing such as the efficient market hypothesis, capital asset pricing model, and arbitrage pricing theory are introduced. The chapter then discusses various factor investment models, such as the Fama-French three-factor model, the Fama-French five-factor model, and the Barra integrated model. The chapter concludes with a detailed description of the process of factor investing, including the construction of a factor library, factor data unification, single-factor test analysis, factor synthesis, and screening of portfolio components.

Chapter 3 delves into portfolio optimization theory and methodology. This chapter introduces the Markowitz mean-variance model and hierarchical risk parity. It also discusses portfolio evaluation metrics such as variance, Sharpe ratio, Calmar ratio, and value-at-risk. The chapter goes on to describe the optimization process based on factor investing, including principal component analysis (PCA), extreme gradient boosting (XGBoost), multiple linear regression (MLR), Monte Carlo simulation (MC), and hierarchical risk parity (HRP).

Chapter 4 applies factor investing and portfolio optimization to current stock market conditions. This chapter provides a detailed description of the data, including data sources, data selection, and descriptive statistics. It then provides an overview of factor investing, including data processing, single-factor return testing, single-factor IC testing, and single-factor trading testing. The chapter concludes with a description of portfolio optimization using various methods such as principal component analysis (PCA), extreme gradient boosting (XGBoost), multiple linear regression (MLR), Monte Carlo simulation (MC), and hierarchical risk parity (HRP). Finally, Chapter 5 summarizes the results of this thesis and provides suggestions for future research on factor investment and portfolio optimization.

2 Factor Investing Theory and Methods

The primary objective of this chapter is to provide the reader with an in-depth understanding of the theoretical underpinnings that underpin factor investing, as well as the various models and techniques used in the approach. It begins with an introduction to the Efficient Market Hypothesis (EMH), the Capital Asset Pricing Model (CAPM), Arbitrage Pricing Theory (APT), and other theories that inform the factor investing framework. Various factor investment models are then presented, including the Fama-French three-factor model, the Fama-French five-factor model, and the Barra integrated model, allowing for a comprehensive understanding of the key characteristics and applications of each model. The chapter then explores the factors themselves, including how they are classified and measured, and how factor exposures and returns are calculated. It concludes with an introduction to the overall process of factor investing, covering topics such as building a factor library, unifying factor data, performing single-factor test analysis, and screening portfolio components.

2.1 Theoretical Foundations of Factor Investing

This section introduces the theoretical foundations of factor investing, including the efficient market hypothesis, capital asset pricing model, arbitrage pricing theory, financial analysis theory and technical analysis theory. These theories provide the basic ideological framework and methodological support for factor investing, which aims to help investors better understand markets, assess investment risks and returns, and optimize portfolios by identifying and investing in the factors that drive asset prices to achieve better risk-adjusted returns. These include the efficient market Hypothesis (EMH), capital asset pricing model (CAPM), arbitrage pricing theory (APT), financial analysis theory, and technical analysis theory.

2.1.1 Efficient Market Hypothesis (EMH)

The efficient market hypothesis (EMH) is a theoretical concept that asserts the market prices of financial securities fully reflect all available information. Fama (1970, p. 383) described it as "*In general terms, the ideal is a market in which prices provide accurate signals for resource allocation: that is, a market in which firms can make production-investment decisions, and investors can choose among the securities that represent ownership of firms' activities under the assumption that security prices at any*

time "fully reflect" all available information. A market in which prices always "fully reflect" available information is called "efficient." "

However, it is worth noting that the efficient market theory still requires many specific and stringent conditions: including that market participants are completely rational economic agents and are not affected by behavioral biases related to emotions. In addition, financial asset prices need to fully reflect market conditions, including changes brought about by potential relevant information, which also need to appear in prices in a timely manner. Market players aim to seek higher returns on investments. According to the degree of market response, the market can be divided into different efficiency levels, and different analytical feasibility can be deduced as shown in *Table 2.1*.

Table 2.1 Market efficiency and investment analysis relationship table

Efficient	Technical analysis	Fundamental analysis	Insider information	Portfolio
Inefficient	Efficient	Efficient	Efficient	Aggressive
Weak-form	Invalid	Efficient	Efficient	Aggressive
Semi-strong	Invalid	Invalid	Efficient	Aggressive
Strong-form	Invalid	Invalid	Invalid	Conservative

Source: Own elaboration

Of course, these three efficiency forms may also be evidence of data inspection, rather than necessarily real existence. Ţiţan (2015, p. 447) concluded in literature review that "*One of the reasons for the markets' possible inefficiecy or prices' responses to event announcements are delayed is that investors are inattentive.*" At the same time, author also admitted that "*The EMH is simple in theory but was proved to be very difficult to test and have a precise result. Because there is no consensus among economists regarding any of the three forms of EMH, some researches and well known scientists issued the hypothesis that the reason the EMH is not validated by models is that the models themselves are biased and may provide erroneous results.*"

So does an efficient market exist? It seems that there is no absolute conclusion at present, but what is certain is that the assumption of absolute strength is too high for the current financial market. In addition, Wang (2022) believes that efficient markets are more closely connected when processing more information through the test of financial data such as metals and energy. And information connectivity is primarily driven by more efficient markets, which process more information and disseminate it to less efficient

markets. This thesis tries to assume from another angle that the verified information of the higher-level efficient market can still be valid during the gradual development of the lower-level efficient market, but the timeliness will be shorter. Financial forecasting methods may have gains in the first users, but get more widely used when the information that is incorporated into the market gradually fails, thus contributing to the innovation race, then what needs to be considered is a rapidly changing model that can subsequently detect possible instances of predictability.(Timmermann, Granger, 2004) This provides a relatively stable theoretical basis for the subsequent single-factor testing and investment portfolio construction in this thesis.

2.1.2 Capital Asset Pricing Model (CAPM)

The CAPM is a forecasting model based on the equilibrium of expected returns on risky assets, expressed using a simple linear relationship. Its main purpose is to measure systematic risk through the covariance of returns in the securities market, which is the so-called market beta. The expected equity premium (excess return) is proportional to the market beta (Chen, 2003). This is consistent with the logic that the riskier the asset, the higher the return. The author also briefly explained the static basic formula:

$$E(R_i) = R_F + [E(R_M) - R_F] \cdot \beta_i, \tag{2.1}$$

where $E(R_i)$ is expected return rate on an asset i; R_F is risk-free rate; $E(R_M)$ is expected market return rate; β_i is measuring the systematic risk of asset i. Similarly, this thesis can also deduce:

$$\beta_i = \frac{E(R_i) - R_F}{E(R_M) - R_F} = \frac{cov(R_i, R_M)}{var(R_M)}. \tag{2.2}$$

But does beta fully explain market risk? According to the hypothesis, risk and return are positively correlated and they should be equal, but this does not explain where the excess return over the average market return comes from. Therefore, Rocciolo, Gheno and Brooks (2022) attempt to obtain a neutral version of alpha by setting up three propositions to test the hypothesis for different scenarios of the CAPM. Two of the assumptions are respectively added that the agent n has the same information and beliefs on the objective joint probability distribution of all individual stock returns; the behavioral factors are introduced into the asset evaluation, which are prospect, risk aversion and agent optimism. The result of which is

$$E\big(R_j\big) - R_F = \big(\rho + \kappa(\gamma)\big) \cdot cov\big(R_j, R_M\big), \tag{2.3}$$

$$E(R_j) - R_F = \alpha_j + (E(R_M) - R_F) \cdot \beta_j^* + \kappa(\gamma) \cdot cov(R_j, R_M), \quad (2.4)$$

$$E_t[\delta_j] = E_t[d_j], \quad (2.5)$$

where ρ and γ are aggregate measures of the agent's absolute risk aversion and degree of optimism respectively; $\kappa(\gamma) = 1 - 2\gamma$; $E_t[\delta_j]$ the average net cross-sectional pricing errors; $E_t[d_j]$ is coincide with the average time-series net intercepts.

Its ultimate conclusion is that there is no rational behavior in the sense of expected utility theory, as with unexplained abnormal returns, attributing the existence of these "anomalies" to the bounded rationality of traders. It also breaks the perfect assumption of efficient markets, which provides some support for the technical analysis in *Table 2.1* and *Section 2.1.5*. In the current state of the market, there is still a need for the use of various types of analysis, and the irrational behavior of the market provides a source of excess returns in the capital markets.

In addition, beta at different times also has a stronger effect on the CAPM model. Cenesizoglu, Reeves (2018, p. 246) demonstrate "*.....that much can be gained when daily and monthly returns are utilized to measure components of systematic risk in an asset pricing framework. Most of the gain occurs through the use of daily returns.....An important implication of this paper for future research in asset pricing, is that at least some caution should be exercised in interpreting an asset pricing anomaly from beta estimation with only a single component*." Therefore, an attempt was made to use days as a time metric in the data selection in *Section 4.2.2* of this thesis. And the single-factor test data is used as an important indicator for the multi-factor model synthesis to avoid "false fit" of the data. However, this thesis still follows several key assumptions of the CAPM, including the existence of a sufficiently diverse market portfolio, the absence of transaction costs, and the availability of perfect information.

2.1.3 Arbitrage Pricing Theory (APT)

The arbitrage pricing theory (APT) is a multi-factor asset pricing model that aims to explain the behavior of asset returns in financial markets. The specific underlying assumptions and return decomposition therein are "*assumes that there are many assets, with returns determined by a small number of factors, and that competitive markets do not permit arbitrage opportunities in equilibrium. Thus returns can be split into two components: a non-diversifiable systematic risk component and an idiosyncratic part*

which can be eliminated in a well diversified portfolio. Assets with similar risk factors are close substitutes so should have similar expected returns. In this linear return generating process, expected excess returns are proportional to systematic risk, measured by factor loadings and risk premia are the coefficients of such loadings." (Pesaran, Smith, 2021, p. 17)

The APT formula proposed by Ross (1976) provides evidence that employment of the multiple index model invariably results in a singular relative pricing model. Priestley (1996) also incorporates analogous formulas in an alternate approach for generating unforeseen components. The specific formula is as follows:

$$R_{it} = \alpha_{it} + \sum_{t=1}^{k} \beta_{ik} F_{kt} + \varepsilon_{it}, \tag{2.6}$$

where R_{it} is the return on asset i in time t; α_{it} is the returns not explained by the k-factor; β_{ik} is the sensitivity of asset i to the kth factor; F_{kt} is the kth factor in time t and ε_{it} is random error term.

This again raises the question of whether the risky component can still be priced efficiently when the market is full of different choice preferences? Kelsey, Yalcin (2007) show by reasoning with different preference propositions that the arbitrage pricing theorem is obtained as a central result of asset pricing and that it will continue to hold when preferences are incomplete or non-transmissible. Their results show that any effect on asset prices works through factor prices and that APT remains robust when preferences may be incomplete or non-transmissible and when uncertainty cannot be expressed in terms of unique subjective probabilities. Based on this, this thesis can get: the APT is based on the principle of arbitrage, which states that when there are mispricings in the market, investors can take advantage of these inefficiencies by buying underpriced securities and selling overpriced securities, thereby eliminating the mispricings and restoring market efficiency.

2.1.4 Theory of Financial Analysis

Financial analysis is a channel of analysis based on the mandatory financial disclosure system of listed companies, which evaluates and calculates financial data indicators and information to understand the overall performance and financial position of the company, the industry and even the market, and is a small part of fundamental analysis. Value investors evaluate profitability, liquidity, solvency, growth potential, etc.

based on financial data, and focus on other factors such as market trends, economic conditions and regulatory changes to assess a company's competitive position and market opportunities in order to select quality assets for long-term investment. However, the thesis suggests that the market can be overvalued or undervalued at any time, so can financial data follow the same trend as stock prices, and thus potentially be a factor in explaining the source of returns?

Xi, Gao and Zhou, et al (2021) conducted an analysis after establishing a similarity network of financial indicators of listed mining companies to conclude that the similarity of similar main businesses is strong and the similarity of different financial indicators has a revelatory effect on stock returns. Moreover, the stronger structural similarity of financial indicators between a company and other companies will lead to a decrease in their returns. Likewise, the data disclosure of financial indicators may affect investor sentiment and thus the overall market investment structure and level. Liu, Wang and Xue (2023) in an analysis of the textual tone of annual reports of listed companies found that the tone of annual reports significantly increases the synergy of returns and has a significant impact on the synchronization of returns in the Chinese stock market. But which specific financial indicators are used to better explain the sources of returns? Wang, Tan (2009) find that the residual sum of squares consistently decreases when more financial variables are added to the model, i.e., the latter selection of factors does not add much useful information. It also indicates that even though different sets of candidate indicators have been studied from different perspectives, researchers have not yet reached a consensus on which factors should be included in the model.

2.1.5 Theory of Technical Analysis

Alhashel, Almudhaf and Hansz (2018, p. 92) describe it as "*Technical analysis is a method used by investors to determine when to buy and sell stocks. This method relies on the analysis of price and volume historical data to determine price trends and future movements. Technical analysis represents a challenge to the efficient market hypothesis (EMH), especially in its weak form*." The essential reason for judging future market trends through the calculation of mathematical indicators such as price and volume is that supporters of technical analysis believe that due to the repeatability of historical trends, economic cycles can be used to predict potential profit opportunities.

As mentioned earlier in this thesis, the complete rationality of the market exists in the EMH theory, but it is too demanding for the current market state. The repeatability of

the market is logical in theory, but can it be proved in practice? Although there are still debates in the academic circles, (Wang, Chiao and Chang, 2012) through the analysis of market order submission behavior, it is believed that the order submission behavior of professional institutional investors is compatible with the strategy suggested by the KD rule, and they are more inclined to buy and sell, the stock thus sends a trading signal. Individual investors seem to have been neglected, and relative to the use of other strategies, technical analysis seems to be associated with higher portfolio concentration, more turnover, less trend betting, more options trading, higher Unsystematic risk is associated with lower total and net returns and lower risk-adjusted returns for total risk. (Hoffmann, Shefrin, 2014)

In addition, since technical analysis may obtain excess returns in theory and practice, how to determine the judgment indicators of technical analysis has become a problem. Investors can often use many indicators in real-world trading, and may even consider some indicators that are combinations of other indicators rather than their counterparts. This selection process can reduce the level of redundant information considered by the model and potentially lead to better predictive outcomes and asset allocations. In addition, investors need to achieve a certain return, at least covering the maximum transaction cost level, before they may be willing to operate in a market below the threshold of this investment process. (Peng, Albuquerque and Kimura, et al, 2021)

2.2 Factor Investing Model

Factor investing is a model-based investment approach designed to systematically capture excess returns associated with specific risk factors, such as value, momentum, size, quality or low volatility. The approach is based on a quantitative model that estimates factor exposures for individual assets and constructs a portfolio of factors that reflects the desired level of factor exposure and diversification. These models are based on multi-factor broad class models, i.e. Fama-French three-factor, Fama-French five-factor and Barra models, among others. The model-based approach to factor investing has become increasingly popular in recent years because it has the potential to provide better risk-adjusted returns than traditional investment approaches. However, it also brings with it certain risks, such as model uncertainty, parameter estimation errors and factor crowding, which need to be carefully managed and monitored.

2.2.1 Fama-French Three-Factor Model

On the basis of the CAPM model, in addition to the market factor, the expected return cross-section has two more important factors, the book-to-market ratio and the stock market value. The famous three-factor model is then proposed in combination with the investment portfolio model, but the mechanism behind it is still controversial regarding the rational pricing of risk or behavioral mispricing (Liu, Gao, 2019). Reference Foye (2018) provides an overview of the tested factor models and separately provides a detailed description of the factor constructions, with the following equations described:

$$R_{it} = R_{Ft} + \beta_1(R_{Mt} - R_{Ft}) + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_{it}, \quad (2.7)$$

where R_i is the return on test portfolio i; R_F is the risk-free rate; β_1 is portfolio CAPM beta; R_M is the return on the market; SMB (small minus big) is formed from market capitalization (size); and HML (high minus low) is the value factor, formed from the book-price ratio (bp); β_2, β_3 are coefficients for SMB and HML, ε_{it} is the random error term.

Although still controversial, the three-factor model proposed by Lin, Wang, Cai (2012) has been empirically tested several times and the results have proved useful for most mature stock markets. Also there are still some difficulties in assessing the adequacy of observed variables as proxies for unobserved factors. The estimation of potential factors by a small number of indicators is not exact, consistent estimation of potential factors is theoretically unachievable under traditional assumptions is large and fixed, and no formal test exists to compare observed and potential factors. In the conclusion, they also demonstrate that the market factor is suitable to represent the risk factors of individual stocks, while other proxies are not so suitable. So finding the full range of potential variables is difficult from a practical test, and only as many potential variables as possible can be tested to approximate the final value that exists in theory.

2.2.2 Fama-French Five-Factor Model

Fama-French extends the three-factor model by adding two additional factors, profitability and investment, to better capture the cross-sectional variation in stock returns. This thesis also refers to the formula in Foye (2018) which is as follows:

$$R_{it} = R_F + \beta_1(R_{Mt} - R_{Ft}) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t + \varepsilon_{it}, \quad (2.8)$$

where RMW (robust minus weak) is the profitability factor, formed from operating profit (op); and CMA (conservative minus aggressive) is the investment factor, formed using year-on-year change in total assets.

Does the five-factor explain market returns better than the three-factor? Logically the model enhancement certainly wants the returns to be more stable. "*.....shows that the five-factor model always outperforms the FF three-factor model on all metrics, including the GRS F-statistic, in that it lessens the average returns that remain unexplained. In addition, the five-factor model leaves only 25%–41% of the cross-section dispersion of average excess returns unexplained, much lower than those in FF (42 – 54%), while the dispersion of average excess returns left unexplained by the traditional three-factor model is more than half.*" (Lin, 2017, p. 159) Based on this result, this thesis provides a hypothesis: the final result would be better if more effective factors of different classifications could be found.

2.2.3 Barra Integrated Model

The Barra integrated model, which explains portfolio risk and return through a common set of factors, is still widely used among institutional investors. If the model described above is the basis, the Barra model is most similar to the factor investment process in this thesis, both of which use a large number of different classes of factors to find sources to explain excess returns. Since the model is still being updated and widely attempted to be used profitably, this section is presented with reference to the USE4 of the Barra model (The Barra US Equity Model, 2011) and publicly available information from Investopedia.

Similar to the above, the Barra model incorporates a wide range of factors, from sector, country, macro, financial, and technology all of which correspond to returns, capturing different sources of risk and return in the market. These fundamental factors include various characteristics such as yield, return growth, volatility, liquidity, momentum, size, P/E ratio, leverage, and growth. By moving from quantitative but unspecified factors to easily identifiable fundamental characteristics, these factors are used to fully describe the risk or return of a portfolio or asset. Barra's method formula is as follows:

$$\tilde{r}_{pj} = \sum_i b_i^j \, \tilde{r}_i = \sum_i b_i^j \left(a_i + b_i^1 \tilde{f}_1 + b_i^2 \tilde{f}_2 + \cdots + b_i^K \tilde{f}_K + \tilde{e}_i\right)$$

$$= a_{pj} + \tilde{f}_1 \sum_i b_i^1 b_i^j + \cdots + \tilde{f}_j \sum_i \left(b_i^j\right)^2 + \cdots + \sum_i b_i^j e_i \,. \tag{2.9}$$

Due to Z-score processing,

$$\sum\nolimits_i \left(b_i^j\right)^2 = 1 \,, \tag{2.10}$$

where $\tilde{r}_{pj}$ is factor mimicking portfolio for factor j, b_i^j is the factor exposure value for the jth factor, $\tilde{f}_j$ is the factor value, a_{pj} is constant term, e_i is the random error term.

On the other hand, the Barra model also has shortcomings, one of the more important ones being that the final selection of factors is expressed as a linear relationship, and interpreting the market in terms of a linear relationship is still controversial, and the robustness, i.e., reliability, of its interpretation is still unknown. Logical judgments do not necessarily fit perfectly with the market's return interpretation. Another problem lies in the time, due to the rapid changes in the market, the data required by the model may have to be more compact, this thesis takes data in days in order to fit the market as much as possible, but its effect is not addressed if it is in minutes.

2.3 Introduction of Factors

In this section, several concepts of factors in factor investing are introduced. Starting with the classification of factors, the classification of factors in this thesis differs from the universal classification due to the fact that it was mentioned above that the lower the usage of the factors searched for the higher the possibility of additional returns. This section then explains the concept of factor exposure, i.e., the sensitivity of a portfolio to a particular factor. Finally the concept of factor returns is introduced, i.e., the excess returns generated by a portfolio as a result of exposure to a particular factor. These are the basis for constructing a successful factor investment strategy that can lead investors to be able to identify and target specific factors that are likely to generate positive returns over time.

2.3.1 Classification of Factors

The current application of factors is mostly based on the Fama Five Factors (see *Section 2.2.2*) for empirical testing of the broad class of factors, but in practice, the main objective of investors is to obtain excess returns and do not care much whether the academic hypotheses are met. Referring to Fons, Dawson and Yau (2021) various factors

of dynamic portfolio allocation and Barra US equity model (USE4). The factors in this thesis mainly use the following subfactors of the broad class factors, the classification is different from the universal classification, and only the broad class factors are described below, and the overview of the subfactors can be found in *Annex 1*.

Base factors refer to the factors that affect a company's basic performance, mainly referring to financial indicators. Base factors are one of the fundamental reasons for changes in asset prices, and in-depth exploration of a company's basic data is the basis for making investment decisions. It is usually a long-term value investment strategy that requires a comprehensive and in-depth analysis of a company's financial condition, industry position, management quality, and other factors, and based on these factors to construct and optimize investment portfolios. In this thesis, the focus is on the main financial indicators based on data consistency and availability.

Quality factor refers to a factor in investment that mainly refers to the quality of a company's fundamentals in equity, aiming to identify companies with high-quality fundamentals. These companies can generate stable returns over a period of time while minimizing the risk of unexpected events that may have a negative impact on the company's financial health. Some key indicators used to evaluate quality factors include EBIT on per share, liabilities per share, and earnings stability.

Growth factor refers to a factor in investment that focuses on a company's potential for earnings or revenue growth, with the main objective of identifying companies with high growth potential. These companies typically reinvest their earnings to expand their business, leading to higher earnings and revenue growth rates. Compared to value investing, which focuses on finding relatively cheap stocks, growth investing emphasizes finding companies with higher growth potential, especially in high-growth industries such as technology, biotechnology, and the internet. When evaluating growth factors, key indicators include operating profit growth ratio and net profit growth ratio.

Market factors attempt to capture the overall behavior and direction of the stock market and explain portfolio or individual stock returns by identifying common risk factors across the market. Its key characteristics include sensitivity to changes in market conditions and its ability to affect portfolio returns. But specific factors may vary depending on the modeling approach used, such as market breadth, volatility and trading volume to capture the overall momentum of the market.

Emotional factors attempt to capture the impact of investor emotion and emotional bias on asset prices. These factors are based on various technical indicators and indicators of market activity that can be influenced by emotions such as fear, greed and optimism. Emotional factor indicators are derived from market data, such as trading volume, price and time. They can be used to identify trends, reversals and momentum in the market and generate trading signals based on these patterns. Traders and investors tend to act in a predictable manner when experiencing certain emotions, such as panic or euphoria. Capturing the irrational and unpredictable behaviour of human beings in financial markets while operating in reverse can lead to more informed decisions.

The momentum factor focuses on the recent performance of a company's stock and is designed to identify companies that have performed strongly in the near term and are likely to continue. The factor is based on the fact that the stocks of companies that have performed well in the past are likely to continue to perform well in the near future and that the stock price trend will continue for some time before reversing. The metrics used to evaluate the momentum factor are the rate of change in stock prices over a given time period and the relative strength or weakness.

Technical factors in quantitative investing refer to a set of technical indicators used to analyze and predict market trends and price movements. These factors are usually based on mathematical calculations using historical market data. They are often used to identify buy and sell signals and to help investors make informed decisions about when to enter or exit the market. The main factors represented are MACD and MA. By combining these factors with other fundamental and market data, investors can develop more sophisticated trading strategies and identify market timing for better returns.

2.3.2 Factor Exposure

Multi-factor models reflect the relationship between the mean of returns and factor exposure in a cross-section, while factor exposure (factor loading) is the extent to which a portfolio or asset is affected by a particular factor in a factor model; in other words, it measures the sensitivity of an investment to changes in the value of a given factor. This exposure is not only limited to specific factors, but also exists in systematic risk. It is an "index" or a "measure" that shows the factors that together make up a whole. It shows the "relative importance" or "magnitude" of a collection of items (characteristics, features) that together form a whole. The formula is as follows:

$$r_n(t) = \sum_k x_{n,k}(t) \cdot b_k(t) + \mu_n(t), \tag{2.11}$$

where decompose the rate of return linearly. $r_n(t)$ is the excess return of asset n from moment t to $t+1$, $x_{n,k}(t)$ is the exposure of asset n to factor k at moment t, and $b_k(t)$ is the factor return of factor k from moment t to $t+1$, and $\mu_n(t)$ is the idiosyncratic rate of return of asset n from moment t to $t+1$.

