

Application and Development of Quantitative Factor Mining in Financial Markets

Tielong Ma^a

Department of Economics, University of Maryland College Park, College Park, Maryland, U.S.A.

Keywords: Quantitative Finance, Factor Mining, Financial Engineering, Market.


Abstract: Quantitative factor mining is a crucial element of modern quantitative investment, focusing on identifying indicators that can predict asset price movements through data analysis. This paper systematically examines five key factors influencing the effectiveness of quantitative factor mining: data quality, factor construction methods, market environment, model selection, and computational resources, and provides targeted optimization strategies. The research highlights that data quality serves as the foundation of factor mining, while factor construction methods and model selection directly influence the predictive accuracy of factors. Market environment shifts may cause factor degradation, and sufficient computational resources are essential for efficient factor mining. The paper explores traditional factor mining approaches, including fundamental analysis, technical analysis, statistical analysis, and macroeconomic analysis, as well as prediction methods based on major influencing factors, such as linear regression models, multi-factor models, time series models, machine learning models, and ensemble learning techniques, summarizing the strengths and limitations of each. Finally, the paper suggests future research directions, such as integrating multiple factor construction methods and leveraging advanced machine learning techniques to enhance factor mining efficiency and accuracy.

1 INTRODUCTION

Quantitative investment, as an important part of modern finance, relies on data-driven decision-making processes. With the rapid development of information technology and the globalization of financial markets, quantitative investment has gradually shifted from traditional qualitative analysis to quantitative analysis based on big data and algorithms (Hull, 2018). Quantitative factor mining is a core component of quantitative investment, aiming to extract indicators that can predict asset price movements through data analysis. These factors can be fundamental factors based on company financial data, technical factors based on market trading data, or macroeconomic factors based on macroeconomic data. The effectiveness of quantitative factor mining directly impacts the returns and risks of investment strategies, making it a critical issue in quantitative research.

In recent years, the complexity and volume of financial markets have exploded. High-frequency

trading, algorithmic trading, and the interconnectedness of global markets have made traditional investment analysis methods inadequate in coping with increasingly complex market environments. At the same time, the rapid development of big data technology, cloud computing, and artificial intelligence has provided new tools and methods for quantitative investment. As a core component of quantitative investment, quantitative factor mining not only helps investors identify potential opportunities in the market but also explains cross-sectional differences in asset returns through the construction of multi-factor models. However, as market environments change and data volumes increase, the difficulty of factor mining continues to rise. Meanwhile, the effectiveness of factor mining is influenced by multiple factors. First, data quality is the foundation of factor mining. High-quality data provides accurate information, while low-quality data may lead to inaccurate factor construction, thereby affecting the predictive power of models. Second, the choice of factor construction

^a <https://orcid.org/0009-0002-7012-7093>

methods directly impacts the predictive power of factors. Traditional factor construction methods such as fundamental analysis and technical analysis are easy to understand but may perform poorly in complex market environments. In recent years, with the development of machine learning technology, more researchers have begun to explore how to use machine learning methods to automatically generate factors to capture nonlinear relationships in the market.

This paper first systematically analyzes the key factors affecting the effectiveness of quantitative factor mining from five aspects: data quality, factor construction methods, market environment, model selection, and computational resources, and proposes corresponding optimization suggestions. Then, this paper explores traditional factor mining methods, such as fundamental analysis, technical analysis, statistical analysis, and macroeconomic analysis, as well as prediction methods based on mainstream influencing factors, such as linear regression models, multi-factor models, time series models, machine learning models, and ensemble learning methods. Finally, this paper summarizes the advantages and disadvantages of different prediction methods and proposes future research directions.

