

RESEARCH

Open Access



A factor score clustering approach to analyze the biopharmaceutical sector in the Chinese market during COVID-19

Jiahui Xi¹, Conghua Wen^{1,2*} , Yifan Tang¹ and Feifan Zhao¹

*Correspondence:
conghua.wen@xjtlu.edu.cn

¹ School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

² Research Institute of Quantitative Finance, Xi'an Jiaotong-Liverpool University, Suzhou, China

Abstract

The biopharmaceutical sector is of considerable interest during the COVID-19 pandemic. This study aims to investigate the biopharmaceutical sector using the Shenwan Industry Classification and provides insights into investment strategies. We combine factor and cluster analyses to reduce data dimensions and detect their latent similarities. Specifically, the biopharmaceutical sector is divided into six categories based on second-level industry classification. It is observed that medical devices, medical services, biological products, and chemical pharmaceuticals maintained their upward tendency, while Chinese medicine and pharmaceutical commerce declined slightly. We also develop optimal investment strategies using various metrics for different investor types.

Keywords: Factor analysis, Cluster analysis, Biopharmaceutical sector, Investment analysis

JEL Classification: C38, C61, G11

Introduction

Since its initial announcement in early 2020, the coronavirus disease 2019 (COVID-19) outbreak has precipitated recessions across major global economies such as China, the European Union, the United Kingdom, and the United States. This period also saw some countries experience significant political turmoil, contributing to considerable instability in financial markets. Notably, the global stock market has exhibited several phenomena indicative of these disruptions. AlAli (2020) identified the negative impact of COVID-19 on abnormal market returns, highlighting marked differences in returns before and after the World Health Organization's declaration in major Asian stock markets, as determined through event study methodologies. Additionally, using a novel time series approach based on a Bayesian structure, Takyi and Bentum-Ennin (2021) estimated that the stock performance of 13 countries (Ghana, Nigeria, South Africa, Kenya, Tanzania, Tunisia, Mauritius, Morocco, Zambia, Namibia, Botswana, Cote D'Ivoire and Uganda) was rarely positively affected by COVID-19. Specifically, since January 2020, the Chinese

stock market has experienced significant downturns, with the Shanghai stock index dropping by 8.7%, reaching a low of 2646 in March.

Esparcia and López (2022) reported that COVID-19 provided the biopharmaceutical industry with unprecedented revenue sources. The sector has been extensively analyzed in the literature (Robke et al. 2020; Ayati et al. 2020; Esparcia and López 2022; Ho et al. 2022). According to the Shenwan Industry Classification, China's biopharmaceutical sector includes medical services, medical devices, biological products, pharmaceutical commerce, Chinese medicine, and chemical pharmaceuticals. This raises critical questions regarding the uniformity of growth across these categories and their investment potential. This article focuses on two main inquiries: First, how have the six categories within China's biopharmaceutical sector changed during the pandemic? We explore this using factor scores before and after the onset of pandemic. Second, which investment strategies are optimal for various investor profiles within this sector? We apply factor and cluster analyses to guide stock selection and offer investment portfolio recommendations based on standard investment metrics.

As financial markets have rapidly evolved in recent decades, multivariate analysis has become integral to managing large, high-dimensional datasets in financial research (Wagenvoort et al. 2011; Vats and Samdani 2019; Seong and Nam 2021). This study combines factor analysis and clustering techniques for stock market analysis. As noted by Gorman and Primavera (1983), the primary purpose of factor analysis is to reduce the number of variables by grouping them into factors based on their correlations. Unlike factor analysis, cluster analysis can homogeneously group variables based on one or more multivariate similarity criteria. The goal is to segment data into clusters that reflect similarities and differences. The strength of factor analysis is the detection of a set of common and underlying dimensions of variables. However, factor analysis is unsuitable for investigating the latent similarities of data, which motivates us to consider using cluster analysis because the strength of cluster analysis is in identifying the different profiles or categories of the data and respondents. Based on the characteristic of the different techniques, we first consider using factor analysis to reduce the number of financial indicators to several components. Subsequently, we apply cluster analysis to homogeneously group the data according to the components.

To examine the impact of COVID-19 on the biopharmaceutical sector, this study analyzes 165 listed A-shared companies in China according to Shenwan Industry Classification. These findings highlight several key trends. First, biopharmaceutical industry stocks maintained an upward momentum during the pandemic. Specifically, medical devices, medical services, biological products, and chemical pharmaceuticals have increased, whereas Chinese medicine and pharmaceutical commerce have declined slightly. Second, based on the results of the cluster analysis, we allocate stock categories to different types of investors. Specifically, we classify investors based on their long- and short-term investments and degree of risk aversion. Finally, we translate the empirical results into economic gains, develop an optimal investment strategy based on the Sharpe ratio and mean–variance utility, and offer additional performance assessments using Treynor's ratio and Jensen's alpha. For short-term investments, the results vary because the short-term volatility of stocks is much higher, whereas for long-term investments, the results are similar.

This study contributes to the literature in three ways. First, it integrates factor and cluster analyses within the context of the Chinese financial market,—a combination seldom used in financial research. We illustrate the effectiveness of these combined techniques in analyzing the Chinese biopharmaceutical sector. Second, we contribute to second-level industry classification research. Most previous related literature focuses on first-level industry classification; in this article, we also discuss second-level industries (six categories) in the biopharmaceutical sector. In the context of the pandemic, investors are confident in the biopharmaceutical industries but not all categories have been boosted by COVID-19. Therefore, research on second-level industry classifications is useful for practical investments. Finally, we extend studies on the effects of COVID-19 on financial markets.

The remainder of this paper is organized as follows. In the next section, we review related literature and background. Section “[Methodology](#)” presents the methodologies used, including factor analysis and K-means clustering. Section “[Empirical study](#)” introduces the data and empirical research. We then provide an investment analysis in Section “[Sector investment analysis](#)” based on the categories derived in the previous section. Finally, we conclude the paper in Section “[Conclusion](#)”.

Literature review

Recent studies have advanced our understanding of stock market dynamics during the COVID-19 pandemic. Cox et al. (2020) found that the Federal Reserve played a role in stock market fluctuations in the early weeks of the pandemic. Daglis et al. (2022) presented another interesting conclusion: a decreasing impact of COVID-19 on the Italian stock market index and increasing volatility, both of which are statistically significant. Cevik et al. (2022) employed multiple analytical methods, including a panel regression with fixed effects, panel quantile regressions, a panel vector autoregression model, and country-specific regressions, to explore the relationship between investor sentiments and stock market returns and volatility across 20 countries. They observed that rising positive investor sentiment boosts stock returns, whereas negative sentiment dampens returns at the lower quantiles. Esparcia and López (2022) developed a global and dynamic ratio to summarize different investor profiles according to their attitudes toward risk and to consider the dynamic nature of the economy and financial markets. They find that Principal Component Analysis enables the first principal component to summarize the information contained in the initial performance rankings. In the context of China, Liu et al. (2020) identified a downturn in Chinese stock markets since the onset of the pandemic, suggesting that stock price movements can serve as indicators of economic performance and future economic trends (Wagner 2020).

Global market sectors exhibited varying dynamics amid the pandemic. Gupta et al. (2022) reported significant adversity inflicted by COVID-19 on India’s manufacturing, agriculture, and service sectors. In contrast, He et al. (2020) observed a negative impact on China’s transportation, mining, and utilities sectors, while noting resilience in manufacturing, IT, education, and healthcare. Piñeiro-Chousa et al. (2022) assessed the stock market responses of two pioneering US biopharmaceutical firms in mRNA vaccine development, highlighting the distinct volatility influences on Pfizer and Moderna’s returns before and during the pandemic. Zou and Wang (2023) revealed an

undervaluation of the medical sector, pinpointing China's medical sector as being significantly valuable for its stable risk premium. In the realm of pharmaceuticals, Ho et al. (2022) employed the Fama–French five-factor model to determine the influence of medical reform announcements and COVID-19 vaccine approvals on Chinese pharmaceutical and healthcare company returns. Their findings indicated negative effects on younger and smaller firms due to medical reforms, whereas vaccine approvals generally boosted stock returns, except for smaller entities. Robke et al. (2020) reviewed the impact of the pandemic on pharmaceutical innovation investment and speculated on future changes in the innovation-sourcing landscape. Ayati et al. (2020) delved into the pandemic's short- and long-term effects on the pharmaceutical sector, ranging from immediate demand shifts and regulatory changes to longer-term industry growth deceleration and supply chain shifts toward self-sufficiency. Despite the focus of existing literature on China's biopharmaceutical sector, there is a gap in second-level industry classification studies. This study aims to address this gap by examining the transformations within the biopharmaceutical sector and its six subdivisions in China throughout the COVID-19 timeline, culminating in tailored investment recommendations based on an innovative factor-score clustering method.

Factor analysis is a statistical tool used to identify a relatively small number of factors that represent the relationships between many interrelated variables. Early studies have used this tool for financial analyses (Morelli 1999; Jones 2006; Bai and Ng 2006; Ludvigson and Ng 2007). Wagenvoort et al. (2011) used factor analysis to reveal that the rates on large and small loans with long fixation periods converged weakly. Through factor analysis, they introduced a new measure to reassess whether retail bank market integration must be present, ongoing, or complete. Few studies have used factor analysis to score stocks; hence, little investment advice is provided. We provide portfolio suggestions by ranking stocks according to their factor scores. Clustering is an unsupervised technique that is used to generate groups of similar objects. K-means clustering has been used in several financial analyses. For instance, Nanda et al. (2010) compared several clustering methods, including K-means, self-organizing maps (SOM), and Fuzzy C-means, to perform stock classification in India. The results showed that K-means clustering can help to build the most compact clusters. Seong and Nam (2021) used K-means clustering and multiple kernel learning techniques to predict stock price movements. Expanding on these methodologies, our study employs K-means clustering to classify 165 Chinese listed companies into distinct categories, from which we devise corresponding investment strategies. Notably, this study merges factor and cluster analyses, which is a rare approach in financial studies, to enhance investment decision-making.

