Open Journal of Statistics, 2018, 8, 14-24
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

The Application of Robust Statistics to China's Stock Market

Xiongying Li[1]*, Yaomin Zhang[1], Bin Yan[2]

[1]School of Mathematics and Statistics, Guangdong University of Finance and Economics, Guangzhou, China
[2]College of Economics, Jinan University, Guangzhou, China
Email: *lixiongying2818@163.com

How to cite this paper: Li, X.Y., Zhang, Y.M. and Yan, B. (2018) The Application of Robust Statistics to China's Stock Market. *Open Journal of Statistics*, **8**, 14-24.
https://doi.org/10.4236/ojs.2018.81002

Received: November 3, 2017
Accepted: January 29, 2018
Published: February 1, 2018

Copyright © 2018 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Portfolio theory is used to measure the expected return and risk on the basis of the return ratio, but in fact there is always excessively high or low return ratio caused by some short-term fundamental good or bad news in the history data of return ratio. We introduce the robust statistic idea into the portfolio theory in this paper, thus reduce outliers' influence on portfolio decision in the history data of return ratios, and bring back the portfolio on its long-term investment value track. We focused on the robust estimate method and apply them to solution processing in the portfolio model and obtained good results.

Keywords

Portfolio, Outlier, Robust Estimate, Robust Regression

1. Introduction

As for outliers in social science data, we cannot simply delete them when we deal with the data on natural science, and robust statistics method (RSM) can overcome the influence on the final result under the condition that data will not be deleted, which is a useful multivariate analysis method for exploratory analysis in social science researches. With the development of RSM, more and more scholars concentrate on using RSM thought to optimize a model, and it also has been a hotspot these years. In reality, among the data of security's history return ratio there are excessively high or low ones incurred by fundamental good or bad news, so when we estimate their expected returns and risks by the history data of return ratio, the portfolio constructed by classic methods will deviate from their its actual investment value to influence the decision on portfolio. How should we do if this happened? We always concern on this focus. Classic statistics for data description or data distribution property are not so representative in many cases that analysis outcome is not inconsistent with the fact [1]. What lead this to happen is due to that classic statistic methods are heavily dependent the assumption of normal distribution of researched data, but when acquired data is not or incompletely from normally distributed population. i.e., there are some outliers [2]. Once classic statistic method is used to describe the researched object, there must be some deviation, sometimes even enormously big deviation. Some researchers tell that normal distribution is theoretical while is often normal that actual data deviate from normality assumption, or utmost approximate to normal distribution [3], there is some skewness existing among normal distribution which causes to an fatal influence on robustness of classic statistic method [4].

Robust statistic method is the statistic one with robust property including two sides [5]. One is with the characteristic of anti-disturbance, that means the method still keep good statistic performance when actual model differs little from theoretical assumption, the other is the estimate performance can still be acquired and not be destructively influenced when actual model differs much form theoretical assumption [6]. The former side means one robust statistic method must perform well in assumed mod el and the around, this guarantee that the statistic model is approximate correct and this approach to the desirable conclusion is the best or nearly when there are few of outliers among the data. The latter side means that some bad cases could be prevented, such as that robust statistic method could not perform poor or lead to completely wrong conclusion when the assumed model differs much from the fact or there are many outliers on database.

2. Literature Review

In the early nineteen century when Gauss proposed normal distribution and ordinary least squares, robust statistic idea spring up, at later some researchers found some actual samples did not follow normal distribution if there are outliers among the collective data. Limited to the complication of the robust statistic method itself and computing technology, robust statistic always underwent its embryonic stage for nearly one and a half century until nineteen-fifties [7]. In 1953, G. E. P. Box introduce robustness concept for the first time, but limited to plain idea and simple method. It was W. Tulay that made the statistic circle concentrated on the robust statistic in the early nineteen-sixties, he researched back and forth the non-robustness of the classic statistic methods since nineteen-forties and started to make certain the good robust property of the estimated method such as trimmed mean and mean absolute deviation. In nineteen sixty-four, P. J. Huber published an innovative paper with the title as *the robust estimated at location parameter* in which he proposed moment estimation as one of robust estimation at location parameter and solved the corresponding problem of asymptotic maximum and minimum [8] [9] [10] [11] [12].