2 RESEARCH ON FACTORS AFFECTING QUANTITATIVE FACTOR MINING

Quantitative factor mining is a core component of quantitative investment, and its effectiveness is influenced by multiple factors (Tortoriello, 2012). This paper systematically analyzes the key factors affecting the effectiveness of quantitative factor mining from five aspects: data quality, factor construction methods, market environment, model selection, and computational resources, and proposes corresponding optimization suggestions. Research shows that data quality is the foundation of factor mining, factor construction methods and model selection directly impact the predictive power of factors, changes in the market environment may lead to factor failure, and computational resources are the hardware guarantee for efficient factor mining. By optimizing these influencing factors, the effectiveness of quantitative factor mining can be significantly improved, thereby providing strong support for the construction of quantitative investment strategies. Extracting indicators that can predict asset price movements through data analysis

directly impacts the returns and risks of investment strategies. As the complexity and volume of financial markets continue to increase, the role of quantitative factor mining in investment decision-making becomes increasingly important. However, the effectiveness of factor mining is influenced by multiple factors, including data quality, factor construction methods, market environment, model selection, and computational resources. This paper aims to systematically analyze these influencing factors, provide references for quantitative researchers, and explore how to improve the effectiveness of factor mining by optimizing these factors.

2.1 Data Quality

Data is the foundation of quantitative factor mining, and data quality directly affects the effectiveness of factor mining. First, the completeness of data is crucial. Missing data can lead to inaccurate factor construction, thereby affecting the predictive power of models. For example, when constructing financial factors, if some companies' financial data is missing, it may lead to poor performance of the factor in backtesting. Second, the accuracy of data is also a key issue. Incorrect data (such as outliers) can introduce noise and reduce the effectiveness of factors. For example, outliers in stock price data may lead to errors in the calculation of technical indicators. Additionally, the frequency and source of data also affect the effectiveness of factor mining. High-frequency data may contain more information but may also introduce more noise. The quality of different data sources (such as exchange data, third-party data) may vary, so choosing reliable data sources is an important step in improving data quality.

2.2 Factor Construction Methods

Factor construction methods are the core of quantitative factor mining, and their choice directly impacts the predictive power of factors. Traditional factors, such as price-to-earnings ratio (PE), price-to-book ratio (PB), etc., are easy to understand but may be outdated. Technical indicators, such as moving averages (MA), relative strength index (RSI), etc., are suitable for short-term predictions. Fundamental factors, such as revenue growth rate, net profit growth rate, etc., are suitable for long-term investments. Sentiment factors are extracted from news and social media through natural language processing and are suitable for event-driven strategies. In recent years, machine learning-generated factors have gradually

become a research hotspot (López de Prado, 2018). By using methods such as deep learning to automatically generate factors, nonlinear relationships may be captured, thereby improving the predictive power of factors. However, machine learning-generated factors also carry the risk of overfitting and thus need to be used with caution.

2.3 Market Environment

Changes in the market environment can affect the effectiveness of factors. First, market style shifts are an important factor. For example, value factors may fail in bull markets, while momentum factors may fail in volatile markets. Second, changes in the macroeconomic environment can also affect the performance of factors. For example, changes in interest rates, inflation, and other macroeconomic factors may lead to the failure of certain factors. Additionally, changes in the structure of market participants are also a factor that cannot be ignored. For example, as the proportion of institutional investors increases, the effectiveness of certain factors may decline. Therefore, in the process of factor mining, it is necessary to fully consider changes in the market environment and dynamically adjust factors and models.

2.4 Model Selection

The selection and optimization of models are crucial to the effectiveness of factor mining (Fabozzi et al., 2010). Linear models, such as multiple linear regression, are simple but may not capture nonlinear relationships. Nonlinear models, such as random forests and support vector machines, can capture complex relationships but may overfit. Deep learning models, such as neural networks, are suitable for high-dimensional data but require significant computational resources. Model ensemble methods, such as stacking and boosting, can improve the robustness of models. In practical applications, it is necessary to select appropriate models based on specific problems and data characteristics, and optimize model parameters through methods such as cross-validation to avoid overfitting.

2.5 Computational Resources

Computational resources are the hardware foundation of quantitative factor mining. First, computing power is a key factor affecting the efficiency of factor mining. High-performance computing equipment can accelerate the process of factor mining and

backtesting, thereby improving research efficiency. Second, storage capacity is also an important issue. Large-scale data requires sufficient storage space, so it is necessary to reasonably plan storage resources. Additionally, algorithm optimization is also an important means to improve computational efficiency. Efficient algorithms can reduce the consumption of computational resources, thereby achieving more efficient factor mining with limited computational resources.