The impetus for our research emerges from the scant literature on second-level industry classifications. Few studies have employed both factor and cluster analyses in tandem. Our study unites these methodologies to reduce the variables and categorizes them based on their intrinsic traits.

Methodology

Factor analysis

The factor analysis method (Kim et al. 1978) recombines the original multidimensional variable indices to identify common factors, namely, the main factors. This reflects

the primary statistical information of the multidimensional index used to achieve the dimension reduction. Factor analysis can be used to analyze multiple statistical variables. After dimension reduction, the main factors were the primary information of the original variables, making the research process simple, effective, and objective. In this study, each stock had 12 financial indicators with obvious multicollinearity (see Fig. 1). To reduce them to fewer variables, we adopted factor analysis to extract the common factors. The main steps of the factor analysis are as follows:

Main Steps:

- Standardization of data and applicability test

First we must deal with the indicators through standardization. Here, X_{ij} is the i th financial indicator in year j , and X_i represents the mean of i th indicator, and S_j represents the standard error of i th indicator. Z_{ij} represents the variable after standardization, which follows a standard normal distribution.

$$Z_{ij} = \frac{X_{ij} - X_i}{S_j} (Z_{ij} \sim N(0,1))$$

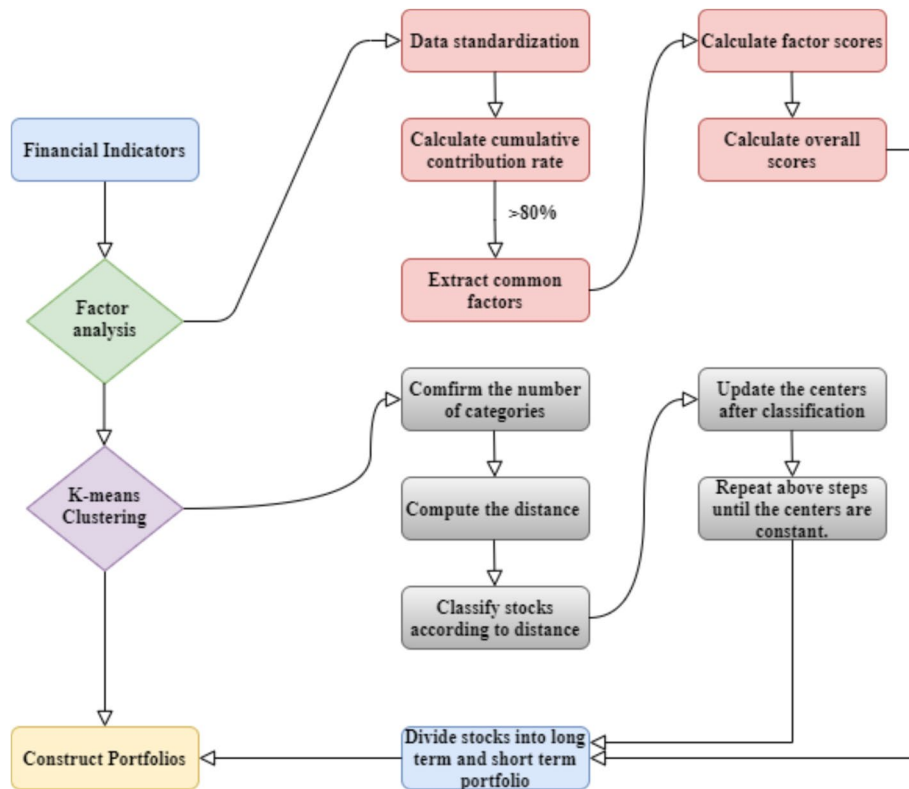


Fig. 1 Flowchart of research methodology. Note: This figure shows the methodology steps of our study. Specifically, we do factor analysis and K-means clustering on financial indicators. After that, we combine the results to select stocks and construct long term and short term portfolios for investors. Detailed description are in Sections “Factor Analysis” and “Cluster analysis(K-means)”

- Factor extraction and naming
Common factors were extracted in alignment with the methodology of Hao et al. (2019) using the cumulative contribution rate. To minimize data loss from common factors and enhance factor analysis utility, factors were chosen when their cumulative variance contribution rates exceeded 80%. Following the factor extraction predicated on the eigenvalues, we computed the variance contribution rate and the corresponding cumulative rates.
- Factor scores and composite scores calculation
The score for each factor was based on the factor coefficient and standardized variables. The common factors were calculated as follows:

$$F_i = \beta_{i1}X_1 + \beta_{i2}X_2 + \beta_{i3}X_3 + \dots + \beta_{in}X_n$$

where β_{ip} ($i = 1, 2, \dots, n$) is the score of factor F_i on variable X_p .

The overall scores can be computed by multiplying the score of each main factor by the contribution rate as follow:

$$F = \alpha_1F_1 + \alpha_2F_2 + \alpha_3F_3 + \dots + \alpha_mF_m$$

where F_i ($i = 1, 2, \dots, m$) is the score of each factor, and α_i ($i = 1, 2, \dots, m$) is the contribution rate of each factor. Factor analysis can efficiently deal with several intercorrelated variables and identify common factors that contain most of the information in the data.

Cluster analysis (K-means)

Clustering is a data-mining technique that divides a dataset into multiple categories by calculating the similarity between the data. K-means clustering (Hartigan and Wong 1979) is a technique in which data are divided into preset K categories, making the data characteristics in the same category more similar. The cluster centers were iteratively updated to optimize the results. We adopted the K-means clustering to classify 165 stocks into several categories, and the steps are as follows (Seong and Nam 2021).

Main steps:

- Step 1: Select $K=7$ initial centers for classification. ($K=7$ is the value determined from the empirical results.)
- Step 2: Calculate the distances between each point and the centers.
- Step 3: Classify the points according to distance using the Euclidean distance metric.
- Step 4: Update the centers by calculating the centroid of different categories.
- Step 5: Repeat Steps 2, 3, and 4 until the data points around each center remain constant.

In addition to its fast convergence speed and excellent clustering performance, K-means clustering requires one-parameter tuning (K) and is more explicable than other clustering techniques. Our methodology steps are plotted in the flowchart shown in Fig. 1.

Empirical study

Data description

We employed factor and cluster analyses in our empirical study. We chose companies in the biopharmaceutical sector according to the latest 2021 Shenwan Industry Classification. Six categories were selected, based on the second-level industry classification of the Shenwan biopharmaceutical sector. We selected annual data from the audited financial statements of six categories of listed companies in the biopharmaceutical industry from 2018 to 2021. All data were obtained from the Wind Database. We eliminated companies with incomplete data, leaving 165 listed companies for the empirical analysis. Information on the 165 listed companies is presented in Table 1. Data preprocessing included truncation and cleansing to derive 12 evaluative indicators selected for their comprehensiveness and relevance. These indicators were chosen to maintain the objectivity, accuracy, and referential integrity of the assessments, ensuring that they encapsulated the development trajectory of the companies' stocks. Following Wang and Lee (2008), who advocated clustering based on financial ratio variability, the chosen indicators and their summary statistics are detailed in Tables 1 and 2.

Table 1 Information of companies and financial indicators

	Chemical pharmaceuticals	Biological products	Medical service	Medical devices	Pharmaceutical commerce	Chinese medicine
Percentage	34% (56/165)	10% (16/165)	7% (12/165)	18% (29/165)	8% (13/165)	24% (39/165)
Average registered capital (Yuan)	1,045,551,145	827,859,633	1,171,901,459	479,815,344	652,019,296	810,789,305
Average number of employees (people)	41,538	16,943	114,089	23,797	95,085	39,326
Indicators	Description					
X1	The ratio of net cash flow from operating activities minus preferred stock dividends to the number of common shares outstanding					
X2	The reflection of the value of the company's net assets as represented by each share of stock, a criterion for judging the company's quality					
X3	The ratio of current year's increase in operating profit to last year's total operating profit reflects the increase or decrease in operating profit of an enterprise					
X4	The number of total assets to total shareholders' equity, reflecting an enterprise's financial leverage					
X5	The working capital ratio also reflects the enterprise's ability to realize assets					
X6	A measure of an enterprise's current assets that can be immediately used to repay current liabilities					
X7	The total liabilities ratio to total owners' equity reflects how a business can borrow to operate					
X8	The growth rate of total assets reflecting the liquidity of assets					
X9	The ratio of the growth of the total assets of an enterprise in the current year to the total assets at the beginning of the year reflects the overall situation of the enterprise's assets in the current period					
X10	The reflection of the growth rate of equity capital					
X11	The profit after tax ratio to total equity reflects the company's profitability					
X12	Recalculating earnings per share based on basic earnings per share, after converting potential ordinary shares into ordinary shares, increasing the total number of common shares					

This table reports information of 165 listed companies including their average registered capital and average number of employees. In addition, this table gives a description of twelve financial indicators and their summary statistics