2.3.3 Factor Returns

Factor returns, which are returns generated by specific factors or systematic sources of financial market risk, are persistent and systematic drivers of returns that can be used to explain a cross-section of expected returns in a diversified portfolio of securities. They are typically estimated by constructing factor portfolios or factor scores that capture the return patterns of individual securities relative to a particular factor or set of factors. These factor portfolios or scores are then used to calculate the returns generated by a particular factor over a given time period. The formula is as follows:

$$r_{i,t} = \vec{f}_t^T \vec{b}_{i,t} + \varepsilon_{i,t}, \tag{2.12}$$

where $r_{i,t}$ is factor return ratio, $\vec{f}_t^T$ is factor vector at time t, $\vec{b}_{i,t}$ is the coefficients of the factors in period t, $\varepsilon_{i,t}$ is random error term.

Compared to the above, the formulas for these concepts are more similar, but not identical, including being fitted and vectors, etc. It is also worth noting that factor returns are not fully guaranteed and can fluctuate significantly over time. Risk tolerance and investment objectives need to be evaluated before incorporating a factor investment strategy into their portfolio.

2.4 Factor Investing Process

Factor investing is a methodical approach to selecting stocks based on certain factors empirically associated with higher returns, with the goal of identifying stocks with higher potential returns while maintaining an appropriate level of risk. This process requires an integrated approach that includes building factor libraries, testing individual factors, synthesizing factors into a model, and screening outperforming stocks. The results of the entire process form the basis for portfolio optimization in *Chapter 3*. It should be noted that due to differences in factors such as capital scale and quality of listed companies, adjustments must be made before making comparative calculations. In

addition, risk control and portfolio adjustment can be achieved by modifying factor exposure, which does not conflict with subsequent chapters. The process is structured in five steps: building factor libraries, factor data unification, single-factor test analysis, factor synthesis, and screening portfolio components. Details are as follows.

2.4.1 Constructing Factor Libraries

Building a factor library involves creating and maintaining a comprehensive database of factors that can be used in asset pricing models and quantitative investment strategies. Factor libraries typically include a broad range of factors, and the various types of factors need to be categorized and selected to determine the factors that are appropriate for one's investment strategy and objectives. Common factor classifications include macroeconomic factors, industry factors, financial factors, technology factors, etc. These factors are considered to capture different aspects of systematic risk to ensure the relevance and robustness of the included factors. In general, effective factors should have strong market predictive power and stable historical performance, and low correlation with other factors that can independently explain the return performance of a security or portfolio. Various techniques such as regression analysis, time series analysis and machine learning algorithms can be used to identify and validate potential factors.

Constructing a factor pool involves identifying various potential factors that may affect asset returns and then testing these factors to determine which factors are statistically significant in explaining returns. Factors that pass this test are included in the pool of valid factors, while those that do not pass are discarded. The process of alternative factor pooling, on the other hand, is an iterative process, meaning that factors are continuously evaluated and refined as new data becomes available or as market conditions change. This allows the factor pool to remain relevant and valid over time. The current factor library construction in this thesis is an alternative factor library, a collection of as many diverse factors as possible, with the ultimate goal of creating a robust set of factors that can be used to identify and explain the sources of risk and return in a portfolio.

2.4.2 Factor Data Unification

Factor data unification is the process of combining multiple data sources related to financial and investment factors (e.g., risk, return, and other statistical indicators) into a single comprehensive data set. This dataset can then be used to analyze and model the relationships between different factors and their impact on investment performance. It is

necessary to first need to winsorize the data to make it smoother, with the following equation:

$$X_i' = \begin{cases} X_{median} + nMAD_e, if\ X_i > X_{median} + nMAD_e, \\ X_{median} - nMAD_e, if\ X_i < X_{median} - nMAD_e, \\ X_i, \qquad if\ X_{median} - nMAD_e < X_i < X_{median} + nMAD_e, \end{cases} \tag{2.13}$$

where X_i' is the result after winsorization, X_{median} is the median of the data X, $nMAD_e$ is the data that deviates n times from the expected median X.

In this thesis, the triplet method is implemented, so n=3. In order to compare different aspects of the factors to enable subsequent machine learning, further Z-score standardization methods are required. Z-score standardization is a statistical method for converting a set of raw data into standardized values or Z-scores for use in data analysis and statistical modeling to compare different data sets with different scales or units of measurement variables or data sets with different scales or units of measurement. Calculated by subtracting the mean of the data from each data point and dividing the result by the standard deviation of the data, this process generates a distribution of values with a mean of zero and a standard deviation of one. The formula is as follows:

$$X_i'' = \frac{X_i' - \bar{X}_i'}{\sigma_{X_i'}}, \tag{2.14}$$

where X_i'' is a standardization factor, $\bar{X}_i'$ is the average value after winsorization, $\sigma_{X_i'}$ is the standard deviation after winsorization. At this point, the processed factor data is obtained in this thesis, and the next step is to test the single factor data.

2.4.3 Single Factor Test Analysis

Single factor test analysis is a tool to identify the impact of specific factors on investment performance in order to determine the impact of specific factors on a particular asset or portfolio. In addition, historical relationships between variables may not hold in the future, thus requiring the use of information coefficients (IC) in an integrated judgment. It is the correlation coefficient between the model's predicted value and the realized returns of the asset or portfolio under analysis. The higher the coefficient, the better the model is in predicting the actual return of the asset or portfolio. Combined with with *Section 2.3.2*, the IC for a given period is the correlation coefficient in the cross-section between the value of the factor exposure for that period and the actual return of the stock for the next period. Its formula is as follows:

$$IC_t = Correlation(\overrightarrow{f_t}, \vec{R}_{t+1}), \tag{2.15}$$

where IC_t represents the IC value of period t, $\overrightarrow{f_t}$ is factor vector represents the factor value of period t, and $\vec{R}_{t+1}$ ret vector represents the return rate of period $t+1$.

In addition, the information ratio (IR), also called adjustmental IC (AIC), is a performance metric that needs to be evaluated. It is a measure of the risk-adjusted performance of an investment or investment strategy. It is calculated by dividing the excess return of an investment or strategy by its tracking error, which is a measure of the volatility of returns relative to the benchmark. On the other hand, the breadth and consistency of factor signals can also be used to measure the predictive power of investment factors. The formula is as follows:

$$IR_t = Adj\ IC_t = \frac{\overline{IC_t}}{\sigma_{IC_t}}, \tag{2.16}$$

where IR_t is the information ratio, $Adj\ IC_t$ is the adjustmental IC, $\overline{IC_t}$ is the single-factor IC mean value, σ_{IC_t} is the single-factor IC standard deviation.

After calculating the two judgment values a scoring method can be used to filter the factors according to universal criteria, and one point is scored for meeting one criterion. However, this method has a shortcoming: the criteria related to the IC value are not fixed. Since the market is constantly changing, the universal single-factor judgment criteria may be judged invalid with different periods, and the judgment parameters can be adjusted according to the situation of the factors themselves, so this thesis gave up using this method. Instead, this thesis adopt the multiple linear regression approach, as shown in the following chapter.

2.4.4 Factor Synthesis

Factor synthesis is the process of creating new investment factors by combining or transforming existing factors, since single factors do not fully explain market excess returns, in this thesis we use multiple linear regression for factor synthesis, as described in *Section 3.3.3*. Before that it is necessary to explain the correlation between factors.

The main correlations currently used include Spearman correlation and Pearson correlation, and the main difference between them is the different types of data used for the analysis. The Pearson correlation coefficient is used to measure the linear relationship between two continuous variables. It assumes that the relationship between two variables

is linear and that the data is normally distributed. The Pearson correlation coefficient has a range of -1 to 1, where -1 indicates a perfectly negative correlation, 0 indicates no correlation, and 1 indicates a perfectly positive correlation. On the other hand, the Spearman correlation coefficient is a nonparametric correlation measure used to measure the monotonic relationship between two variables regardless of whether the relationship is linear or not. It is often used when the data are ordered, non-normal, or when there are outliers. The Spearman correlation coefficient takes values from -1 to 1, where -1 indicates a perfectly negative monotonic correlation, 0 indicates no correlation, and 1 indicates a perfectly positive monotonic correlation.

From this it can be obtained that Pearson correlation is more sensitive to outliers than Spearman correlation, while Spearman correlation is more robust to outliers. The Spearman correlation matrix between different factors or different factor IC needs to be derived, and the results can be seen in *Annex 4*. Overly correlated and uncorrelated factors need to be excluded, and in addition Pearson correlation is still used in the other steps of this t-investment process.

2.4.5 Screening Portfolio Components

After the factors have been screened and synthesized, we can trade the market based on them, but the problem is that the individual investor has limited capital, plus the minimum individual stock investment limit of 100 shares, so this thesis screens the portfolio components by trading frequency and holding time. The holding stock is the historical period that matches the synthetic factor, the holding stock benchmark is fixed, and the portfolio optimization can be continued next.

3 Portfolio Optimization Theory and Methods

Portfolio optimization theory and methods are critical for investors seeking to maximize returns while managing risk. In this chapter, this thesis introduces the theoretical foundations of portfolio optimization and explores various methods for achieving optimal factor portfolio allocations. Central to portfolio optimization is the Markowitz mean-variance model, which forms the basis of modern portfolio theory. This is followed by an examination of the Hierarchical Risk Parity (HRP) approach, which aims to divide assets into clusters and manage risk at each level of the hierarchy.

In order to assess portfolio performance, various portfolio evaluation metrics are also discussed, including variance, Sharpe ratio, Calmar ratio and value at risk. These metrics can help investors assess the risk and return characteristics of their portfolios. In addition, this thesis explores the use of factor investing in the optimization process, which involves identifying and exploiting systematic sources of risk and return in financial markets, using various factor feature identification and synthesis methods, including Principal Component Analysis (PCA), Extreme Gradient Boosting (XGBoost), Multiple Linear Regression (MLR).

It is worth noting that factor-based stock selection strategies can also be optimized through portfolio construction. Specifically, the weights of the factors can be adjusted by controlling the factor exposure. As a result, the portfolio components will evolve according to the specific multi-factor portfolio approach adopted and evolve accordingly as the stock pool is screened. The combination of factor eigenvalues and returns will vary due to different controls over investment risk and return, leading to multiple optimizations of the portfolio. The optimization process is guided by different constraints in order to satisfy the realities and preferences of different investors.

3.1 Theoretical Foundations of Portfolio Optimization

In this section, this thesis examines the theoretical foundations of portfolio optimization, namely the Markowitz mean-variance model and the hierarchical risk parity (HRP) model, respectively. These models provide a framework for constructing efficient portfolios that balance risk and return. The Markowitz model is the classical approach that considers covariance among assets, while the HRP model uses a hierarchical clustering approach to portfolio optimization.

3.1.1 Markowitz Mean-Variance Model

The Markowitz mean-variance model aims to maximize expected returns while minimizing portfolio risk. By assuming that investors are risk averse, meaning that they prefer less risk and are perfectly rational when choosing investments, decisions will be made based on expected returns and the risks associated with those returns. If an agency is used for investment there are "*Markowitz assumed that, from an agent's customary or normal level of wealth, her reference point, the agent was initially risk loving then risk averse over gains, whilst initially risk averse then risk seeking over losses, and that the value function is bounded from above and below. Markowitz also assumed his representative agent is loss averse, and that individuals did not exhibit probability distortion, although he did not rule out that possibility.*" (Georgalos, Paya and Peel, 2021, p. 528) The underlying formula is as follows:

Minimum risk objective

$$\sigma_p \rightarrow min. \tag{3.1}$$

Constraints

$$\sum_i x_i = 1, \quad x_i \geq 0, for\ i = 1, 2, \dots, N. \tag{3.2}$$

Equation

$$\sigma_p = \sqrt{\sum_i \sum_j x_i \cdot \sigma_{ij} \cdot x_j} = \sqrt{\vec{x}^T \cdot C \cdot \vec{x}} \tag{3.3}$$

Maximum return objective

$$E(R_p) \rightarrow max. \tag{3.4}$$

Constraints

$$\sum_i x_i = 1, \quad x_i \geq 0, for\ i = 1, 2, \dots, N. \tag{3.5}$$

Equation

$$E(R_p) = \sum_i x_i \cdot E(R_i) = \vec{x}^T \cdot E(\vec{R}) \tag{3.6}$$

where σ_p is portfolio standard deviation, x_i is single asset weight, C is horizon expected return and the covariance matrix C, $E(R_i)$ is the single asset expected return, $E(R_p)$ is the portfolio expected return. Reference to Markowitz (1952), this section leads to the following conclusion, and the formula can be abbreviated as:

$$E = E(X_1, X_2), \tag{3.7}$$

$$V = V(X_1, X_2), \tag{3.8}$$

$$X = X_1 \geq 0, X_2 \geq 0, 1 - X_1 - X_2 \geq 0, \tag{3.9}$$

by using relations *E, V, X*, can be work with two-dimensional geometry as follows:

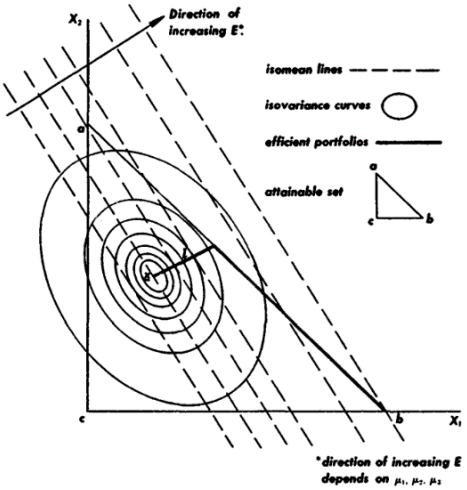


Figure 3.1 Markowitz Mean-Variance Composition Range

Source: Markowitz (1952)

The set of portfolios satisfies the constraint X. The triangle in the figure represents the achievable set of X_1, X_2. The lower and left sides of the x and y axes cannot be reached because the constraint is not satisfied. isomean curve is defined as the set of all points (portfolios) with a given expected return, and isovariance line is defined as the set of all points (portfolios) with a given variance of returns. According to the formulae for E and V it follows that the isomean curve is a system of parallel straight lines; the isovariance curve is a system of concentric ellipses. X is the center of the isovariance elliptic system, can fall inside or outside the reachable set. Due to

$$X_2 = \frac{E - \mu_1}{\mu_2 - \mu_3} - \frac{\mu_1 - \mu_3}{\mu_2 - \mu_3} \cdot X_1, \tag{3.10}$$

thus, the isovariance slope associated with $E = E_0$ is $-\frac{(\mu_1 - \mu_3)}{(\mu_2 - \mu_3)}$ and its intercept is $\frac{E_0 - \mu_1}{\mu_2 - \mu_3}$. If E is changed, it is the intercept of the isometric line that is changed, not the slope of the isometric line.

In addition to pure mean variance, socially responsible factors can also be used, along with risk-free assets, to create a capital allocation plane in a three-dimensional return/risk/socially responsible space. This is achieved by performing an aggressive

screening process of the assets included in the portfolio (i.e. excluding all assets without positive social responsibility ratings) and then optimizing for μ/σ. Implementing the model on an individual asset basis does have the potential to incorporate investors' preferences for their investments to be socially responsible. (Gasser, Rammerstorfer and Weinmayer, 2017) This provides imagination for an extension of the Markowitz model.

3.1.2 Hierarchical Risk Parity (HRP)

Hierarchical risk parity (HRP) is a portfolio optimization approach designed to achieve a more balanced and diversified risk allocation across multiple classes or asset groups, created by De Prado M L (2016). It starts with a hierarchical clustering of pairs of correlations or distances of assets. Clustering divides assets into a set of groups or clusters that reflect their similarities or dependencies. The number and size of clusters can vary depending on the chosen clustering method and parameters. This effectively takes into account the correlation between multiple assets and integrates them into a unified optimization framework.

The HRP approach is divided into three main steps. The first step is to calculate the correlation coefficient or covariance matrix between assets. The second step is to transform the correlations between assets into a distance matrix and then use a hierarchical clustering algorithm to divide the assets into different subgroups. The third step is to achieve risk balance within each subgroup by certain optimization algorithms and calculate the weights of each asset with better interpretation and operability.

The HRP model has several advantages over traditional mean-variance optimization and risk parity methods. It allows for a more detailed and fine-grained allocation of risk across multiple levels or asset groups, resulting in a more efficient and stable portfolio allocation. It also takes into account the hierarchy of assets and their interdependencies, which can capture some of the systematic risks and tail events that are often overlooked in traditional models. However, the problem is that it affects the robustness of the results and requires the estimation of risk parameters, constraints or preferences.

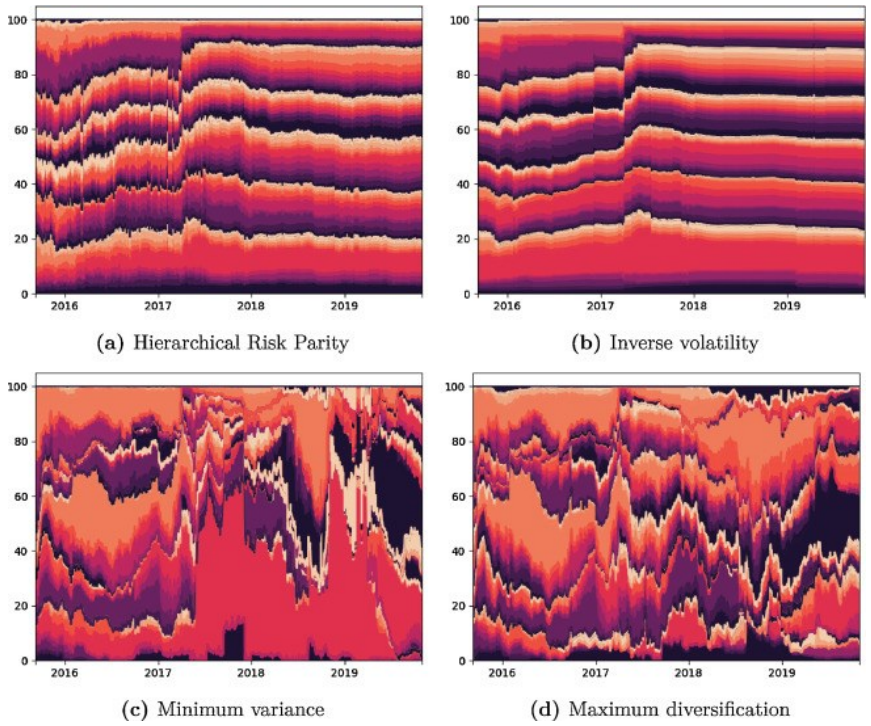


Figure 3.2 Portfolio weight decomposition of asset allocation strategies

Source: Burggraf (2021)

The *Figure 3.2* compares the weighting of different cryptocurrency asset allocation methods, and it is more obvious that HRP has an advantage, as the asset allocation to the stock pool is required later in the thesis, but the stocks picked out have been relatively stable in terms of obtaining returns over the history and the transaction costs can be saved. Their empirical evidence on cryptocurrencies has led to some conclusions: "*……our results show that HRP better navigates volatility and tail risk compared to traditional risk-based strategies. In addition, HRP has the most desirable diversification properties – while IV portfolios tend to be too static, MV results in too concentrated portfolios. Our results survive many robustness tests, including different estimation windows, covariance estimation methodologies, and rebalancing periods. Thus, HRP provides a meaningful alternative to traditional asset allocation approaches and an important risk management tool for cryptocurrency investors.*" (Burggraf, 2021, p. 6) This thesis tries to explain the whole process equation by referring to Lohre, Rother and Schäfer (2020); Millea, Edalat (2022) as follows:

Assuming N assets $X_{i=1,2\ldots,N}$, each asset has a return sequence (column vector) of length T, and these N assets (column vector) need to be constructed into a tree structure. First calculate the $N \cdot N$ correlation matrix, whose elements are $\rho = (\rho_{i,j})_{i,j=1,2,\ldots,N}$,

where $\rho_{i,j} = \rho[X_i, X_j]$ represents the correlation between two assets. Define the distance measure d as:

$$d(X_i, X_j) \subseteq B \to R \in [0,1], d_{i,j} = d[X_i, X_j] = \sqrt{0.5 \cdot (1 - \rho_{i,j})}, \tag{3.11}$$

where B is the Cartesian product of assets, and the distance matrix $D = \{d_{i,j}\}_{i,j=1,2,\dots,N}$ between assets can be calculated from the distance measure between assets. D is a complete metric space that satisfies non-negativity, consistency, symmetry and subadditivity. Second, calculate the Euclidean distance between any two column vectors in the space D as follows:

$$\tilde{d}(D_i, D_j) \subseteq B \to R \in [0, \sqrt{N}], \tilde{d}_{i,j} = \tilde{d}[D_i, D_j] = \sqrt{\sum_{n=1}^{N} (d_{n,i} - d_{n,j})^2}, \tag{3.12}$$

note that $d_{i,j}$ is defined on the asset correlation matrix, while $\tilde{d}_{i,j}$ is defined on the distance space D between assets, which represents the distance of the distance, which is a function defined on the entire correlation matrix. Next, gather the column vector (i^*, j^*) satisfying the following relationship into a category and define it as a cluster $u[1]$:

$$(i^*, j^*) = argmin_{(i,j), i \neq j} \{\tilde{d}_{i,j}\}, \tag{3.13}$$

then, define the distance between the newly constructed category $u[1]$ and the N independent elements in the original D space. When $i = 1, j = 2$ the distance between two elements in D space is $\tilde{d}_{1,2}$, now it is necessary to find the distance $\bar{d}_{i,u[1]}$ of $i = 1,2, \dots, N, j \in u[1]$, which is defined as follows:

$$\bar{d}_{i,u[1]} = min\left[\{\tilde{d}_{i,j}\}_{j \in u[1]}\right], \tag{3.14}$$

update matrix $\tilde{d}_{i,j}$: first add $\bar{d}_{i,u[1]}$ to $\tilde{d}_{i,j}$, then remove the rows and columns contained in $u[1]$. Finally, formulas (3.13) and (3.14) are iterated, and a new cluster $(N-1)$ is found each time until all original categories are included in a large cluster, and the clustering algorithm ends.[1]

[1] Example information can be seen: https://hudsonthames.org/an-introduction-to-the-hierarchical-risk-parity-algorithm/

3.2 Portfolio Evaluation Indicators

In this section, various portfolio evaluation metrics are discussed, including variance, Sharpe ratio, Calmar ratio, and value at risk (VaR). These metrics are used to evaluate the performance of portfolios and to compare the risk and return characteristics of different portfolios. The variance measures the deviation of returns from the mean, while the Sharpe ratio adjusts for risk and measures excess returns relative to the risk-free rate. The Calmar ratio compares the average annual return to the maximum drawdown, while the VaR estimates the potential loss that could occur in a worst-case scenario.

3.2.1 Variance

Variance is a statistical measure of the dispersion or dispersion of a set of values. In a portfolio optimization scenario, variance represents the level of risk associated with a portfolio. The variance of a portfolio is calculated as the sum of the squared deviations of each asset's return from its expected return multiplied by their respective weights in the portfolio. A portfolio with high variance indicates that the portfolio's returns are spread over a broader range of values, indicating a higher level of risk. Conversely, a portfolio with low variance indicates that the returns are less volatile and more similar to the expected value, resulting in a lower level of risk. The formula is expressed as follows:

$$E(R_i) = \sum_{i=1}^{N} p_i \cdot R_i \,, \tag{3.15}$$

$$var(R_i) = \sum_{i=1}^{N} p_i \cdot [R_i - E(R_i)]^2, \tag{3.16}$$

where $E(R_i)$ is the expected return of a single asset, p_i is the asset's likelihood in period i, R_i is the expected return of a single asset in period i for duration N, and $var(R_i)$ is the variance of a single asset. Based on the above formula this thesis can find the variance of individual equity assets, followed by the variance of the portfolio as described below:

$$cov(R_i; R_j) = \sum_{i,j} p_{i,j} \cdot [R_i - E(R_i)] \cdot [R_i - E(R_j)], \tag{3.17}$$

$$var(R_p) = \sum_{i} \sum_{j} w_i \cdot w_j \cdot cov(R_i; R_j) \,, \tag{3.18}$$

$$\sigma_p = \sqrt{var(R_p)}, \tag{3.19}$$

where R_j is the expected return of a single asset in period j, the rest are the same as R_i, $cov(R_i; R_j)$ is the covariance between the two assets and $var(R_p)$ is the variance of the portfolio, σ_p is the standard deviation of the portfolio.

