Stock Price Trend Prediction: Based on Dimension Reduction Techniques and Cluster Analysis of Multi-Dimensional Financial Data

Qijia He[1], Ruiwei Liang[1], Yuru Feng[1], and Xinyi Zhu[1]

[1]Department of Statistics and Data Science, Email:12111211@mail.sustech.edu.cn
[1]Department of Statistics and Data Science, Email:12111219@mail.sustech.edu.cn
[1]Department of Statistics and Data Science, Email:12112243@mail.sustech.edu.cn
[1]Department of Statistics and Data Science, Email:12112944@mail.sustech.edu.cn

June 15, 2024

Abstract

This report explores the utilization of factor analysis and clustering methods to predict stock price trends by analyzing financial data from listed companies. By applying advanced statistical techniques to reduce dimensionality and classify companies based on their financial health, the study aims to provide investors with robust tools for making informed decisions. Data from Baostock and various financial indicators from the last quarter of 2023 were analyzed to identify underlying relationships and group companies into categories reflecting similar financial characteristics. The findings suggest that specific financial metrics can predict stock performance, aiding investment strategies.

Keywords
Factor Analysis, K-Means, PCA, Multivariate Regression

1 Introduction

1.1 Background Information

In the ever-evolving landscape of financial markets, investors constantly seek reliable and sophisticated methods to make informed decisions. Financial statements, which encompass a company's income statement, balance sheet, and cash flow statement, provide crucial insights into a company's financial health and performance. However, interpreting these statements to make sound investment decisions can be complex and challenging. This necessitates the development of robust analytical tools to extract meaningful patterns and provide clear investment recommendations.

Factor analysis and clustering methods have emerged as powerful techniques in financial data analysis. Factor analysis helps in identifying underlying relationships among various financial indicators, reducing dimensionality, and highlighting the most influential factors affecting a company's performance. Clustering methods, on the other hand, allow for the grouping of companies into distinct categories based on their financial characteristics, facilitating easier comparison and investment decision-making.

Previous research has extensively explored the use of these methods in various domains. Studies have shown that factor analysis can effectively reduce the complexity of financial data, while clustering can reveal hidden patterns and groupings that are not immediately apparent.

1.2 Research Objects

The primary objective of this study is to develop a robust framework for analyzing and classifying companies based on their financial health, which will serve as a reliable tool for investors aiming to make informed decisions. This framework will employ advanced statistical methods, particularly factor analysis and clustering techniques, to systematically evaluate and categorize companies.

- Application of Factor Analysis and Clustering

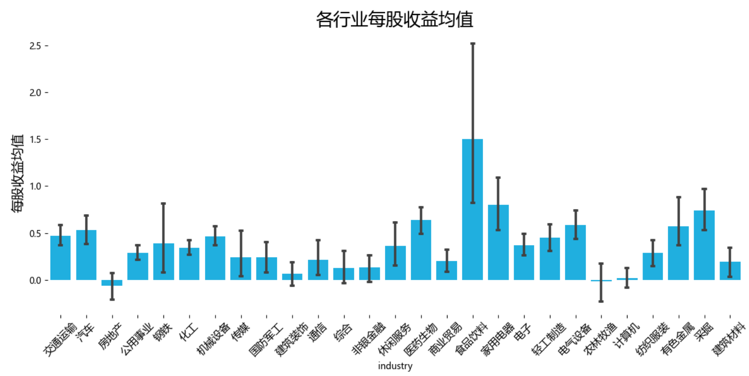


Figure 1: Industry Comparison: Average Earnings per Share

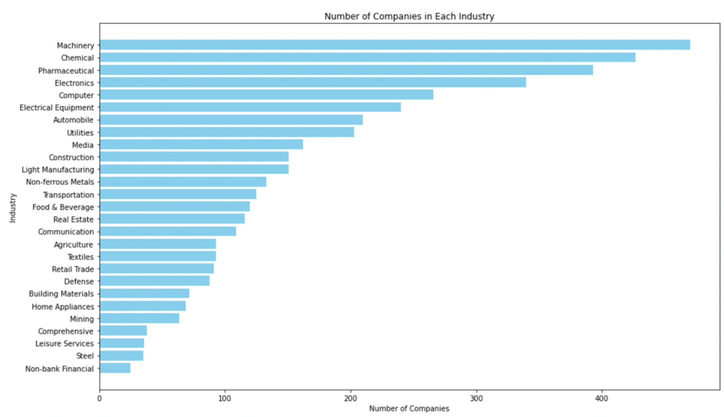


Figure 2: Number of Companies in Each Industry

Methods for Company Rating Classification

- Provision of Investment Advice through Analysis of Company Financial Statements

By achieving these goals, the study aims to bridge the gap between complex financial data and actionable investment strategies, thereby empowering investors with sophisticated, data-driven tools for better decision-making.

1.3 Data Source

The data for this project is sourced from www.baostock.com, a comprehensive platform that offers financial data services. Through the Python API, Treasure Data provides users with convenient access to data, enabling them to retrieve historical and real-time financial market information such as stocks, funds, indexes, futures, and more.

We crawled data from this website, obtaining the operation status of 5156 listed companies in China for the fourth quarter of 2023, with a total of 6 sets of data, approximately 40 50 columns in total. The 6 sets of data include:

- Quarterly earnings capability
- Quarterly operating capability
- Quarterly growing capability
- Quarterly debt repayment capability
- Quarterly cash flow
- Quarterly Dupont index

1.4 Data Description

The dataset used in this study comprises various financial and market indicators across multiple industries. Each column in the dataset provides specific information about the companies and their performance. A detailed description of the dataset columns is provided in Table 6.

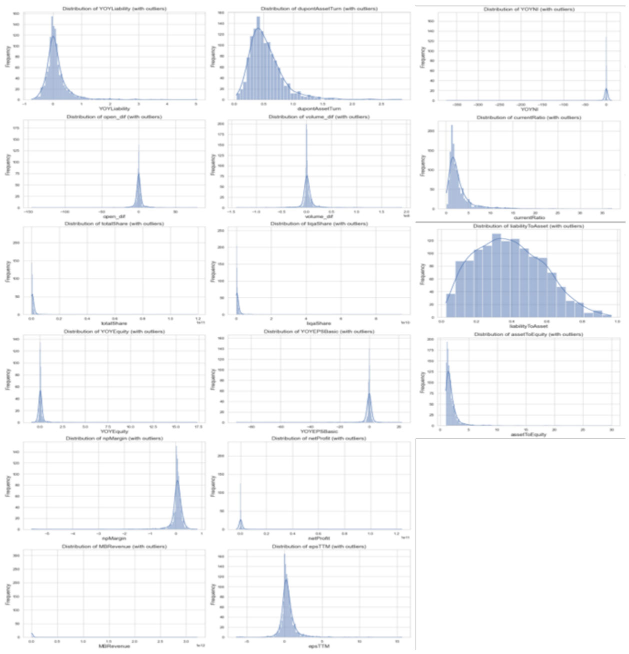


Figure 3: Data Distribution With Outliers

2 Exploratory Data Analysis

As Figure1 shows, the prices and returns of stocks show obvious fluctuation trends among different industries.

