Open Journal of Social Sciences, 2014, 2, 12-18
Published Online July 2014 in SciRes. http://www.scirp.org/journal/jss
http://dx.doi.org/10.4236/jss.2014.27003

Scientific Research

Robust Factor Analysis and Its Applications in the CSI 100 Index

Yingying Zhang

Department of Statistics and Actuarial Science, Chongqing University, Chongqing, China
Email: robertzhangyying@qq.com

Received May 2014

Abstract

We apply the object-oriented robust factor analysis R package robustfa to the 28 financial indicators of the 100 listed companies in China's Chinese Securities Index (CSI) 100 index in the first quarter of 2013. First of all, according to the size of the data, we automatically choose a robust estimator, the robust Ogk estimator. By the Mahalanobis distances which are computed by the robust Ogk estimator, greater than the critical value, we find a total of 47 abnormal points. This paper discovers that the results of the sample correlation matrix, the rotated factor loading matrix, the contribution of the factors to the original variables, the contribution rate, the cumulative contribution rate, the screeplot of the eigenvalues of the sample correlation matrix, the scatter plot of the first two factor scores, factor scores, and the sorted scores according to factor scores etc. computed by the classical estimator and the robust Ogk estimator are quite different. Finally, we condense the 28 financial indicators to 5 factors by combining the principal component analysis method and the robust Ogk estimator: Provident fund market value factor, profit factor, market value profit rate factor, value per share factor, and asset liability factor. Finally, we sort the 5 factor scores from high to low of each factor, and also get some special stocks according to the factor scores. The robust factor analysis results provide a good basis for investors to choose the stocks.

Keywords

Robustness, Factor Analysis, R Software, CSI 100 Index, Financial Indicators

1. Introduction

The stock market is an important part of the rapid development of China's socialist market economic system. In order to achieve long-term stability and healthy development of stock market, we must strengthen the domestic stock market research and regulation. Listed companies are the foundation and objects of research and supervision of stock markets. Make an objective evaluation for operating performance of listed companies is one of the main bases for operation of securities companies and investors in investment decisions. So making a scientific and fair evaluation for listed companies is one of the main contents of the research on stock market.

The 100 stocks of the CSI 100 Index are the largest constituent stocks that are selected from the CSI 300 Index. It is a comprehensive reflection of overall state of the Shanghai and Shenzhen stock markets' influential capitalization companies. Therefore, it is of great practical significance to effectively evaluate the 100 listed

How to cite this paper: Zhang, Y.Y. (2014) Robust Factor Analysis and Its Applications in the CSI 100 Index. *Open Journal of Social Sciences*, **2**, 12-18. http://dx.doi.org/10.4236/jss.2014.27003

companies.

Factor analysis [1]-[4] is a dimension reduction method of multivariate statistical analysis. It is a model used to analyze the hidden factors in the phenomena. Factor analysis studies internal dependent relationships of the correlation matrix or the covariance matrix. It condenses the original variables into a few factors and displays the relationship between the original variables and factors.

Outliers virtually exist in any datasets of any application field. To avoid the impact of outliers, we need to use robust estimators. Classical estimators of multivariate mean and covariance matrix are the sample mean and the sample covariance matrix. Outliers will affect the sample mean and the sample covariance matrix, and thus they will affect the classical factor analysis which depends on the classical estimators [5]. So it is necessary to use the robust estimators of the sample mean and the sample covariance matrix. There are several robust estimators in the literature: MCD, OGK, MVE, M, S, and Stahel-Donoho. The most direct way to robustify multivariate analysis is to replace the sample mean and the sample covariance matrix of the classical estimators to robust estimators [6] [7], which is our choice of robust factor analysis. A base package for robust multivariate analysis is **robustbase** [8]. The package **robust** [9] has a large overlap with **robustbase**. The package **rrcov** [10] builts upon **robustbase**. The package **rrcov** provides many methods based on the S4 method, contains almost a complete set of estimators to compute the robust sample mean and robust sample covariance matrix, and also contains robust principal component analysis, robust linear discriminant analysis, and robust quadratic discriminant analysis. The package **robustfa** [11] used in this paper follows **rrcov**, and it is a package created to solve robust factor analysis.

2. Empirical Analysis

2.1. Sample Selection and Data Source

This paper selects 28 financial indicators of the 100 listed companies in China's Chinese Securities Index (CSI) 100 index in the first quarter of 2013 to do the empirical research. The data is downloaded from the great wisdom free software.

2.2. Roubst Factor Analysis

Here we use *FaCov*() function to calculate robust factor analysis. Its usage is:
FaCov(x, factors = 2, cor = FALSE, cov.control = CovControlMcd(), method = c("mle", "pca", "pfa"),
scores Method = c ("none", "regression", "Bartlett"), ...)

Where *x* is a numeric matrix or an object that can be coerced to a numeric matrix (such as a data frame), *factors* is the number of factors, the default value is 2. *cor* is a logical value indicating whether the calculation should use the covariance matrix (*cor = FALSE*, the default) or the correlation matrix (*cor = TRUE*). *cov.control* specifies which covariance estimator to use by providing a *CovControl-class* object. The default is *CovControlMcd-class* which will indirectly call *CovMcd*(). If *cov.control = NULL* is specified, the classical estimates will be used by calling *CovClassic*(). *method* is the method of factor analysis, one of "*mle*" (the default), "*pca*", and "*pfa*". *scoresMethod* specifies which type of scores to produce. The default is "*none*", "*regression*" gives Thompson's scores, and "*Bartlett*" gives Bartlett's weighted least-squares scores.