The calculation of volatility by variance is a measure of risk, the dispersion of returns near the expected return, which can measure the potential loss of its portfolio. Logically equity returns have the potential for excess returns only in volatility, but it is also possible to capture low-profit firms. "*The aggregate volatility risk explanation of the profitability anomaly relies on the fact that option-like equity of unprofitable/distressed firms benefits from increases in idiosyncratic volatility, which tend to coincide with increases in market volatility.*" (Barinov, 2022, p. 17)

3.2.2 Sharpe Ratio

The Sharpe ratio is used as a performance indicator to calculate the risk-adjusted return of an investment or portfolio to assess the efficiency of the investment strategy and whether the additional risk taken is worth the additional return generated. It is calculated by dividing the excess return of an investment or portfolio over the risk-free rate by the standard deviation of the excess return, which is usually a government bond, but is recorded as 0 in the fixed value calculation in this thesis since it is used as a benchmark through an index. The formula is as follows:

$$SR = \frac{R_p - R_I}{\sigma_p}, \tag{3.20}$$

where SR is the Sharpe ratio, R_p is the portfolio return, R_I is the index return, σ_p is the portfolio return standard deviation.

So can the Sharpe ratio empirically measure the risk ratio of investors expected instantaneous returns? Sharpe ratios can also show an increasing trend in times of world crisis, but the problem is that, given the limitations of the model, in some countries where there may also be a risk of default on long-term government debt, there is no linear trend in Sharpe ratio changes, which may be difficult to predict. (Vukovic, Vyklyuk and Matsiuk, et al, 2020) Sharpe ratios with long-term information tend to be lower than Sharpe ratios with short-term information, on the other hand: regardless of the investment style, Sharpe ratios are sufficiently robust for assessment in the classical Markowitz setup, i.e., in the mean-variance framework. (Guo, Ou, 2021)

3.2.3 Calmar Ratio

Calmar ratio is a performance metric that calculates the ratio of an investment's or portfolio's average annualized return to its maximum drawdown. Similar to the Sharpe ratio, the former is more concerned with the ability to control the risk of maximum drawdown, while the latter focuses more on the volatility of the portfolio. The Calmar ratio is particularly useful for evaluating investments or portfolios that have experienced significant drawdowns, as it penalizes investments with high risk but low returns, again assuming a normal distribution of returns. Its formula is as follows:

$$MDD = \frac{T - P}{P}, \tag{3.21}$$

$$CR = \frac{R_p - R_I}{MDD}, \tag{3.22}$$

where MDD is the maximum drawdown, T is the trough value, P is the peak value; CR is the Calmar ratio, R_p is the portfolio return, R_I is the index return.

3.2.4 Value at Risk

VaR (Value at Risk) is a method of measuring the risk of a financial asset or portfolio, using statistical analysis to estimate future changes in the value of the asset or portfolio and mapping it to a probability distribution to obtain a value at risk. It can give the maximum possible loss of an asset or portfolio in a certain period in the future at a certain confidence level (such as 95% or 99%). Referring to Petneházi (2021); Ahmed, Soleymani and Ullah, et al (2021) the formula is as follows:

$$VaR_\theta(X) = F_{-X}^{-1} \cdot (1-\theta) = -\inf\{x \in R : \theta \le F_X(x)\}, \tag{3.23}$$

$$VaR_\alpha(X) = \min\{z \epsilon R | F_X(z) \ge \alpha\}, \tag{3.24}$$

where $F_X(\cdot)$ is the cumulative distribution function (CDF), X stands for a random quantity, and the pre-determined α stands for confidence level. Sometimes, min in (3.25) is switched by inf and $1-\alpha$ is called the confidence level and α as the risk/tail level. In the context of the portfolio the following is obtained:

$$VaR = \Phi^{-1}(1-\alpha) \cdot \sigma_{\Delta\tilde{\Pi}} - E(\Delta\tilde{\Pi}), \tag{3.25}$$

where $E(\Delta\tilde{\Pi})$ is the mean value of the portfolio increment, $\sigma_{\Delta\tilde{\Pi}}$ is the standard deviation of portfolio increments, and $\Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$ quantile of cumulative distribution function of standardized normal probability.

VaR can be calculated using various methods, the most common of which are the historical simulation method, the parametric method and the Monte Carlo simulation method. The historical simulation method is to sort the historical return series by time, select a confidence level and a look-back period, and calculate the maximum possible loss during the look-back period at the confidence level based on the historical return series. Parametric method is to calculate VaR by analyzing the statistical characteristics of assets or portfolios and constructing probability distribution models, including normal distribution, t-distribution, log-normal distribution, etc. Monte Carlo simulation method is a stochastic simulation method based on probability theory, which calculates the maximum possible loss in a certain period in the future at a certain confidence level by simulating the future price changes of an asset or portfolio randomly several times.

But no matter which method is used, VaR has some limitations. First, VaR only considers the maximum possible loss of an asset or portfolio without considering other possible risks, such as the risk of extreme events, which may result in VaR underestimating risk. Second, VaR's calculation results are affected by the quality of data and the choice of time window, which may not be accurate enough if there is insufficient historical data or the time window is too short. Finally, VaR provides only a single risk metric rather than a comprehensive assessment of overall risk; therefore, VaR should be used in conjunction with other risk metrics.

3.3 Optimization Process Based on Factor Investing

This section describes the optimization process based on factor investment. Factor optimization is first performed by principal component analysis (PCA) and extreme gradient boosting (XGBoost) for factor feature identification, followed by synthetic screening of stocks with high holding time and trading frequency using multiple linear regression (MLR), and then portfolio optimization using Monte Carlo simulation (MC) and hierarchical risk parity (HRP). These techniques and methods are widely used in financial modeling and portfolio management. Through this series of optimization techniques, stocks and weights for the portfolio are thus screened and optimized.

3.3.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is used to reduce the dimensionality of a data set while retaining as much variance as possible (the more variance retains more information about the data). The reduced dimensionality transforms multiple sets of

variables into a few (smaller) sets of uncorrelated variables. These sets of variables explain most of the variance in the original data, which is called the principal component. The aim is to reduce the number of variables in the dataset and to identify patterns and relationships among them or to identify the most important variables in the dataset, with the caveat that these variables are linearly uncorrelated and are ranked according to the amount of variance explained. "..... *PCA, while retaining trends and patterns, transforms the data into fewer dimensions, which act as summaries of features. This transformation is defined in such a way that the first few principal components capture the largest possible variance. The first component has maximum variance, the second component has maximum variance in the orthogonal direction to the first one and so on for the rest of the components, and these components are ordered sequentially with the first component describing the maximum variance.*" (Shah, Chauhan and Chaudhury, 2021, p.3)

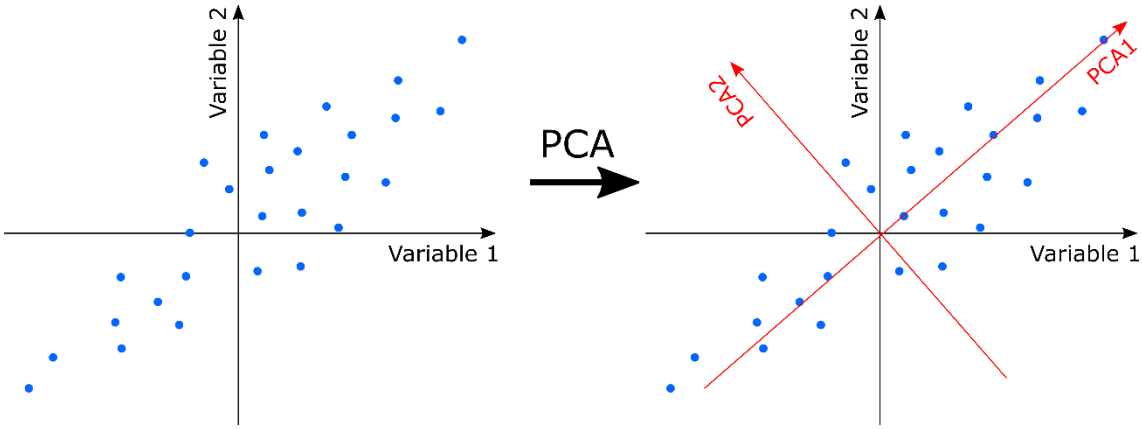


Figure 3.3 Principal component analysis process diagram

Source: https://ourcodingclub.github.io/tutorials/ordination/

So does the PCA method of practical use or not? The answer may be found based on some extensions of the PCA approach. Zhang, Wang (2023) applied a combined PCA-ASR approach to conclude that its forecasts are robust to alternative combination methods, and in this case for forecasting oil futures returns over longer periods, the PCA-ASR approach is still the preferred choice among all models used. Beyond that, perhaps the role of PCA would be more imaginative. When forecasting volatility sentiment for the Chinese stock market, it uses from the scaled PCA method, which extracts more useful information from proxies and target variables than existing methods. And future research may shift to large-scale empirical investigations to apply investor sentiment constructed

by scaled PCA methods, and it is also interesting to employ more dimensionality reduction techniques for supervised learning. (Song, Gong and Zhang, et al, 2023)

3.3.2 Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting (XGBoost) is an efficient integrated learning algorithm, improved by Chen, Guestrin (2016) on the basis of decision tree-based gradient boosting algorithm, which has become a more outstanding achievement in parallel computational efficiency, missing value handling, prediction performance and automatic feature selection, and has become a popular and effective tool for solving various XGBoost can also combine multiple weak models to create strong models, use a series of decision trees and apply gradient boosting techniques to minimize the overall loss function. A series of decision trees are built iteratively to minimize the loss function, which measures the difference between predicted and actual values, and during each iteration, errors are evaluated and the weights of each instance are adjusted to focus on misclassified data points.

In this thesis, based on the stock index direction prediction of Han, Kim and Enke, (2023) and the machine learning trading system of Deng, Huang and Zhu, et al, (2023), the following formula is obtained:

Assuming that K trees have been trained, the final predicted value for the i-th sample is equal to:

$$\tilde{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{3.26}$$

where x_i represents the sample characteristics, $f_k(x_i)$ is the prediction result of the k-th tree for sample x_i, and finally all these values are added together to get the final result $\tilde{y}_i$, and F is the set of all decision trees. Combined with the true result label (target variable) is y_i the loss function (to the current k trees cumulative loss function) can be constructed as follows:

$$Obj = \sum_{i=1}^{n} l(y_i, \tilde{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3.27}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2, \tag{3.28}$$

where l represents the loss function, and common loss functions are used. The latter item Ω is used to control complexity and prevent overfitting, also called model regularization.

γ represents the parameter of complexity, ω is the fraction of leaf nodes, T represents the total number of leaf nodes, and λ is the regularization factor, $\frac{1}{2}$ is to facilitate the subsequent calculation of the derivative. Regularization refers to reducing the complexity of the model by adding penalty items during the training process of the machine learning model. The penalty item is an extra item of the regularized model. The optimization process will try to make the value of the model parameters close to 0, so as to achieve the purpose of reducing the complexity of the model. Split into associated formulas:

$$\begin{aligned} Obj &= \sum_{i=1}^{n} l(y_i, \tilde{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ &= \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{<k>}\right) + \sum_{k=1}^{K} \Omega(f_k) \\ &= \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{<k-1>} + f_k(x_i)\right) + \sum_{j=1}^{K-1} \Omega(f_j) + \Omega(f_K). \end{aligned} \tag{3.29}$$

Because the final prediction result is the cumulative result of all models (trees), $\tilde{y}_i$ can be written as $\tilde{y}_i^{<k>}$ (a prediction result accumulated up to the k-th tree). When training the K-th tree, minimize the following loss function:

$$Obj = \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{<k-1>} + f_k(x_i)\right) + \Omega(f_K). \tag{3.30}$$

According to the Taylor expansion:

$$f(x + \Delta x) \approx f(x) + f'(x) \cdot \Delta x + \frac{1}{2} f''(x) \cdot \Delta x^2, \tag{3.31}$$

follow by treating $\tilde{y}_i^{<k-1>}$ as x, and $f_k(x_i)$ as Δx:

$$\begin{aligned} f(x) &= l\left(y_i, \tilde{y}_i^{<k-1>}\right), \\ f(x + \Delta x) &= l\left(y_i, \tilde{y}_i^{<k-1>} + f_k(x_i)\right), \end{aligned} \tag{3.32}$$

$$\begin{aligned} Obj &= \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{<k-1>} + f_k(x_i)\right) + \Omega(f_K), \\ &= \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{<k-1>}\right) + \partial_{\tilde{y}_i^{<k-1>}} l\left(y_i, \tilde{y}_i^{<k-1>}\right) \\ &\quad + \frac{1}{2} \partial^2_{\tilde{y}_i^{<k-1>}} l\left(y_i, \tilde{y}_i^{<k-1>}\right) \cdot f_k^2(x_i) + \Omega(f_K), \end{aligned} \tag{3.33}$$

the current objective function is the function when training the k-th tree, where the $l\left(y_i, \tilde{y}_i^{<k-1>}\right)$ item is the loss of the real value and the accumulated prediction results up

to the $k-1$-th tree, which can be regarded as known and does not participate in the optimization process.

Make g_i and h_i are the gradient statistics of the loss function, expressed as:

$$g_i = \partial_{\tilde{y}_i^{<k-1>}} l\left(y_i, \tilde{y}_i^{<k-1>}\right), h_i = \partial^2_{\tilde{y}_i^{<k-1>}} l\left(y_i, \tilde{y}_i^{<k-1>}\right), \tag{3.34}$$

can be further optimized as:

$$\begin{aligned} Obj &= \sum_{i=1}^{n} \left[g_i \cdot f_K(x_i) + \frac{1}{2} h_i \cdot f_K^2(x_i) \right] + \Omega(f_K) \\ &= \sum_{j=1}^{n} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T, \end{aligned} \tag{3.35}$$

when the structure of the tree is fixed, that is, if $q(x)$ is fixed, the optimal ω_j is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_j + \lambda}, \tag{3.36}$$

and the best objective function value under the current tree structure:

$$Obj^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_j + \lambda} + \gamma T. \tag{3.37}$$

Iterative addition of branches to the tree is generally required. Assuming that I_L and I_R are the instance sets of the left and right nodes after splitting, then the loss reduction value can be written as:

$$Gain = \frac{1}{2} \left(\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right) - \gamma. \tag{3.38}$$

Finally, this thesis uses rank-pairwise to sort the importance feature scores of the loss reduction, and the results can be seen in *Figure 4.20.*

Whether XGBoost is useful or not, by exploring Deng, Huang and Zhu, et al (2023, p. 14) mentioned above, the following results can be found "*……an explainable XGBoost based approach was adopted for direction forecasting and simulation trading. The experimental results show that the XGBoost model outperforms OLS, KNN, ANN, throughout the testing period, SVM, and RF, among other traditional methods. The experimental results show that the proposed method can be applied as a reliable method for direction forecasting and simulation trading of the Shanghai Composite Index and Shenzhen Composite Index. In addition, the relative importance score and SHAP results*

provide the most influential sentiment factors for market participants to predict the directional changes of the Shanghai Composite and Shenzhen Composite." This gives evidential support to the importance level ranking (see *Section 4.4.2*) in this thesis.

3.3.3 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) quantitatively portrays the linear correlation between a response variable and multiple independent variables using regression equations to estimate the relationship between the dependent variable and each independent variable by fitting a linear equation that explains the change in the response variable based on the change in the predictor variable. It is important to note that regression analysis can only be carried out if a causal relationship between multiple independent variables and the dependent variable is theoretically satisfied. The independent variables are not highly correlated with each other, and even if the regression analysis is significant, the causal relationship is only statistically significant as a possibility. In MLR, the linear equation is usually expressed as:

$$y_t = \beta_0 + \beta_i x_t + \varepsilon_t, \tag{3.39}$$

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{\mathrm{i}} x_{\mathrm{t}} + \varepsilon_t, \tag{3.40}$$

the parameters β_0, β_1,..., β_{i} are to be estimated in the model, where ε_t are independent random variables with a normal distribution $N\,(0, \sigma^2)$. y_t represents the dependent variable, the variable that is wanted to predict or explain. And x_{t} represents the independent variable, the variable or variables that are used to predict or explain y_t the regression coefficient, β_{i} is called the regression coefficient, which characterizes the degree of influence of the independent variable on the dependent variable.

3.3.4 Monte Carlo Simulation (MC)

Monte Carlo simulation (MC) complex systems by generating random samples to estimate statistical probabilities and assess risk, and then using statistical analysis to estimate the behavior or properties of the system. Since a large number of random samples are drawn from a probability distribution to create a distribution of possible outcomes, thus representing the model built without certainty of inputs or parameters, its easier to apply to problems where it is difficult or impossible to obtain an analytical or deterministic solution, a flexible and powerful method to explore test hypotheses and to assess the robustness and sensitivity of the system to changes in inputs and parameters. As mentioned above, the validity of financial markets is still in doubt and the pattern of

market volatility remains elusive, using estimating risk in a financial portfolio and simulating the behavior of financial markets, especially when there is a minimum purchase is applicable.

The impact that Monte Carlo simulation can have on a portfolio is multifaceted. The monthly values corresponding to each portfolio are defined as random variables with specified means and standard deviations to incorporate uncertainty into the optimization process, and then the probabilistic optimization problem can take into account the uncertainty of the risky assets.(Shadabfar, Cheng, 2020) Alternatively, the use of parametric constraints generates portfolio structures with the same joint distribution and can well preserve the nonlinear correlation between individual stocks. Seyfi, Sharifi and Arian (2021) performs the portfolio selection problem in a multiple-input multiple-output setup, which is based on a statistical approach to the problem, generates appropriate stochastic portfolios and estimates non-convex efficient bounds. Its results are more accurate and robust to numerical accidents. It is also good to analyze the effect of sample size with simulated data, but on the other hand also focus on different aspects of sample size. (Nalpas, Simar and Vanhems, 2017)

But Monte Carlo simulation also has some limitations, requiring the selection of probability distributions and sampling methods to ensure a representative and independent sample. Sufficient computing power and time to generate enough samples and run the model multiple times. The accuracy and reliability of the results depend on the model settings and parameters, the validity of the assumptions, and the consistency of the data used. If available, sensitivity analysis, uncertainty quantification, and procedure validation are still required.

3.3.5 Hierarchical Risk Parity (HRP)

Hierarchical risk parity (HRP) literally means that assets need to be judged hierarchically. First calculate the similarity matrix: Calculate the similarity between assets by choosing a similarity measure, such as Pearson correlation coefficient or Spearman rank correlation coefficient. Then construct the clustering tree, use the similarity matrix, construct the clustering tree through the hierarchical clustering algorithm. The ward algorithm is usually used, its advantage is that the resulting clustering retains the original data structure to the greatest extent possible, and does not merge isolated data points together. Then the clusters are generated, and the clusters are obtained according to the pruning of the clustering tree.

After obtaining the clustering tree, it is still necessary to calculate the covariance matrix of the clusters. For each cluster, calculate the covariance matrix of its component stocks. Then the covariance matrix of the top-level clusters is generated, and the covariance matrix of the entire investment portfolio is obtained by clustering the covariance matrix of the clusters again. Finally, the portfolio weights are optimized, using the selected risk measure (such as variance) and constraints (such as the sum of asset weights being 1) to optimize the portfolio weights. The goal of HRP is to group assets and reduce inter-group correlation while maintaining high similarity within the group, thereby reducing overall risk and possible duplication.

4 Application of Factor Investing and Portfolio Optimization

The main purpose of this chapter is to provide a practical guide to applying factor investing to portfolio optimization from the perspective of the individual investor. It should be noted in advance that the key to factor investing, as an investment approach to systematically capture returns associated with different types of risk factors, is to capture factor characteristics, but different capture methods will promote completely different factor synthesis results. And due to the variability of equity markets, the final results are not always better.

The chapter begins with an overview of current stock market conditions, which provides the context for the analysis in subsequent chapters. The chapter then describes the data used in the analysis, including data sources, selection criteria and descriptive statistics. It then proceeds to test factor investing, which is the core of the empirical evidence, before introducing single-factor returns, information coefficients (IC) and trade tests for assessing the effectiveness of individual factors in generating alpha. Finally, portfolio optimization is discussed, which is the process of constructing an optimal portfolio that maximizes returns while minimizing risk. This section discusses several popular optimization methods to capture factor characteristics through Principal Component Analysis (PCA), Extreme Gradient Boosting (XGBoost), Multiple Linear Regression (MLR) for fitting, Monte Carlo simulation (MC) and Hierarchical Risk Parity (HRP) for further optimization of screened stocks. By providing practical insights and tools, investors can use these to make better investment decisions and achieve better risk-adjusted returns.

4.1 Current Stock Market Status

"The Chinese stock market poses an interesting study, as it has obvious differences from the conventional markets in North America and Europe, particularly in terms of market structures, government regulations, information asymmetry, and investor composition." (Huang, 2019, p. 278) In terms of the trading system alone, the Chinese stock market is vastly different from the European and American stock markets, and it is necessary to compare it to the more familiar U.S. stock market in order to make the factor investment process run smoothly. The main differences are summarized in *Table 4.1*.

Table 4.1 China and U.S. stock market comparison

Aspects	U.S.	CHINA
Trading Time	9:30-16:00	9:30-11:30; 13:00-15:00
Trading Code	Company Abbreviation	Number Code
Trading Unit	None	Lot. (1 Lot = 100 Shares)
Trade Sell Time	T+0	T+1
Market Circuit Breaker	10%,30%; 5 min	20%, -36%-44%,10%; 1 day
Shorting Tools	Enrichment	Scarcity

Source: Own elaboration

As can be seen from *Table 4.1*, although the trading objectives are the same, there are still many differences in trading methods between the China and U.S. stock markets. Further explanation is needed: the numeric codes of Shanghai A-shares are prefixed with 600, 601 or 603, the codes of Shenzhen A-shares are prefixed with 000, and the codes of SME board are prefixed with 002. ST* is a company that has lost money for two consecutive years or its net assets are below the face value of the stock, and * means the risk of delisting. As for the conditions of market circuit breaker, the U.S. stocks are index components fluctuations of 10%, more than $1 stock fluctuations of 30%, pause time for 5 minutes; Chinese stocks are 20% of the Growth Enterprise Market, the first five trading days of the new shares listed, the daily trading fluctuations can be between -36%-44%, the main board daily fluctuations of 10%, pause time for 1 day. The time series changes of Shanghai Composite Index and Shenzhen Composite Index from 2000-2023 Q1 can be seen in *Figure 4.1*.

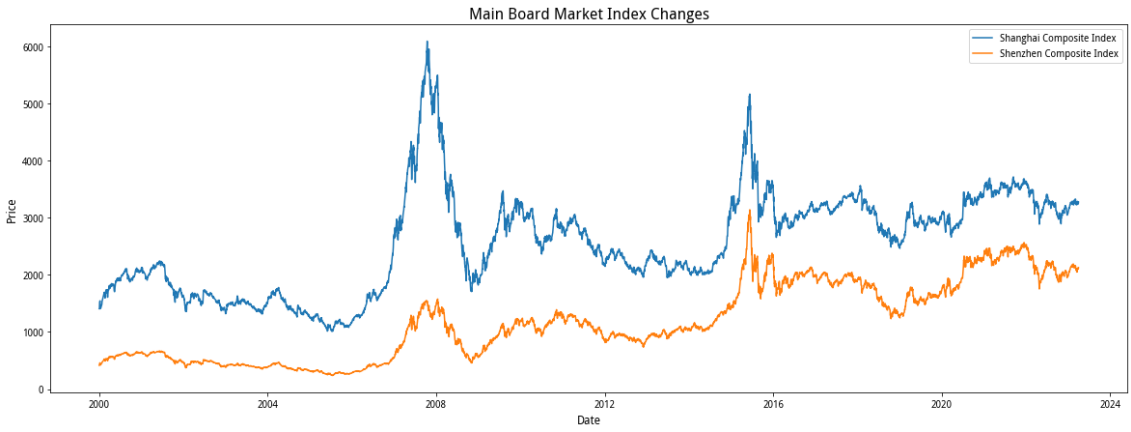


Figure 4.1 Main Board Market Index Changes

Source: Ricequant

First, in terms of the overall trend, both the Shanghai Composite Index and the Shenzhen Composite Index show a gradual upward trend. Although the indices may

fluctuate or fall during certain periods of history, the overall trend of both indices is upward in the long run. This reflects the rapid development of China's economy and the long-term investment value of the stock market.

Second, in terms of seasonal and cyclical changes, there are also significant seasonal and cyclical changes in the Shanghai Composite Index and the Shenzhen Composite Index. For example, in the first quarter of each year, the indices usually exhibit an upward trend, while in the fourth quarter they usually exhibit a downward trend.

Finally, event-driven, the Shanghai Composite Index and Shenzhen Composite Index are also influenced by political, economic and social events. Mainly in 2008 and 2016, China experienced financial crisis volatility, the Wenchuan earthquake and the opening of the Olympic Games in 2008, which were far more volatile than usual; in 2015-2016, macroeconomic pressures peaked and the regulatory policies of the Chinese stock market changed significantly, most importantly the triple impact of the proposed supply-side reform policy had a significant impact on the stock market. In addition, events such as the adjustment of government policies and the development of international trade can also have an important impact on the stock market.

4.2 Data Description

In this section, this thesis provides an overview of the data used in the study and the testing time. First, the source of the data and the process of selecting it are described. Second, descriptive statistics of the dataset are provided to provide insight into its characteristics, which are essential for interpreting the results obtained from the subsequent analysis.