We selected three industries with the highest number of listed companies as representative examples from the entire spectrum of industries. The sorted results are illustrated in Figure 2. Therefore, we have chosen **Machinery**, **Chemical**, and **Pharmaceutical** sectors for our analysis.

2.1 Outlier Detection

Outliers are defined using the Interquartile Range (IQR) method, where the original data is divided into two parts: normal data and outliers. It's important to note that outliers hold their own significance, warranting special attention in the subsequent analysis section. In this section, we present the distribution of the three industries we selected earlier, both before and after filtering out the outliers. The distributions are illustrated in Figure 3 and Figure 4, respectively.

2.2 Variable Selection

2.2.1 Correlation Analysis

To explore the relationships among variables, we computed the correlation coefficients between each pair of variables. The correlation coefficients quantify the strength and direction of linear relationships between variables. We used the heatmap function from the seaborn library to visually represent these correlations in Figure5.

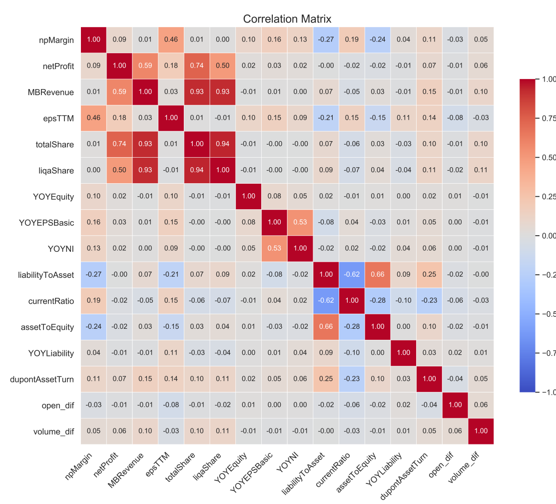


Figure 5: Correlation Matrix of Variables

After filtering out column pairs with correlation coefficients greater than 0.6, we investigated whether these pairs exhibit linear relationships. To do so, we performed linear regression on each pair and plotted the fitted curves in Figure6.

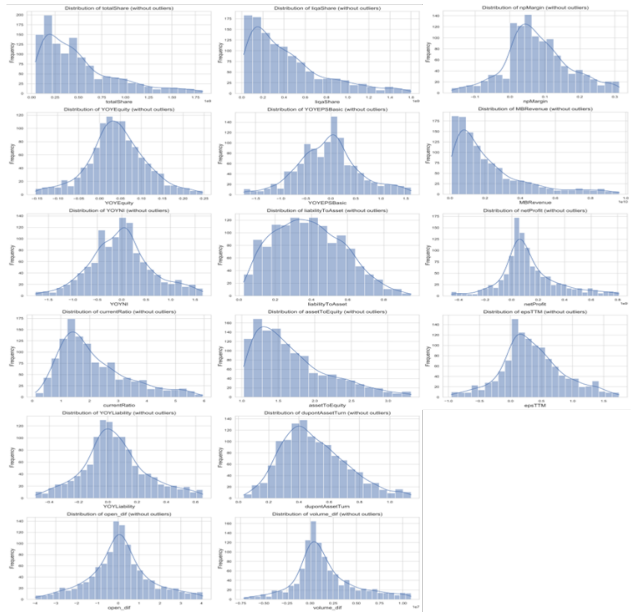


Figure 4: Data Distribution Without Outliers

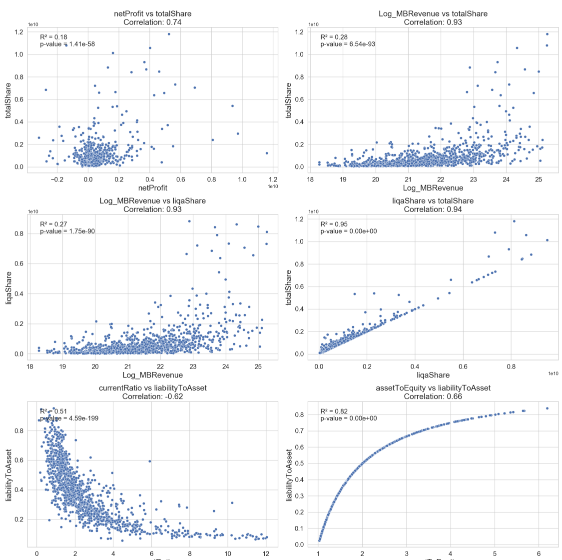


Figure 6: Linearity Check of Strong Correlated Variables

The analysis of these plots shows varying degrees of linear and non-linear relationships among the financial metrics.

Particularly notable are the very strong correlations between liqShare and totalShare ($R^2 = 0.95$, p-value < 0.05), and the significant impact of MBRevenue(log trasformed) on both liqShare and totalShare, as indicated by their high R^2 values and very low p-values. The inverse relationship between currentRatio and liabilityToAsset ($R^2 = 0.51$, p-value $<$ 0.05), along with the strong, albeit non-linear, relationship between assetToEquity and liabilityToAsset ($R^2 = 0.82$, p-value < 0.05), suggests complex dynamics in financial health indicators that are crucial for financial analysis.

Each of these relationships, proven statistically significant through their p-values, highlights specific areas where financial metrics interact significantly, influencing each other in predictable and important ways that can be leveraged for financial planning and risk assessment.

2.2.2 Comparisons with One-Way ANOVA

Next, we investigated which variables exhibit significant differences in means among the three industries. To accomplish this, we conducted one-way ANOVA test.

ANOVA allows us to assess whether the means of three or more groups differ significantly. The results, presented in Table 1, reveal numerous financial indicators with significant differences across industries.

Consequently, separate analysis and predictions are warranted based on industry distinctions.