2.6 Optimization Suggestions

To improve the effectiveness of quantitative factor mining, optimization can be carried out from the following aspects. First, improve data quality. Improve data completeness through data cleaning and interpolation, and choose reliable data sources. Second, diversify factor construction methods. Combine traditional factors and machine learning-generated factors to improve the predictive power of factors. Third, dynamically adjust models. Dynamically adjust factors and models based on changes in the market environment. Finally, optimize computational resources. Use distributed computing and efficient algorithms to improve computational efficiency.

2.7 Conclusion on Factor Mining

The effectiveness of quantitative factor mining is influenced by multiple factors, including data quality, factor construction methods, market environment, model selection, and computational resources. By systematically analyzing these influencing factors and taking corresponding optimization measures, the effectiveness of factor mining can be improved, thereby enhancing the performance of quantitative investment strategies. Future research can further explore how to combine multiple factor construction methods and how to use more advanced machine learning techniques to improve the effectiveness of factor mining.

3 PREDICTION METHODS

3.1 Research on Traditional Factor Mining Methods

3.1.1 Fundamental Analysis

Fundamental analysis is one of the most traditional methods of factor mining, mainly through the

analysis of company financial data and operating conditions to construct factors (Chincarini & Kim, 2006). Commonly used fundamental factors include price-to-earnings ratio (PE), price-to-book ratio (PB), dividend yield, revenue growth rate, and net profit growth rate. These factors reflect a company's profitability, growth potential, and valuation level, and are suitable for long-term investment strategies. For example, stocks with low price-to-earnings ratios and high dividend yields are usually considered value stocks with high investment value. However, the limitation of fundamental analysis is its strong reliance on financial data, which may be lagging and easily affected by accounting policies. Additionally, fundamental analysis usually assumes that the market can effectively reflect a company's intrinsic value, but in actual markets, this assumption may not hold.

3.1.2 Technical Analysis

Technical analysis is a method of predicting future price trends by analyzing historical price and trading volume data. Commonly used technical indicators include moving averages (MA), relative strength index (RSI), Bollinger Bands, and MACD (moving average convergence divergence). These indicators capture price trends, market sentiment, and overbought or oversold conditions to construct factors. For example, moving averages can be used to identify trends, while the relative strength index can be used to judge overbought or oversold conditions in the market. The advantage of technical analysis is that it is suitable for short-term trading strategies and responds quickly to market changes. However, the limitation of technical analysis is its strong reliance on historical data and its susceptibility to market noise. Additionally, technical analysis usually assumes that historical price patterns will repeat, but in actual markets, this assumption may not hold.

3.1.3 Statistical Analysis

Statistical analysis is a method of extracting factors from data through mathematical and statistical methods. Commonly used statistical methods include principal component analysis (PCA), factor analysis, and regression analysis. Principal component analysis extracts the main features of data through dimensionality reduction, factor analysis extracts latent variables to explain the correlation between observed variables, and regression analysis constructs factors by establishing the relationship between dependent and independent variables. For example, regression analysis can be used to construct multi-factor models, such as the Fama-French three-factor

model and the Carhart four-factor model (Fama & French, 1993). The advantage of statistical analysis is that it can handle high-dimensional data and extract latent patterns. However, the limitation of statistical analysis is its strong assumption about data distribution and its susceptibility to multicollinearity and overfitting. Additionally, statistical analysis methods usually assume linear relationships between variables, making it difficult to capture complex nonlinear relationships.