This paper marked the beginning of systematic research on robust statistic. In nineteen eighty-one, Huber published another statistic book named as *robust statistic* in which he defined the robust statistic formally, robust statistic theory just grew up until now. Since that, research on robust statistic progressed much further. On board, researchers focused on constructing multivariate location and scatter, high break-down point and high-efficiency estimate in linear regression and test's break-down property. Since robust statistic has extensive field, it could progress further in terms of classic statistic method in case the fact deviates from the assumption, so it become necessary to used robust statistic method to improve classic one.

3. Fast-MCD Robust Estimate Model

In nineteen eight-four, Rousseeuw suggested minimum covariance determinant as multivariate robust estimate method, but limited to its complicate algorithm and computing technology, this method did not prevail even though it had strong robustness. After that, Rousseeuw & Van Driessen (1999) improved minimum covariance determinant and suggested Fast-MCD, sped up computing largely. We will estimate robust expected return ratio vector and covariance matrix on basis of Fast-MCD.

Fast-MCD constructs on robust covariance matrix estimator by iteration and mahalanob is distance. This progress can be as follows: on a matrix $X_{n\times p}$ with p lines and n columns, *i.e.* return ratio data of p pieces of stocks in n periods, draw h samples and compute its mean T_1 and covariance matrix S_1, then reckon mahalanob is distance from n samples to their center T_1 by the formula $d_1(i)=\sqrt{(x_i-T_1)^{\mathrm{T}} S_1^{-1}(x_i-T_1)}$, choose the smallest h distances, get the sample mean T_1 and covariance matrix S_1 by these h samples, it can be proved that there is $\text{delta}(S_1)\le\text{delta}(S_2)$, they will equate if and only if $T_1=T_2, S_1=S_2$. Likewise, the iteration in this process go on and on until $\det(S_m)=\det(S_{m-1})$.

Specifically speaking, it is on the basis that we construct robust mean vector and covariance matrix by Fast-MCD through the following procedure.

1) Make certain h value by $h=n*a$, a is drawing ratio with a value range from 0.5 to 1. more smaller is a, more stronger is it to resist outliers, but it is not smaller than 50% because outliers and normal values can not be differentiated at this critical point, so its default value is 0.75 generally, otherwise 0.9 if sample quantity is not enough.

2) Compute covariance matrix and its determinant by randomly drawing p + 1 samples from n samples. If the determinant value is zero, the another random sample shall be jointed into the previous drawn sample until its determinant is not zero. At this time, computed covariance matrix become initial covariance matrix S_0, we can get initial sample mean T_0 by that random sampling.

3) If n is smaller than 600, we can get mahanobis distance from n sample data to their center T_0 by the formula $d_1(i)=\sqrt{(x_i-T_1)^{\mathrm{T}} S_1^{-1}(x_i-T_1)}$, find the smallest h distance value as the initial h, then compute samples mean T_1 cova-

riance matrix S_1 of these h sample data, and get S_3 after two iterations through C-step process.

4) Repeat the above procedure 500 times to get 500 values of S_3, among them we choose 10 groups of h values of the smallest $\text{delta}(S_3)$,and go on with iteration until convergence through C-step process, then go back to T and S of the group in which h made $\det(S_m)$ become the smallest, remark them as T_{MCD} and S_{MCD} respectively.

5) If n is bigger, we can classify n samples into parts. For example, n samples can be divided into 5 sub-sample groups if n is 1500, so each sub-sample group has 300 samples. We get T_1 & S_1 from T_0 & S_0 in each group and start at T_0 & S_0 to iterate for 2 times to acquire S_3 through C-step process. Thus, one hundred of S_3 can be gotten after repeating 100 times for each sub-sample group. After the ten smallest S_3 are selected from each sub-sample groups, all sub-sample groups are incorporated into one full sample size while ten S_3 from sub-samples are done so to get fifty S_3, then iterate twice by 50 groups of h samples matched at 50 S_3, keep 10 groups of h which make the determinants of covariance matrices the smallest after iteration, and keep up with iterating until convergence, at last return to T & S of which group h make $\text{delta}(S_m)$ the smallest and mark them up as T_{MCD} and S_{MCD}.