3 Data Analysis and Results

The same analysis process and methods can be conducted on three industries, here we take chemicals as an example to illustrate the detailed process. The

Table 1: Comparison of Financial Metrics Across Industries

Feature	Chemical Mean	Pharmaceutical Mean	Machinery Mean	F-Statistic	p-Value	Significant
epsTTM	0.341	0.638	0.463	7.272	0.001***	Yes
liabilityToAsset	0.412	0.322	0.418	33.222	0.0***	Yes
currentRatio	2.429	3.632	2.619	18.663	0.0***	Yes
assetToEquity	2.093	1.657	2.027	10.602	0.0***	Yes
YOYLiability	0.174	0.06	0.147	7.693	0.0***	Yes
dupontAssetTurn	0.651	0.504	0.485	40.073	0.0***	Yes
open_dif	-0.521	1.089	0.049	5.791	0.003**	Yes
npMargin	0.028	0.026	0.054	1.114	0.329	No
netProfit	7.96E+08	4.50E+08	2.53E+08	1.976	0.139	No
MBRevenue	1.68E+10	6.20E+10	4.00E+10	2.412	0.09	No
totalShare	1.28E+09	7.87E+08	7.01E+08	2.871	0.057	No
liqaShare	1.03E+09	7.00E+08	5.88E+08	2.68	0.069	No
YOYEquity	0.059	0.11	0.088	0.904	0.405	No
YOYEPSBasic	-0.77	-0.296	-0.34	1.291	0.275	No
YOYNI	-1.047	-1.553	-0.373	0.947	0.388	No
volume_dif	3.21E+06	2.25E+06	1.95E+06	0.799	0.45	No

Significance levels: ***p<0.001, **p<0.01, *p<0.05

results of the other two industries can be seen in Appendix.

3.1 Factor Analysis

Regarding determining the factor dimension, we determine the appropriate factor dimension by considering the proportion of the factor in interpreting the total variance of the data set, as shown in the figure7. [1]

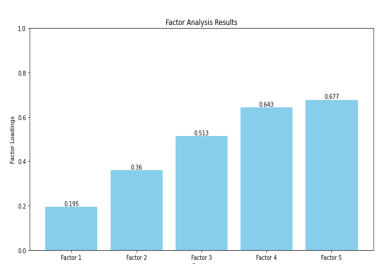


Figure 7: Factor Analysis Results

When the number of factors is 5, the cumulative percentage of total explained variance reaches 0.667, so it can be considered that the selection of factors in this dimension can ensure the good explanatory ability of the model.

Thus we can get the heatmap of factor loading matrix, which displays the factor loadings of various financial metrics against five identified factors. Each cell's color intensity and sign (positive or negative) indicate the strength and direction of the association between the variable and the factor.

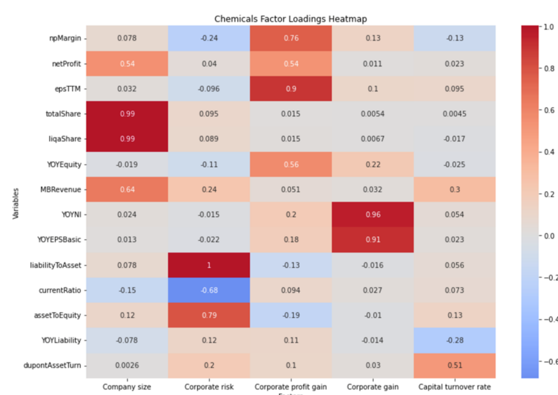


Figure 8: Factor Loading Heatmap

- Factor 1: **Company Size**

  Strong Positive Loadings: TotalShare and LiqaShare both have high loadings (0.99) on this factor, suggesting that Factor1 might represent the total shares and floating shares of a company. Moderate Positive Loading: MBRevenue with a loading of 0.64 also supports the interpretation that this factor relates to main business income of a company.

- Factor 2: **Corporate Risk**

Strong positive loadings for liabilityToAsset (1) and assetToEquity (0.79) suggest that this factor includes considerations of a company's leverage and capital structure. Meanwhile, the strong negative loading for currentRatio (-0.68) indicates that this factor inversely relates to Factor 2, potentially supporting the interpretation that this factor relates to a company's risk.

- Factor 3: **Corporate Profit Gain**

  Strong Positive Loadings: epsTTM(0.9), npMargin (0.76) and netProfit (0.54) strongly load on this factor, pointing to profitability and margins as defining elements. This factor likely captures aspects related to financial performance and profit efficiency.

- Factor 4: **Corporate Year-on-year growth**

  Strong Positive Loadings: YOYNI (0.96) and YOYEPSBasic (0.91) indicate that this factor represents year-over-year gains, reflecting growth in net income and earnings per share, which are critical for assessing year-to-year business performance.

- Factor 5: **Capital Turnover Rate**

  Strong Positive Loading: dupontAssetTurn (0.51) on this factor suggests it relates to how efficiently a company utilizes its assets to generate sales, indicative of operational efficiency.

3.2 K-Means Clustering

In this section, we explore the process of identifying the optimal number of clusters for k-means clustering[2] through the evaluation of the silhouette coefficient[3]. The silhouette score is a metric that assesses how similar an object is to its own cluster compared to other clusters. Essentially, a higher silhouette score suggests that clusters are well-defined and distinct from each other.

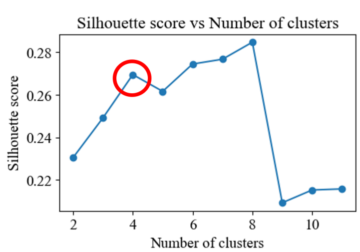


Figure 9: Silhouette Coefficient of Chemical Industry

Figure 10 presents the silhouette coefficient curve, which is instrumental in determining the optimal cluster count. From the chart, it is evident that the highest silhouette score is achieved when the number of clusters is eight. This peak suggests that eight clusters provide the most distinct and well-separated grouping according to the dataset's inherent structures.

However, we chose k=4 as our clustering number because the silhouette score at four clusters, while not the peak, remains comparatively high. This indicates that the clustering at this level still maintains satisfactory quality. Unlike the sharp decline observed after 8 clusters, the stability and performance at k=4 are comparatively better and provide a more meaningful interpretation.

Figure10 is the scatter matrix diagram of factor analysis and clustering results. The two populations numbered 0 and 1 have been able to show a relatively clear difference in the scatter distribution of the five-factor combination. These two clusters show different aggregation characteristics in the factor space, which indicates that they can better classify the data into unique categories according to the different values of these five factors.

From Table 2, we observe the K-Means cluster centers for five factors.