4.2.1 Data Source

Considering the cost, difficulty and accessibility for the individual investor, Ricequant[2] and Joinquant[3] platforms are selected as data sources in this thesis, which can directly provide multiple types of financial information such as stock history, real-time prices and various factor data. The efficient and easy-to-use nature of Python in processing data selection also makes it one of the most popular languages in the field of quantitative investment, while the interactive computing environment based on the main use of Jupyter has the advantage of providing an interactive programming environment

[2] Ricequant: https://www.ricequant.com/
[3] Joinquant: https://www.joinquant.com/

that facilitates data visualization and analysis for investors, with each calculations can directly output graphical data and other results, which are also integrated on the web side in the Ricequant and Joinquant platforms. Its activity level is shown in *Figure 4.2*.

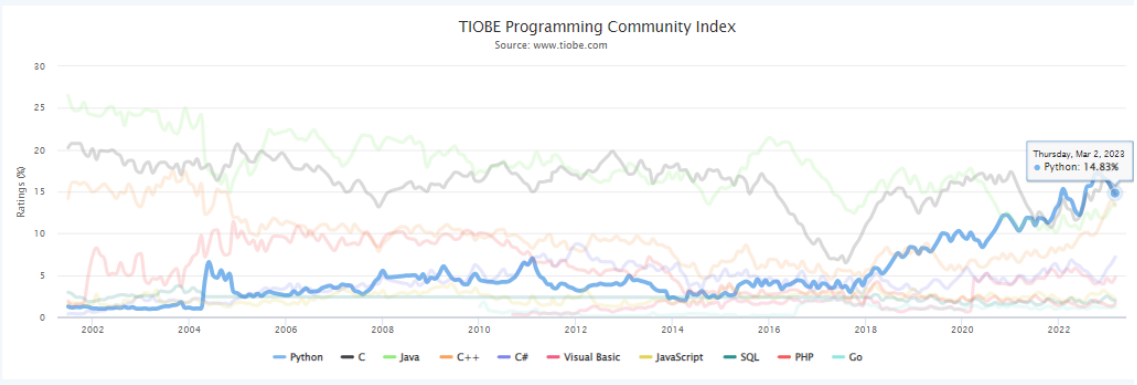


Figure 4.2 Python global community usage proportion

Source: https://www.tiobe.com/tiobe-index/

This allows investors to analyze data and build trading strategies more effectively. As technology evolves and data continues to grow, the role of these computer language kernels and platforms in the quantitative investment space will become more important and may be replaced by other platforms or languages, but the overall trend will remain one of ease of use and integration.

4.2.2 Data Selection

In order to ensure the normal operation of the entire investment process, the model needs to be tested for overall stress. Factor investment process model training, historical data collection time from 01/01/2017 – 01/01/2021, with about 240 million or more pieces of raw model data. Factor back testing, model validation time is 01/01/2021 – 01/01/2022, data volume of about 60 million or more. The subsequent portfolio optimization time for assets with higher screening frequency is 01/01/2022 – 31/10/2022, and the backtesting time is 01/11/2022 – 31/03/2023. The overall timeline is shown in *Figure 4.3*.

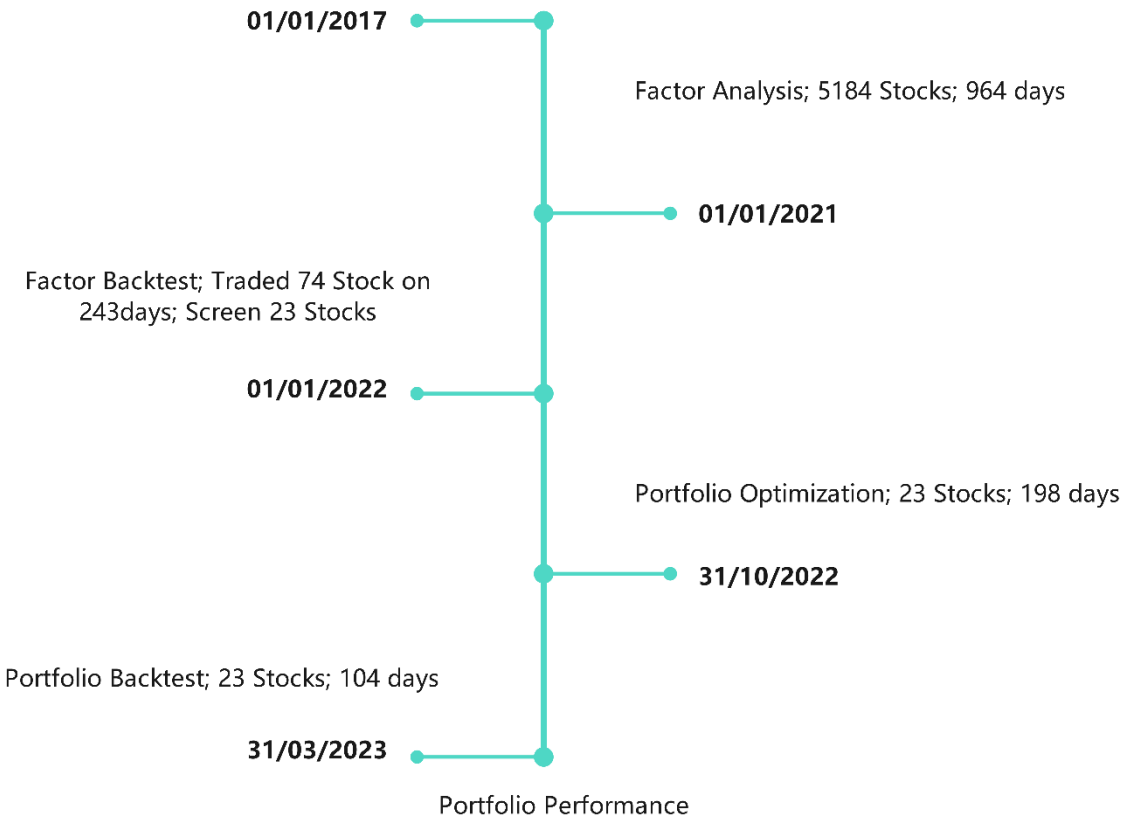


Figure 4.3 Investment process timeline

Source: Own elaboration

It should be noted that the use of the model is time-sensitive, so the selection of time needs to follow the principle of longest to shortest. That is, the factor testing time is the longest and the final portfolio optimization time is the shortest, thus enhancing the effectiveness. In addition, the choice of time point can be adjusted by oneself according to different situations. The range of stocks used in this thesis refers to all stocks in the Chinese stock market, including but not limited to stocks in Shanghai and Shenzhen stock markets; sectors include major sector stocks, small and medium-sized enterprise sector stocks, technological innovation enterprise sector stocks and ST* sector stocks. The following is a brief overview of the stock and factor data.

Another issue to note is subsequent single-factor calculations are calculated between individual stock prices and factors every trading day, and finally merged to obtain *Annex2*. But the single factor test data calculation time for each factor is about 15 minutes. Subsequent machine learning models such as increasing the number of stocks or extending the depth to more than 50, or using more complex models recommend can only use GPU, thereby saving time and reducing the possibility of crashes.

4.2.3 Data and Process Description

The overall idea applied in this thesis is to screen stocks and factors from more to less, the further the process goes the more practical and operational it is for the individual investor, but it is important to note in advance that automated trading using quantitative trading platforms for factor identification is also feasible, but the risks for factor signal identification can be seen in *Figure 4.28* and need to be treated with extreme caution.

First, in this thesis, the validity of the factors after processing the data is tested, positively and negatively, and this step is interpreted in the form of a single-factor test. For a single factor, there is still the possibility of an excess return, but the probability is like "finding a needle in a haystack". For the construction of a factor pool, any data related to the financial markets may yield a return, but the scope is wide and the common economic data plays a minimal role because of the need to judge the judgment of others. However, this step is particularly important for the comprehensive judgment of factors, whether or not they are in the valid factor pool.

Second, after all single-factor tests, valid factors are identified using XGBoost and multi-factor synthesis is performed using MLR. It should be noted that this thesis does not use PCA for synthesis because after the PCA process, new factor interpretation data will be generated and it is not possible to use historical data for simulated trading, but PCA, as an important dimensionality reduction method, will be one of the first choices for factor synthesis if it breaks the limitation of using historical data.

Then, this thesis simulate trading with multiple factors and use one year to screen "high quality" stocks with holding time and trading frequency as the stock pool for portfolio optimization. Finally, the stock weights are obtained by Monte Carlo simulation and HRP, and the corresponding portfolio weights will be backtesting in the next time period. The general description is shown in *Figure 4.4*.

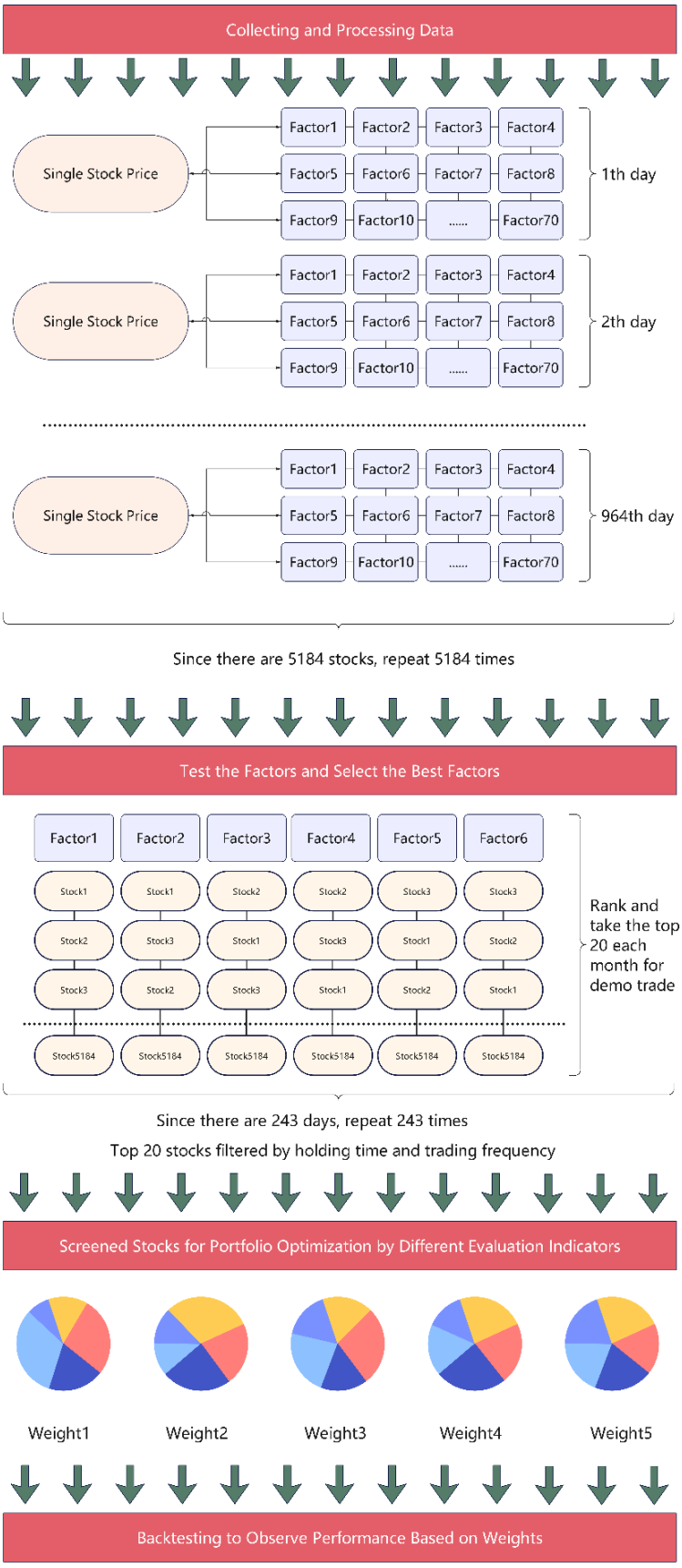


Figure 4.4 A brief description of the data and modeling process

Source: Own elaboration

It is still important to note that this thesis does not use full optimization, but rather Monte Carlo simulation weights instead. There are two reasons for this approach: first, the individual investor needs a clearer understanding of market probability, i.e., whether he or she will be able to make a profit after investing in a stock, and how likely it is that he or she will not be able to make a profit, and the visualization of probability is an important factor for the individual investor to judge the market and decide to invest; second, fully optimization can only be performed through fixed evaluation metrics, and there are many other strategies outside of it that cannot be considered, although these strategies are not addressed in this thesis, but it does exist and used in the market, for example: small capitalization strategy, double mean strategy, bank stock rotation and undervaluation, etc. The combined judgment of these strategies can be found the corresponding possible points in Monte Carlo simulation, but will be ignored in the full optimization.

4.3 Factor Investing

Due to the high volume of data, this thesis only uses factor parameter basic_earnings_per_share as an example for single-factor analysis. The same processing method applies to all factors, and the final results can be found in *Annex 3*. In addition, the process of constructing the factor library is omitted in this thesis, and the specific factors description can be found in *Section 2.3.1* and *Annex 1*.

4.3.1 Data Processing

Firstly, this thesis addresses the need to winsorize factor data to eliminate the influence of outliers on data analysis. The winsorization method used in this thesis is the three-median winsorization method. Due to the large amount of data, just a straight line in the display of the original data, so take the first 100,000 data for further display, where basic_earnings_per_share is the original data (blue), basic_earnings_per_share_3mad is the data after winsorization (orange). Since the data is three-dimensional, after reducing to two-dimensional, the data is formatted as factor values for the first stock from 2017-2021, then factor values for the second stock from 2017-2021, and so on until the last stock, with the main purpose of showing the importance of winsorization. Results after the winsorization can be seen in *Figure 4.5*.

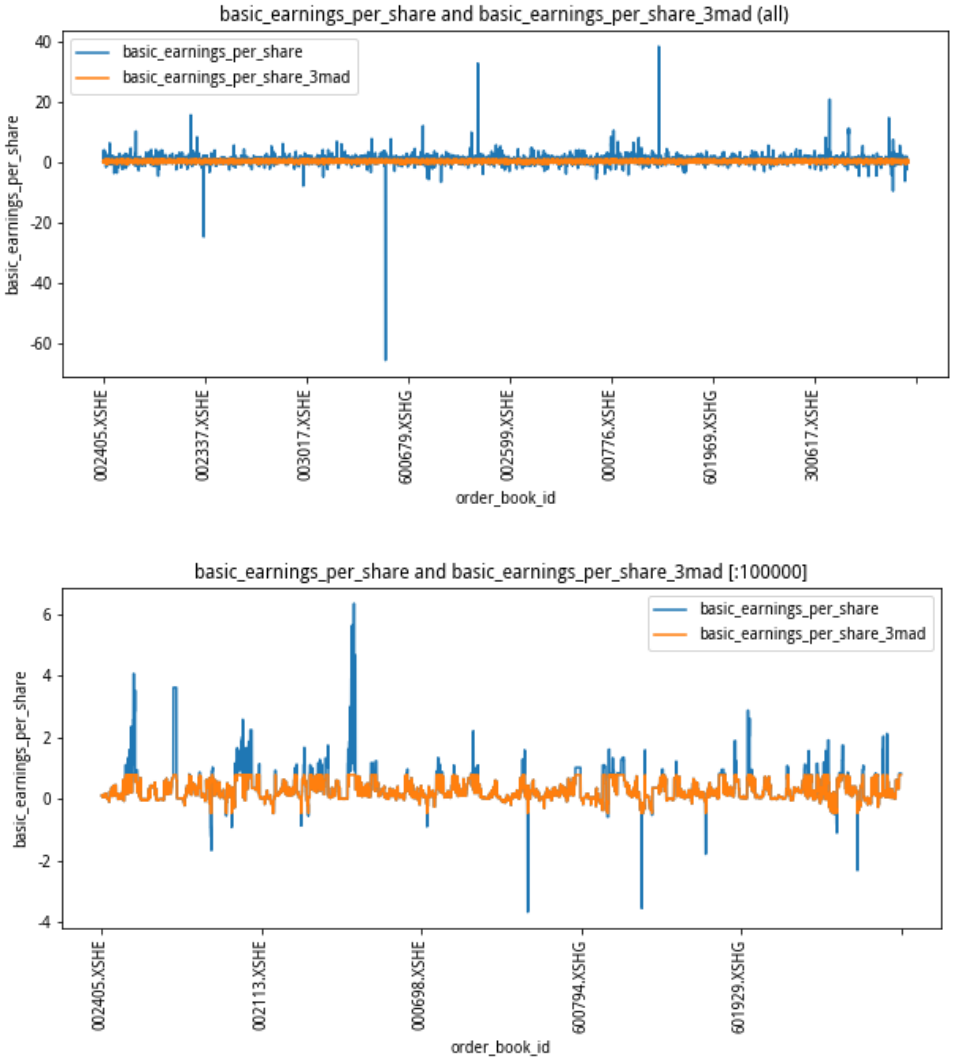


Figure 4.5 Comparison of basic_earnings_per_share original data and winsorized data

Source: Own calculations

The winsorization graph is mainly used to compare the difference between the winsorization processed data and the original data. Typically, winsorization replaces extreme values in the original data with values at the edge of the dataset or reduced to a specific range. Therefore, it can be seen from *Figure 4.5* that after the winsorization process, the extreme values in the data set will be eliminated or weakened, thus making the data more representative and comparable. The extreme values in the first graph exceed -60, -20, and 20; the values in the second graph (orange) are mainly concentrated between -1 and 1. so that it is more obvious that the result of the triple median de-extremum method of the data forms a stable interval, which forms the foundation for the subsequent data fitting process.

Then this thesis continues to standardize the data through Z-score, which can facilitate the comparison of variables with different units and scales. The data format is the same as winsorized, where basic_earnings_per_share_3mad (orange) is the data after winsorization, basic_earnings_per_share_stand (blue) is the data standardized according to the original data after Z-score transformation. The standardized data is the data used for single-factor test, the standardized results of which are displayed as *Figure 4.6*.

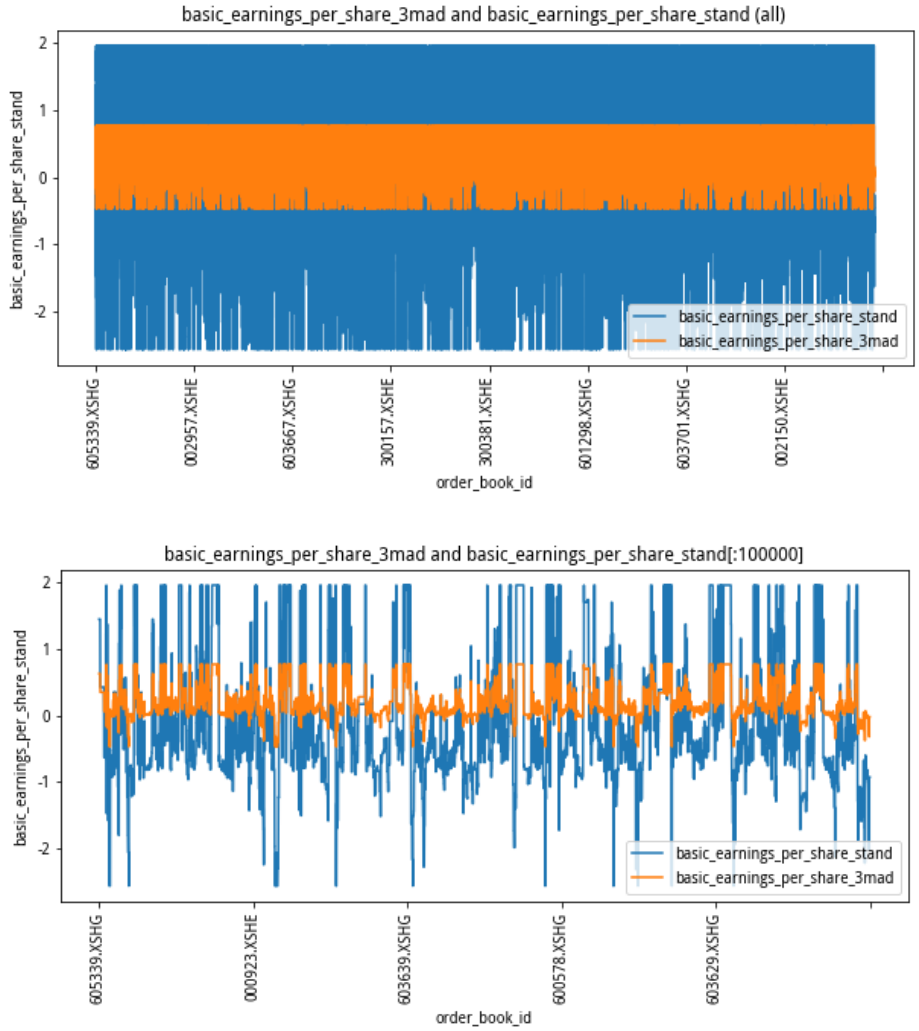


Figure 4.6 Comparison of basic_earnings_per_share winsorized data and standardized data

Source: Own calculations

Similar to the previous discussion, it becomes apparent that the de-trended data is centered around the range of -0.5 to 0.5, while the standardized data is concentrated within the range of -2.5 to 2.0. From the overall dataset, it appears that the values have transformed into a nearly rectangular distribution. Examining the first 100,000 data points,

it becomes apparent that the fluctuations in the data follow a process of expanding variance, but the overall state of the dataset has not undergone significant changes. The primary purpose of Z-score standardization is to facilitate comparability across different features, whereby the distribution of the transformed features remains unchanged. For convenience, this thesis will continue to call it basic_earnings_per_share later. After processing, the following single factor test can be performed and compared.

4.3.2 Single Factor Return Test

The most important thing in multifactor analysis is to analyze the relationship between factors and stock returns. One of these methods is to divide the variable into multiple quantiles or groups, so as to observe the changes in returns within different groups. In this thesis, it is divided into 5 quantiles, which can judge whether the relationship between the factor and the response variable will change with different levels of the response variable.

factor_quantile	min	max	mean	std	count	count %
1	-2.564580	-0.572883	-1.102991	0.458632	703528	20.391826
2	-0.968347	-0.158414	-0.615616	0.169672	692150	20.062034
3	-0.815937	0.426375	-0.303496	0.297293	682237	19.774705
4	-0.662795	1.517373	0.191867	0.524622	688085	19.944209
5	-0.364918	1.961206	1.205709	0.780416	684049	19.827226

Figure 4.7 basic_earnings_per_share basic information under different quantile intervals

Source: Own calculation

Figure 4.7 shows the statistical properties of variables classified into five quantiles. Indicates the minimum, maximum, mean, standard deviation, counts, and percentage of counts, respectively. The first quantile has the lowest mean (-1.102991) and highest standard deviation (0.458632), indicating that the values in this group are more widely distributed than those in the other quantiles. The percent count for this group is 20.391826%, which is slightly higher than the other quantiles. The other grouping analyses are basically similar, showing the basic grouping information of the factors.

After the underlying quantile statistics, the return analysis is an initial description of the value of the factor, and the returns expressed in monetary terms show the power of

the factor. The basic way to see the predictive power of a factor is to look at the value of the average return factor for different quartiles of the factor.

Next, it is necessary to look at the parameters for calculating the rate of return for different time windows on time series. These parameters represent the time frame considered when calculating the average yield. Generally, the larger the value of these parameters, the smoother the calculated rate of return and the smaller the volatility. But in the Chinese stock market, due to the 10% circuit breaker limit (see *Section 4.1*), the performance is reversed, and the results are as shown in *Figure 4.8*.

	1D	5D	10D
Ann. alpha	0.157	0.130	0.112
beta	-0.030	-0.070	-0.085
Mean Period Wise Return Top Quantile (bps)	6.776	5.614	4.820
Mean Period Wise Return Bottom Quantile (bps)	-5.393	-4.484	-4.043
Mean Period Wise Spread (bps)	12.170	10.068	8.828

Figure 4.8 Single factor return analysis of basic_earnings_per_share

Source: Own calculation

The first row of the table shows positive annualized alpha, indicating that the annual returns of the factor are interpretable, and the higher 1-day alpha than the 10-day indicates that the interpretation is limited to a shorter time horizon. In contrast, negative beta volatility indicates that the single factor is less volatile than the market as a whole and may have a negative return, but note that when market returns fall, assets with negative beta return greater than 0. The next three rows show the average periodic returns for the highest quartile, the lowest quartile, and the spread between the two quartiles. The top quartile is the positive average periodic return, while the bottom quartile is the negative average periodic return. The difference between the top and bottom quartiles is positive in all three times range, indicating that the single factor has positive alpha and outperforms the market.

This thesis classifies its single-factor returns into 1-day, 5-day and 10-day periods. It is important to note that all subsequent factors will use the 1-day period, and the 5-day and 10-day periods are used as a reference for comparison during the single-factor analysis. As shown in *Figure 4.8*, the 1-day period will capture the market efficiency more finely, but the performance may be more volatile and affected by noise, but more

information will provide the basis for subsequent machine learning models to make judgments.