Table 2 Summary statistics of financial indicators

Year	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
(Min, Max)	2021 (- 3.21,15.63)	(0.67,47.94)	(- 517.88,872.79)	(1.04,4.86)	(0.35,14.2)	(0.18,13.34)	(0.04,4.85)	(17.15,92.34)	(- 20.31,1.5895)	(- 70.47,533.51)	(- 0.94,697.38)	(- 0.94,13.98)
	2020 (- 7.06,9.36)	(- 0.63,32.58)	(- 1652.89,47.37,6.4)	(- 1.04,4.86)	(- 0.45,14.2)	(- 0.19,13.34)	(- 0.04,4.85)	(- 0.11,92.34)	(- 31.31,1.5895)	(- 63.98,533.51)	(- 40.91,697.38)	(- 2.03,13.98)
	2019 (- 1.98,9.04)	(- 1.21,39.95)	(- 12,830.21,3266.81)	(- 1.03,4.86)	(- 0.73,18.52)	(- 0.29,16.46)	(- 0.04,5.14)	(- 17.26,91.45)	(- 33.49,432.24)	(- 58.46,1957.45)	(- 2.11,10.27)	(- 2.11,10.27)
	2018 (- 1.98,9.04)	(- 1.21,39.95)	(- 12,830.21,3266.81)	(- 1.03,4.86)	(- 0.73,18.52)	(- 0.29,16.46)	(- 0.04,5.14)	(- 17.26,91.45)	(- 33.49,432.24)	(- 58.46,1957.45)	(- 2.11,10.27)	(- 2.11,10.27)
Mean	2021 0.31	8.00	32.20	1.61	3.10	2.47	0.67	54.22	16.54	62.28	1.01	1.00
	2020 0.49	7.04	311.30	1.62	2.86	2.26	0.68	44.35	21.78	48.54	21.43	0.91
	2019 0.26	6.49	- 51.37	1.59	3.01	2.28	0.66	54.66	13.02	81.86	0.67	0.67
	2018 0.26	6.49	- 51.37	1.59	3.01	2.28	0.66	54.66	13.02	81.86	0.67	0.67
Skewness	2021 4.89	2.56	1.82	3.15	4.38	4.56	3.11	0.09	5.37	4.67	6.95	6.97
	2020 2.11	1.96	12.47	2.67	2.42	2.50	2.91	- 0.45	1.40	3.24	5.91	4.95
	2019 4.17	2.75	- 9.56	2.65	2.91	3.24	3.18	0.08	8.34	7.89	4.24	4.26
	2018 4.17	2.75	- 9.56	2.65	2.91	3.24	3.18	0.08	8.34	7.89	4.24	4.26
Kurtosis	2021 33.39	8.35	12.90	12.63	28.96	30.42	11.71	- 0.49	43.03	27.53	62.89	63.32
	2020 14.00	4.21	155.69	9.08	7.72	8.52	10.56	- 0.62	2.81	13.30	37.36	31.19
	2019 25.54	10.63	110.58	9.79	11.80	15.17	13.81	- 0.51	87.95	80.25	34.09	34.24
	2018 25.54	10.63	110.58	9.79	11.80	15.17	13.81	- 0.51	87.95	80.25	34.09	34.24
Std	2021 1.83	7.47	152.44	0.66	2.91	2.66	0.75	16.72	31.21	119.43	1.92	1.91
	2020 1.61	5.87	369.92	0.61	2.19	1.93	0.74	25.27	27.52	83.80	87.33	1.73
	2019 1.12	5.50	1093.87	0.57	2.50	2.12	0.71	16.54	37.94	173.58	1.10	1.10
	2018 1.12	5.50	1093.87	0.57	2.50	2.12	0.71	16.54	37.94	173.58	1.10	1.10

This table reports the summary statistics of 12 financial indicators

Factor analysis

The data were first analyzed using SPSS software to measure the company’s financial performance at different levels. The correlation heat maps of the economic indicators for 2018 to 2021 are in Fig. 2. The tables of correlation coefficients of the financial indicators for 2018, 2019, 2020, and 2021 are reported in Annexed Tables 23, 24, 25, 26 in the Appendix.

Common factor formation

According to the requirements of the factor analysis method for the correlation degree of variables, we adopted Bartlett’s spherical test and the Kaiser–Meyer–Olkin (KMO) test to verify the suitability of the factor analysis. Bartlett’s spherical test was used to compare the correlation matrix of the data with the identity matrix. A significant Bartlett’s test (p -value below 0.05) indicated the suitability of the data for factor analysis. The KMO test was used to compare the simple and partial correlation coefficients. If the KMO value approaches 1, the correlation among the variables is significant, and factor analysis is suitable for the variables. The results obtained after processing using SPSS are presented in Table 3.

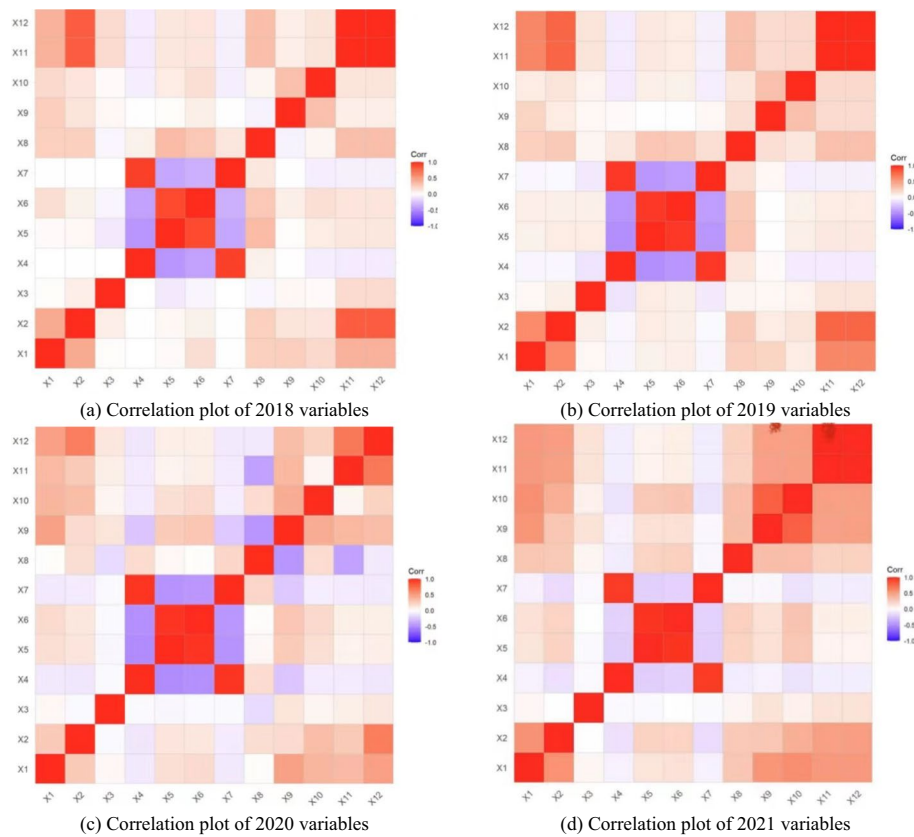


Fig. 2 Correlation heat map plots of financial indicators. *Note* The correlation heatmaps of 12 financial indicators from year 2018–2021 are plotted. A red box stands for a high positive correlation and a purple box stands for a low negative correlation. We can find that in each year, there is a high positive correlation between X4 (equity multiplier) and X7 (equity ratio), X5 (current ratio) and X6 (quick ratio), and there is an obvious negative correlation between X7 (equity ratio) and X5 (current ratio), X7 (equity ratio) and X6 (quick ratio)

Table 3 KMO and Bartlett test

	2018	2019	2020	2021
KMO metric for sampling adequacy	0.655	0.676	0.603	0.644
Bartlett's test for sphericity <i>p</i> -value	0.000	0.000	0.000	0.000

KMO and Bartlett tests results are used to check if the factor analysis can be adopted in the dataset. With KMO metric larger than 0.6 and Bartlett test *p*-value less than 0.05, factor analysis can be adopted

Table 4 Total variance explained

Year	Common factor	Variance contribution rate	Cumulative contribution rate
2018	F ₁	25.748%	25.748%
	F ₂	18.479%	44.226%
	F ₃	16.709%	60.935%
	F ₄	12.238%	73.172%
2019	F ₁	27.265%	27.265%
	F ₂	18.574%	45.839%
	F ₃	17.491%	63.330%
	F ₄	11.527%	74.857%
2020	F ₁	24.601%	24.601%
	F ₂	18.963%	43.564%
	F ₃	14.927%	58.490%
	F ₄	13.973%	72.463%
2021	F ₁	23.522%	23.522%
	F ₂	18.678%	42.200%
	F ₃	18.283%	60.483%
	F ₄	17.306%	77.789%

The variance contribution rates and cumulative contribution rates of four common factors from 2018 to 2021 are reported in Table 3. The final cumulative contribution rates for four years are respectively **73.172%**, **74.857%**, **72.463%** and **77.789%**

From the test results above, all KMO values were larger than 0.6, which means that there were more similar factors among the variables; therefore, the selected variables were suitable for the molecular factor test. The *p*-values of Bartlett's spherical test were all 0.00, below the threshold of 0.05, signifying robust interrelations among the financial indicators and endorsing their suitability for factor analysis.

After extracting common factors from the 12 financial indicators, the total variance explained degree, gravel plot, common factor variance results, and component score coefficient matrix were obtained by analyzing the relationships among the financial indicators and revealing the primary information contained in the common factors.

As the loadings of the common factors F₁, F₂, F₃, and F₄ on some of the initial variables did not differ significantly, an explanatory relationship between the variables and the common factors could not be observed; therefore, rotation of the component matrix was required.

From the results in Table 4, four common factors were extracted from the 12 selected indicators. The cumulative variance contribution of F₁, F₂, F₃ and F₄ were 73.172%, 74.857%, 72.463%, and 77.789% respectively for 2018–2021. Analysis of the gravel plot in Fig. 2 also shows that the inflection point occurred at the fourth root of the feature; therefore, the first four factors could be retained.