- cluster1(ID = 0): Represent a stable, low-risk investment option primarily due to their effective management and efficient capital utilization, despite their smaller size and modest financial gains.

- Cluster 2 (ID = 1): Companies in this cluster are characterized by their smaller size and higher risk but distinguish themselves with strong profit performance despite challenges in capital efficiency.

- Cluster 3 (ID = 2): Represents very large, established businesses that, despite their significant size and ability to generate profits, face challenges in terms of stock market performance and capital efficiency.

- Cluster 4 (ID = 3): Companies in this cluster are of approximately average size but are currently facing a spectrum of challenges that place them at moderate to high risk.

- cluster4(ID = 3): Companies are of approximately average size but are currently facing a spectrum of challenges that place them at moderate to high risk.

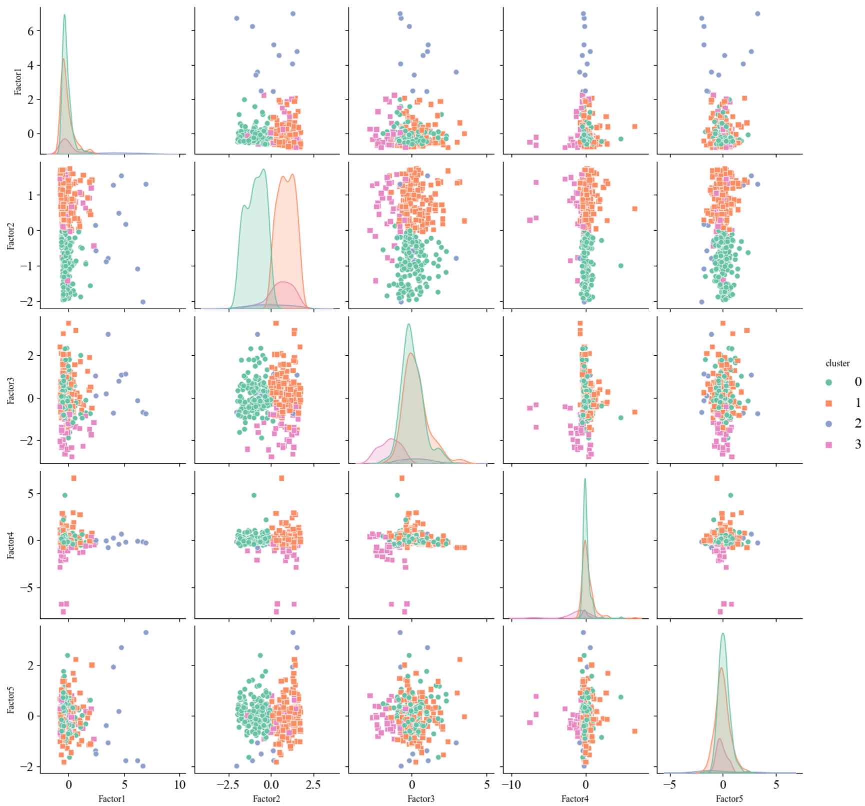


Figure 10: Scatter matrix diagram of factor analysis and clustering results

Table 2: K-Means Cluster Centers of Five Factors

Cluster ID	Factor 1 Company Size	Factor 2 Corporate Risk	Factor 3 Profit Gain	Factor 4 Annual Growth	Factor 5 Turnover Rate
0	-0.1640	-0.9310	0.0399	0.7000	0.8085
1	-0.1535	0.8803	0.2909	0.2439	-0.0473
2	4.5766	-0.0405	0.4462	-0.1430	-0.1642
3	0.0181	0.5834	-1.4768	-1.2507	-0.1196

3.3 Multiple Linear Regression

In this part, we utilized Ridge regression to construct a multiple linear regression model. Cross-validation was employed to determine the optimal regularization parameter (alpha). The selected alpha value was then utilized to train the final Ridge regression model, which was subsequently evaluated on the test set to assess its predictive performance. The model's ability to generalize to new data was assessed using the mean squared error (MSE). Here we get the best alpha as 300, and MSE as 780.33 for the current instance.

The final regression equation obtained is as follows:

$$\begin{aligned} y = &-3.914x_1 - 5.325x_2 + 11.737x_3 + 4.716 \times 10^{-10}x_4 \\ &- 3.664 \times 10^{-9}x_5 - 0.432x_6 + 9.572 \times 10^{-11}x_7 \\ &- 0.032x_8 - 0.436x_9 + 3.488x_{10} + 0.866x_{11} \\ &- 0.753x_{12} + 7.291x_{13} - 4.609x_{14} + 13.390 \end{aligned}$$

A comparison between actual and predicted values generated by our regression model, as illustrated in Figure 11.

In the visualization, points where the predicted values exceed the actual values are distinctly marked in red. This color coding facilitates easy identification of overestimations by the model. Conversely, points where the predicted values are less than or equal to

Table 3: Potential Stocks Information in Chemical Industry

Stock Code	Company Name	Opening Price X	Opening Price Y
sh.600096	云天化	15.62	18.85
sh.600346	恒力石化	13.18	14.06
sh.601163	三角轮胎	14.31	16.28
sh.603299	苏盐井神	8.53	8.85
sh.603599	广信股份	14.50	14.62
sh.603639	海利尔	15.73	14.75
sz.000707	双环科技	8.01	7.01
sz.000822	山东海化	6.84	6.64
sz.002360	同德化工	7.25	6.25
sz.002986	宇新股份	15.61	13.85

the actual values are colored green. This color scheme allows us to quickly assess the accuracy of the model's predictions relative to real-world outcomes.

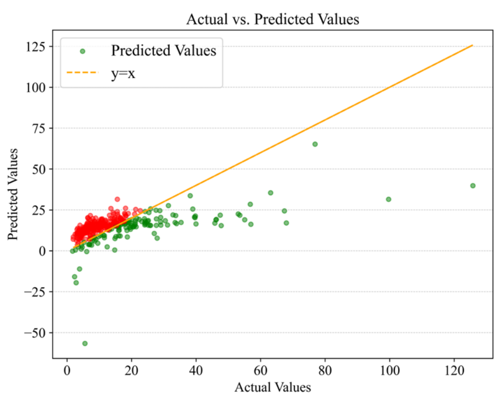


Figure 11: Actual vs Prediction Values

3.4 Potential Stock Comparison

To leverage the insights obtained from previous analyses, including factor analysis, k-means clustering, and multiple linear regression, we introduce the definition of **"potential stock"** to support our stock market predictions.

Here comes a novel definition: we define a **"potential stock"** as a stock that exhibits characteristics suggesting it may experience significant growth or positive performance in the future.