6) Based in T_{MCD} and S_{MCD}, compute each sample's robust mahanob is distance $d(i)$. Since computed $d(i)$ follows approximately Chi-square distribution with p freedom degree, remark $\omega_i = 0$, otherwise $\omega_i = 1$ when $d(i) > \sqrt{x^2_{p,0.975}}$, then reckon T value according to ω_i:

$$T = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i}, \quad S = \frac{\sum_{i=1}^{n} \omega_i (x_i - T)(x_i - T)'}{\sum_{i=1}^{n} \omega_i - 1}$$

At that time, T & S are respectively final robust mean vector and robust covariance matrix.

Thus we get robust mean vector and covariance matrix by Fast-MCD, then if we substitute respectively acquired robust mean vector T and robust covariance matrix S into Markovits mean-variance portfolio model as μ & $\sum$ in the formula, Markovits mean-variance portfolio model become as:

$$\begin{aligned} &\min_X X'SX \\ s.t.\quad &\mu_p = X'T \geq r \\ &X'I = 1 \\ &X \geq 0 \end{aligned}$$

4. Empirical Research

1) The comparison of variance, contribution rate and factor loading matrix

To prove that robust factor analysis method's results are more accurate than traditional method's when there exists outliers in the data, we choose two groups of enterprise annual financial index data of China's listed companies on December 31, 2016, with a set of enterprises in good financial condition (called normal group), sample No. 1 to No. 32, another set of financial data of bankrupt enterprises (outliers), sample No. 33 to No. 36. Seven major indexes are selected: the X_1 (flow rate), X_2 (ratio of working capital), the X_3 (working capital to total assets ratio) and X_4 (operating profit margin), X_5 (sales net interest rates), X_6 (total assets net profit margin), X_7 (net income), the data is shown in Table 1.

When doing factor analysis on these variables, we hope that there are a certain degree of correlation between these variables, for either too high or too low correlation is not conducive to doing factor analysis. In the first case, a high correlation always leads to an obvious multicollinearity, thus the obtained factor's structure is not stable. The variables are not suitable for doing factor analysis. In another case, it's difficult to extract a set of stable factors in the condition of too low correlation between the variables, and the variables are also not suitable for doing factor analysis. Based on this understanding, we use the KMO and Bartlett test firstly to determine whether these variables are suitable for doing factor analysis.

KMO (Kaiser-Meyer-Olkin) test statistic, who values between 0 and 1, is used to compare the simple correlation coefficient and partial correlation coefficient between variables. A great KMO indicates that the correlation between variables

Table 1. The enterprises' annual financial index data.

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	1.40	0.28	0.27	0.20	0.15	0.04	0.20	19	2.24	0.55	0.49	0.07	0.06	0.01	0.03
2	1.71	0.41	0.34	0.27	0.20	0.07	0.18	20	2.31	0.57	0.39	0.01	0.05	0.00	0.02
3	1.39	0.28	0.23	0.26	0.20	0.09	0.25	21	1.22	0.18	0.15	0.31	0.21	0.04	0.15
4	1.48	0.32	0.29	0.10	0.08	0.02	0.05	22	2.39	0.58	0.28	0.08	0.06	0.02	0.02
5	1.58	0.37	0.35	0.25	0.17	0.04	0.14	23	1.03	0.03	0.02	0.01	0.00	0.00	0.03
6	2.27	0.56	0.46	0.13	0.10	0.03	0.07	24	1.29	0.22	0.19	0.07	0.06	0.05	0.17
7	1.56	0.36	0.33	0.13	0.09	0.02	0.11	25	1.13	0.11	0.10	0.02	0.02	0.02	0.12
8	3.21	0.69	0.55	0.33	0.25	0.02	0.03	26	1.19	0.16	0.12	0.02	0.02	0.01	0.09
9	1.46	0.32	0.31	0.06	0.01	0.00	0.01	27	1.86	0.46	0.35	0.14	0.11	0.02	0.06
10	1.63	0.39	0.32	0.32	0.23	0.07	0.17	28	1.35	0.26	0.21	0.05	0.04	0.03	0.15
11	2.30	0.56	0.48	0.21	0.15	0.05	0.14	29	1.37	0.27	0.22	0.30	0.21	0.04	0.11
12	1.33	0.25	0.17	0.16	0.10	0.03	0.14	30	1.77	0.43	0.41	0.32	0.24	0.13	0.24
13	4.17	0.76	0.74	0.22	0.18	0.03	0.09	31	1.34	0.25	0.21	0.05	0.12	0.02	0.19
14	1.19	0.16	0.14	0.08	0.09	0.04	0.18	32	1.39	0.28	0.26	0.05	0.04	0.01	0.07
15	1.33	0.25	0.18	0.08	0.08	0.03	0.08	33	0.68	−0.12	−0.21	−0.72	−0.42	−0.19	−0.72
16	1.78	0.44	0.36	0.08	0.03	0.00	0.00	34	1.72	0.42	0.28	−1.30	−1.46	−0.07	−0.11
17	1.77	0.43	0.34	0.21	0.16	0.04	0.11	35	0.75	−0.33	−0.11	−1.30	−1.46	−0.07	−0.11
18	1.54	0.35	0.29	0.33	0.29	0.06	0.21	36	0.83	−0.21	−0.10	−0.62	−0.62	−0.09	−0.53