Table 5 Relationship between indicators and common factor

Common factors	indicators	Relevance			
		2018	2019	2020	2021
F ₁	X11	0.964	0.964	0.744	0.973
	X12	0.964	0.946	0.921	0.973
	X2	0.884	0.882	0.746	0.779
F ₂	X9	0.817	0.801	0.531	0.916
	X1	0.449	0.105	0.590	0.839
	X10	0.739	0.794	0.801	0.708
	X3	− 0.029	0.173	0.150	0.148
F ₃	X7	0.950	0.947	− 0.098	0.962
	X4	0.949	0.938	− 0.122	0.949
F ₄	X5	0.854	0.900	0.815	0.908
	X6	0.827	0.897	0.798	0.905
	X8	0.659	0.620	− 0.125	0.577

The relationships between indicators and common factors from 2018 to 2021 are reported in Table 4. The most related indicators for each common factor are listed. A larger absolute value in the table means a closer relationship between two certain factors

Since we aimed to examine the attribution of each variable, the original component matrix was rotated for ease of naming. This rotation ensures that each variable has a more extensive loading on one common factor and a smaller loading on the remaining common factors, and the results are shown in Table 5.

The rotated factor-loading matrix is presented in Table 5. In the first common factor, diluted earnings per share, basic earnings per share, and net assets per share have more significant loadings, indicating that these three indicators are more correlated with each other and reveal the enterprise’s current profits, losses, and future earnings expectations. Common factor F_1 is the operating income factor. The second common factor, net cash flow per share, operating profit growth, total assets growth rate, and growth rate of shareholders’ equity, have extensive loadings, indicating that these four indicators are highly correlated and reveal the company’s development capability in 2021. Common factor F_2 is the development potential factor. In the third common factor, equity multiplier and equity ratio have high loadings, indicating that these two indicators are strongly correlated and reveal a company’s assets and equity. Common factor F_3 is the asset structure factor. Finally, the fourth common factor, with substantial loadings on the current ratio, quick ratio, and proportion of current assets to total assets, indicates a robust correlation among indicators that reflect a firm’s short-term liquidity and is aptly labeled the “Solvency Factor.”

Composite score of the company

A matrix of component score coefficients was obtained using SPSS (Table 6).

The standardized values of the initial indicators were substituted into the factor score function to calculate the factor scores of each sample, and a further comprehensive evaluation of the observed indicators was performed. A complete evaluation model was established using the variance contribution of the four common factors extracted as weights, combined with each factor score.

Table 6 Component score coefficient matrix

	2018				2019			
	F1	F2	F3	F4	F1	F2	F3	F4
X1	0.110	0.076	0.047	0.271	0.241	-0.012	-0.035	-0.021
X2	0.296	-0.004	-0.062	-0.005	0.291	-0.010	-0.032	-0.098
X3	0.122	-0.112	-0.203	-0.043	0.016	-0.075	-0.016	0.119
X4	-0.018	0.448	0.065	0.030	-0.009	0.439	0.054	-0.021
X5	-0.050	-0.064	0.419	-0.028	-0.048	-0.013	0.437	-0.034
X6	-0.053	-0.048	0.403	0.065	-0.048	-0.002	0.441	-0.040
X7	-0.019	0.460	0.109	0.015	-0.008	0.452	0.081	-0.019
X8	0.082	0.242	0.397	-0.114	0.029	0.324	0.399	0.067
X9	-0.069	0.008	-0.068	0.588	-0.073	0.029	-0.030	0.615
X10	-0.058	0.004	0.010	0.521	-0.080	-0.039	-0.032	0.611
X11	0.327	-0.029	-0.038	-0.070	0.298	-0.028	-0.042	-0.022
X12	0.328	-0.029	-0.038	-0.070	0.299	-0.028	-0.041	-0.023
	2020				2021			
	F1	F2	F3	F4	F1	F2	F3	F4
X1	-0.052	0.088	0.305	0.075	-0.117	0.443	0.033	-0.040
X2	0.011	0.392	-0.005	-0.311	0.271	0.013	0.017	0.013
X3	-0.064	-0.114	0.136	0.303	-0.013	0.064	-0.061	0.004
X4	-0.329	-0.068	0.198	0.010	-0.005	0.015	0.454	0.066
X5	0.268	-0.094	0.160	-0.085	-0.039	-0.058	-0.019	0.454
X6	0.256	-0.115	0.194	-0.055	-0.040	-0.049	-0.011	0.454
X7	-0.325	-0.067	0.211	-0.006	0.001	0.010	0.465	0.084
X8	-0.022	-0.002	0.183	-0.509	0.073	-0.028	0.275	0.362
X9	-0.003	-0.071	0.310	0.382	-0.127	0.480	0.012	-0.044
X10	-0.058	-0.092	0.501	-0.005	0.066	0.311	-0.047	-0.114
X11	-0.020	0.345	-0.154	0.166	0.394	-0.130	-0.009	-0.019
X12	-0.024	0.445	-0.066	-0.075	0.394	-0.130	-0.009	-0.019

The component score coefficient matrix of four common factors based on 12 financial indicators from 2018 to 2021 is reported

$$F = \frac{23.522\%}{77.789\%}F_1 + \frac{18.678\%}{77.789\%}F_2 + \frac{18.283\%}{77.789\%}F_3 + \frac{17.306\%}{77.789\%}F_4$$

$$\text{i.e. } F = 0.302F_1 + 0.240F_2 + 0.235F_3 + 0.222F_4$$

The composite score for each firm was calculated, and the scores were ranked separately in descending order, with only the top 20 stocks with positive composite scores shown in this paper, as shown in Table 7.

The literature review, including studies by He et al. (2020) and Gupta et al. (2022), indicated that industry firms broadly confronted the adverse consequences of the COVID-19 pandemic, leading to economic downturns. From Tables 7 and 8, the composite scores of the top three leading enterprises uniformly show upward trends, with the top four enterprises' composite scores being 2.04, 1.84, 1.8, and 1.44, respectively, in 2018 before the outbreak, and 2.45, 1.85, 1.73, and 1.64 in 2021 after the outbreak. The positive impact of the pandemic was offset by the adverse effects of the economic

Table 7 Top 20 scoring stocks and scores for 2018 and 2019

Rank	Stock codes (2018)	Scores (2018)	Stock codes (2019)	Scores (2019)
1	000028.SZ	2.04	000661.SZ	2.74
2	000661.SZ	1.84	000538.SZ	1.57
3	000411.SZ	1.80	000411.SZ	1.48
4	600,713.SH	1.44	600,713.SH	1.48
5	603,590.SH	1.27	000028.SZ	1.29
6	300,760.SZ	1.22	603,368.SH	1.29
7	688,739.SH	1.21	002462.SZ	1.14
8	600,332.SH	1.14	300,760.SZ	1.02
9	002462.SZ	1.06	002793.SZ	0.92
10	603,368.SH	1.00	600,511.SH	0.92
11	000538.SZ	0.97	600,129.SH	0.85
12	601,607.SH	0.93	600,332.SH	0.82
13	002940.SZ	0.88	688,739.SH	0.80
14	300,753.SZ	0.82	601,607.SH	0.78
15	000423.SZ	0.81	300,519.SZ	0.75
16	600,129.SH	0.78	600,829.SH	0.75
17	600,511.SH	0.78	300,485.SZ	0.72
18	300,723.SZ	0.69	300,759.SZ	0.71
19	603,259.SH	0.69	000710.SZ	0.62
20	000710.SZ	0.59	600,276.SH	0.61

The top 20 scoring stocks and scores in 2018 and 2019 are reported. Ranks, stock codes and scores are provided. **000028.SZ** and **000661.SZ** are respectively two best stocks in 2018 and 2019 with scores: **2.04** and **2.74**. The top 3 stocks both in 2018 and 2019 are all from Shenzhen Stock Exchange

Table 8 Top 20 scoring stocks and scores for 2020 and 2021

Rank	Stock Codes (2020)	Scores (2020)	Stock Codes (2021)	Scores (2021)
1	688,399.SH	2.26	688,399.SH	2.45
2	300,347.SZ	2.23	688,739.SH	1.85
3	603,301.SH	1.87	002821.SZ	1.73
4	002821.SZ	1.13	603,127.SH	1.64
5	300,639.SZ	1.01	600,713.SH	1.26
6	300,595.SZ	0.93	000661.SZ	1.22
7	300,760.SZ	0.90	000411.SZ	1.06
8	002030.SZ	0.89	002524.SZ	0.79
9	300,519.SZ	0.85	600,129.SH	0.74
10	002382.SZ	0.79	000028.SZ	0.73
11	688,739.SH	0.79	002462.SZ	0.66
12	300,685.SZ	0.78	601,607.SH	0.61
13	300,753.SZ	0.78	300,573.SZ	0.56
14	300,702.SZ	0.73	603,368.SH	0.55
15	300,529.SZ	0.72	000538.SZ	0.53
16	300,485.SZ	0.59	300,760.SZ	0.52
17	603,259.SH	0.57	832,735.BJ	0.51
18	000661.SZ	0.57	600,332.SH	0.51
19	002901.SZ	0.57	600,829.SH	0.49
20	300,206.SZ	0.53	600,511.SH	0.47

The top 20 scoring stocks and scores in 2020 and 2021 are reported. Ranks, stock codes and scores are provided. **688,399.SH** and **688,399.SH** are respectively two best stocks in 2020 and 2021 with scores: **2.26** and **2.45**. Besides, 300,347.SZ also has a good score at 2.23 in 2020

environment, resulting in a slight increase in the overall scores of the leading companies. However, owing to the weaker capability of small medium enterprises to withstand economic downturns, even though the pandemic boosted their growth to a certain extent, the overall score still shows a slight decline. This result was consistent with the findings of Thukral (2021).

In 2021, the composite scores of the top seven companies rose by approximately 10% compared to 2020, reflecting sustained governmental support for biotechnological innovation and bio-industry development, coupled with increased consumer demand for medical supplies like alcohol and masks. This support has propelled leading biopharmaceutical stocks upward during the pandemic.