3.1.4 Macroeconomic Analysis

Macroeconomic analysis is a method of constructing factors by analyzing macroeconomic indicators. Commonly used macroeconomic factors include interest rates, inflation rates, GDP growth rates, and unemployment rates. These factors reflect the impact of the macroeconomic environment on asset prices and are suitable for asset allocation and long-term investment strategies. For example, a low-interest-rate environment is usually favorable for the stock market, while high inflation rates may lead to rising bond yields. The advantage of macroeconomic analysis is that it can capture the systemic impact of the macroeconomic environment on the market. However, the limitation of macroeconomic analysis is its high requirement for data timeliness, and changes in macroeconomic indicators are usually slow, making it difficult to use for short-term trading strategies. Additionally, macroeconomic analysis usually assumes a stable relationship between macroeconomic indicators and asset prices, but in actual markets, this assumption may not hold.

3.1.5 Advantages and Disadvantages of Traditional Methods

Traditional methods have the following advantages in factor mining. First, traditional methods are usually based on economic and financial theories, with clear logic and easy understanding and interpretation. Second, the computational complexity of traditional methods is low, making them suitable for large-scale data processing. Finally, many traditional factors have shown stable performance in long-term backtesting, with high reliability. However, traditional methods also have the following limitations: First, traditional methods have high requirements for data completeness and accuracy, and data quality issues may affect the effectiveness of factor mining. Second, traditional methods usually assume linear relationships between variables, making it difficult to capture complex nonlinear relationships. Finally, traditional methods may fail in

the face of market style shifts and structural changes, with poor adaptability.

3.2 Prediction Methods Based on Mainstream Influencing Factors

3.2.1 Linear Regression Models

Linear regression models are one of the most basic prediction methods, predicting by establishing a linear relationship between dependent and independent variables. In linear regression models, influencing factors are used as independent variables, and asset prices or their returns are used as dependent variables. For example, price-to-earnings ratio (PE), price-to-book ratio (PB), and other fundamental factors can be used as independent variables to predict future stock returns. The advantage of linear regression models is that they are simple, easy to understand, computationally efficient, and the results are easy to interpret. However, the limitation of linear regression models is that they assume linear relationships between variables, making it difficult to capture complex nonlinear relationships. Additionally, linear regression models are sensitive to outliers and susceptible to multicollinearity.

3.2.2 Multi-Factor Models

Multi-factor models predict asset prices by combining multiple influencing factors. Commonly used multi-factor models include the Fama-French three-factor model and the Carhart four-factor model. The Fama-French three-factor model explains differences in stock returns through market factors, size factors, and value factors, while the Carhart four-factor model adds momentum factors to the Fama-French three-factor model (Carhart, 1997). The advantage of multi-factor models is that they can capture the combined effects of multiple influencing factors and are suitable for explaining cross-sectional differences in asset returns. However, the limitation of multi-factor models is their strong reliance on factor selection and weight allocation, and they make it difficult to capture nonlinear relationships and dynamic changes.

3.2.3 Time Series Models

Time series models predict future price trends by analyzing time series data (Mengxia, 2023). Commonly used time series models include autoregressive models (AR), moving average models (MA), autoregressive moving average models (ARMA), and autoregressive integrated moving

average models (ARIMA). The advantage of time series models is that they can capture trends and periodicity in time series data and are suitable for short-term predictions. However, the limitation of time series models is their strong reliance on historical data and difficulty in handling high-dimensional data and external influencing factors. Additionally, time series models usually assume that time series are stationary, but in actual markets, this assumption may not hold.

3.2.4 Machine Learning Models

Machine learning models automatically learn the relationship between influencing factors and asset prices through training data (Jansen, 2020). Commonly used machine learning models include decision trees, random forests, support vector machines (SVM), and neural networks. The advantage of machine learning models is that they can capture complex nonlinear relationships and are suitable for handling high-dimensional data. For example, random forest models can be used to combine multiple influencing factors to predict future stock returns. However, the limitation of machine learning models is their high requirement for data volume and computational resources, and they are susceptible to overfitting (McLean & Pontiff, 2016). Additionally, the results of machine learning models are usually difficult to interpret, which may affect their application in investment decision-making.