is strong, the variables are suitable for doing factor analysis. While a smaller KMO means that the correlation between variables is weaker, and the original variables are not that suitable for doing factor analysis. Generally we think variables who's KMO values greater than 0.6 are suitable for doing factor analysis. Bartlett test is used to test out whether a group of variables is related. If the overall correlation matrix is a unit matrix, then we accept the null hypothesis, suggesting that these variables are not suitable for doing factor analysis.

It can be seen from Table 2 that KMO values 0.78 in the correlation coefficient matrix $\boldsymbol{R}$, and the significant rate of Bartlett ball test chi-square statistic is 0.00, so we think these variables are suitable for doing factor analysis.

At first, we can obtain these variables' characteristic value and characteristic vector by doing traditional factor analysis on normal operated companies' financial data. Then we add in a bankrupted company's financial data (sample 33) and four bankrupted companies' financial data (sample 33 - 36) respectively, and do factor analysis on them separately. For normal operated companies' financial data, those bankrupted companies' financial data should be outliers. By doing factor analysis with traditional method and robust method respectively, we find that there are a certain gap between the results of traditional factor analysis and the results based on original data. On the other hand, the results of robust factor analysis method show off the characteristics of the original data better than that of traditional method, ignoring that there is a little discrepancy between the results of robust factor analysis method and the results based on original data. When we raise the number of outliers to 4 (sample 33 to 36), the above results remain valid. The specific results are shown in Table 3 and Table 4.

According to the judgment standard of choosing principal components which needing their eigenvalues are greater than 1 or cumulative contribution rate reaches more than 85%,we extracted two principal components which containing more than 85% information and representing the most information from indexes with the traditional factor analysis method and the robust factor analysis method. And their eigenvalues are greater than 1(basically achieved 2 above).

As shown in Table 3, when the samples don't contain outliers, the variance, the variance contribution ratio and the cumulated variance contribution ratio have little difference with rotation under the methods of the traditional factor analysis method and the robust factor analysis method: with the traditional factor analysis method, the variances of the two factors were 3.46 and 2.8, the variance contribution rate were 49.5% and 40.1%. And with the robust factor analysis

Table 2. KMO test and Bartlett's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.78
Bartlett's Test of Sphericity	Approx. Chi-Square	1162.9
	df	21
	Sig	0.00

Table 3. Traditional and robust factor analysis' variance, variance contribution rate and cumulative variance contribution rate before and after rotation.

Data	method	Before rotation			After rotation		
		variance	variance contribution rate %	accumulation%	variance	variance contribution rate %	accumulation %
Normal set of data	Traditional factor analysis	3.46	49.5	49.5	3.14	44.8	44.8
		2.80	40.1	89.5	3.13	44.7	89.5
	Robust factor analysis	3.61	51.6	51.6	3.26	46.6	46.6
		2.81	40.2	91.8	3.16	45.2	91.8
Data with an outlier	Traditional factor analysis	4.55	64.9	64.9	3.58	51.2	51.2
		2.00	28.6	93.5	2.96	42.3	93.5
	Robust factor analysis	3.35	47.9	47.9	3.26	46.6	46.6
		2.98	42.6	90.5	3.16	45.2	91.8
Data with four outliers	Traditional factor analysis	4.65	66.4	66.4	3.47	49.6	49.6
		1.62	23.2	89.6	2.80	39.9	89.6
	Robust factor analysis	3.61	51.6	51.6	3.26	46.6	46.6
		2.81	40.2	91.8	3.16	45.2	91.8

Table 4. Load matrix of traditional and robust factor analysis before and after rotation.