After understanding the two important parameters of quantile and time window, this thesis begins to formally conduct single-factor analysis. The first thing to observe is "top minus Bottom quantile mean return", which is used to evaluate the performance of the stock portfolio. Measures the difference in average return between the best-performing fraction of stocks in the high quantile and the worst-performing fraction stocks in the low quantile. Its time series figures are as *Figure 4.9*.

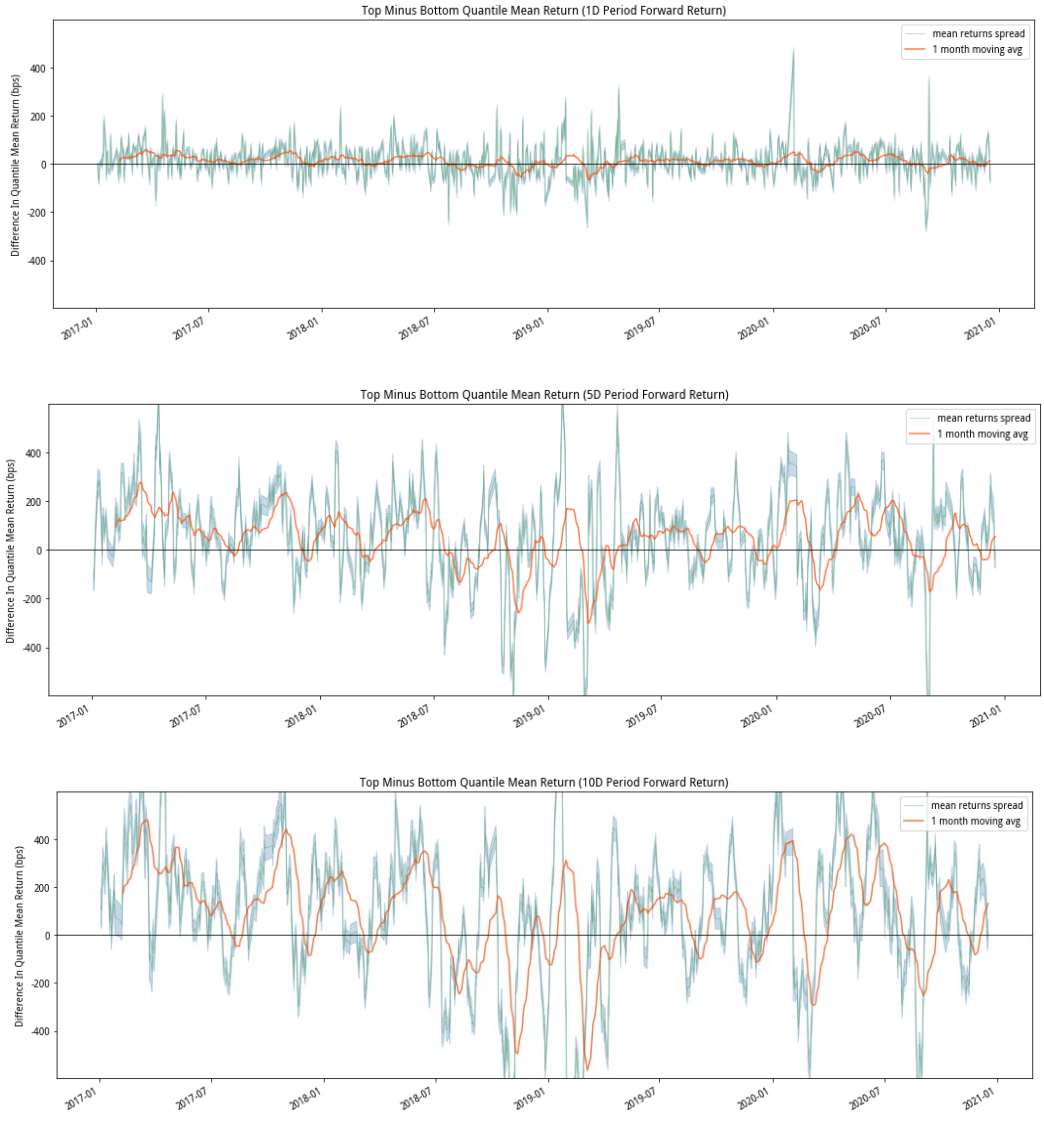


Figure 4.9 Time series graph of top minus bottom quantile mean return

Source: Own calculations

According to *Figure 4.9* the volatility of the highest minus the lowest quartile average return over different time horizons is basically supported by 0, reflecting from

the side the validity of the efficient market hypothesis that the market always returns to a reasonable state., this suggests that the ability of the single factor explanation to generate alpha will fluctuate in magnitude with regularity as the holding time increases, due to the fact that the market is not in a steady state. In addition, the 1-month average in the figure shows more clearly that the volatility of returns increases with longer holding times. The 1-day return is the least volatile, followed by the 5-day return, while the 10-day return is more volatile and exceeds the limits of the chart. This indicate that longer holding periods are associated with greater uncertainty and more significant return volatility. Then look at the total return quantile in *Figure 4.10*.

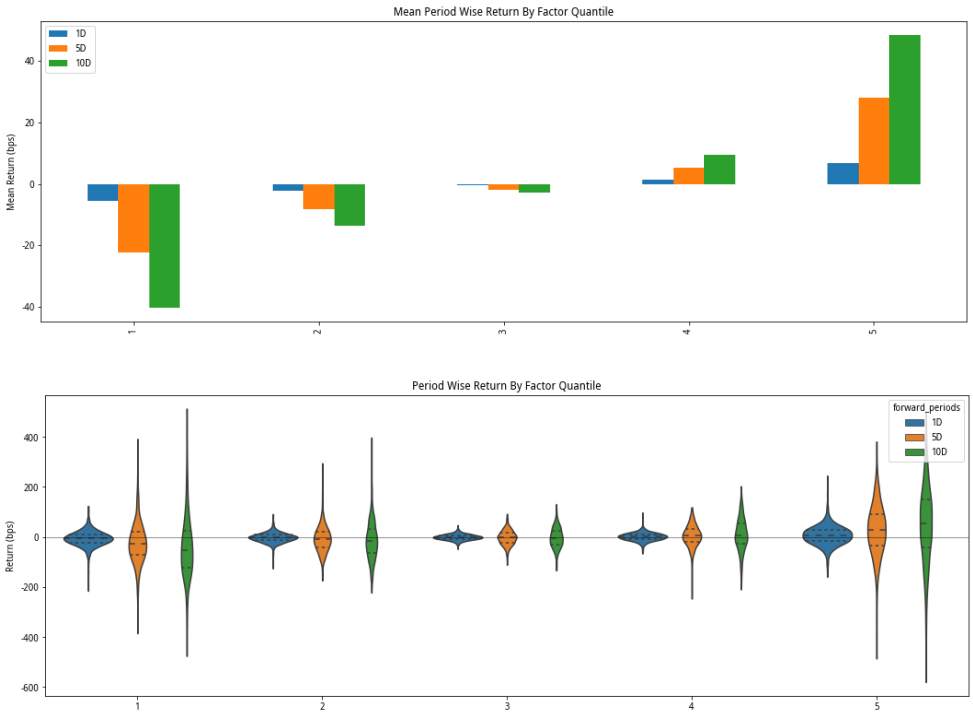


Figure 4.10 Mean period wise return by factor quantile

Source: Own calculations

The bars of each quantile group represent the factor returns at different time periods (1, 5 and 10 days), and the direction of the factor is judged by the change in the returns of the five groups: the larger the factor quantile position the larger, the return the first group (the worst return) to the fifth group (the best return). In *Figure 4.10*, it can be seen that as the quantile group changes from, the average return changes as well. In the first group, the returns are negative and have the highest negative returns in the longer time period (10 days), but as the quantile group increases, the average return becomes progressively more positive, finally reaching its highest point in the fifth group.

Within each time period in the violin plot, the 10-day return is the widest, indicating the widest distribution of returns for that factor, along with the widest range of upper and lower quartiles. The 5-day period is the second widest, and the 1-day period is the narrowest, indicating the narrowest range of returns for that factor. The middle line of each violin indicates the median return, which corresponds to the distribution bar, indicating that the longer the holding period, the wider the range of return fluctuations for this factor. In addition, it is evident that the direction of this factor is positive and should be retested under different market environments and time conditions. Other possible factor directions are shown in *Figure 4.11*.

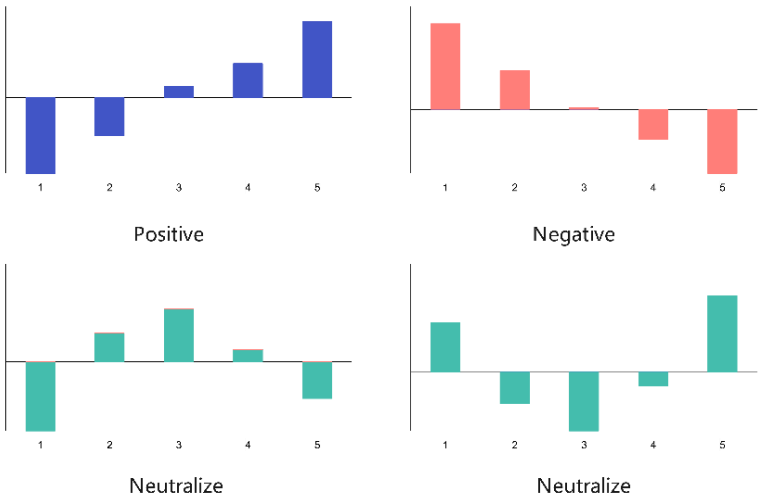


Figure 4.11 The examples of performance of different factor quantiles

Source: Own elaboration

After calculating the factors, several situations as shown in *Figure 4.11* may be obtained. Among them, the one-way rise is the largest in the fifth quartile, indicating that the factor direction is positive, and a high factor value means a higher stock return rate. One-way downward, the factor direction is negative, the lower the factor value, the higher the return. The key is if the direction of the factor is neutral, that is, there is no obvious positive or negative relationship between the factor value and the stock return. This single factor does not have much explanatory power to the return rate of stocks. It can be considered to exclude this factor from investment decisions, but at the same time, it can be considered to combined with other factors. This is an important reason for subsequent factor synthesis.

4.3.3 Single Factor IC Test

Information content analysis gives us a way to assess the predictability of factor values without worrying about the impact of transaction costs, and the main approach taken in this thesis is IC analysis. The descriptive statistics of other factor IC data can be found in *Annex 2*.

	1D	5D	10D
IC Mean	0.022	0.035	0.043
IC Std.	0.106	0.126	0.135
Risk-Adjusted IC	0.208	0.276	0.321
t-stat(IC)	6.457	8.554	9.956
p-value(IC)	0.000	0.000	0.000
IC Skew	-0.096	-0.322	-0.358
IC Kurtosis	0.425	0.205	-0.042

Figure 4.12 Descriptive statistics of IC values for basic_earnings_per_share

Source: Own calculation

Figure 4.12 shows that the mean IC increases with increasing holding period and the predictive power of the factor improves over a longer time horizon. However, in contrast, the standard deviation of IC also increases, indicating that the factor's predictions become more uncertain over a longer time horizon. This is further confirmed by the risk-adjusted IC, which indicates that the risk-adjusted performance of the factor is highest over the 10-day holding period.

Value of the t-statistics and p-values for IC indicate that for all three holding periods, the factor's predictions are significantly different from zero. The IC skew is negative for all holding periods, indicating that the distribution of IC values is slightly skewed to the left. The positive kurtosis of the IC for the 1-day and 5-day holding periods indicates that the tails of the distribution are heavier than the normal distribution, while the negative kurtosis for the 10-day holding period indicates that the tails of the distribution are lighter. Forecasts become increasingly uncertain over time.

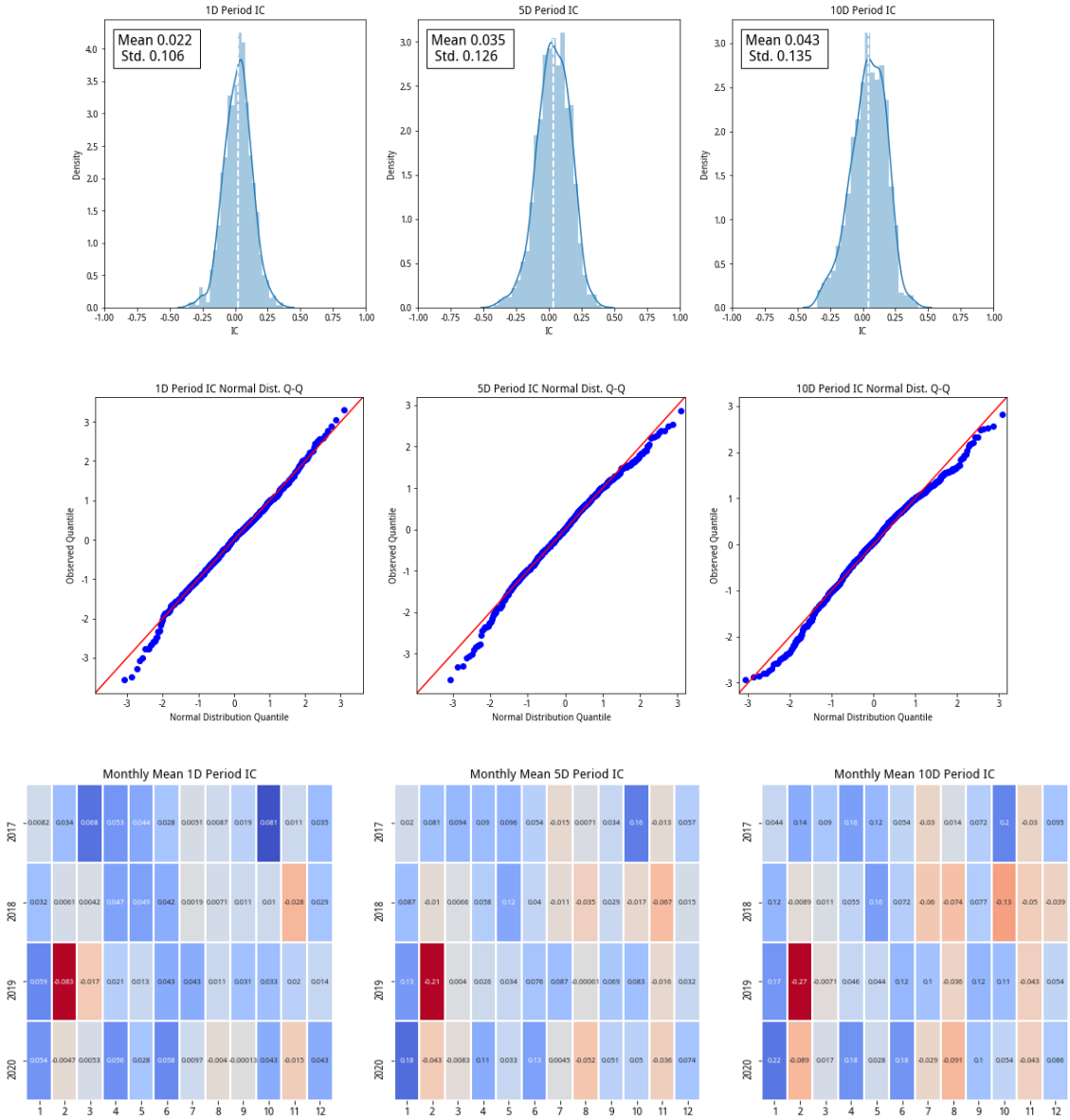


Figure 4.13 Normal and monthly distributions of the factor IC

Source: Own calculations

The normality distribution of the factor IC values are judged, and the first set of plots shows the period IC of the factor. Each subplot shows the distribution of IC values for a given time period, along with their mean and standard deviation. The mean IC values for 1D, 5D, and 10D are 0.022, 0.035, and 0.043, respectively, while their standard deviations are 0.106, 0.126, and 0.135, respectively. The distribution of IC values appears to be approximately normal, with the mean located near the center of the distribution.

The second set of plots shows the Q-Q plots of the IC distributions for different time windows, which are used to assess the normality of the IC distributions. The three subplots show the IC values for each of the three loops and compare them to the expected

values of a normal distribution, with the straight lines representing the theoretical normal distribution. Looking at the plot as a whole, the IC values in the subplots appear to be close to the theoretical normal distribution, but the lower left and upper right corners of the plot deviate slightly.

The third set of plots shows the monthly average IC of the factors for the period 2017-2020. Each small plot in the main plot corresponds to a different year. The color of the bars represents the magnitude of the IC, with darker colors indicating higher magnitudes. The graph shows the monthly average IC variation over the years, with the largest magnitude occurring in February 2019, and can be used to identify patterns and trends in the monthly performance of the factor over time. Overall the factor passes conforming to a normal distribution and requires attention to the factors behind February 2019. A look at the SSE Index reveals that the SSE Index rose rapidly in February 2019, with from 2584.57 to 2940.95 points. The rise was 13.79%, a side note that the factor is effective.

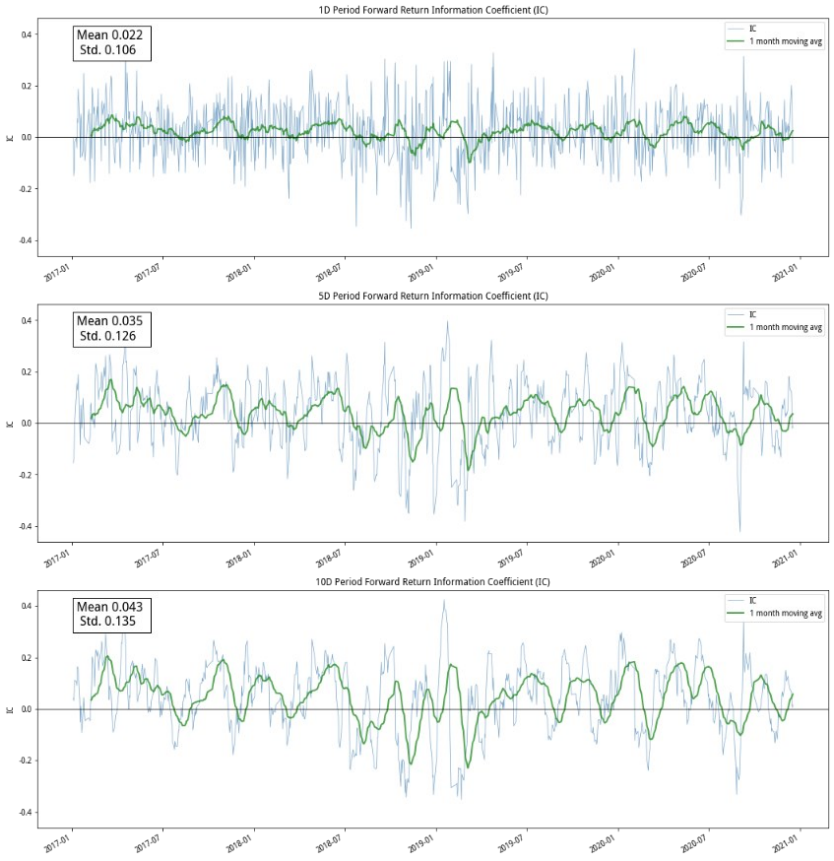


Figure 4.14 Time series of forward return information coefficients during single factor IC

Source: Own calculations

From *Figure 4.14* moving averages, the IC value is mostly positive, and with the increase in holding time, the volatility increases, without further explanation. However, it should be noted that when the IC value is negative, need to pay close attention to the market trend, to determine whether it is short term fluctuations or intermittent fluctuations. For example, in the second half of 2018 and the first half of 2019, there was a period when the value was relatively low to consider shorting, which is difficult to accomplish due to the restrictions on shorting tools for individual investors in the Chinese market. However, the current reform period is underway, and individual investors can be pay attention to the latest shorting tool changes.

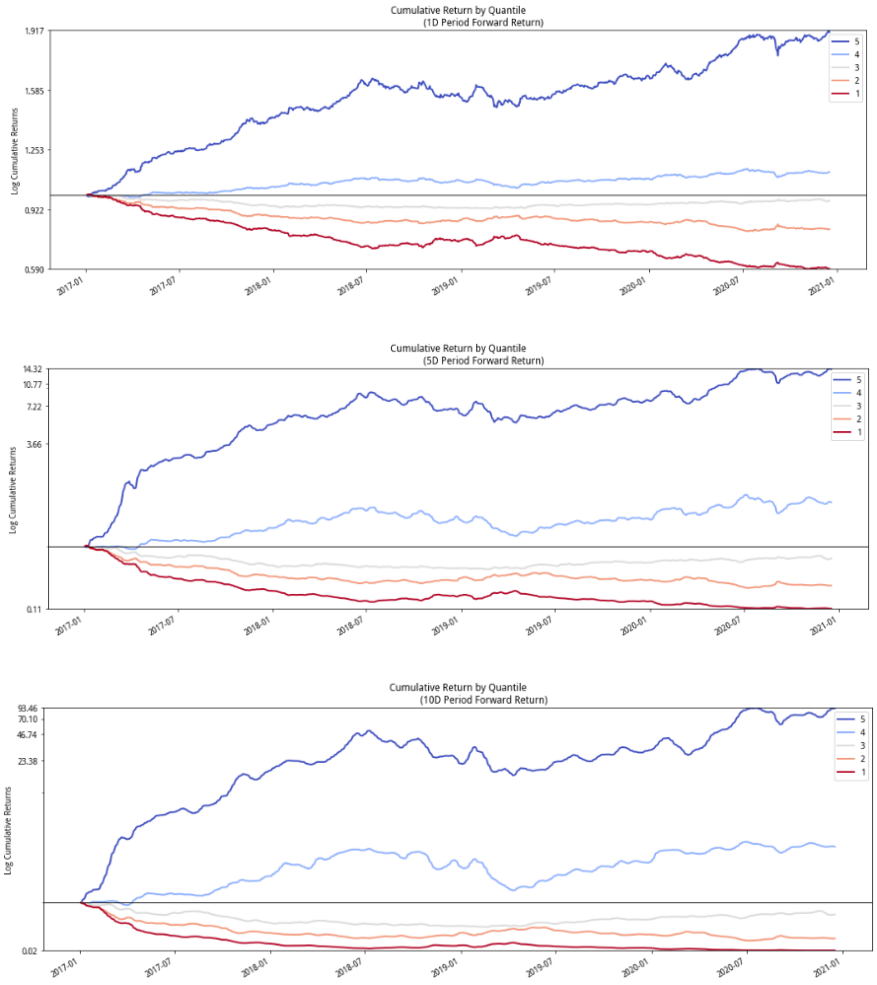


Figure 4.15 Quantile cumulative return time series figure

Source: Own calculations

When obtained from *Figure 4.15* one needs to focus on the values of the vertical coordinate, as well as the vertical axis for log cumulative returns, focusing on the increasing trend of cumulative returns rather than the magnitude of absolute returns. In

the 1-day period, the values are small and very close in quartiles 1 to 3, while quartiles 4 and 5 are relatively large. The same trend is seen in the 5 and 10 day periods and appears to be striking in the 5-quartile of the 10-day period at 93.46. The returns in the different quartiles have similar trends over time, but the strength of the returns varies widely across the quartiles, suggesting that the single factor appears to be prone to high yielding stocks in these periods and more suitable for longer term investments.

4.3.4 Single Factor Trade Test

Turnover analysis illustrates how these factors are composed and how these components change. Factor transformations reveal the integration of new information and the formation of signal extremes. Looking at the new members of the head and tail bits array shows how many parts of the factor change from day to day.

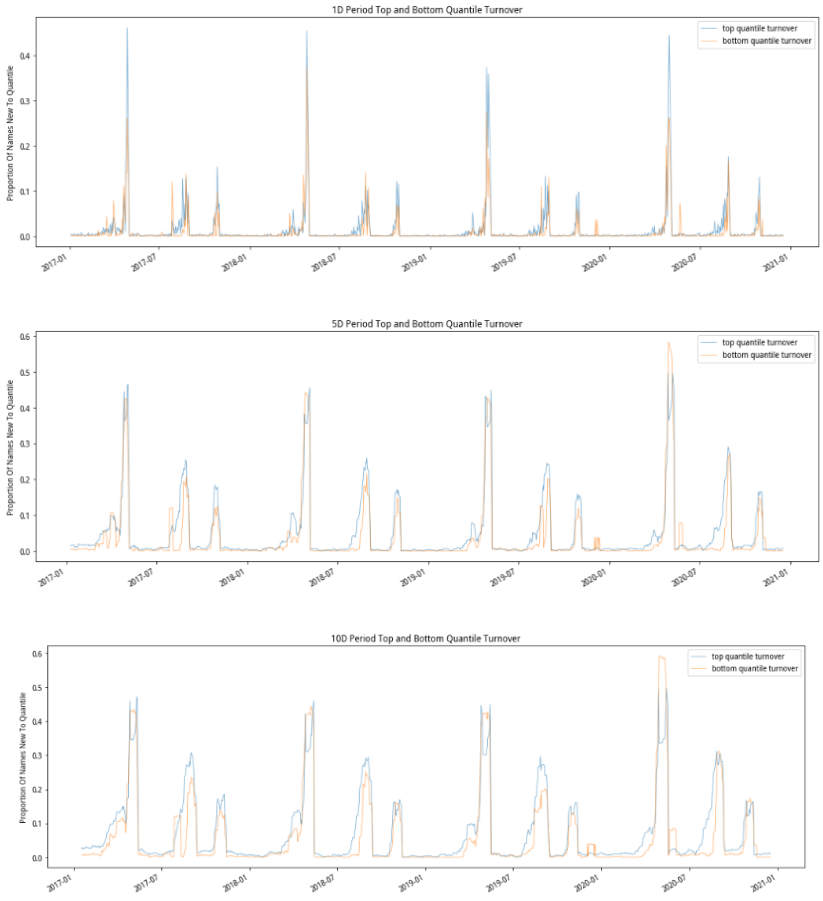


Figure 4.16 Top and bottom quartile turnover rates during the single factor period

Source: Own calculations

From *Figure 4.16*, charts show the percentage of turnover in the top and bottom quartiles of the factor distribution as a function of holding period. The figure shows the

cyclicality of the proportion of turnover, with the 1-day holding period showing the greatest degree of volatility. The percentage of new names for the 1-day holding period ranges from 0 to approximately 0.45 (almost half of the names in the quartile change). The 5 to 10 day holding periods fluctuate even more, with high changeover periods corresponding to a new name ratio of around 0.45 and low changeover periods corresponding to a new name ratio of around 0.25-0.3.