The classification and indicator score ranking of stocks in 2020 and 2021 have undergone a significant reshuffle compared to 2018 and 2019. Therefore, the specific development of the different types of pharmaceutical stocks must be discussed further. Table 9 presents the scores of the top 20 stocks in 2021 over the last 4 years.

The results in this table will be used to calculate the stock score in Tables 7 and 8.

Table 9 Stock score comparison

Rank	Categories	Stock Codes	Scores (2018)	Scores (2019)	Scores (2020)	Scores (2021)
1	Medical devices	688,399.SH	0.44	0.44	2.26	2.45
2	Biological products	688,739.SH	1.21	0.80	0.79	1.85
3	Medical services	002821.SZ	0.11	0.04	1.13	1.73
4	Medical services	603,127.SH	0.01	− 0.05	0.00	1.64
5	Pharmaceutical commerce	600,713.SH	1.44	1.48	− 1.32	1.26
6	Biological products	000661.SZ	1.84	2.74	0.57	1.22
7	Pharmaceutical commerce	000411.SZ	1.80	1.48	− 1.06	1.06
8	Medical services	002524.SZ	− 0.21	− 0.64	− 1.25	0.79
9	Chinese medicine	600,129.SH	0.78	0.85	− 1	0.74
10	Pharmaceutical commerce	000028.SZ	2.04	1.29	− 0.5	0.73
11	Pharmaceutical commerce	002462.SZ	1.06	1.14	− 0.53	0.66
12	Pharmaceutical commerce	601,607.SH	0.93	0.78	− 0.46	0.61
13	Chemical pharmaceuticals	300,573.SZ	− 0.6	− 0.46	− 0.09	0.56
14	Pharmaceutical commerce	603,368.SH	1.00	1.29	− 0.20	0.55
15	Chinese medicine	000538.SZ	0.97	1.57	0.31	0.53
16	Medical devices	300,760.SZ	1.22	1.02	0.9	0.52
17	Chemical pharmaceuticals	832,735.BJ	− 0.15	− 0.01	0.05	0.51
18	Chinese medicine	600,332.SH	1.14	0.82	− 0.30	0.51
19	Pharmaceutical commerce	600,829.SH	− 0.03	1.71	0.66	0.49
20	Pharmaceutical commerce	600,511.SH	− 0.07	0.32	0.5	0.47

The top 20 scoring stocks and their categories in 2021 are reported in Table 8. Moreover, their scores in 2018, 2019 and 2020 are also provided for comparison. The top one stock, 688,399.SH's score increased year by year starting from 0.44 in 2018. Among the top 20 stocks, 8 stocks come from pharmaceutical commerce and 5 stocks come from Medical services

The top 20 stock categories in the composite score for 2021 are more evenly dispersed, with biopharmaceutical companies in all six categories. Of these, 80% of the companies in the four categories—medical devices, medical services, biological products, and chemical pharmaceuticals—showed an upward trend in their composite scores after the outbreak. As the response to the pandemic necessitated medical personnel and equipment support, the medical market experienced a significant boost in capital inflows. Additionally, advancements in vaccine research and development prompted a re-evaluation of the biopharmaceutical industry’s significant growth potential. Several stocks exhibited multiple upward movements. For companies in the two categories of Chinese medicine and pharmaceutical commerce, the composite score tended to decline slightly or remained stable after the outbreak. The increased popularity of disinfection products post-pandemic led to a decline in the incidence of other infectious diseases such as influenza, resulting in a reverse impact on commercial businesses and Chinese medicine.

Cluster analysis

This study used the K-means clustering method to cluster the stocks in the sample and refine the similarities among them, analyzing the characteristics and commonalities of the 165 listed stocks. From the gravel plot (Fig. 3), it can be seen that the curve remains flat at $K=7$, while it suddenly increases at $K=8$; therefore, K is taken as 7, that is, the sample is divided into seven categories. We also show the results of the clustering data when the number of clusters is six and seven in Annexed Tables 27 and 28 in the Appendix.

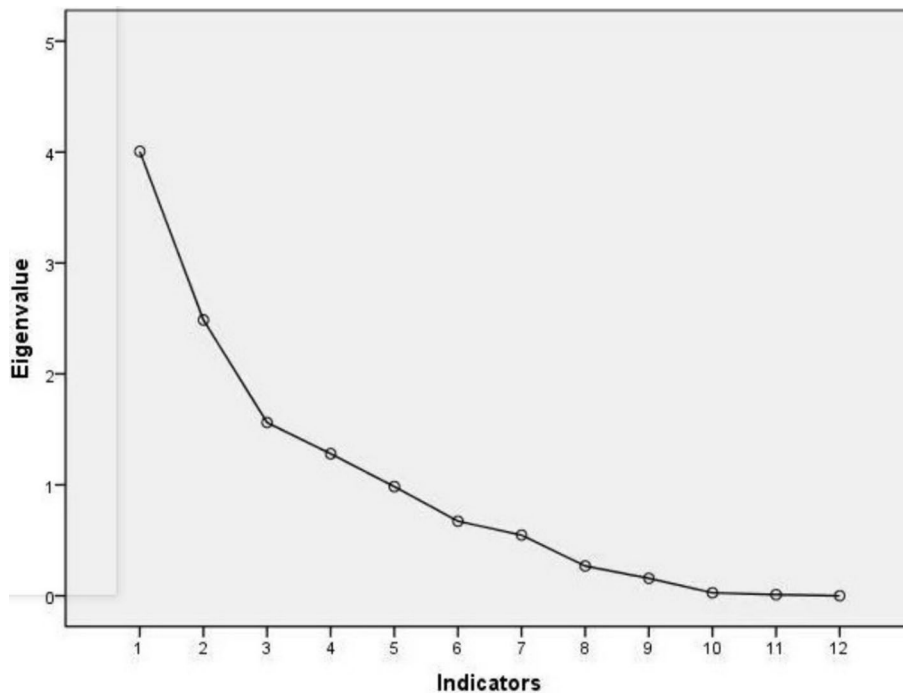


Fig. 3 Gravel Plot. Note: Gravel plot of common factors is shown in this figure. The 4th point turns out to be an inflection point and the exact variance contribution rate is shown in Table 4

Table 10 Ranking and factor scores for Category 1 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	688,399.SH	9.54	-0.49	-0.76	-0.61

Table 11 Ranking and factor scores for Category 2 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	688,739.SH	-0.04	1.82	-1.05	7.54

Table 12 Ranking and factor scores for Category 3 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	300,573.SZ	0.19	2.61	-0.45	-0.09
2	000538.SZ	1.32	0.30	-0.40	0.68
3	300,760.SZ	2.91	-0.91	-0.33	-0.27
4	832,735.BJ	-0.26	2.31	-0.60	0.77
5	600,436.SH	1.64	-0.43	-0.71	1.01
6	300,347.SZ	1.72	0.57	-0.85	-0.21
7	300,294.SZ	-0.16	0.63	-0.81	2.06
8	300,485.SZ	-0.47	-0.93	-0.95	4.22
9	300,529.SZ	-0.19	0.47	-0.38	1.61
10	300,639.SZ	0.84	0.66	-0.67	0.08

Table 13 Ranking and factor scores for Category 4 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	002821.SZ	1.69	5.57	-0.42	-0.1
2	000661SZ	4.18	0.39	-0.41	-0.15

Table 14 Ranking and factor scores for Category 5 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	603,127.SH	-0.68	9.1	-0.74	-0.74

Based on the scores derived from the factor analysis, the stocks in each of the seven categories in 2021 are ranked in descending order of total scores. The top ten scoring stocks were screened and analyzed for their four public factor scores, as shown in Tables 10, 11, 12, 13, 14, 15, 16. Baker and Haslem (1974) divided investors into two types according to a decision-orientation criterion, which refers to investors' confidence in their decision-making abilities. This study divided investors into confidence groups based on these criteria.

In the first category, the stocks score exceptionally well on the operating income factor, while the other types perform at average levels. Operating earnings represent the existing earnings of a business, suggesting that this category is suitable for

Table 15 Ranking and factor scores for Category 6 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	600,332.SH	0.94	0.12	0.62	0.22
2	600,511.SH	0.98	-0.07	0.32	0.50
3	300,759.SZ	0.56	1.25	0.14	-0.53
4	300,171.SZ	0.07	0.63	0.61	0.29
5	300,244.SZ	0.44	0.17	0.72	0.01
6	002727.SZ	0.19	0.77	0.74	-0.56
7	002793.SZ	-0.22	1.16	0.35	-0.41
8	600,976.SH	0.54	0.14	0.17	-0.19
9	002030.SZ	0.49	1.00	-0.45	-0.45
10	603,301.SH	1.31	-0.09	-0.26	-0.70

Table 16 Ranking and factor scores for Category 7 stocks

Rank	Stock Codes	Scores (F_1)	Scores (F_2)	Scores (F_3)	Scores (F_4)
1	600,713.SH	-0.17	0.05	4.53	1.07
2	000411.SZ	0.02	0.10	3.60	0.84
3	002524.SZ	-0.82	0.04	5.02	-0.67
4	600,129.SH	-0.73	-0.58	4.6	0.08
5	000028.SZ	2.11	-0.87	1.10	0.21
6	002462.SZ	0.46	0.01	1.67	0.58
7	601,607.SH	0.80	-0.24	1.67	0.15
8	603,368.SH	0.72	-0.44	1.40	0.51
9	600,829.SH	-0.16	-0.03	1.71	0.66
10	300,358.SZ	-0.14	0.73	1.35	-0.19

investors aiming for short-term profits. However, due to their moderate growth potential, long-term investments are less advisable. This category is recommended for short-term holdings, particularly for investors with a lower risk tolerance.