Data	Variable	Traditional factor analysis				Robust factor analysis			
		Before rotation		After rotation		Before rotation		After rotation	
		Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Normal set of data	X_1	0.665	−0.696	−0.0124	0.9623	0.687	−0.701	0.0505	0.9801
	X_2	0.714	−0.666	0.0439	0.9750	0.715	−0.688	0.0805	0.9890
	X_3	0.761	−0.595	0.1270	0.9577	0.678	−0.695	0.0478	0.9701
	X_4	0.840	0.395	0.8762	0.3059	0.873	0.332	0.8739	0.3294
	X_5	0.859	0.412	0.9016	0.3068	0.864	0.415	0.9217	0.2614
	X_6	0.614	0.685	0.9182	−0.0591	0.689	0.641	0.9408	−0.0234
	X_7	0.333	0.853	0.8352	−0.3761	0.433	0.820	0.8680	−0.3278
Data with an outlier	X_1	0.602	0.764	0.00473	0.9727	0.606	−0.772	0.0505	0.9801
	X_2	0.796	0.570	0.27701	0.9395	0.673	−0.730	0.0805	0.9890
	X_3	0.783	0.586	0.25685	0.9434	0.650	−0.740	0.0478	0.9701
	X_4	0.923	−0.285	0.90261	0.3434	0.830	0.413	0.8739	0.3294
	X_5	0.935	−0.292	0.91678	0.3449	0.852	0.403	0.9217	0.2614
	X_6	0.783	−0.505	0.92830	0.0834	0.725	0.589	0.9408	−0.0234
	X_7	0.772	−0.571	0.96012	0.0251	0.413	0.792	0.8680	−0.3278
Data with four outliers	X_1	0.655	0.717	0.0657	0.969	0.687	−0.701	0.0505	0.9801
	X_2	0.849	0.478	0.3664	0.903	0.715	−0.688	0.0805	0.9890
	X_3	0.837	0.514	0.3351	0.923	0.678	−0.695	0.0478	0.9701
	X_4	0.875	−0.361	0.9094	0.262	0.873	0.332	0.8739	0.3294
	X_5	0.850	−0.369	0.8952	0.240	0.864	0.415	0.9217	0.2614
	X_6	0.845	−0.434	0.9312	0.187	0.689	0.641	0.9408	−0.0234
	X_7	0.771	−0.401	0.8536	0.166	0.433	0.820	0.8680	−0.3278

method, the variances of the two factors were 3.61 and 2.81, the variance contribution rate were 51.6% and 40.2%. After rotation, with the traditional factor analysis method, the variances of the two factors were 3.14 and 3.13, the variance contribution rate were 44.8% and 44.7%, with the robust factor analysis method, the variances of the two factors were 3.26 and 3.16, the variance contribution rate were 46.6% and 45.2%. In a word, the results of those two methods are indifferent to rotation.

When an outlier exists in the sample data (sample 33), there are two different results between two methods. Before rotation, the variance of two factors is 4.55, 2.00 by using traditional method, and the variance contribution rate is 64.9%, 28.6% respectively. While the variance of two factors is 3.35, 2.98 by using robust method, and the variance contribution rate 47.9%, 42.6% respectively. That is to say, when using traditional method, the variance of factor 1 reaches 4.55, and that of factor 2 only 2.00, meanwhile the corresponding variance contribution rates have deviation. Bu contrast, the variance and contribution rate, which are calculated by using robust factor analysis, are almost the same whether there are outliers or not. However, the results by using two method after rotation are all better than the results before rotation.

When four outliers exist in the sample data (sample 33 - 36), comparing with the above two cases (sample data with no outlier and sample data with just one outlier), there is just a subtle difference between results before and after rotation by using robust factor analysis method, while a notable difference between results by using traditional method. For example, the variance of two factors calculated by using traditional method before rotation is 4.65 and 1.62, the variance of factor 2 is merely 1.62, and the variance contribution rate reaches only 23.2%, which is much different from the value in the condition of no outlier (the variance of two factors is 3.46, 2.80, and the variance and the variance contribution rate of factor 2 is 2.80 and 40.1% respectively). In addition, the variance contribution rate of factor 1 reaches 66.4%. This suggests that outliers affect a lot before rotation when using traditional method, leading to a difference between the calculated results and the real value. While after rotation, the results obtained from traditional are closed to that with no outlier. It can be said that the results based on robust factor analysis method really can have certain resistance to outliers, and this method is stable.