These again indicate similarities in shape and trend across holding periods, and the cyclicality of the new name ratio is consistent across holding periods, with higher turnover periods coinciding with higher new name ratios. The 10-day holding period fluctuates slightly more significantly than the 5-day period, but the overall trend remains the same. Provides insight into the temporal stability of a factor's performance and the extent to which its component names change over time, which is used to assess the robustness and persistence of the factor signal.

4.4 Portfolio Optimization

After the single-factor test, we start the factor synthesis. The portfolio optimization in this thesis has two parts, one is the optimization of the factor synthesis and the other is the optimization of the synthetic factor screening stock pool. In order to unify the results, the optimized factor IC uses a period of 1 day.

4.4.1 Principal Component Analysis (PCA)

In this thesis, the factors are transformed into 5 principal components, the next view of the change in principal component contribution for each factor is the proportion of each principal component's contribution to the total variance. The height of each bar indicates the size of the contribution of the corresponding principal component, and the bars are sorted from highest to lowest contribution. From *Figure 4.17*, can be see that if the principal components involved with the original contribution degree look more complicated, after reducing to 5 principal components, especially the degree of factors involved in PC1 is about 30%, which is not a high proportion but still covers most of the factors. The reduced cumulative explained variance can be seen in *Figure 4.18*.

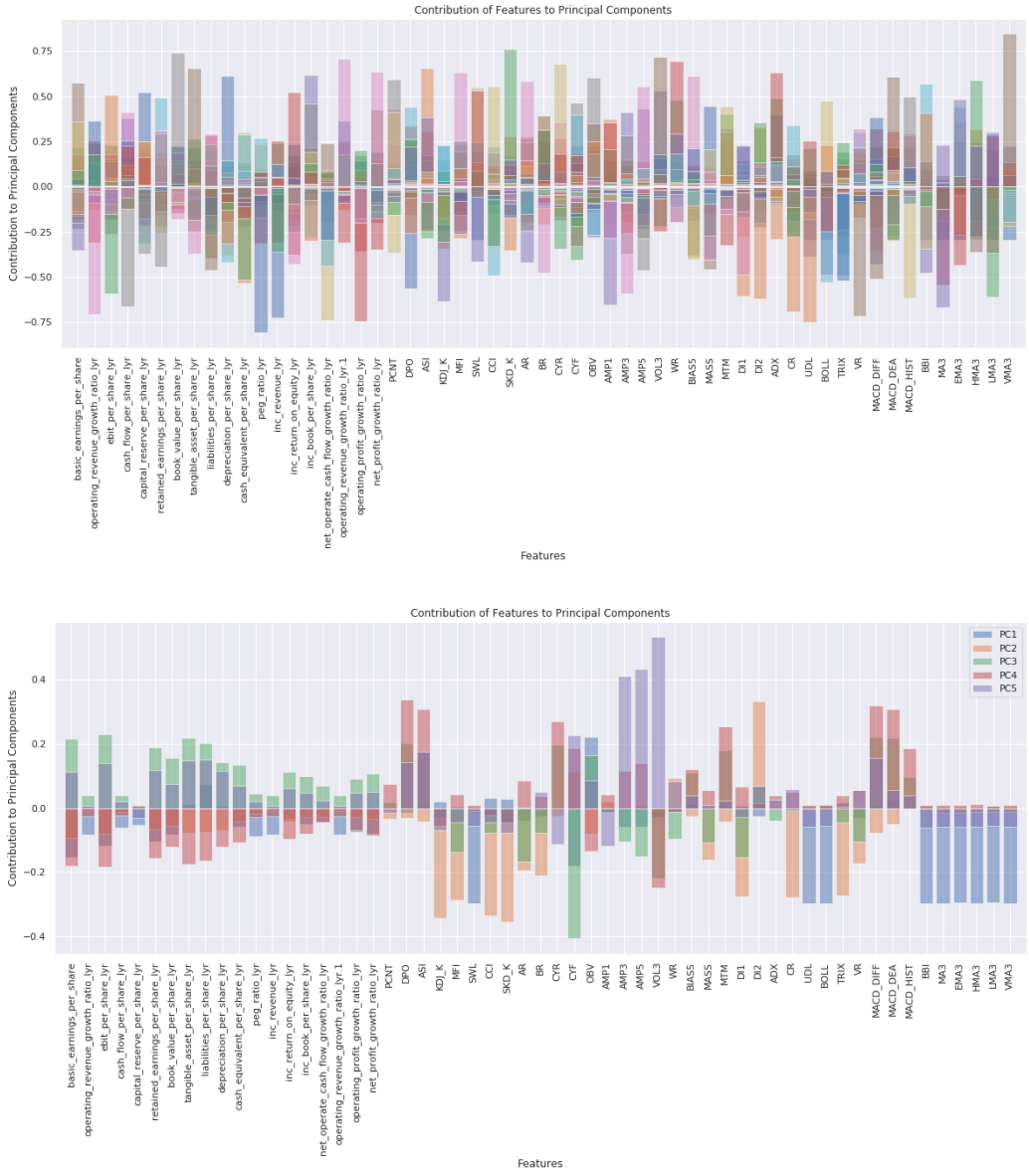


Figure 4.17 Change in the contribution of principal components

Source: Own calculations

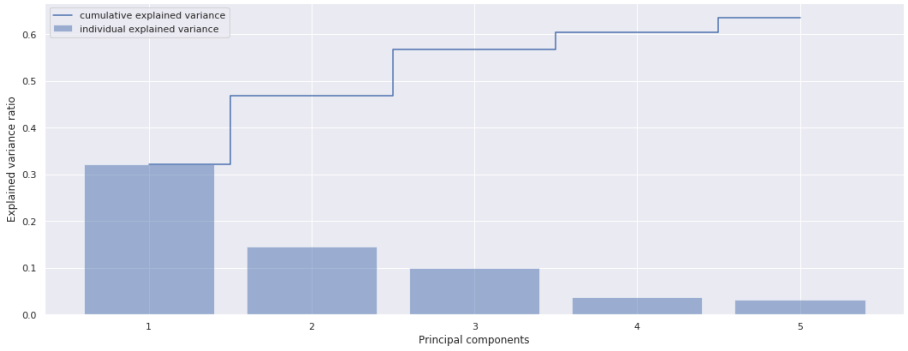


Figure 4.18 Cumulative explained variance of the 5 principal components

Source: Own calculation

It can be seen from *Figure 4.18* that the cumulative variance (degree of explainability) of the principal components reaches more than 60%, which makes it possible to use the 5 principal component factors to explain most of the factors described and thus make simplified judgments. However, the problem is that if the 5 principal components used to perform the PCA process are newly generated variables, the next action cannot be performed with the factor history data, and if the prediction method is utilized it will further expand the parameter variables to produce errors, so this section is only for illustration. If the problem of historical data is solved or other methods are used, various indications suggest that PCA is one of the important methods suitable for the simplification of this investment process. In order to make the investment process to continue effectively, this thesis continues the ranking of factor importance using XGBoost in the next section.

4.4.2 Extreme Gradient Boosting (XGBoost)

XGBoost's learning curve plots the trend of the training and test set errors with the number of training rounds, thus showing how the model's performance changes as the number of training rounds increases. The error of the model drops sharply as the model begins to learn the characteristics of the data. As training continues, the model's performance will gradually stabilize and eventually stop improving. At this point the optimal number of training rounds can be determined or training can be stopped to avoid overfitting. The result is shown in *Figure 4.19*.

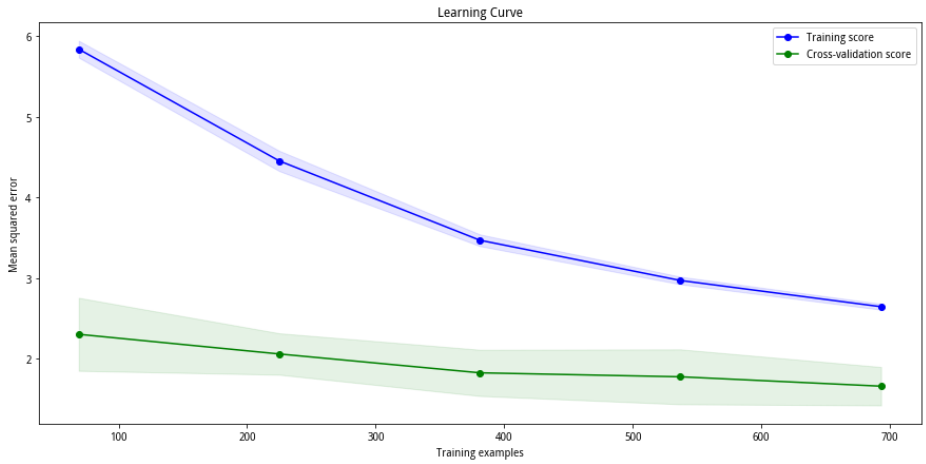


Figure 4.19 XGboost factor learning curve

Source: Own calculation

XGBoost's learning curve plots the training set (20%) and validation set (80%) errors as the number of samples increases. According to *Figure 4.19*, the training and validation set errors converge and show a stabilizing trend, x-axis represents the number of samples in the training set or the hyperparameter value representing the complexity of the model, and y-axis represents the error or evaluation metric of the model on the training and validation sets. The model does not suffer from underfitting or overfitting and does not require more training data or a more complex model. The training score refers to the performance of the model on the training set, while the Cross-validation score is the performance on the validation set. The light green areas on both sides of the green line indicate the uncertainty of the model, which can be seen to fit more stably.

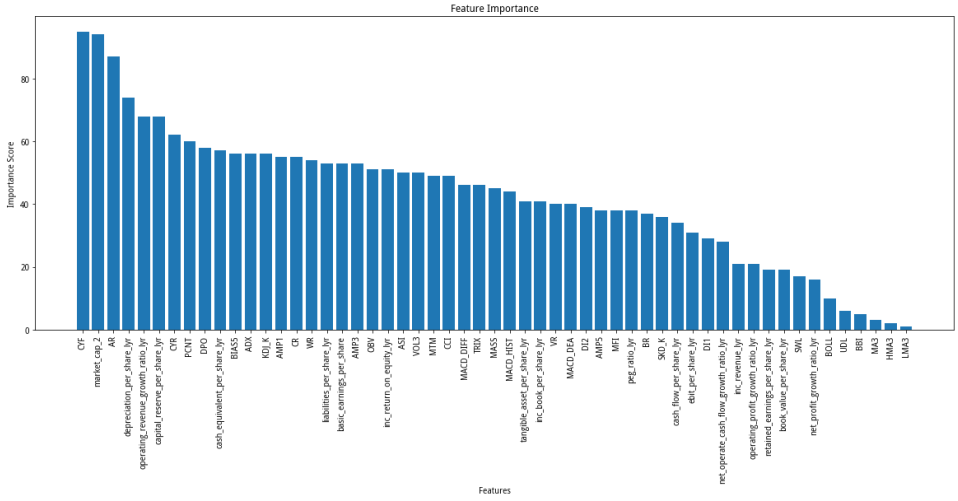


Figure 4.20 XGBoost factor feature importance ranking

Source: Own calculation

As can be seen from *Figure 4.20*, the differences in the results of the degree of importance of the factors vary widely, and in this thesis the top 6 ranked factors are taken as the basis for use in the next section. Because of the need to take important features while needing to ensure that the interpretability is sufficiently diverse, but it is also possible to adjust the reference value on a personal level. Their names and scores are Market Force (CYF), 95; Total market value of tradable shares, 94; Accumulation/ Distribution, 87; Depreciation and amortization per share, 74; YoY growth rate of operating income, 68; Capital reserve per share, 68.

4.4.3 Multiple Linear Regression (MLR)

Then use the formula in *Section 3.3.3* to bring it into the calculation coefficient to get the formula as:

$$y_t = \beta_0 + (-5.53843576e-02)x_1 + (2.96580823e+11)x_2 + (-6.28662109e-03)x_3 + (1.81045532e-02)x_4 + (8.53729248e-03)x_5 + (8.02612305e-03)x_6 + \varepsilon_t, \quad (4.1)$$

where y_t is next day return ratio of stock, x_1 is the factor CYF, x_2 is the factor market_cap_2, x_3 is the factor AR, x_4 is the factor depreciation_per_share_lyr, x_5 is the factor operating_revenue_growth_ratio_lyr, x_6 is the capital_reserve_per_share_lyr. Note that this stage uses not the IC value of the factor, but rather the value of the factor itself. Then we run the multi-factor synthetic backtest to get the following *Figure 4.21*.

Figure 4.21 Multi-factor synthetic backtesting

Source: Own calculation

After the multi-factor synthesis, its validity needs to be verified, and this thesis uses the full-year data from 01/01/2021-01/01/2022 for validation. The initial capital is set to 1 million RMB and the benchmark is CSI 300 and CSI 500. In *Figure 4.21*, it can be observed that blue is the benchmark return time series, red is the synthetic multi-factor return time series, and orange is the excess return time series. The detailed indicators are shown in the figure, it is obvious that the synthetic multi-factor model is valid and the return can be more than 20%, but it is important to note that this does not mean that all returns can be maintained in the future, it is only valid for this time.

The results of its validation are formed based on simulated trading, by identifying factor signals for sorting to achieve the purpose of trading hold for profit. At the same time, in order to make the number of stocks holding the portfolio stable, the results of each buy and sell will be held at the close of only 20 stocks, finally forming a list of stock changes and the amount invested. In order to satisfy the medium and long-term investment needs of individual investors, keep portfolio assets fixed and reduce transaction costs, this thesis uses holding time and trading frequency as criteria for stock screening. The screened stocks were extracted and merged, and the results can be seen in *Annex 5*.

4.4.4 Monte Carlo Simulation (MC)

Once the stocks have been screened, the next step to be performed is to use Monte Carlo to generate random numbers to simulate the stock weights, which are displayed using the mean-variance framework to give individual investors a clearer figure of the returns and the risk from their investments, the results of which are shown in *Figure 4.22*.

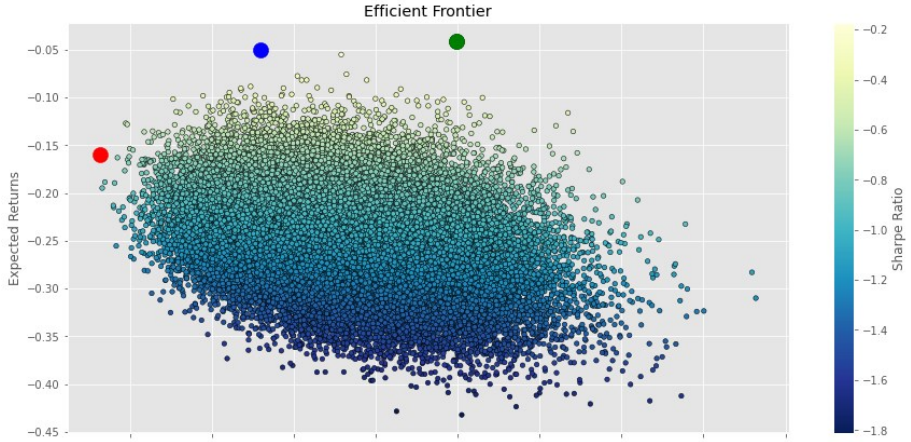


Figure 4.22 Monte Carlo simulation of portfolio weights for Markowitz model

Source: Own calculation

The red point is the minimum variance portfolio, the green is the minimum Calmar ratio portfolio, the blue is the minimum VaR portfolio, and there is a yellow point that is the maximum Sharpe ratio portfolio, but it coincides with the minimum Calmar ratio portfolio point. For convenience continue, this thesis called them as Variance, Sharpe, Calmar and VaR. In addition, the entire generated portfolio effective boundaries are different from the standard ones as seen in *Figure 4.22*.

Even from the vertical coordinates, the expected returns are mostly negative, with the maximum expected loss reaching -0.45. The reason for doing this can be seen in *Section 4.1*, and the negative expected return is due to the overall decline of the Chinese stock market during this period. The global stock market with the factors of interest rate hikes is contracted. This is the normal phenomenon of using historical data for simulation, as individual investors do not need too much simulation more comprehensive. Just need to reach a certain number of simulations to find out the weight of the expected stock return evaluation indicators. The weights of its four portfolios are shown in *Figure 4.23*.

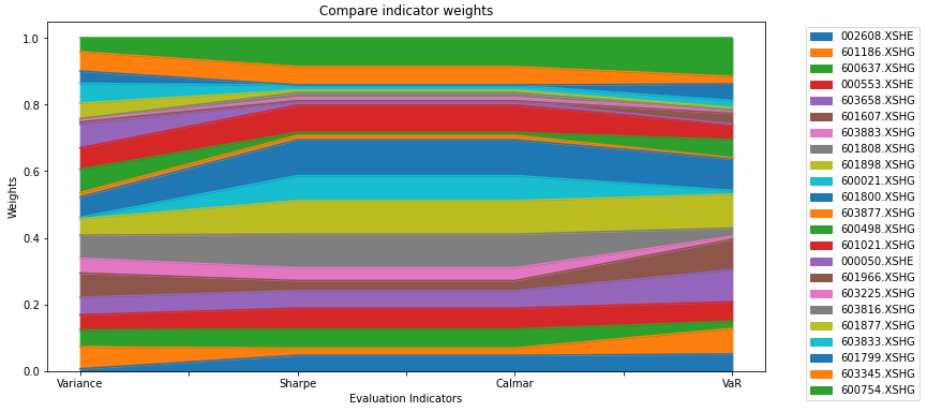


Figure 4.23 Compare evaluation index weights

Source: Own calculation

In *Figure 4.22*, only three points are shown because Sharpe and Calmar have the same weights for the investment stocks, but in other situations (changing stocks), it is still possible to have four points that correspond to each of the respective portfolios. And in *Figure 4.23* can see the weights of these four portfolios and clearly show the changes in the weights. It is worth notice that certain stocks have relatively stable weights in the chart and can be considered as relatively high performing stocks selected for medium to long term investment, and if one can go further, one can try other strategies used in the market, thus expanding the weights of these stocks. It is also observed that the optimized weights

found by the indicator allow the participation of all stocks and no stocks are excluded, thanks to the 50,000 simulations performed. However, if the investor's capital facing these stocks still does not meet the minimum of 100 shares, the investment assets with lower weights can be dropped.

4.4.5 Hierarchical Risk Parity (HRP)

In addition to Monte Carlo simulations, this thesis performs hierarchical risk parity by dividing equity assets into different clusters and iteratively generating trees, the result is shown in *Figure 4.24.*

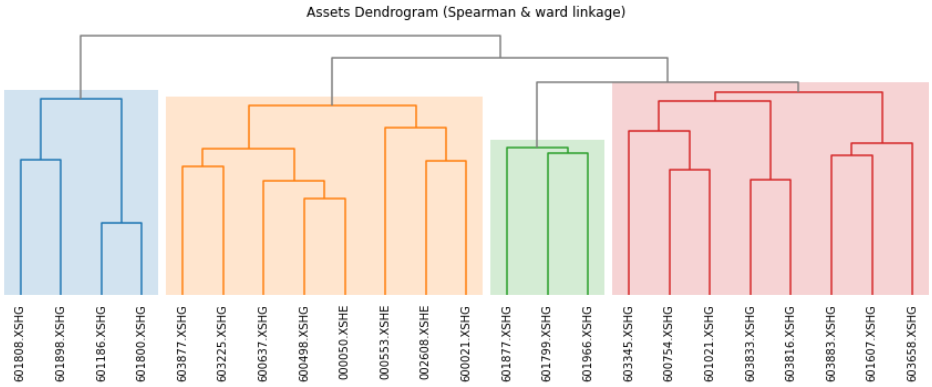


Figure 4.24 Assets dendrogram (spearman and ward linage)

Source: Own calculation

Visible in *Figure 4.24* is the result of using a tree diagram to represent the clustering between different stocks. One of the color blocks represents a cluster where the samples in the cluster are more similar to each other, while the samples between different clusters are less similar. The line height represents the similarity or distance between them, and the Spearman rank correlation coefficient is used in this thesis to calculate the correlation between the variables. ward linkage is the ward's minimum variance method, represents the minimization of within-group variance after clustering, where the two closest clusters are combined into a larger cluster at each step and iterated until all clusters are combined into one large cluster. After carrying out this process the portfolio can be obtained with its weights as shown in *Figure 4.25*.

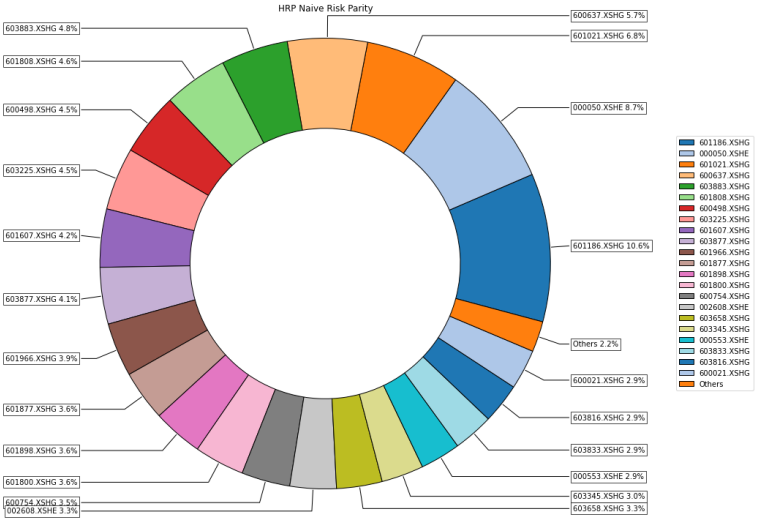


Figure 4.25 HRP portfolio stock wight allocation

Source: Own calculation

Figure 4.25 shows the portfolio weighting results after hierarchical risk parity evaluation. For the convenience of continued, this thesis still uses HRP instead. Clustering will divide the portfolio into multiple levels, each containing similar assets, and then assign weights within each level using a risk parity approach. By grouping assets with high correlation, the variance within the group is optimized to obtain a better portfolio. Also, clustering can control the risk of the portfolio, such as avoiding too many assets concentrated in the same group and avoiding too much volatility in one group. In addition, the clusters are linked differently, some appear only in the same cluster with better performance of weights and backtesting, but in this thesis only ward linkage is used for better presentation. The main risks involved are shown in *Figure 4.26*.