In the second category, the stocks are more prominent in the solvency factor score. The scores of the other three factors are relatively stable, and the stock types are concentrated in biopharmaceuticals. The outbreak of the pandemic has led to a rapid and steady growth in new product R&D expenditures, and the scale of new product production and sales of large- and medium-sized enterprises in China's biopharmaceutical industry is on a faster growth trend. Therefore, this category represents an ideal option for investors seeking medium- to long-term holdings, particularly for those with lower risk tolerance.

In the third category, the stocks have stable scores in operating income, solvency, and development potential, reaching a positive value of approximately 60%. However, the asset structure factor scores are all negative. This finding indicates that the pandemic did not significantly affect the development of these A-share listed companies. There is a commonality of weak stability in funding sources, making it challenging to achieve high short-term returns on stocks. It is suitable for cautious investors to hold in the medium- or long-term.

In the fourth category, the stocks have positive scores for operating income and development potential factors and negative scores for asset structure and solvency factors. Operating income represents the existing earnings of a business, suggesting that this category is suitable for confident investors seeking short-term profitability. It is recommended that investors hold forward contracts for this category of stocks, combined with the excellent performance of their growth potential.

In the fifth category, this group of stocks score particularly well on the development potential factor, whereas the other factors are relatively flat. Owing to the long duration of the impact of the coronavirus pandemic on China’s biopharmaceutical industry, the potential of this category is stable in the long term. This is suitable for long-term holdings of cautious investors.

In the sixth category, the stocks do not score well on the operating income, solvency, asset structure, and development potential factors; all are relatively “medium.” The distribution of stocks in this category is relatively even, and the risk profile is medium. Although there are some variations in share prices, stocks appear to be suitable for cautious investors.

In the seventh category, stocks stand out, with positive scores on the asset structure factor, whereas the other three factors have relatively flat scores. The leading asset structure indicator indicates the relative relationship between the funding sources provided by creditors and those offered by investors, reflecting the stability of the underlying financial structure of the business. Hence, stocks in this category are suitable for confident investors with a short-term horizon.

Sector investment analysis

We have offered stock-selection guidance for various investor profiles. However, translating these empirical findings into tangible economic benefits requires specific investment portfolio recommendations. Moving forward, we aim to devise optimized strategies for both long- and short-term investments, integrating the insights from Section “3.3” with return-risk analyses. In this study, we address the optimization challenge by employing Markowitz’s model, using the Sharpe ratio (Sharpe 1998) as a performance metric. Markowitz’s model is designed to assist rational investors in maximizing returns for a given level of risk, or minimizing risk for a specified return level.

Long and short term portfolio

Suppose that investment P has a set of N variable assets in the market. Let r_p be the expected rates of return and δ_p be the risks. We will have expected return $r_p = \sum_{i=1}^N w_i r_i$, where w is the weight factor with values between 0 and 1, and the variance $\delta_p^2(w) = \sum_{i=1}^N \sum_{j=1}^N w_i w_j \delta_{ij} = w^T \Delta w$, where w_i and w_j are the weights assigned to stock i and j respectively; Δ is the covariance matrix of the stocks, and δ_{ij} is the covariance between the stock price of i and j . Then we define:

$$\text{risk} : \delta_p = \sqrt{w^T \Delta w}$$

$$\text{Sharpe Ratio} : \frac{r_p - r_f}{\delta_p}$$

where r_f is the risk-free interest rate. We choose benchmark 1-year and 5-year deposit rates as proxies. The optimization problem can be described as follows:

$$\begin{aligned} & \text{Max } \frac{r_p - r_f}{\delta_p} \\ & \text{s.t. } \sum_{i=1}^N w_i = 1 \end{aligned}$$

We consider a mean–variance investor and define the quadratic utility function for investing in this portfolio (Wang et al. 2016) as:

$$U = E(r_p) - \frac{1}{2}\gamma var(r_p)$$

where U is the mean–variance utility, $E(r_p)$ is the portfolio’s mean return, and $var(r_p)$ is the portfolio’s variance, which is a proxy for the portfolio risk. γ is the risk aversion coefficient, and we set γ to be 3 for the short-term investment and 6 for the long-term investment.

For long-term investment, we selected stocks from Categories 3 and 5, which include five stocks. Due to extensive missing data, stock 832,735, BJ (stocks from Beijing Exchange) was excluded from the portfolio analysis. We ran the optimization on the remaining four stocks: 300,573.SZ, 000538.SZ, 300,760.SZ, and 603,127. SH. We set r_f as the benchmark five-year deposit rate for 2021 (2.75%). Optimization simulations used stock prices from the start of 2021 to the end of 2021. R programming was used to perform optimization according to the above formula. Figure 4a shows the Sharpe ratio curve of this portfolio against the expected return with a minimum α . For demonstration, we display the nearest ten results around the maximum point. As shown in Table 17, 0.03452 had the largest Sharpe ratio, and the weights of 000523. SZ and 300,760. The SZ in this portfolio is 0 with a weight of 300,573.SZ and 603,127.SH were 24.8% and 75.2%, respectively, with an expected return of 7.179% and a standard deviation of 1.283. Figure 4a shows the Sharpe ratio curve against the expected return with a minimum α . Table 17 also reports the mean–variance utility values for each portfolio. An optimal value of 4.7134 was achieved when the weights of 300,573. SZ and 603,127.SH were 26.99% and 73.01%, respectively, with an expected return of 7.148% and standard deviation of 1.274.

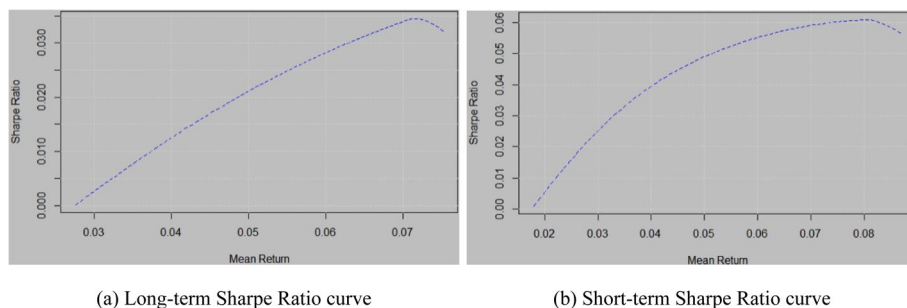


Fig. 4 Sharpe Ratio curves of long- and short-term investment

Table 17 Long-term investment of stocks

ER (%)	SD	Sharpe ratio	Utility	000538.SZ (%)	300,760.SZ (%)	300,573.SZ (%)	603,127.SH (%)
7.023	1.255	0.03405	4.660	0.58	0.00	32.35	67.07
7.054	1.258	0.03420	4.6802	0.19	0.00	32.45	67.36
7.085	1.262	0.03435	4.6960	0.00	0.00	31.37	68.63
7.116	1.267	0.03446	4.7081	0.00	0.00	29.18	70.82
7.148	1.274	0.03451	4.7134	0.00	0.00	26.99	73.01
7.179	1.283	0.03452	4.7099	0.00	0.00	24.80	75.20
7.210	1.293	0.03449	4.7022	0.00	0.00	22.61	77.39
7.241	1.305	0.03441	4.6865	0.00	0.00	20.42	79.58
7.272	1.319	0.03429	4.6624	0.00	0.00	18.23	81.77
7.303	1.334	0.03413	4.6337	0.00	0.00	16.04	83.96
7.334	1.350	0.03395	4.6003	0.00	0.00	13.85	86.15

This table reports the nearest 10 points around the maximum point of Sharpe Ratio and mean–variance utility in the situation of long term portfolio strategy. In the first line, ER and SD respectively stands for expected return and standard deviation of an investment, and the values in the column of stocks represent the weight of this stock in an investment. The largest Sharpe ratio and utility value are highlighted in bold

Table 18 Short-term investment of stocks

ER (%)	SD	Sharpe ratio	Utility	002821.SZ (%)	600,713.SH (%)	600,129.SH (%)
7.848	1.004	0.05076	4.8240	29.66	2.16	68.18
7.905	1.013	0.05091	4.8265	29.83	1.25	68.92
7.962	1.021	0.05106	4.8347	29.99	0.34	69.67
8.019	1.029	0.05119	4.8425	28.57	0.00	71.43
8.076	1.039	0.05124	4.8374	26.23	0.00	73.77
8.133	1.051	0.05123	4.8192	23.88	0.00	76.12
8.190	1.064	0.05114	4.7937	21.53	0.00	78.47
8.248	1.078	0.05100	4.7617	19.19	0.00	80.81
8.305	1.094	0.05079	4.7145	16.84	0.00	83.16
8.362	1.110	0.05054	4.6657	14.49	0.00	85.51
8.419	1.128	0.05024	4.6018	12.14	0.00	87.86

This table reports the nearest 10 points around the maximum point of Sharpe Ratio and mean–variance utility in the situation of short-term investment strategy. In the first line, ER and SD respectively stands for expected return and standard deviation of an investment, and the values in the column of stocks represent the weight of this stock in an investment. The largest Sharpe ratio (which is 0.05123 in this table) is highlighted in bold

For short-term investment, we selected the best stock from Categories 1, 4, and 7, which comprise 12 stocks in total. In this situation, we set α to be 1.75%, the one-year risk-free interest rate in 2021. Similarly, Fig. 4b shows the Sharpe ratio curve for the 12 stocks against the expected returns. Similar to Table 17, the optimization results with the nearest ten points are listed in Table 18. This table excludes the weights of 688,399.SH, 601,607.SH, 000661.SZ, 603,368.SH, 600,829.SH, 002524.SZ, 000028.SZ, 000411.SZ, and 002462.SZ because their weights are approximately 0.00% when a minimum of 0.0175 expected return is required. The remaining three stocks are 002821.SZ, 600,713.SH, and 600,129.SH. From Table 18, 0.05123 is the largest Sharpe ratio, and for this investment, the weights of 600,173.SH and 601,607.SH were both 0 and the weights of 002821.SZ and 600,129.SH were 23.88% and 76.12%, respectively, with an expected return of