The criteria for identifying potential stocks are as follows:

1. The predicted price of the stock is higher than the actual price
2. The return of the stock is positive overall (i.e. higher than the red line, can only be screened, not ranked)
3. Rank = Stock Yield * residual (residual between yield and red line)

Thus potential stock candidates can be identified based on the defined criteria. The detailed information for these candidates is presented in Table 3.

The average gains for all stocks in the chemical industry stand at -1.849, while for the recommended stocks within this sector, it rises to +0.158. This significant discrepancy underscores the reliability of our definition for potential stocks.

The same analysis process was applied to the other two industries. The results are presented in the Appendix under the section "Data Analysis Complements."

After gathering all the information, we calculated the average gains for all stocks in a specific industry and compared them with the recommended stocks within the same sector. Figure 12 shows the comparison of average returns for potential stocks versus the overall industry.

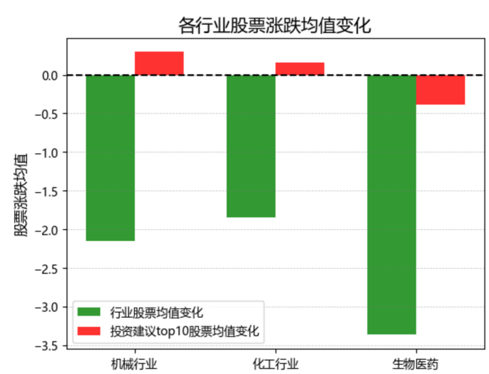


Figure 12: Comparison of Average Returns: Potential Stocks vs. Overall Industry

From this figure, we can conclude that the potential stocks we selected significantly outperform the average performance of other companies in their respective industries.

4 Outlier Exploration

The above analysis is applicable to the type of company in general, but for some companies with less common operating conditions (such as industry giants), such analysis may not be used. In this section, we conduct a separate analysis of the outlier companies identified in the previous article and draw some valuable conclusions.

4.1 Principle Component Analysis

After removing the outliers, the data approximately follows a multivariate normal distribution; Therefore, as previously mentioned, we employ factor analysis for data reduction to identify the main components.[4] In the case of outlier data, since there is no clear distribution pattern, we use PCA for the analysis. After filtering out the companies with outliers, we first use the elbow method to determine the optimal number of clusters for PCA.

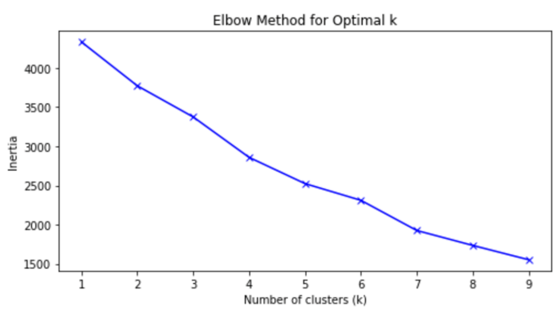


Figure 13: Elbow Method

From Figure 13 we can find that the decline rate of inertia value at 4 is significantly slower, so we choose 4 number of clusters when implementing PCA.

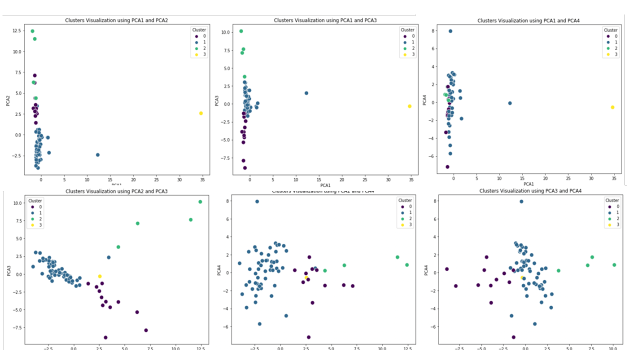


Figure 14: Pair-wise PCA Scatter Plot

Figure 14 displays multiple scatter plots that visualize data clusters using combinations of principal components (PCs) obtained through Principal Component Analysis (PCA). In all plots, each color represents a different cluster (Cluster 0, Cluster 1, Cluster 2, Cluster 3), as indicated in the legend. These visualizations are useful for examining how different combinations of principal components can affect the understanding of data structure and clustering.

To elucidate the underlying structure of the dataset, an extraction of the constituent variables for the four principal components obtained from the Principal Component Analysis (PCA) was performed. The methodology involved isolating and identifying variables within each principal component where the absolute weights exceeded a threshold of 0.3. This threshold was selected to ensure that only the most influential variables, those with substantial contributions to the direction and magnitude of the components, were considered. A detailed table of these variables is constructed to provide a clear view of the components' makeup, thereby offering insights into the dominant dimensions of variance in the data. Table 4 shows variables for each components.

Table 4: Principal Component Analysis Weights for Selected Variables

Component	Selected Columns	Weights
PCA1 Components		
	netprofit	0.4050
	totalShare	0.5318
	liqaShare	0.4995
	MBRevenue	0.5230
PCA2 Components		
	npMargin	-0.4003
	YOYNI	-0.4935
	YOEPSBasic	-0.4449
	assetToEquity	0.3020
PCA3 Components		
	npMargin	0.3342
	YOYNI	-0.4831
	YOEPSBasic	-0.4865
	assetToEquity	-0.4998
PCA4 Components		
	currentRatio	-0.3103
	YOLiability	0.3444
	open_dif	0.7356
	volume_dif	0.3443

The interpretation for each component:
PCA1: This component is heavily weighted by

'Net Profit,' 'Total Equity,' 'Circulating Equity,' and 'Main Operating Income,' which collectively represent the overall volume of the companies. These variables are indicative of the firms' fundamental economic scale and operational scope, suggesting that PCA1 captures aspects related to the size and core financial health of the businesses.

PCA2: Composed of variables such as 'Net Sales Profit Margin' (-), 'Net Profit Year-on-Year Growth Rate' (-), 'Basic Earnings Per Share Year-on-Year Growth Rate' (-), and 'Equity Multiplier' (+). This component signifies challenges in management efficiency, an increasing financial burden, a decline in revenue, and extensive reliance on financing. PCA2, therefore, reflects underlying issues in operational management and financial strategies that may hinder sustainable growth.