3.2.5 Ensemble Learning Methods

Ensemble learning methods improve prediction accuracy by combining the prediction results of multiple models. Commonly used ensemble learning methods include bagging, boosting, and stacking. Bagging reduces variance by training multiple models in parallel and averaging their results, boosting reduces bias by training multiple models in series and weighting their results, and stacking constructs meta-models by combining the prediction results of multiple models. The advantage of ensemble learning methods is that they can improve prediction accuracy and model robustness, making them suitable for complex prediction tasks. However, the limitation of ensemble learning methods is their high computational complexity and strong requirement for model diversity. Additionally, the results of ensemble learning methods are usually difficult to interpret, which may affect their application in investment decision-making.

3.2.6 Comparison of Different Prediction Methods

Different prediction methods are suitable for different market environments and data types. Linear regression models are suitable for simple linear relationships, multi-factor models are suitable for explaining cross-sectional differences in asset returns, time series models are suitable for short-term predictions, machine learning models are suitable for capturing complex nonlinear relationships, and ensemble learning methods are suitable for improving prediction accuracy and model robustness. In practical applications, it is necessary to select appropriate prediction methods based on specific problems and data characteristics, and optimize model parameters through methods such as cross-validation to avoid overfitting.

4 CONCLUSIONS

The effectiveness of quantitative factor mining is influenced by various factors, including data quality, factor construction methods, market environment, model selection, and computing resources. By systematically analyzing these influencing factors and adopting appropriate optimization measures, the effectiveness of factor mining can be significantly improved, thereby providing strong support for the development of quantitative investment strategies. This study shows that data quality is the foundation of factor mining, factor construction methods and model selection directly affect the predictive power of factors, changes in the market environment may lead to factor failure, and computing resources serve as the hardware guarantee for efficient factor mining.

Traditional factor mining methods, fundamental analysis, technical analysis, statistical analysis, and macroeconomic analysis each have their own strengths and weaknesses. Fundamental analysis is suitable for long-term investment strategies but relies heavily on financial data; technical analysis is suitable for short-term trading strategies but is highly dependent on historical data; statistical analysis can handle high-dimensional data but relies heavily on assumptions about data distribution; and macroeconomic analysis can capture the systemic impact of the macro environment on the market but requires high timeliness of data.

For predictive methods based on mainstream influencing factors, linear regression models, multi-factor models, time series models, machine learning models, and ensemble learning methods each have

their applicable scenarios. Linear regression models are simple and easy to understand but struggle to capture complex nonlinear relationships; multi-factor models can capture the combined effects of multiple influencing factors but rely heavily on factor selection and weight allocation; time series models are suitable for short-term forecasting but depend heavily on historical data; machine learning models can capture complex nonlinear relationships but require large amounts of data and computational resources (Gu et al., 2020); and ensemble learning methods can improve predictive accuracy and model robustness but have high computational complexity.

Future research could further explore how to combine multiple factor construction methods and how to leverage more advanced machine learning techniques to improve the effectiveness of factor mining (Binhui, 2024). Additionally, as computing resources continue to advance, how to use them more efficiently for factor mining is also a worthwhile research direction. By continuously optimizing each aspect of factor mining, the performance of quantitative investment strategies will be further enhanced.

REFERENCES

- Binhui, W. 2024. Research on LSTM stock price prediction model based on financial comments and functional data. CNKI.
- Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57–82.
- Chincarini, L., & Kim, D. 2006. *Quantitative equity portfolio management: An active approach to portfolio construction and management*. McGraw-Hill Education.
- Fama, E. F., & French, K. R. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.
- Gu, S., Kelly, B., & Xiu, D. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.
- Hull, J. C. 2018. *Financial data analysis* (2nd ed.). Pearson Education.
- Jansen, S. 2020. *Machine learning for algorithmic trading: Predictive models to extract insights from data and build trading strategies* (2nd ed.). O'Reilly Media.
- López de Prado, M. 2018. *Advances in financial machine learning*. John Wiley & Sons.
- McLean, R. D., & Pontiff, J. 2016. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1), 5–32.
- Mengxia, L. 2023. *Pattern mining and correlation analysis methods for financial time series*. CNKI.

Tortoriello, R. 2012. Quantitative strategies for achieving alpha: The standard and poor's approach to testing your investment choices. McGraw-Hill Education.