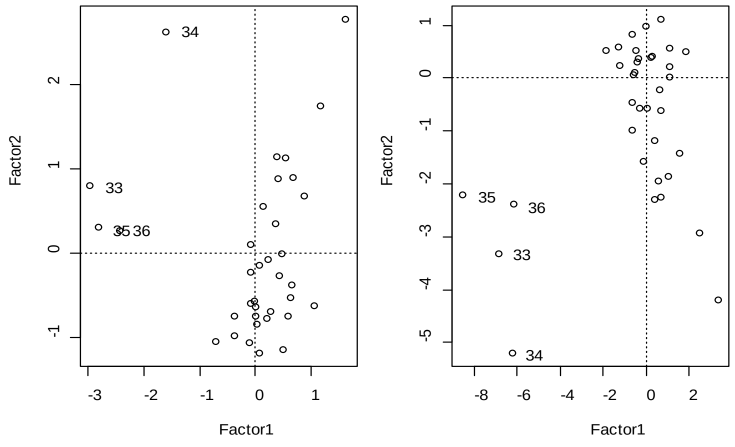
According to the Table 4, the results with no outlier are much different from the results with outliers when using traditional factor analysis method, while there is no significant difference between these two results when using robust method. Before rotation, there is just a subtle difference between the results of traditional method and robust method in the condition of no outlier, while an outlier (sample 33) is added in the sample data, the results by using traditional method is much different from the results with no outlier. For example, the load values of factor 2 in the X_1, X_2 and X_3 are all positive, while turning into negative in the condition of no outlier in the data. The load values of factor 2 in the X_4, X_5,

X_6 and X_7 are all negative, while becoming positive in the condition of no outlier in the data. By contrast, there is no significant difference between the load values of X_1 to X_7 with no outlier and the load values with outliers by using robust method. When there exists four outliers (sample data 33 to 36), the results by using traditional method is much different from the results with an outlier. The symbol of load value of factor 2 in every variable has changed. Similar to the above analysis, after rotation, there is certain deviation between the results with outliers and the results with no outlier by using traditional factor analysis method. While when using robust method, the outliers really do not have a significant impact on the result. This further illustrates that robust factor analysis has strong anti-interference ability, and can resist the influence of outliers effectively.

2) The analysis of factor score map

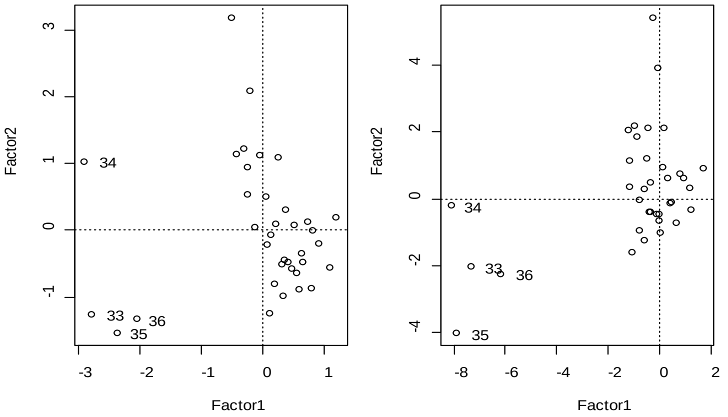
Now we can obtain the factor score map of the above 36 enterprises' annual financial indicators, which is based on Factor1 and Factor2 coordinate axis, by using the traditional factor analysis method and robust factor analysis method. The specific results are shown in Figure 1 and Figure 2.

In general, considering of liquidity ratio, ratio of working capital, working capital to total assets ratio, operating profit margin, sales net interest rate, total assets, net profit margin and net assets yield. Most of them are positive in a normal enterprise, at the same time, the vast majority of these values of a bankrupt enterprise is negative. As can be seen from Figure 1(a), before factor rotation, the financial data of four bankrupt enterprises is in the second quadrant, and the score of data 34 on the second factor is relatively large positive. The score of financial data of other three bankrupt enterprises on Factor2 is also positive. The above score doesn't tally with the actual situation. According to the Figure 1(b), before factor rotation, the financial data of four bankrupt enterprises is in the third quadrant, namely, the score on the factor 1, as well as on the factor 2 is negative. In addition, the financial data of these four enterprises



(a) (b)

Figure 1. (a) The traditional factor score chart before rotating; (b) The solid factor score chart before rotating.