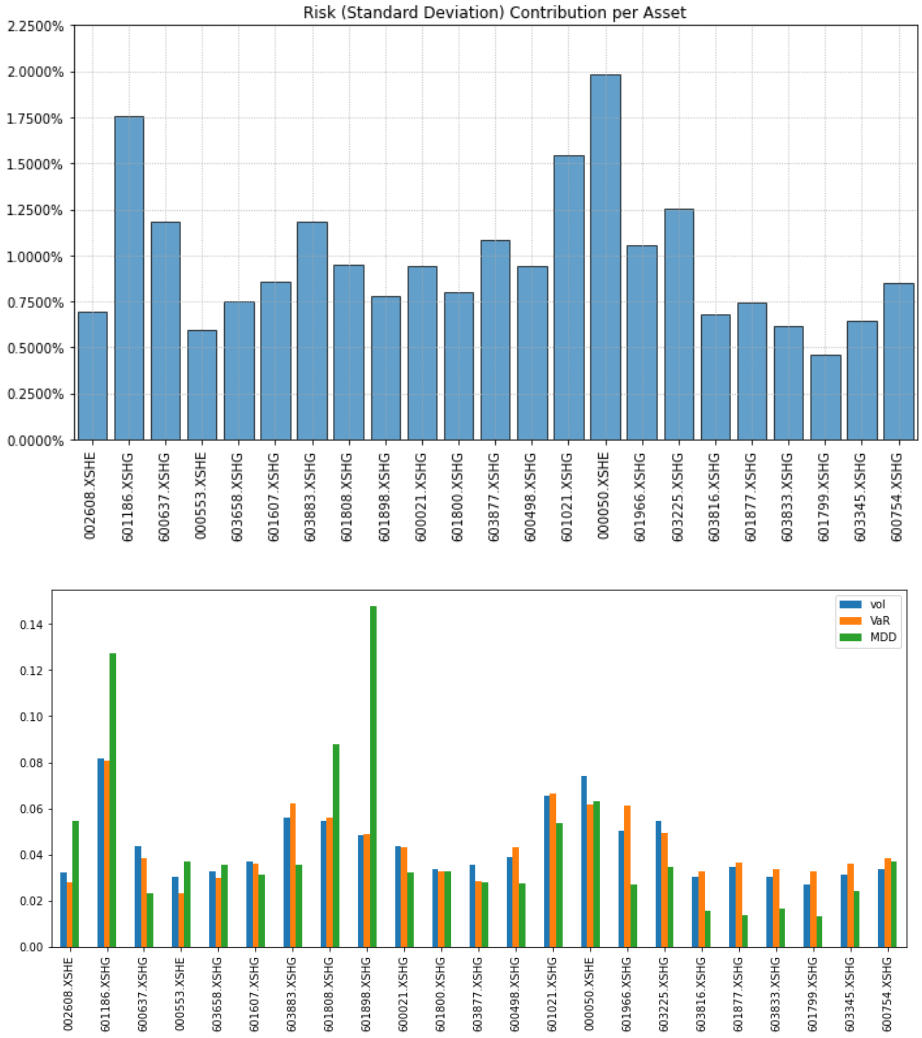


Figure 4.26 Per asset risk measure comparison

Source: Own calculations

It can be seen from *Figure 4.26* that the risk values of each stock are relatively close to each other. The two standard deviations higher are 601186.XSHG and 000050.XSHE in the first graph, and the two maximum drawdown higher are 601898.XSHG and 601186.XSHG in the second graph. In general, although there are still some stocks with higher value at risk, the risk is generally balanced and manageable. The standard deviation is mostly maintained around 1.25%, while volatility, value at risk and maximum drawdown are maintained at around 6%. The higher risk is due to higher weights and unidentified variance indicators in the weighting calculation. If the risk is manageable, it is possible to proceed (can be consider investing), but if there are multiple risk indicators that are above average, one needs to consider ways to construct the portfolio, for example using different strategies as mentioned earlier. Then we need to check the performance of the backtest.

4.5 Backtest and Interpretation

The backtest requires a completely different time interval than the previous one, need notice again that in general the testing time needs to be gradually decreasing. In this thesis, the five portfolios mentioned above are backtesting and compare with the index, using the period 01/11/2023-31/03/2023, and the results are shown in *Figure 4.27*.

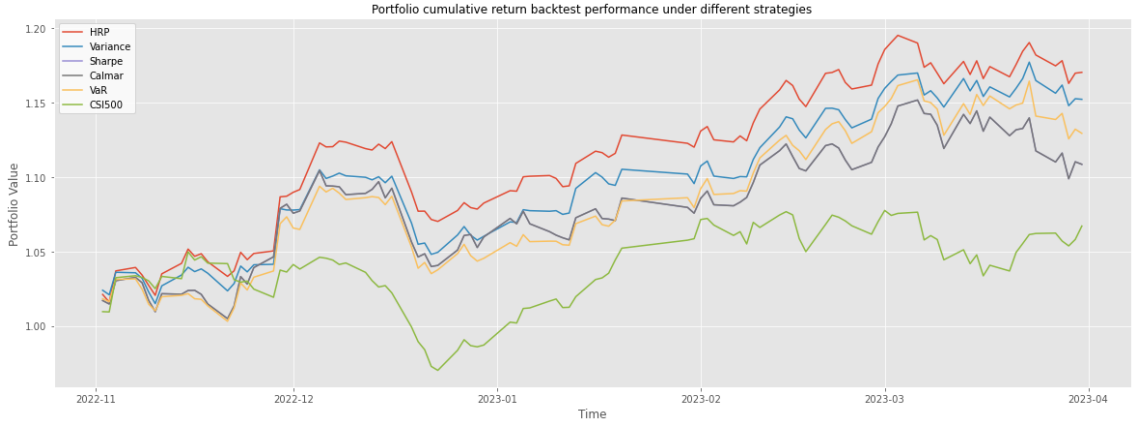


Figure 4.27 Portfolio cumulative return backtest performance under different strategies

Source: Own calculation

From *Figure 4.21*, we can see that CSI500 has a better return than CSI300, so this thesis uses CSI500 as the backtest benchmark and invests according to the optimized weights in *Section 4.4.4* and *4.4.5* for the period 01/11/2022-31/03/2023 to obtain *Figure 4.27*. What we can see is that all the optimization methods achieve better returns. It can be seen that all optimization methods can achieve good returns. The best two are the investment portfolios constructed by the two indicators of HRP and Variance. The portfolios composed of Sharpe, Calmar and VaR also have good returns, but there is a brief close to the benchmark index in early February 2023. Considering that there are only more than 20 stocks screened in this thesis, the overall effect is still good.

It is also important to note that since the screening is done on all stocks, there are sometimes ST* stocks, or briefly identified rising characteristics. In this thesis found during testing that it would be possible to identify the following stock trends as shown in *Figure 4.28*.

Figure 4.28 Time series trend of stock 003023

Source: https://xueqiu.com/S/SZ003023

Although this stock does not appear in the strategy described in this thesis, it is still important to note. As seen in *Figure 4.28* in just over half a month from the end of September 2022 to mid-October 2022 the stock price rises rapidly, and the subsequent fall back is just half a month long. And because of the market circuit breakers rules (see *Section 4.1*), the gap in the fallback is the opening of the fallback down, in other words it is simply impossible to sell during this period. This is very alarming, but in the multi-factor model will recognize this rally and thus automatically buy and sell when other stock factors increase, so logically there will still be some gain. But if it is entering a fixed stock pool, it is equivalent to a “big loss”.

Therefore, it is necessary to take care of the timing or use of index constituents to avoid possible losses in multi-factor backtesting. In addition, there will also be ST* (see *Annex 5*) stocks that need to be avoided or have been delisted in the backtesting. Consequently, unless signal recognition is automated and maximum loss limits are set, stock screening requires considerable care to avoid taking unnecessary risks.

5 Conclusion

This thesis creates a process around factor investing and portfolio optimization for the individual investor, from factor explanation of return sources, multi-factor synthesis and construction of stock pools to portfolio optimization, and back-testing against corresponding index investments. The thesis proposes a different classification of factors from the prevailing ones, validates a broader range of factors and evaluates their performance under the Chinese stock market. The empirical results show that PCA has the ability to explain most of the factors, and the results of ranking the factors by XGBoost and performing multi-factor synthesis outperform the major index fund investments, demonstrating the effectiveness of factor investing in improving portfolio efficiency.

Then, this thesis takes the position of an individual investor considering the minimum investment requirements and further optimizes the portfolio through a pool of screened stocks using mean-variance optimization and hierarchical clustering-based optimization. The empirical results show that hierarchical clustering-based optimization using the ward linkage approach, combined with Spearman's rank correlation as a distance measure, yields the most robust and superior performance across time and market conditions.

In addition, the thesis repeatedly emphasizes the importance of setting realistic risk constraints and appropriately diversifying factor exposures to mitigate risk in factor-based portfolios. Potential challenges of factor investing are also discussed, including the risk of overfitting, the difficulty of accurately identifying and measuring factors, the choice of backtesting time, and the prudence of automated trading versus stock pool selection.

Despite the encouraging results, the model has some limitations. For example, the factor and portfolio optimization analysis are based on historical data and the results may not be representative of future performance. In addition, the effects of transaction costs and market liquidity are not considered, which may affect the performance of the portfolios. Therefore, further research is needed to investigate the impact of these factors.

In summary, the evidence provided in this thesis supports the use of factor investing as a promising approach to portfolio selection and optimization that has the potential to provide superior returns to investors. By integrating various risk factors and utilizing advanced optimization techniques, factor investing can help investors achieve

better risk-adjusted returns and has the potential to improve long-term portfolio performance. Future research could focus on improving methodologies and exploring new factors, as well as examining the impact of transaction costs and factor trading signal identification.

Bibliography

Professional book

GRINOLD, R. and R. KAHN. *Advances in Active Portfolio Management.* New York: McGraw-Hill, 2019. ISBN: 978-1-26-045372-0.

HILPISCH, Y. J. *Python for Finance: Mastering Data-Driven Finance.* 2nd ed. Sebastopol, CA: O'Reilly, 2018. ISBN 978-1-492-02433-0.

REILLY, F. K., BROWN, K. C. and S. J. LEEDS. *Investment Analysis and Portfolio Management*. Eleventh edition. Boston: Cengage, 2019. ISBN 978-1-305-26299-7.

An thesis in a journal (periodical) or in proceedings

Ahmed D, Soleymani F, Ullah M Z, et al. Managing the risk based on entropic value-at-risk under a normal-Rayleigh distribution. *Applied Mathematics and Computation*, 2021, 402: 126129, ISSN 0096-3003.

Alhashel B S, Almudhaf F W, Hansz J A. Can technical analysis generate superior returns in securitized property markets? Evidence from East Asia markets. *Pacific-Basin Finance Journal*, 2018, 47: 92-108, ISSN 0927-538X.

Barinov A. Profitability anomaly and aggregate volatility risk. *Journal of Financial Markets*, 2022: 100782, ISSN 1386-4181.

Burggraf T. Beyond risk parity–A machine learning-based hierarchical risk parity approach on cryptocurrencies. *Finance Research Letters*, 2021, 38: 101523, ISSN 1544-6123.

Cenesizoglu T, Reeves J J. CAPM, components of beta and the cross section of expected returns. *Journal of Empirical Finance*, 2018, 49: 223-246, ISSN 0927-5398.

Chen M H. Risk and return: CAPM and CCAPM. *The Quarterly Review of Economics and Finance*, 2003, 43(2): 369-393, ISSN 1062-9769.

Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016: 785-794.

De Prado M L. Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management*, 2016, 42(4): 59-69, ISSN 0095-4918.

Deng S, Huang X, Zhu Y, et al. Stock index direction forecasting using an explainable eXtreme Gradient Boosting and investor sentiments. *The North American Journal of Economics and Finance*, 2023, 64: 101848, ISSN 1062-9408.

Fama, E. F. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 1970, 25(2): 383-417, ISSN 0022-1082.

Fons E, Dawson P, Yau J, et al. A novel dynamic asset allocation system using Feature Saliency Hidden Markov models for smart beta investing. *Expert Systems with Applications*, 2021, 163: 113720, ISSN 0957-4174.

Foye J. A comprehensive test of the Fama-French five-factor model in emerging markets. *Emerging Markets Review*, 2018, 37: 199-222, ISSN 1566-0141.

Gasser S M, Rammerstorfer M, Weinmayer K. Markowitz revisited: Social portfolio engineering. *European Journal of Operational Research*, 2017, 258(3): 1181-1190, ISSN 0377-2217.

Georgalos K, Paya I, Peel D A. On the contribution of the Markowitz model of utility to explain risky choice in experimental research. *Journal of Economic Behavior & Organization*, 2021, 182: 527-543, ISSN 0167-2681.

Guo M, Ou-Yang H. Alpha decay and Sharpe ratio: Two measures of investor performance. *Economic Modelling*, 2021, 104: 105558, ISSN 0264-9993.

Han Y, Kim J, Enke D. A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost. *Expert Systems with Applications*, 2023, 211: 118581, ISSN 0957-4174.

Hoffmann A O I, Shefrin H. Technical analysis and individual investors. *Journal of Economic Behavior & Organization*, 2014, 107: 487-511, ISSN 0167-2681.

Huang T L. Is the Fama and French five-factor model robust in the Chinese stock market?. *Asia Pacific Management Review*, 2019, 24(3): 278-289, ISSN 1029-3132.

Kelsey D, Yalcin E. The arbitrage pricing theorem with incomplete preferences. *Mathematical Social Sciences*, 2007, 54(1): 90-105, ISSN 0165-4896.

Lin J, Wang M, Cai L. Are the Fama–French factors good proxies for latent risk factors? Evidence from the data of SHSE in China. *Economics Letters*, 2012, 116(2): 265-268, ISSN 0165-1765.

Lin Q. Noisy prices and the Fama–French five-factor asset pricing model in China. *Emerging Markets Review*, 2017, 31: 141-163, ISSN 1566-0141.

Liu C, Wang F F, Xue W. The annual report tone and return Comovement—Evidence from China's stock market. *International Review of Financial Analysis*, 2023: 102610, ISSN 1057-5219.

Liu H, Gao Y C. The impact of corporate lifecycle on Fama–French three-factor model. *Physica A: Statistical Mechanics and Its Applications*, 2019, 513: 390-398, ISSN 0378-4371.

Lohre H, Rother C, Schäfer K A. Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi‐asset Multi‐factor Allocations. *Machine Learning for Asset Management: New Developments and Financial Applications*, 2020: 329-368, ISBN: 978-1-119-75117-5.

Markowitz H. Portfolio selection. *Journal of Finance*, 1952, 7(1): 77-91, ISSN 0022-1082.

Millea A, Edalat A. Using Deep Reinforcement Learning with Hierarchical Risk Parity for Portfolio Optimization. *International Journal of Financial Studies*, 2022, 11(1): 10, ISSN 2227-7072.

Nalpas N, Simar L, Vanhems A. Portfolio selection in a multi-moment setting: A simple Monte-Carlo-FDH algorithm. *European Journal of Operational Research*, 2017, 263(1): 308-320, ISSN 0377-2217.

Peng Y, Albuquerque P H M, Kimura H, et al. Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators. *Machine Learning with Applications*, 2021, 5: 100060, ISSN 2666-8270.

Pesaran M H, Smith R P. Arbitrage pricing theory, the stochastic discount factor and estimation of risk premia from portfolios. *Econometrics and Statistics*, 2021,26: 17-30, ISSN 2452-3062.

Petneházi G. Quantile convolutional neural networks for Value at Risk forecasting. *Machine Learning with Applications*, 2021, 6: 100096, ISSN 2666-8270.

Priestley, R. The arbitrage pricing theory, macroeconomic and financial factors, and expectations generating processes. *Journal of Banking & Finance*, 1996, 20(5): 869-890, ISSN 0378-4266.

Rocciolo F, Gheno A, Brooks C. Explaining abnormal returns in stock markets: An alpha-neutral version of the CAPM. *International Review of Financial Analysis*, 2022, 82: 102143, ISSN 1057-5219.

Ross, S. A. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 1976, 13(3): 341-360, ISSN 0022-0531.

Seyfi S M S, Sharifi A, Arian H. Portfolio Value-at-Risk and expected-shortfall using an efficient simulation approach based on Gaussian Mixture Model. *Mathematics and Computers in Simulation*, 2021, 190: 1056-1079, ISSN 0378-4754.

Shadabfar M, Cheng L. Probabilistic approach for optimal portfolio selection using a hybrid Monte Carlo simulation and Markowitz model. *Alexandria engineering journal*, 2020, 59(5): 3381-3393, ISSN 1110-0168.

Shah A, Chauhan Y, Chaudhury B. Principal component analysis based construction and evaluation of cryptocurrency index. *Expert Systems with Applications*, 2021, 163: 113796, ISSN 0957-4174.

Song Z, Gong X, Zhang C, et al. Investor sentiment based on scaled PCA method: A powerful predictor of realized volatility in the Chinese stock market. *International Review of Economics & Finance*, 2023, 83: 528-545, ISSN 1059-0560.

Timmermann, A., Granger, C. W. J. Efficient market hypothesis and forecasting. *International Journal of forecasting*, 2004, 20(1): 15-27, ISSN 0169-2070.

Ţiţan, A. G. The efficient market hypothesis: Review of specialized literature and empirical research. *Procedia Economics and Finance*, 2015, 32: 442-449, ISSN 2212-5671.

Vukovic D, Vyklyuk Y, Matsiuk N, et al. Neural network forecasting in prediction Sharpe ratio: Evidence from EU debt market. *Physica A: Statistical Mechanics and its Applications*, 2020, 542: 123331, ISSN 0378-4371.

Wang Z M, Chiao C, Chang Y T. Technical analyses and order submission behaviors: Evidence from an emerging market. *International Review of Economics & Finance*, 2012, 24: 109-128, ISSN 1059-0560.

Wang Z, Tan S. Identifying idiosyncratic stock return indicators from large financial factor set via least angle regression. *Expert Systems with Applications*, 2009, 36(4): 8350-8355, ISSN 0957-4174.

Wang, X. Efficient markets are more connected: An entropy-based analysis of the energy, industrial metal and financial markets. *Energy Economics*, 2022, 111: 106067, ISSN 0140-9883.

Xi X, Gao X, Zhou J, et al. Uncovering the impacts of structural similarity of financial indicators on stock returns at different quantile levels. *International Review of Financial Analysis*, 2021, 76: 101787, ISSN 1057-5219.

Zhang Y, Wang Y. Forecasting crude oil futures market returns: A principal component analysis combination approach. *International Journal of Forecasting,* 2023, 39(2): 659-673, ISSN 0169-2070.

Electronic documents and others

Aditya Vyas, The Hierarchical Risk Parity Algorithm: An Introduction, [online]. 2023. [cite. 29/03/2023]. Available on: https://hudsonthames.org/an-introduction-to-the-hierarchical-risk-parity-algorithm/

Introduction to ordination, finding patterns in your data. [online]. 2023. [cite. 07/04/2023]. Available on: https://ourcodingclub.github.io/tutorials/ordination/

Rainbow Group (SZ:003023) [online]. 2023. [cite. 12/04/2023]. Available on: https://xueqiu.com/S/SZ003023

The Barra US Equity Model (USE4), Methodology Notes. [online]. 2023. [cite. 20/03/2023]. Available on: http://faculty.baruch.cuny.edu/lwu/890/USE4.pdf

TIOBE Index for March 2023. [online]. 2023. [cite. 12/03/2023] Available on: https://www.tiobe.com/tiobe-index/

List of Abbreviations

AIC	Adjustmental Information Coefficient
APT	Arbitrage Pricing Theory
CAPM	Capital Asset Pricing Model
CDF	Cumulative Distribution Function
CMA	Conservative Minus Aggressive
CR	Calmar Ratio
EMH	Efficient Market Hypothesis
HML	High Minus Low
HRP	Hierarchical Risk Parity
IC	Information Coefficient
IR	Information Ratio
KD	Stochastic
MA	Moving Average
MACD	Moving Average Convergence Divergence
MAD	Expected Median
MC	Monte Carlo Simulation
MDD	Max Drawdown
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
RMW	Robust Minus Weak
SMB	Small Minus Big
SR	Sharpe Ratio
VaR	Value at Risk
XGBoost	Extreme Gradient Boosting

List of Annexes

Annex 1 Factor Code Parameters and Description

Factor Type	Factor Parameter	Factor Code	Meaning
Base Factor	operating revenue	operating_revenue	The total revenue obtained by the company from operating its main business
	inventory	inventory	Finished goods or merchandise held for sale, work in process, materials and supplies used in the production process or in the provision of services in the ordinary course of an enterprise's activities
	Operating profit	profit_from_operation	The profit achieved by an enterprise in its total sales business, also known as operating profit, operating profit, which includes the profit from the main business
	Selling expenses	selling_expense	Refers to the enterprise in the sale of products, self-made semi-finished products and industrial services and other costs incurred in the process
	Total operating expenses	total_expense	The costs incurred by a business for its operational activities
	Management expenses	ga_expense	The costs incurred by the administration of the enterprise for the management and organization of operations
	Cash equivalent	cash_equivalent	Cash equivalents include bank accounts and marketable securities
	Total market value of tradable shares	market_cap_2	Total market value of outstanding shares = Outstanding share capitalization * Unrevised closing price of A shares
	Total assets	total_assets	Economic resources owned or controlled by an enterprise that can be measured in monetary terms, including all kinds of property, claims and other rights
	Basic earnings per share	basic_earnings_per_share	Basic earnings per share is calculated by dividing the net profit for the period attributable to common shareholders by the weighted average number of common shares outstanding
Quality Factor	Gross operating income per share	operating_total_revenue_per_share_lyr	Total operating income lyr / Total equity
	EBIT per share	ebit_per_share_lyr	EBIT lyr / total share capital
	Cash flow per share	cash_flow_per_share_lyr	(Net cash flows from operating activities lyr + Net cash flows from investing activities lyr + Net cash flows from financing activities lyr) / Total equity
	Capital reserve per share	capital_reserve_per_share_lyr	Capital reserve lyr / Total share capital
	Retained earnings per share	retained_earnings_per_share_lyr	Retained earnings lyr / total equity
	Net assets per share	book_value_per_share_lyr	Total shareholders' equity attributable to the parent lyr / total share capital
	Tangible assets per share	tangible_asset_per_share_lyr	(Total assets lyr - intangible assets lyr - goodwill lyr) / Total equity
	Liabilities per share	liabilities_per_share_lyr	Total liabilities lyr / Total equity
	Depreciation and amortization per share	depreciation_per_share_lyr	(Depreciation of fixed assets lyr + Amortization of intangible assets lyr + Amortization of long term amortization lyr) / Total equity
	Cash equivalent per share	cash_equivalent_per_share_lyr	Monetary fund balance lyr / total equity
Growth Factor	PEG value	peg_ratio_lyr	P/E ratio lyr / Average growth rate of net profit attributable to parent company in the past year *100 lyr
	YoY growth rate of total operating revenue	inc_revenue_lyr	Total operating revenue lyr / Previous year's total operating revenue lyr - 1
	YoY growth rate of ROE	inc_return_on_equity_lyr	Diluted return on net assets lyr / Last year's diluted return on net assets lyr - 1
	YoY growth rate of net assets per share	inc_book_per_share_lyr	Net assets per share lyr / Net assets per share last year lyr - 1
	YoY growth rate of net operating cash flow	net_operate_cash_flow_growth_ratio_lyr	Net cash flow from operating activities lyr / Net cash flow from operating activities last year lyr - 1
	YoY growth rate of net assets	net_asset_growth_ratio_lyr	Total equity attributable to the parent company lyr / Total equity attributable to the parent company last year lyr - 1
	YoY growth rate of operating income	operating_revenue_growth_ratio_lyr	(Operating income lyr - Last year's operating income lyr) / Last year's operating income lyr
	YoY growth rate of operating profit	operating_profit_growth_ratio_lyr	(Operating profit lyr - Last year's operating profit lyr) / Last year's operating profit lyr
	YoY growth rate of net profit	net_profit_growth_ratio_lyr	(Net income lyr - Last year's net income lyr) / Last year's net income lyr
Risk Factor	Price Change Percentage	PCNT	PCNT = (CLOSE-REF(CLOSE,1))/CLOSE*100;
	Detrended Price Oscillator	DPO	DPO = CLOSE - REF(MA(CLOSE, M1), M2)
	Marginal Cost of Short-Term	MCST	MCST = DMA(AMOUNT / VOLUME, 100 * VOLUME / CAPITAL)
	Accumulative Swing Index	ASI	ASI = SUM(SI, M1)
	Stochastic Oscillator %K%D	KDJ_K	RSV = (CLOSE - LLV(LOW, N)) / (HHV(HIGH, N) - LLV(LOW, N)) _ 1I K = EMA(RSV, (M1 _ 2 - 1))
	Money Flow Index	MFI	TYP = (HIGH + LOW + CLOSE) / 3 V1 = SUM(IF(TYP REF(TYPE, 1), TYP _ VOLUME, 0), N) / SUM(IF(TYP < REF(TYP, 1), TYP _ VOLUME, 0), N); MFI = 100 – (100 / (1+V1))
	Sea Water Level	SWL	SWL = (EMA(CLOSE,5)7+EMA(CLOSE,10)3)/10
	Commodity Channel Index	CCI	CCI = (TYP – MA(TYP, N)) / (0.015 * AVEDEV (TYP, N)) TYP = (HIGH + LOW + CLOSE) / 3
	Slow Stochastic Oscillator D%K	SKD_K	LOWV = LLV(LOW, N)HIGHV = HHV(HIGH, N)RSV = EMA((CLOSE – LOWV) / (HIGHV – LOWV) * 100, M)SKD_K = EMA(RSV , M)