Table 19 Short-term investment with optimal TR and JA

Stock name/expected return	Treynor ratio	Jensen Alpha
002821.SZ	9.72%	0.04%
600,713.SH	10.37%	0.00%
603,368.SH	2.75%	0.00%
601,607.SH	28.27%	0.00%
600,829.SH	3.01%	0.00%
600,129.SH	8.35%	99.6%
002462.SZ	35.31%	0.00%
000028.SZ	2.21%	0.00%
Expected return	1.795%	8.704%
Performance metric	0.002	0.156

This table reports the two short term investment respectively with the optimal performance in Treynor Ratio and Jensen Alpha

Table 20 Long-term investment with optimal TR and JA

Stock name/expected return	Treynor ratio	Jensen Alpha
000538.SZ	9.53%	0.19%
300,760.SZ	0.00%	0.00%
300,573.SZ	29.84%	32.45%
603,127.SH	60.63%	67.36%
Expected Return	6.308%	7.054%
Performance Metric	24.575	0.036

This table reports the two long term investment respectively with the optimal performance in Treynor Ratio and Jensen Alpha

8.133% and a standard deviation of 1.051. Table 18 also reports the mean–variance utility values for each portfolio. The optimal value of 4.8425 was achieved when the weights of 002821.SZ and 600,129.SH were 28.57% and 71.43%, respectively, with an expected return of 8.019% and standard deviation of 1.029.

To provide investment suggestions from another risk–return perspective, we calculated the portfolios (which have returns above the risk-free rate), Treynor Ratio (TR), and Jensen Alpha (JA); the optimal investment strategy using these two performance evaluation metrics is shown in Tables 19 and 20. The Treynor Ratio measures the risk premium earned per unit of systematic risk, whereas Jensen Alpha assesses investment performance by quantifying the deviation of a portfolio’s average return from its expected return, based on the Capital Asset Pricing Model. The formulae for these performance metrics are as follows:

$$\text{investment beta} : \beta_p = w_1\beta_1 + w_2\beta_2 + \dots + w_N\beta_N$$

$$\text{Treynor Ratio} : \frac{r_p - r_f}{\beta_p}$$

$$\text{Jensen Alpha} : r_p - (r_f + \beta_p(r_m - r_f))$$

Table 21 Robustness check results for the long-term portfolio

ER	SD	Sharpe Ratio	Utility	000538.SZ	300,760.SZ	300,573.SZ	603,127.SH
7.179%	1.283	0.03452	4.7099	0.00%	0.00%	24.80%	75.20%
EWP							
0.916%	1.002	− 0.00183	− 1.496				

Table 22 Robustness check results for the short-term portfolio

ER	SD	Sharpe Ratio	Utility	002821.SZ	600,713.SH	600,129.SH
8.133%	1.051	0.05123	4.8192	23.88%	0.00%	76.12%
EWP Q1						
− 6.239%	0.570	− 0.14027	− 0.5489			
EWP Q2						
4.058%	0.507	0.04556	− 0.3442			
EWP Q3						
− 8.926%	0.539	− 0.19821	− 0.5245			
EWP Q4						
6.662%	0.474	0.10364	− 0.2704			

where β_i is the beta of the holding and r_m is the average expected return of the market. From Table 19, the short-term optimal portfolios determined by the Treynor Ratio and Jensen Alpha differ considerably from those based on the Sharpe Ratio and mean–variance utility from Table 18. In Table 20, the long-term optimal portfolios determined by the Treynor Ratio and Jensen Alpha are consistent with the results in Table 17. In summary, the optimal long-term portfolio decisions do not vary significantly if different metrics are applied, whereas the optimal short-term portfolio can be completely diverse based on various portfolio profitability evaluation methods.

Robustness check

In addition to constructing long- and short-term portfolios, we conducted a robustness check by comparing our results in Section “Long and short term portfolio” and with traditional equally weighted portfolios. Specifically, for the short-term portfolio, we partitioned 2021 into four quarters and compared them for each quarter. Tables 21 and 22 display the robustness check results for the long- and short-term portfolios, respectively. From the tables, our results from Section “Long and short term portfolio” are clearly better than those from the equally weighted portfolio (EWP). Although EWP had a higher Sharpe ratio, our results lead to higher portfolio returns and utility.

Conclusion

This study employed factor and cluster analyses to examine 165 listed companies from 2018 to 2021. Based on these results, we assessed the overall performance of various biopharmaceutical sector stocks, estimated the impact of COVID-19 on different stock types, and offer recommendations to investors.

The main results of this study are as follows. First, the biopharmaceutical industry stocks maintained an upward momentum during the pandemic. Specifically, medical devices, medical services, biological products, and chemical pharmaceuticals increased, whereas Chinese medicine and pharmaceutical commerce declined. Unlike previous studies, we explored the dynamics of second-level classification stocks during the COVID-19 pandemic. Second, regarding the clustering results, we conclude: Category 1 stocks are suitable for short-term holdings by unconfident investors. Categories 2 and 3 stocks are medium to long-term holdings for cautious investors. The fourth category is ideal for confident investors investing in forward contracts. Category 5 is suitable for long-term holdings by cautious investors. Category 6 is suitable for cautious stockholders. Category 7 is ideal for short-term holdings by confident investors. We considered both the degree of risk aversion and holding period of stock investors. Finally, we developed an optimal investment strategy using the Sharpe ratio and mean–variance utility, and provided alternative performance metrics, including the Treynor Ratio and Jensen Alpha, for comparison. For long-term investments, we recommend that investors allocate 24.8% of their wealth in 300,573.SZ and 75.2% in 603,127.SH to achieve the highest Sharpe ratio. We suggest a portfolio allocation of 26.99% in 300,573.SZ and 73.01% in 603,127.SH for the best mean–variance utility. For short-term investments, we recommend a portfolio allocation of 23.88% in 600,173.SH and 76.12% in 601,607.SH for the highest Sharpe ratio, and 28.57% in 600,173.SH and 71.43% in 601,607.SH for the best mean–variance utility. We provide investment suggestions tailored to investors with different preferences for portfolio analysis metrics.

This study had some limitations. For example, this study focused only on the stock market during the pandemic. Other financial assets such as bonds and futures were not included. Second, the analysis was restricted to data from China, despite the global impact of COVID-19, suggesting the potential relevance of examining financial markets in other countries.

Future research could broaden the scope to include additional financial assets as the pandemic progresses in China. In addition, given the significant effects of the pandemic on countries such as the US and India, a comparative study of their biopharmaceutical sectors relative to China's may yield valuable insights. Furthermore, advances in research methodologies may allow the application of sophisticated machine learning and deep learning techniques, such as DBN (Deep Belief Network), LSTM (Long Short-Term Memory), and support vector machines, to explore nonlinear relationships in the Chinese stock market. While some studies, such as Leippold et al. (2022), have applied nonlinear machine learning methods to the Chinese stock market, few have specifically addressed nonlinear dynamics within sectors such as the biopharmaceutical industry.

Appendix A: Correlation matrix of financial indicators

In this section, we show the correlation coefficient matrix of financial indicators in 2018, 2019, 2020 and 2021.

See Tables 23, 24, 25, 26.

Table 23 Correlation coefficient matrix of financial indicators in 2018

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	1.000											
x_2	0.441	1.000										
x_3	0.012	0.093	1.000									
x_4	0.003	-0.015	-0.002	1.000								
x_5	0.017	0.03	-0.087	-0.46	1.000							
x_6	0.174	0.068	-0.044	-0.4115	0.895	1.000						
x_7	-0.011	-0.012	0.005	0.934	-0.384	-0.342	1.000					
x_8	0.243	0.245	-0.034	0.073	0.348	0.288	0.12	1.000				
x_9	0.253	0.136	0.039	-0.007	0.009	0.08	-0.015	-0.052	1.000			
x_{10}	0.202	0.134	0.021	-0.075	0.100	0.162	-0.074	0.051	0.317	1.000		
x_{11}	0.402	0.809	0.187	-0.083	0.117	0.133	-0.064	0.332	0.091	0.145	1.000	
x_{12}	0.402	0.809	0.187	-0.082	0.117	0.133	-0.063	0.332	0.091	0.145	1.000	1.000

The table shows the correlation coefficients matrix for the 12 financial indicators in 2018. The correlation coefficients above 0.7 are highlighted

Table 24 Correlation coefficient matrix of financial indicators in 2019

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	1.000											
x_2	0.597	1.000										
x_3	0.028	0.083	1.000									
x_4	-0.041	-0.032	-0.111	1.000								
x_5	0.069	0.095	0.096	-0.491	1.000							
x_6	0.085	0.082	0.081	-0.46	0.961	1.000						
x_7	-0.029	-0.02	-0.091	0.962	-0.451	-0.422	1.000					
x_8	0.236	0.271	0.054	0.125	0.299	0.281	0.162	1.000				
x_9	0.217	0.099	0.045	0.022	0.008	-0.007	0.027	0.165	1.000			
x_{10}	0.104	0.144	0.065	-0.076	0.081	0.070	-0.070	0.12	0.317	1.000		
x_{11}	0.618	0.777	0.138	-0.079	0.106	0.101	-0.060	0.322	0.19	0.196	1.000	
x_{12}	0.619	0.778	0.138	-0.079	0.107	0.102	-0.06	0.323	0.188	0.195	1.000	1.000

The table shows the correlation coefficients matrix for the 12 financial indicators in 2019. The correlation coefficients above 0.7 are highlighted