PCA3: This component includes 'Net Sales Profit Margin' (+), 'Net Profit Year-on-Year Growth Rate' (-), 'Basic Earnings Per Share Year-on-Year Growth Rate' (-), and 'Equity Multiplier' (-). It points to a paradoxical scenario where companies are profitable yet experiencing a decline in income for the quarter, alongside limited financing options. PCA3 highlights firms that, despite being profitable, face challenges in revenue growth and financial leverage, indicating a cautious or conservative approach in financial structuring.

PCA4: Characterized by 'Current Ratio' (-), 'Total Debt Year-on-Year Growth Rate' (+), 'Stock Price Increase' (+), and 'Trading Volume Increase' (+). This component suggests that companies operating with significant levels of debt are likely to see short-term growth, possibly driven by speculative trading or temporary market conditions. PCA4 can be seen as reflecting a scenario where higher debt levels are associated with aggressive growth strategies, which may boost stock market performance in the short run.

4.2 Identification strategy

Based on the interpretation of each component of PCA, we propose the following identification strategy:

1. Companies with PCA1 $>$ 10 are categorized as large-scale enterprises.

2. Companies meeting the criteria PCA3 $>$ 0.5, PCA2 $<$ -1, and PCA4 $>$ -0.5 are considered as other well-performing companies.

This strategy provides a clear guideline for categorizing companies based on their PCA component values, aiding in the analysis and interpretation of their characteristics and performance. Implementing above theory to chemical industry, we can get results shown in Table 5.

Table 5: PCA Metric In Chemical Industry

code name	Open Price Difference
PCA1 $>$ 10	
中国石化	0.81
中国海油	8.31
PCA3 $>$ 0.5, PCA2 $<$ -1, PCA4 $>$ -0.5	
中国海油	8.31
藏格矿业	5.80
大庆华科	-1.83
道明光学	-1.55
科思股份	16.19

This analytical method can also be applied to the other two industries, with results detailed in the appendix.

5 Summary

The report details a comprehensive analysis of financial data from 5156 listed companies in China, focusing on their performance in the fourth quarter of 2023. The study employs factor analysis to reduce the complexity of financial data and clustering methods to group companies based on similar financial characteristics. This analytical approach is intended to assist investors in understanding complex financial data and making informed investment decisions.

Data Description and Source: Financial data for the analysis was sourced from Baostock, encompassing various financial and market indicators across multiple industries, with detailed performance metrics provided for each company.

Exploratory Data Analysis: Initial explorations revealed significant fluctuations in stock prices and returns across different industries. Outlier detection and variable selection through correlation analysis were performed to refine the data for subsequent analyses.

Factor Analysis: This technique helped identify key factors affecting financial performance, which included company size, corporate risk, profit gain, annual growth, and capital turnover rate. The analysis demonstrated how these factors could explain the variance in financial data effectively.

Clustering (K-Means): The optimal number of clusters for the data was determined using the silhouette coefficient method. Clusters were then analyzed to show how companies could be grouped based on financial health and performance indicators.

Multiple Linear Regression: This model was applied to predict stock price trends, using ridge regression to optimize the fit and prevent overfitting. The model's predictive accuracy was evaluated based on its performance against actual data.

Potential Stock Identification: Using the results from the regression model and cluster analysis, potential stocks for investment were identified. These stocks were predicted to outperform in their respective industries based on the analysis.

PCA and Outlier Analysis: Principal Component Analysis (PCA) was used to further understand the data structure, particularly for identifying and analyzing outliers.

The findings illustrate the practical application of statistical methods in financial analysis, offering insights into the financial health and potential performance of stocks in different industries. By leveraging factor analysis, clustering, and regression models, the study provides a robust framework for investors aiming to make data-driven decisions in the stock market.

References

[1] L.R. Fabrigar and D.T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, New York, 2011.

[2] J. MacQueen. *Some Methods for Classification and Analysis of Multivariate Observations*, volume 1. University of California Press, Berkeley, Calif., 1967.

[3] T.M. Kodinariya and P.R. Makwana. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95, 2013.

[4] I.T. Jolliffe. Principal component analysis. *Springer Series in Statistics*, 1986.

6 Appendix

6.1 Data Description

The description of our data is shown in Table 6.

6.2 Data Analysis Complements

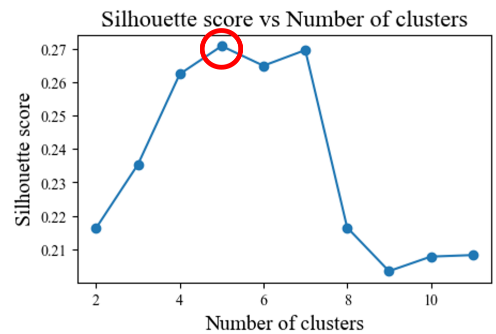


Figure 15: Silhouette Coefficient of Machinery Industry

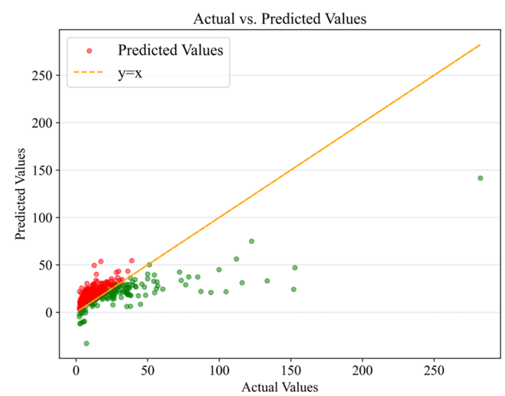


Figure 16: Actual vs Prediction Values of Machinery Industry

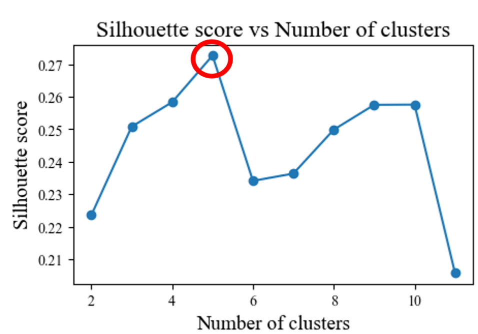


Figure 19: Silhouette Coefficient of Pharmaceutical Industry

Column	Description
code	Stock code, used to uniquely identify each stock.
npMargin	Net profit margin, representing the percentage of net profit to total revenue.
netProfit	Net profit, the company's after-tax profit.
MBRevenue	Main business revenue, revenue from the company's core business.
epsTTM	Earnings per share (trailing twelve months), measuring net profit per share for shareholders.
totalShare	Total shares, the total number of shares issued by the company.
liqaShare	Circulating shares, the number of shares that can be traded on the market.
YOYEquity	Year-over-year equity growth rate, the percentage increase in equity compared to the previous year.
YOYEPSBasic	Year-over-year basic earnings per share growth rate, the percentage increase in basic EPS compared to the previous year.
YOYNI	Year-over-year net income growth rate, the percentage increase in net income compared to the previous year.
liabilityToAsset	Debt-to-asset ratio, the percentage of total liabilities to total assets.
currentRatio	Current ratio, the ratio of current assets to current liabilities.
assetToEquity	Asset-to-equity ratio, the ratio of total assets to shareholder's equity.
YOYLiability	Year-over-year liability growth rate, the percentage increase in liabilities compared to the previous year.
dupontAssetTurn	Dupont asset turnover, used to assess the efficiency of the company in generating revenue from assets.
code_name	Company name.
industry	Industry, the category of industry the company belongs to.
open	Opening price, the stock's price at the beginning of the trading day.
volume	Trading volume, the number of shares traded during the trading day.