(a) (b)

Figure 2. (a) The traditional factor score chart after rotating; (b) The robust factor score chart after rotating.

deviate far from other normal business enterprises', this result is in line with the actual situation. In the same way, after the rotation, as we can see from the factor score map based on the robust factor analysis method, the financial data of these four enterprises is on the third quadrant (**Figure 2(b)**), however, according to the factor score map based on the traditional factor analysis method, one of the bankrupt enterprises' financial data (34) is on the second method (**Figure 2(a)**), which is obviously inconsistent with the facts. Comparing with these two factor score maps, we know that we can detect the outliers effectively by using the robust factor analysis method, and the result will not be affected by outliers.

5. Conclusions

The upper empirical comparison based on both traditional and robust factor analysis method shows that the existence of outliers can affect our judgment of economic phenomenon and trend seriously. It's necessary for us to detect outliers in data preprocessing stage, so that the economic theoretical model can conform to the laws that most of the data shows. In factor analysis stage, it tends to cause that model fitting results do not consistent with actual situation, and even lead to large deviation if we use the traditional method. Therefore, it's essential to construct a robust statistic to overcome the influence of the outliers; this article's contribution to the factor analysis is established on this point.

In this paper, we advance a robust algorithm based on traditional factor analysis method, the empirical comparison show that, when there exists outliers in the sample data before and after factor rotation, the characteristic value and factor loading based on traditional factor analysis method will change according to the number of outliers. Hence, the latter method has a better effect in dealing with outliers in the sample data. The reason is that the covariance matrix is easily affected by outliers, and the eigenvalue and eigenvector, which are calculated according to covariance matrix, is sensitive to outliers too, thus leading to deviation in the results. While when we constructs a robust covariance matrix firstly by using robust factor analysis method, thus reducing the influence of outliers. The eigenvalue and eigenvector calculated by that are less sensitive to outliers, thus affecting less to the results.

Funding

This project is funded by "Youth innovation talents program of education department of Guangdong province (humanities and social sciences) (approval Number: 2016WQNCX046). The project of the 13th five-year plan of the philosophical society of Guangdong province (approval number: GD16XYJ16), The project of the 13th five-year plan of the Guangzhou philosophy and social sciences (approval number: 2017GZQN11)".

References

[1] Guo, Y.F. (2007) Robust Statistic and Comparative Analysis of the Robustness of the Statistic. *Statistical Research*, **9**, 82-85.

[2] Sun, X.H. (2003) Robust Statistics in the Application of Economic Indicators and Its Revelation. *Modern Finance and Economics*, **12**, 36-38.

[3] Xie, Z.Z. (2013) The Single Index Model Robust Regression and Empirical. *Statistics and Decision*, **5**, 27-30.

[4] Wang, B.H. and Chen, Y.F. (2006) A Robust Principal Component Analysis Based on MCD Estimator and Its Empirical Study. *Application of Statistics and Management*, **25**, 462-468.

[5] Wang, B.H. (2007) Robust Principal Component Analysis Method and Its Application. *Statistical Research*, **24**, 72-76.

[6] Xie, Z.Z. (2013) Application of Robust Statistical Inference in Mean-Variance Model. *Journal of Toaghua Normal University*, **4**, 3-6.

[7] Zhao, L.M. (1992) Robust Statistics Application in Samples Containing Outliers. *The Environmental Monitoring Management and Technology*, **4**, 58-59.

[8] Huber, P.J. (1985) Projection Pursuit. *The Annals of Statistics*, **13**, 435-475. https://doi.org/10.1214/aos/1176349519

[9] Rousseeuw, P.J. (1984) Least Median of Squares Regression. *Journal of the American Statistical Assocaition*, **79**, 871-880.

[10] Rousseeuw, P.J. and Driessen, K.V. (1999) A Fast Algorithm for the Monimum Covariance Determinant Estimator. *Technimetrics*, **41**, 212-223.

[11] Hubert, M., Rousseeuw, P.J. and Vanden Branden, K. (2005) ROBPCA—A New Approach to Robust Principal Component Analysis. *Technometrics*, **147**, 64-79.

[12] Huber, P.J. (1964) Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, **35**, 73-101. https://doi.org/10.1214/aoms/1177703732