Factor Type	Factor Parameter	Factor Code	Meaning
Emotional Factor	Accumulation/ Distribution	AR	AR = SUM(HIGH - OPEN, M1) / SUM(OPEN - LOW, M1) _ 100
	Buyer's/ Bullish Ratio	BR	BR = SUM(MAX(0, HIGH - REF(CLOSE, 1)), M1) / SUM(MAX(0, REF(CLOSE, 1) - LOW), M1) _ 100
	Psychological line	SY	SY = COUNT(CLOSE>REF(CLOSE,1),N)/N*100
	Market Strength (CYR)	CYR	DIVE = 0.01*EMA(AMOUNT,N)/EMA(VOLUME,N); CYR = (DIVE/REF(DIVE,1)-1)*100
	Market Force (CYF)	CYF	CYF = 100 – 100 / (1 + EMA(HSL, N))
	On-Balance Volume	OBV	OBV=REF(OBV, 1) + sgn × VOLUME; where sgn is a symbolic function whose value is determined by the followin sgn=1 , CLOSE>REF(CLOSE, 1); sgn=0, CLOSE = REF(CLOSE, 1); sgn=-1 , CLOSE< REF(CLOSE, 1)
	Aroon indicator	AROON_UP	AROON_UP = [(number of days in the calculation period - number of days after the highest price)/number of days in the calculation period]*100
		AROON_DOWN	AROON_DOWN = [(number of days in the calculation period - number of days after the minimum price)/number of days in the calculation period]*100
	Amplitude	AMP1	
		AMP3	AMP1,3,5... = (HHV(HIGH,N)-LLV(LOW,N))/REF(CLOSE,N)
		AMP5	
	Average turnover Volume	VOL3	HSL =100* VOLUME / CAPITAL VOL3, 5, 10... = MA(HSL, N) HSL stands for turnover rate CAPITAL represents the outstanding share capital
	Williams %R	WR	WR = (HHV(HIGH, N) - CLOSE) / (HHV(HIGH, N) - LLV(LOW, N)) * 100
Momentum Factor	Bias Ratio	BIAS5	(CLOSE - MA(CLOSE, L1)) / MA(CLOSE, L1) * 100 BIAS L1 = 5, ...
	Mass Line	MASS	MASS = SUM(MA(HIGH-LOW,N1)/MA(MA(HIGH-LOW,N1),N1),N2)
	Momentum Line	MTM	MTM = CLOSE – REF(CLOSE, N)
	Trend Indicator	DI1	TR = SUM(MAX(MAX(HIGH - LOW, ABS(HIGH - REF(CLOSE, 1))), ABS(LOW - REF(CLOSE, 1))), M1); HD = HIGH - REF(HIGH, 1); LD = REF(LOW, 1) - LOW;DMP = SUM(IF((HD 0) & (HD LD), HD, 0), M1); DI1 = DMP _ 100 / TR
		DI2	DMM = SUM(IF((LD 0) & (LD HD), LD, 0), M1) DI2 = DMM _ 100 / TR
		ADX	ADX = MA(ABS(DI2 - DI1) / (DI1 + DI2) * 100, M2)
	CR Indicator	CR	MID = REF(HIGH+LOW, 1) / 2 CR = SUM(MAX(0,HIGH-MID),N)/SUM(MAX(0,MID-LOW),N)*100
	Ultimate Divergence Line	UDL	UDL = (MA(CLOSE,N1)+MA(CLOSE,N2)+MA(CLOSE,N3)+MA(CLOSE,N4))/4
Technology Factor	Bolinger bands	BOLL	BOLL = MA(CLOSE, N)
	Triple exponential average moving average	BOLL_UP	BOLLUP = BOLL + STD(CLOSE, N) * P
	Volume ratio	BOLL_DOWN	BOLLDOWN = BOLL - STD(CLOSE, N) * P
	Moving average convergence divergence	MACD_DIFF	DIFF = EMA(CLOSE, SHORT) - EMA(CLOSE, LONG)
		MACD_DEA	DEA = EMA(DIFF, M)
		MACD_HIST	HIST = (DIFF - DEA) * 2
	Bull and Bear index	BBI	BBI = (MA(CLOSE, M1) + MA(CLOSE, M2) + MA(CLOSE, M3) + MA(CLOSE, M4)) / 4
	Moving average	MA3	MA3, 5, 10... = MA(CLOSE, N)
	Exponential moving average	EMA3	EMA3, 5, 10... = EMA(CLOSE, N)
	High moving average	HMA3	HMA3, 5, 10... = MA(HIGH, N)
	Low moving average	LMA3	LMA3, 5, 10... = MA(LOW, N)...
	Variable moving average	VMA3	VV = (HIGH+OPEN+LOW+CLOSE)/4 VMA3, 5, 10... = MA(VV, N)...

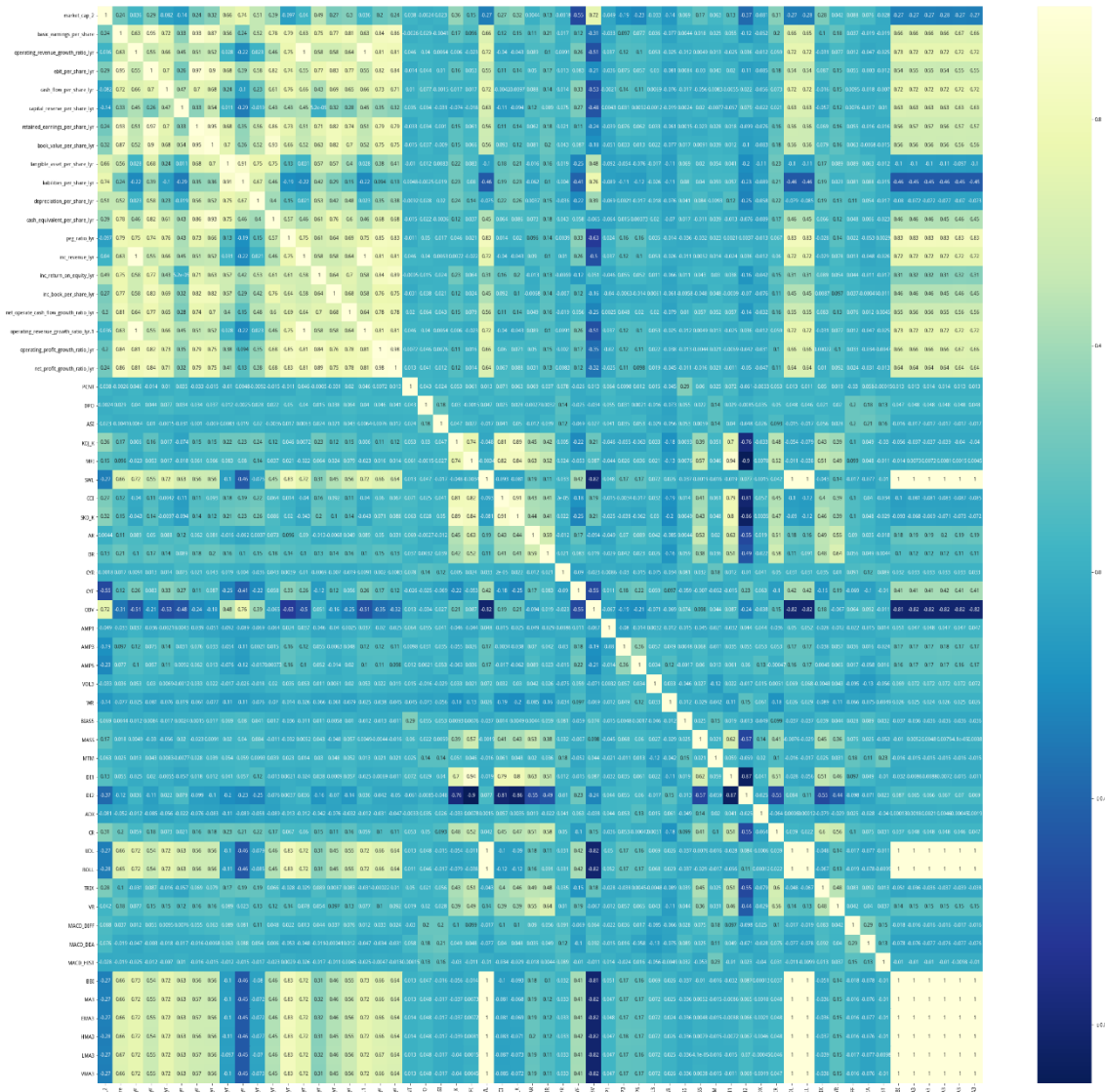
Annex 2 Single Factor IC Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
market_cap_2	964.0	-0.002361	0.147292	-0.423516	-0.101318	-0.006844	0.085056	0.458665
basic_earnings_per_share	964.0	0.022030	0.105928	-0.354611	-0.046772	0.025383	0.090403	0.370666
operating_revenue_growth_ratio_lyr	964.0	0.005317	0.052859	-0.182325	-0.026759	0.005432	0.039695	0.171887
ebit_per_share_lyr	964.0	0.017154	0.096022	-0.364950	-0.048294	0.016085	0.082729	0.341546
cash_flow_per_share_lyr	964.0	0.004954	0.039128	-0.150259	-0.021245	0.005804	0.031296	0.135349
capital_reserve_per_share_lyr	964.0	0.004229	0.040484	-0.213242	-0.020125	0.003227	0.030902	0.183223
retained_earnings_per_share_lyr	964.0	0.019889	0.083702	-0.302452	-0.036398	0.022023	0.073383	0.302165
book_value_per_share_lyr	964.0	0.016667	0.067637	-0.307671	-0.025368	0.014524	0.057454	0.259505
tangible_asset_per_share_lyr	964.0	0.012437	0.076526	-0.211818	-0.034085	0.012837	0.062704	0.270269
liabilities_per_share_lyr	964.0	0.005249	0.081038	-0.216161	-0.050149	0.003006	0.060071	0.231088
depreciation_per_share_lyr	964.0	0.012026	0.064888	-0.192684	-0.030414	0.014313	0.057569	0.223385
cash_equivalent_per_share_lyr	964.0	0.011771	0.057809	-0.242923	-0.021863	0.011937	0.047544	0.204701
peg_ratio_lyr	964.0	0.005680	0.049065	-0.166466	-0.027388	0.004774	0.041493	0.161452
inc_revenue_lyr	964.0	0.005327	0.052838	-0.182527	-0.026915	0.006156	0.039253	0.171944
inc_return_on_equity_lyr	964.0	0.010485	0.048131	-0.133769	-0.018979	0.009911	0.041636	0.178088
inc_book_per_share_lyr	964.0	0.008676	0.047558	-0.181496	-0.022822	0.007924	0.037982	0.164734
net_operate_cash_flow_growth_ratio_lyr	964.0	0.006710	0.035784	-0.106138	-0.016835	0.007229	0.031000	0.122984
operating_revenue_growth_ratio_lyr.1	964.0	0.005317	0.052859	-0.182325	-0.026759	0.005432	0.039695	0.171887
operating_profit_growth_ratio_lyr	964.0	0.009139	0.051835	-0.159516	-0.024559	0.010151	0.044677	0.201875
net_profit_growth_ratio_lyr	964.0	0.009456	0.056225	-0.172370	-0.025077	0.009398	0.047422	0.196086
PCNT	964.0	-0.016400	0.124234	-0.477521	-0.095729	-0.020641	0.051630	0.387802
DPO	964.0	-0.014290	0.150860	-0.548460	-0.116381	-0.019882	0.091399	0.413019
ASI	964.0	-0.015521	0.143621	-0.542002	-0.109793	-0.022455	0.078480	0.405292
KDJ_K	964.0	-0.023882	0.133068	-0.535884	-0.109869	-0.026491	0.060495	0.465152
MFI	964.0	-0.026811	0.106999	-0.494583	-0.099735	-0.027704	0.045500	0.290154
SWL	964.0	0.001503	0.142887	-0.609825	-0.090864	0.004722	0.101207	0.376364
CCI	964.0	-0.028605	0.130252	-0.462704	-0.112286	-0.036587	0.060617	0.430872
SKD_K	964.0	-0.022738	0.131013	-0.538596	-0.107146	-0.026406	0.061614	0.464071
AR	964.0	-0.032673	0.102021	-0.418779	-0.098913	-0.035536	0.031892	0.328404
BR	964.0	-0.023193	0.109251	-0.557031	-0.098067	-0.024186	0.049401	0.373821
CYR	964.0	-0.010356	0.149578	-0.489889	-0.113041	-0.011898	0.097971	0.463560
CYF	964.0	-0.049115	0.181185	-0.504581	-0.171241	-0.063520	0.073879	0.465928
OBV	964.0	-0.007293	0.127807	-0.432112	-0.087013	-0.009020	0.072736	0.406602
AMP1	964.0	-0.068657	0.127310	-0.441022	-0.154120	-0.079310	0.012504	0.340835
AMP3	964.0	-0.054997	0.136991	-0.470354	-0.148896	-0.067264	0.034937	0.395267
AMP5	964.0	-0.055852	0.142883	-0.490572	-0.154952	-0.070038	0.039342	0.484048
VOL3	964.0	-0.057274	0.169788	-0.489570	-0.165730	-0.079222	0.053315	0.439830
WR	964.0	0.026092	0.131181	-0.445260	-0.058167	0.028787	0.110499	0.464691
BIAS5	964.0	-0.032286	0.136389	-0.448591	-0.122680	-0.038929	0.054130	0.456848
MASS	964.0	-0.025467	0.092742	-0.365449	-0.085088	-0.028291	0.034484	0.290713
MTM	964.0	-0.015737	0.134808	-0.558342	-0.100223	-0.016162	0.091236	0.424040
DI1	964.0	-0.037776	0.106481	-0.460521	-0.107991	-0.042036	0.033489	0.282831
DI2	964.0	0.029458	0.120760	-0.409293	-0.045662	0.033047	0.103575	0.557754
ADX	964.0	-0.003215	0.076787	-0.321565	-0.048461	-0.003950	0.038716	0.253106
CR	964.0	-0.029557	0.130334	-0.566143	-0.118873	-0.033683	0.053021	0.374400
UDL	964.0	0.000492	0.143045	-0.613650	-0.092747	0.005726	0.098979	0.375655
BOLL	964.0	0.001875	0.143554	-0.606644	-0.091998	0.007978	0.100580	0.380720
TRIX	964.0	-0.020776	0.135350	-0.512024	-0.107334	-0.020752	0.069022	0.389581
VR	964.0	-0.026504	0.097842	-0.465723	-0.091083	-0.025275	0.036814	0.310270
MACD_DIFF	964.0	-0.001371	0.146136	-0.517837	-0.101223	-0.013495	0.095142	0.418501
MACD_DEA	964.0	0.003179	0.141169	-0.485752	-0.090227	-0.002878	0.089759	0.382487
MACD_HIST	964.0	-0.016750	0.116383	-0.514307	-0.093919	-0.012825	0.071486	0.476859
BBI	964.0	0.000897	0.142747	-0.612215	-0.091988	0.006046	0.099462	0.376589
MA3	964.0	0.000747	0.142742	-0.611793	-0.092528	0.003364	0.100403	0.373399
EMA3	964.0	0.001182	0.142558	-0.611651	-0.091210	0.003750	0.100576	0.373990
HMA3	964.0	0.000346	0.143274	-0.612943	-0.093713	0.003964	0.100395	0.373976
LMA3	964.0	0.001834	0.142496	-0.610538	-0.090119	0.004761	0.101729	0.375123
VMA3	964.0	0.001084	0.142875	-0.611548	-0.092377	0.004265	0.100691	0.374642

Annex 3 Single Factor Test Results

Factor Type	Factor Parameter	Average return	IC mean	IC std	IC>0.02	t-stat(IC)	IC skew	IC kurtosis	Risk-adjust IC
Base Factor	operating_revenue	max_loss (35.0%) exceeded 60.8%							
	inventory	max_loss (35.0%) exceeded 59.1%							
	profit_from_operation	max_loss (35.0%) exceeded 56.6%							
	selling_expense	max_loss (35.0%) exceeded 63.2%							
	total_expense	max_loss (35.0%) exceeded 63.1%							
	ga_expense	max_loss (35.0%) exceeded 55.4%							
	cash_equivalent	max_loss (35.0%) exceeded 40.4%							
	market_cap_2	-0.000097	-0.002	0.147	0.408	-0.498	0.236	0.133	-0.016
	total_assets	max_loss (35.0%) exceeded 62.7%							
	basic_earnings_per_share	0.000576	0.022	0.106	0.526	6.456	-0.096	0.425	0.208
Quality Factor	operating_total_revenue_per_share_lyr	0.000216	0.013	0.064	0.469	6.452	-0.128	0.249	0.208
	ebit_per_share_lyr	0.000357	0.017	0.096	0.483	5.546	-0.07	0.326	0.179
	cash_flow_per_share_lyr	0.000111	0.005	0.039	0.353	3.931	-0.144	0.388	0.127
	capital_reserve_per_share_lyr	-0.000043	0.004	0.040	0.322	3.244	-0.088	1.267	0.104
	retained_earnings_per_share_lyr	0.000299	0.020	0.084	0.511	7.378	-0.057	0.416	0.238
	book_value_per_share_lyr	0.000160	0.017	0.068	0.464	7.651	-0.086	1.180	0.246
	tangible_asset_per_share_lyr	0.000084	0.012	0.077	0.464	5.046	-0.011	-0.035	0.163
	liabilities_per_share_lyr	-0.000003	0.005	0.081	0.412	2.011	0.071	-0.325	0.065
	depreciation_per_share_lyr	0.000169	0.012	0.065	0.473	5.754	-0.179	-0.103	0.185
	cash_equivalent_per_share_lyr	0.000124	0.012	0.058	0.438	6.322	-0.174	0.893	0.204
Growth Factor	peg_ratio_lyr	0.000095	0.006	0.049	0.388	3.594	-0.071	-0.066	0.116
	inc_revenue_lyr	0.000177	0.005	0.053	0.394	3.130	-0.064	0.268	0.101
	inc_return_on_equity_lyr	0.000209	0.010	0.048	0.406	6.763	0.076	0.164	0.218
	inc_book_per_share_lyr	0.000080	0.009	0.048	0.409	5.664	-0.039	0.585	0.182
	net_operate_cash_flow_growth_ratio_lyr	0.000156	0.007	0.036	0.370	5.822	-0.094	0.046	0.188
	net_asset_growth_ratio_lyr	max_loss (35.0%) exceeded 40.6%							
	operating_revenue_growth_ratio_lyr	0.000175	0.005	0.053	0.392	3.123	-0.060	0.271	0.101
	operating_profit_growth_ratio_lyr	0.000188	0.009	0.052	0.415	5.474	-0.029	0.230	0.176
	net_profit_growth_ratio_lyr	0.000197	0.009	0.056	0.427	5.222	-0.049	0.263	0.168
Risk Factor	PCNT	0.001950	-0.018	0.126	0.353	-4.280	0.329	0.517	-0.140
	DPO	0.000055	-0.032	0.144	0.342	-6.689	0.047	0.319	-0.225
	MCST	max_loss (35.0%) exceeded 76.1%							
	ASI	-0.000215	-0.032	0.134	0.330	-7.129	-0.016	0.440	-0.237
	KDJ_K	0.000353	-0.024	0.133	0.364	-5.616	0.032	0.627	-0.181
	MFI	6.014733	-0.027	0.107	0.326	-7.780	0.019	0.225	-0.251
	SWL	0.000178	0.002	0.143	0.454	0.327	-0.296	0.270	0.011
	CCI	0.000170	0.001	0.143	0.451	0.163	-0.294	0.275	0.005
	SKD_K	0.000293	-0.023	0.131	0.364	-5.389	-0.003	0.665	-0.174
Emotional Factor	AR	-0.000301	-0.033	0.102	0.282	-9.999	0.147	0.517	-0.322
	BR	0.000138	-0.024	0.109	0.335	-6.392	0.050	0.940	-0.224
	SY	max_loss (35.0%) exceeded 88.4%							
	CYR	0.000971	-0.017	0.150	0.391	-3.53	0.075	0.130	-0.116
	CYF	-0.000860	-0.045	0.184	0.360	-7.329	0.112	-0.442	-2.42
	OBV	-0.000294	-0.007	0.128	0.405	-1.772	0.037	0.103	-0.057
	AROON_UP	max_loss (35.0%) exceeded 55.2%							
	AROON_DOWN	max_loss (35.0%) exceeded 88.8%							
	AMP1	-0.001077	-0.068	0.129	0.236	-16.284	0.298	0.020	-0.531
	AMP3	-0.000161	-0.054	0.138	0.288	-12.180	0.303	0.063	-0.395
	AMP5	-0.000481	-0.056	0.144	0.291	-12.036	0.383	0.228	-0.389
	VOL3	-0.000671	-0.048	0.176	0.342	-8.099	0.186	-0.470	-0.274
	WR	-0.000529	0.027	0.132	0.527	6.326	-0.158	0.529	0.205
Momentum Factor	BIAS5	0.001172	-0.035	0.137	0.318	-7.852	0.262	0.327	-0.256
	MASS	-0.000340	-0.025	0.093	0.302	-8.526	0.124	0.472	-0.275
	MTM	0.000183	-0.028	0.134	0.342	-6.188	0.038	0.516	-0.208
	DI1	-0.000089	-0.038	0.106	0.286	-11.015	0.091	0.114	-0.355
	DI2	-0.000061	0.029	0.121	0.548	7.574	-0.043	0.649	0.244
	ADX	-0.000095	-0.003	0.077	0.354	-1.300	0.114	0.836	-0.042
	CR	-0.000165	-0.032	0.129	0.333	-7.757	0.009	0.462	-0.252
	UDL	0.000045	0.000	0.143	0.456	0.107	-0.303	0.273	0.003
Technology Factor	BOLL	0.000057	0.002	0.144	0.460	0.405	-0.305	0.255	0.013
	TRIX	0.000140	-0.022	0.135	0.356	-4.970	0.019	0.192	-0.160
	VR	-0.000224	-0.028	0.098	0.303	-8.728	0.046	0.548	-0.282
	MACD_DIFF	0.000338	-0.019	0.136	0.349	-4.181	0.026	0.333	-0.140
	MACD_DEA	0.000165	-0.015	0.131	0.377	-3.046	-0.010	0.238	-0.114
	MACD_HIST	0.000517	-0.025	0.117	0.343	-6.241	-0.107	0.892	-0.209
	BBI	5.479628	0.001	0.143	0.463	0.195	-0.307	0.276	0.006
	MA3	0.000171	0.001	0.143	0.451	0.163	-0.294	0.275	0.005
	EMA3	0.000208	0.001	0.143	0.454	0.257	-0.295	0.277	0.008
	HMA3	0.000160	0.000	0.143	0.452	0.075	-0.289	0.268	0.002
	LMA3	0.000183	0.002	0.142	0.462	0.400	-0.296	0.274	0.013
	VMA3	0.000171	0.001	0.143	0.453	0.236	-0.293	0.272	0.008

Annex 4 Feature Engineering-Spearman Coefficient between Factors



Annex 5 Multi-factor Synthetic Screening of Stocks

Rank by Trading frequency				Rank by Holiding period			
Stock Symbol	Stock Name	Trading frequency	Holiding period (days)	Stock Symbol	Stock Name	Trading frequency	Holiding period (days)
002608.XSHE	Jiangsu Guoxin	10	223	603833.XSHG	Oppein Home Group	3	243
601186.XSHG	China Railway Construction	9	192	002608.XSHE	Jiangsu Guoxin	10	223
600637.XSHG	Oriental Pearl	9	177	601799.XSHG	Xingyu Automotive Lighting Systems	3	221
000553.XSHE	ADAMA A	8	198	603345.XSHG	Anjoy Foods Group	2	205
603658.XSHG	Autobio Diagnostics	8	142	601607.XSHG	Shanghai Pharmaceuticals Holding	8	204
601607.XSHG	Shanghai Pharmaceuticals Holding	8	204	000553.XSHE	ADAMA A	8	198
603883.XSHG	Lbx Pharmacy	6	163	601186.XSHG	China Railway Construction	9	192
601808.XSHG	China Oilfield Services Limited	6	119	600637.XSHG	Oriental Pearl	9	177
601898.XSHG	China Coal Energy	5	102	603883.XSHG	Lbx Pharmacy	6	163
600021.XSHG	Shanghai Electric	5	102	603658.XSHG	Autobio Diagnostics	8	142
601800.XSHG	China Communications Construction Company	5	104	000050.XSHE	Tianma Microelectronics A	3	122
603877.XSHG	Peacebird Group	4	80	601021.XSHG	Spring Airlines	4	122
002024.XSHE	ST Suning	4	78	601808.XSHG	China Oilfield Services Limited	6	119
600498.XSHG	FiberHome Telecommunication Technologies	4	100	600754.XSHG	Shanghai Jin Jiang International Hotels	3	105
601021.XSHG	Spring Airlines	4	122	601877.XSHG	Zhejiang Chint Electrics	3	105
000050.XSHE	Tianma Microelectronics A	3	122	601800.XSHG	China Communications Construction Company	5	104
601966.XSHG	Shandong Linglong Tyre	3	63	603816.XSHG	Jason Furniture	3	103
603225.XSHG	Xinfengming Group	3	82	601898.XSHG	China Coal Energy	5	102
603816.XSHG	Jason Furniture	3	103	600021.XSHG	Shanghai Electric	5	102
601877.XSHG	Zhejiang Chint Electrics	3	105	600498.XSHG	FiberHome Telecommunication Technologies	4	100