Table 6: Description of the dataset columns

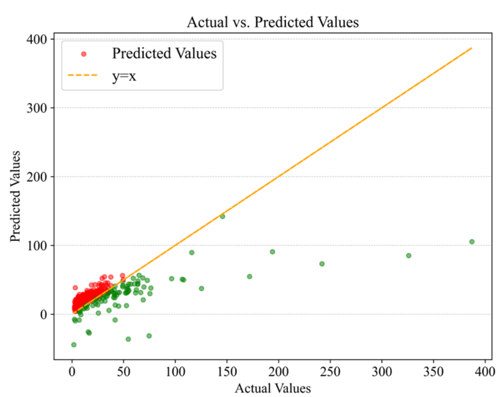


Figure 20: Actual vs Prediction Values of Pharmaceutical Industry

From Table 8, we observe that the PCA metric in the pharmaceutical industry has a significantly negative value. Upon further investigation, we identified that this company is experiencing financial issues, as depicted in Figures 21 and 22. Although our data covers only one quarter, the selected training area shows a temporary increase, while the overall trend in the figures indicates a decline.

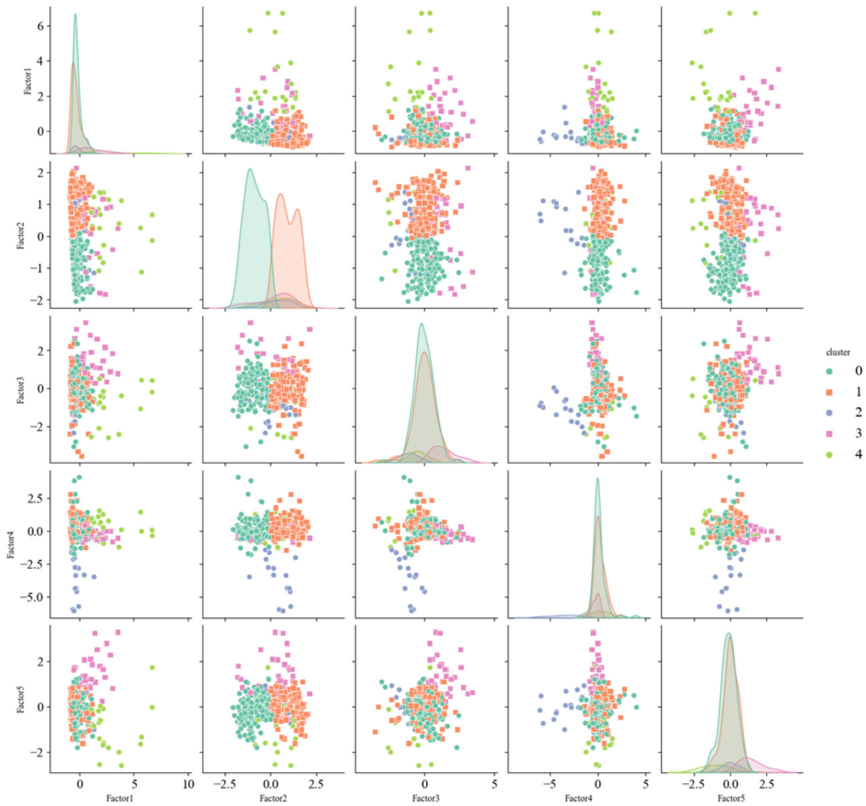


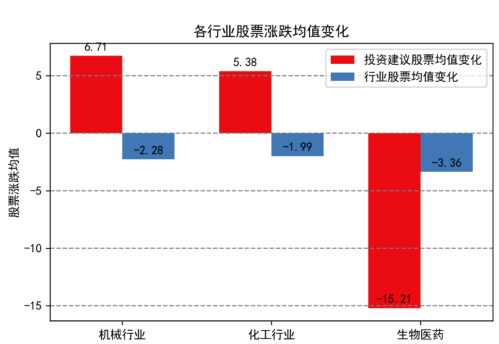
Figure 17: Scatter matrix diagram of factor analysis and clustering results of Machinery Industry

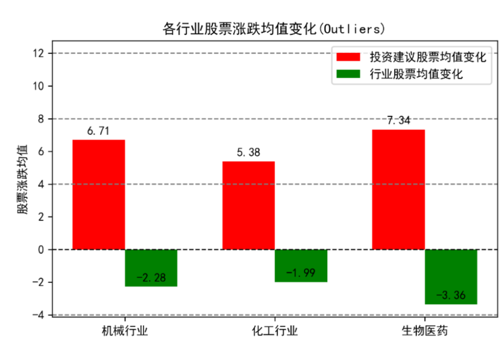
Table 7: PCA Metric In Machinery Industry

code name	Open Price Difference
PCA1 > 10	
中国中车	1.57
徐工机械	0.94
PCA3 > 0.5, PCA2 < -1, PCA4 > -0.5	
浙江鼎力	6.12
石头科技	62.63
铂力特	-34.8
高测股份	-7.99
浩洋股份	7.60

Table 8: PCA Metric In Pharmaceutical Industry

code name	Open Price Difference
PCA1 > 10	
*ST太安	-1.31
PCA3 > 0.5, PCA2 < -1, PCA4 > -0.5	
国药股份	4.50
奕瑞科技	-105.43
惠泰医疗	40.81
迈瑞医疗	-9.14
药易购	-6.83