Table 25 Correlation coefficient matrix of financial indicators in 2020

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	1.000											
x_2	0.27	1.000										
x_3	0.03	-0.016	1.000									
x_4	-0.092	-0.100	-0.018	1.000								
x_5	0.164	0.127	-0.026	-0.500	1.000							
x_6	0.191	0.119	-0.021	-0.48	0.973	1.000						
x_7	-0.085	-0.08	-0.019	0.967	-0.472	-0.451	1.000					
x_8	0.010	0.168	-0.147	0.184	0.044	0.015	0.213	1.000				
x_9	0.479	0.186	0.141	-0.240	0.265	0.305	-0.234	-0.460	1.000			
x_{10}	0.384	0.340	0.049	-0.079	0.158	0.196	-0.066	0.184	0.429	1.000		
x_{11}	0.361	0.266	0.079	-0.079	0.064	0.076	-0.079	-0.413	0.366	0.051	1.000	
x_{12}	0.490	0.663	0.123	-0.101	0.093	0.082	-0.092	-0.094	0.341	0.228	0.689	1.000

The table shows the correlation coefficients matrix for the 12 financial indicators in 2020. The correlation coefficients above 0.7 are highlighted

Table 26 Correlation Coefficient Matrix of Financial Indicators in 2021

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	1.000											
x_2	0.434	1.000										
x_3	0.098	0.051	1.000									
x_4	-0.079	-0.103	-0.127	1.000								
x_5	0.174	0.164	0.087	-0.408	1.000							
x_6	0.179	0.163	0.104	-0.385	0.989	1.000						
x_7	-0.070	-0.084	-0.111	0.971	-0.395	-0.372	1.000					
x_8	0.246	0.252	0.071	0.072	0.275	0.267	0.124	1.000				
x_9	0.668	0.335	0.148	-0.088	0.252	0.267	-0.091	0.209	1.000			
x_{10}	0.459	0.499	0.024	-0.155	0.143	0.158	-0.143	0.186	0.726	1.000		
x_{11}	0.198	0.715	0.091	-0.095	0.074	0.080	-0.085	0.233	0.243	0.243	1.000	
x_{12}	0.197	0.714	0.092	-0.096	0.075	0.081	-0.086	0.234	0.243	0.243	1.000	1.000

The table shows the correlation coefficients matrix for the 12 financial indicators in 2021. The correlation coefficients above 0.7 are highlighted

Appendix B: Determination of the number of clusters

When performing cluster analysis, parameter calls are first taken to determine the number of clusters. The inflection point is often used as the basis for the selection of the number of clusters, and as the number of clusters K increases, the sample is divided more finely and the degree of aggregation of each cluster is gradually increased.

Taking K less than 4, the classification is relatively coarse and less referential. Taking $K=6$, 141 out of 165 stocks belong to the same category and there are four subgroups with only a small number of stocks. This classification method leads to less referential results. Therefore $K=7$ is taken.

The tables below show the clustering data when the number of clusters is 6 and 7.

See Tables 27, 28.

Table 27 Classification with a clustering number of 6

Number of classifications	Number of cases
1	141
2	19
3	1
4	1
5	2
6	1

Table 28 Classification with a clustering number of 7

Number of classifications	Number of cases
1	1
2	1
3	36
4	2
5	1
6	112
7	12

Abbreviations

COVID-19	Coronavirus disease 2019
SOM	Self organizing maps
KMO test	Kaiser–Meyer–Olkin test
R&D expenditure	Research and development expenditure
xxxxx.SH	Stock from Shanghai Exchange
xxxxx.SZ	Stock from Shenzhen Exchange
TR	Treynor ratio
JA	Jensen Alpha
EWP	Equally weighted portfolio
ER	Expected return
SD	Standard deviation
DBN	Deep belief network
LSTM	Long short-term memory

Author contributions

Study conception and design: Conghua wen; data collection: Yifang Tang, Feifan Zhao; analysis and interpretation of results: Jiahui Xi, Conghua Wen; draft manuscript preparation: Jiahui Xi, Yifan Tang, Conghua Wen. All authors reviewed the results and approved the final version of the manuscript.

Funding

XJTLU Postgraduate Research Scholarship-PGRS2012016.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in Wind Database.

Declaration**Competing interests**

No competing interests exist for the publication.

Received: 22 February 2023 Accepted: 30 May 2024

Published online: 03 August 2024

References

- AlAli MS (2020) The effect of who COVID-19 announcement on Asian Stock Markets returns: an event study analysis. *J Econ Business* 3(3):1051–1054
- Ayati N, Saiyarsarai P, Nikfar S (2020) Short and long term impacts of COVID-19 on the pharmaceutical sector. *DARU J Pharm Sci* 28:799–805
- Baker HK, Haslem JA (1974) The impact of investor socioeconomic characteristics on risk and return preferences. *J Bus Res* 2(4):469–476
- Bai J, Ng S (2006) Evaluating latent and observed factors in macroeconomics and finance. *J Econ* 131(1–2):507–537
- Cevik E, Kirci Altinkeski B, Cevik El, Dibooglu S (2022) Investor sentiments and stock markets during the COVID-19 pandemic. *Financ Innov* 8(1):69
- Cox J, Greenwald DL, Ludvigson SC (2020) What explains the COVID-19 stock market? (No. w27784). National Bureau of Economic Research
- Daglis T, Melissaropoulos IG, Konstantakis KN, Michaelides PG (2022) The impact of COVID-19 on global stock markets: early linear and non-linear evidence for Italy. *Evol Inst Econ Rev* 19(1):485–495
- Esparcia C, López R (2022) Outperformance of the pharmaceutical sector during the COVID-19 pandemic: global time-varying screening rule development. *Inf Sci* 609:1181–1203
- Gorman BS, Primavera LH (1983) The complementary use of cluster and factor analysis methods. *J Exp Educ* 51(4):165–168
- Gupta V, Santosh KC, Arora R, Ciano T, Kalid KS, Mohan S (2022) Socioeconomic impact due to COVID-19: an empirical assessment. *Inf Process Manag* 59(2):102810
- Hao Y, Liu H, Chen H, Sha Y, Ji H, Fan J (2019) What affect consumers' willingness to pay for green packaging? Evidence from China. *Resour Conserv Recycl* 141:21–29
- Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C (appl Stat)* 28(1):100–108
- He P, Sun Y, Zhang Y, Li T (2020) COVID-19's impact on stock prices across different sectors—an event study based on the Chinese stock market. *Emerg Mark Financ Trade* 56(10):2198–2212
- Ho KC, Chen C, Yang D, Gao Y (2022) Medical reform, Covid-19 vaccine and stock returns: the case of Chinese listed pharmaceutical and healthcare companies. *Appl Econ Lett* 31:832–839
- Jones CS (2006) A nonlinear factor analysis of S&P 500 index option returns. *J Financ* 61(5):2325–2363
- Kim JO, Mueller CW (1978) Introduction to factor analysis: What it is and how to do it, vol 13. Sage, England
- Leippold M, Wang Q, Zhou W (2022) Machine learning in the Chinese stock market. *J Financ Econ* 145(2):64–82
- Liu H, Wang Y, He D, Wang C (2020) Short term response of Chinese stock markets to the outbreak of COVID-19. *Appl Econ* 52(53):5859–5872
- Ludvigson SC, Ng S (2007) The empirical risk–return relation: a factor analysis approach. *J Financ Econ* 83(1):171–222

- Morelli D (1999) Tests of structural change using factor analysis in equity returns. *Appl Econ Lett* 6(4):203–207
- Nanda SR, Mahanty B, Tiwari MK (2010) Clustering Indian stock market data for portfolio management. *Expert Syst Appl* 37(12):8793–8798
- Piñero-Chousa J, López-Cabarcos MÁ, Quiñoá-Piñero L, Pérez-Pico AM (2022) US biopharmaceutical companies' stock market reaction to the COVID-19 pandemic. Understanding the concept of the 'paradoxical spiral' from a sustainability perspective. *Technol Forecast Soc Chang* 175:121365
- Robke L, Pont LB, Bongard J, Wurzer S, Smietana K, Moss R (2020) Impact of COVID-19 on pharmaceutical external innovation sourcing. *Nat Rev Drug Discov* 19(12):829–830
- Seong N, Nam K (2021) Predicting stock movements based on financial news with segmentation. *Expert Syst Appl* 164:113988
- Sharpe WF (1998) The sharpe ratio. *Streetwise Best J Portfolio Manag* 3:169–185
- Takvi PO, Bentum-Ennin I (2021) The impact of COVID-19 on stock market performance in Africa: a Bayesian structural time series approach. *J Econ Bus* 115:105968
- Thukral E (2021) COVID-19: small and medium enterprises challenges and responses with creativity, innovation, and entrepreneurship. *Strateg Chang* 30(2):153–158
- Vats P, Samdani K (2019) Study on machine learning techniques in financial markets. In: 2019 IEEE international conference on system, computation, automation and networking (ICSCAN). IEEE, pp 1–5
- Wagenvoort RJ, Ebner A, Borys MM (2011) A factor analysis approach to measuring European loan and bond market integration. *J Bank Financ* 35(4):1011–1025
- Wagner AF (2020) What the stock market tells us about the post-COVID-19 world. *Nat Hum Behav* 4(5):440–440
- Wang Y, Ma F, Wei Y, Wu C (2016) Forecasting realized volatility in a changing world: a dynamic model averaging approach. *J Bank Financ* 64:136–149
- Wang YJ, Lee HS (2008) A clustering method to identify representative financial ratios. *Inf Sci* 178(4):1087–1097
- Zou Z, Wang X (2023) Research on the investment value of China's medical sector in the context of COVID-19. *Econ Res-Ekonomska Istraživanja* 36(1):614–633

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.