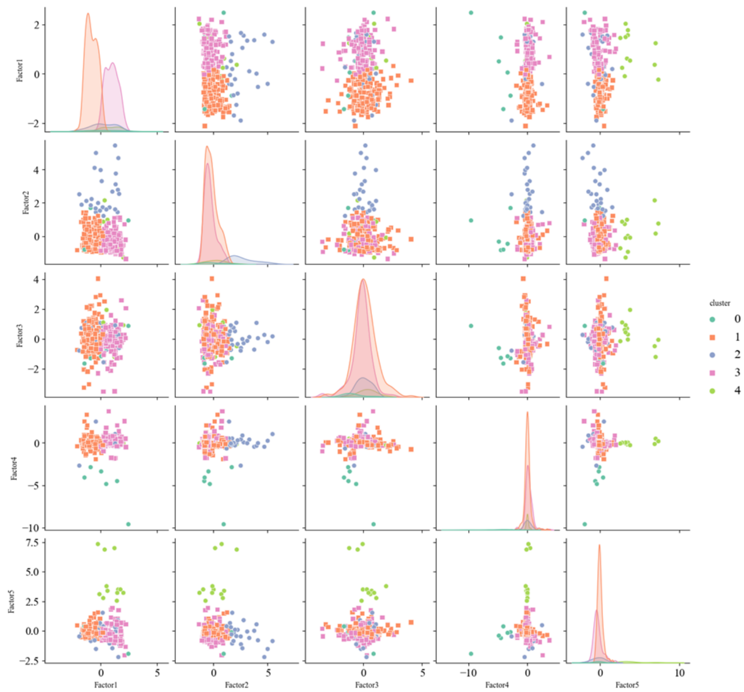


Figure 18: Scatter matrix diagram of factor analysis and clustering results of Pharmaceutical Industry

Table 9: Potential Stocks Information in Machinery Industry

Stock Code	Company Name	Opening Prize X	Opening Price Y
sh.600499	科达制造	10.51	10.55
sh.600835	上海机电	11.88	12.10
sh.603298	杭叉集团	24.73	27.60
sh.603611	诺丽股份	18.93	19.22
sh.688057	金达莱	14.22	11.51
sh.688556	高测股份	38.94	30.95
sz.002353	杰瑞股份	28.12	30.25
sz.002564	*ST天沃	3.92	3.89
sz.002595	豪迈科技	29.78	36.10
sz.002884	凌霄泵业	17.36	19.18

Table 10: Potential Stocks Information in Pharmaceuticals Industry

Stock Code	Company Name	Opening Prize X	Opening Price Y
sh.600211	西藏药业	48.95	44.13
sh.600511	国药股份	28.50	33.00
sh.600566	济川药业	31.55	37.43
sh.603368	柳药集团	18.90	21.20
sz.000028	国药一致	29.02	30.74
sz.000661	长春高新	145.80	120.43
sz.000915	众望达	29.21	35.41
sz.002393	力生制药	26.66	24.80
sz.002432	九安医疗	37.65	44.05
sz.002737	赛诺医疗	26.12	27.35

奕瑞科技(688301):上海市方达（深圳）律师事务所关于上海奕瑞光电子科技股份有限公司2023年年度股东大会的法律意见书

原标题:奕瑞科技:上海市方达（深圳）律师事务所关于上海奕瑞光电子科技股份有限公司2023年年度股东大会的法律意见书上海市方达（深圳）律师事务所（以下简称"本所"）是具有中华人民共和国法律执业资格的律师...

法律意见书 股东大会 公司章程 中国法律 CFi.CN 中财网 | 2024-05-24

注意！奕瑞科技将于5月24日召开股东大会

每经AI快讯，奕瑞科技（SH 688301，收盘价：206.98元）4月29日发布公告称，2024年5月24日14点30分，公司将在上海浦东新区金海路1000号45幢上海奕瑞光电子科技股份有限公司一楼...

科技创新 探测器 AI 浦东新区 议案 | 2024-04-29

奕瑞科技申请图像处理专利，提高了图像压缩处理的效率

金融界2024年4月27日消息，据国家知识产权局公告，上海奕瑞光电子科技股份有限公司申请一项名为"图像处理方法及系统、解压缩方法、设备及介质"，公开号CN117939172A，申请日期为2023年12...

专利申请 专利 知识产权 图像处理 搜狐 | 2024-04-27

奕瑞科技公布2023年年度权益分配预案 拟10转4股派20元

同花顺（300033）财经讯 奕瑞科技于4月20日发布公告，公司2023年年度权益分配预案内容如下：以总股本10182.50万股为基数，向全体股东每10股派发现金红利人民币20.00元，合计派发现金红...

科技创新 探测器 同花顺 股东 同花顺iFinD | 2024-04-19

上海奕瑞光电子科技股份有限公司关于子公司涉及诉讼的公告

● 是否会对上市公司损益产生负面影响：本次诉讼的结果尚存在不确定性，对公司本期及期后损益的影响具有不确定性，最终影响以法院判决为准上海奕瑞光电子科技股份有限公司（以下简称"公司"）全资子公司iRay...

探测器 首尔 韩国 hs编码 搜狐 | 2024-04-09

魅丽纬叶完成数亿元D轮融资，中平资本领投

Figure 21: Problems About Yirui Company

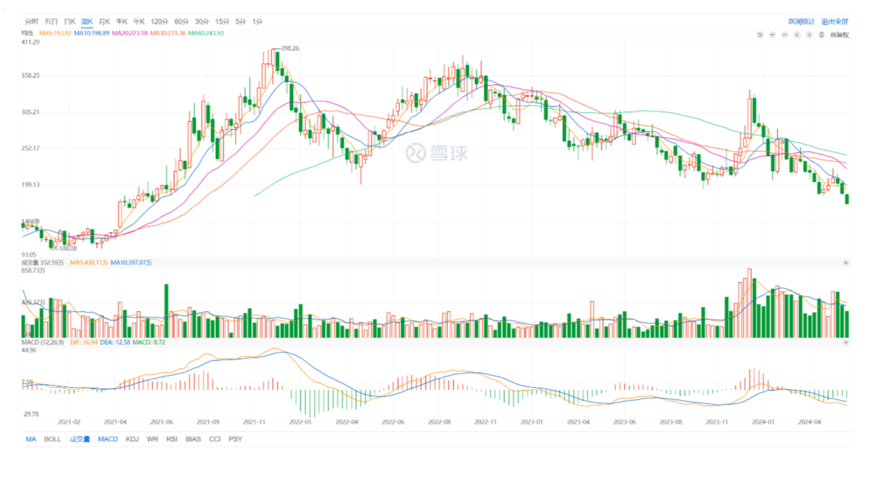


Figure 22: Problems About Yirui Company