First we choose the value of *cov.control*, the function *getMeth*() in **rrcov** can be used to choose one of the robust estimators according to the size of the dataset.

R> *getMeth*(*CovRobust*(*scale*(*ZhongZheng*100_28))).

[1] "*Orthogonalized Gnanadesikan-Kettenring Estimator*".

Therefore, we take *cov.control = CovControlOgk*(). Then we choose the *method* value, the criterion is to choose the smallest sum of squares $Q(E)$ of the residual matrix. Where

$$E = R - \left(AA^T + D\right) = \left(e_{ij}\right)_{p\times p},$$

$$Q(E) = \sum_{i=1}^{p}\sum_{i=1}^{p} e_{ij}{}^2.$$

The results are given in **Table 1**. At the same time, we also calculated $Q(E)$ of three classical factor analysis methods. In **Table 1**, for $Q(E)$, the smallest combination is (pfa, Classic), but the sample mean and the

Table 1. The Q(E) values of classical and robust factor analysis.

	Classic	Ogk
pca	2.131621	2.568465
pfa	1.412886	1.946182
mle	NA	NA

sample covariance matrix of classical estimators have been severely affected by outliers, thus we select the minimum $Q(E)$ from the remaining robust estimators. It is the combination (pfa, Ogk), and $Q(E)=1.946182$.

We can also calculate the difference of the sample correlation matrices of the robust Ogk estimator and the classical estimator. Due to limitted space, we do not show it here. Readers can run the program to see the result in the R software. Their difference is big, which means that the outliers seriously influence the classical sample correlation matrix.

Now we use the *myplotDD*() function in the package **robustfa** to plot a *Cov-class* object, and the figure is omitted due to limitted space. Here the robust estimators is the Ogk estimator. We find that the robust mahalanobis distance is much larger than the classical mahalanobis distance. Outliers have large robust mahalanobis distance. In *myplotDD*(), *id.n* and *ind* are shown. Here *id.n* is the number of observations to identify by a label. By default, the number of observations with robust distances larger than *cutoff* is used. By default *cutoff* = *sqrt*(*qchisq*(0.975, *p*)). *ind* is the index of robust distances whose values are larger than *cutoff*.

cutoff = 6.667893
id.n = length(*which*(*rd* > *cutoff*))
id.n = 47
Here y is the robust distance (*rd*).
sort.y = (*To save space, only the smallest five and largest five elements of sort.y$x and sort.y$ix are shown.*)
$x
Omitted to save space
$ix
[1] 20 19 72 4 60 … 74 98 71 96 77
ind =
[1] 32 49 41 40 2 45 43 59 34 23 44 78 94 42 29 25
[17] 3 90 76 70 84 22 80 28 75 52 63 69 30 39 33 1
[33] 82 92 93 73 68 79 100 26 31 37 74 98 71 96 77

From the above results we see that *cutoff* = *6.667893*. There are *id.n* = 47 observations with robust distance larger than *cutoff*. *sort.y* is a list containing the sorted values of *y* (the robust distance). *sort.y$x* is arranged in increasing order. *sort.y$ix* contains the indices. *ind* shows the indices of the outliers.

Next, we try to draw the scatter plot of the first two factor score and 97.5% confidence ellipses of the two estimators. Get an error message:

Error in solve.default(*cov*, ...) :
System is computationally singular: *reciprocal condition number* = 8.61033e−09.

The reasons may be the determinant (det) of the sample correlation matrices of the two estimators are close to 0, namely the sample correlation matrices are singular; The condition numbers (kappa) of the sample correlation matrices of the two estimators are very large, namely the sample correlation matrices are ill-conditioned. The determinants and condition numbers of the sample correlation matrices of the two estimators are shown in **Table 2**. In this case, the principal factor analysis method still iterates about the sample correlation matrix and the resulting factor loading matrix is very bad, and thus the factor scores are not good. Therefore, in this dataset, we do not pursue the smallest Q(E), instead see results of method = “pca”.

Table 3 shows the rotated factor loading matrix, the communality, the contribution of the factors to the original variables, the contribution rate, and the cumulative contribution rate that are calculated by the classical estimator and the robust Ogk estimator by the method = “pca”. In **Table 3**, in each row of the factor loading matrix, the element with the largest absolute value and its value is greater than or equal to 0.4 is displayed in red font, thus there is at most one element which is displayed in red font in each row, these elements have the effect of

Table 2. Determinants (det) and condition numbers (kappa) of the sample correlation matrices of two estimators.

	det	kappa
R_classical	7.153293e−38	2.920069e+15
R_robust	1.766092e−32	2.502002e+13

Table 3. The rotated factor loading matrices and other results of classical estimator and robust Ogk estimator.

Variables	Classical estimator						Robust Ogk estimator					
	Factor loading					Communality	Factor loading					Communality
	F1	F2	F3	F4	F5	h_i^2	F1	F2	F3	F4	F5	h_i^2
x1	0.053	0.901	−0.155	−0.155	0.161	0.889	−0.272	0.617	0.149	0.632	−0.108	0.888
x2	−0.124	0.897	0.102	0.196	0.149	0.891	0.252	0.024	0.087	0.909	−0.125	0.914
x3	0.263	0.586	−0.168	−0.507	0.161	0.723	−0.444	0.737	0.171	0.205	−0.032	0.813
x4	−0.035	0.619	0.459	0.121	0.129	0.627	−0.110	0.531	−0.161	0.467	−0.082	0.545
x5	−0.238	0.397	0.562	0.374	0.145	0.691	0.566	−0.321	0.082	0.487	−0.025	0.667
x6	−0.052	0.882	−0.229	0.016	0.063	0.837	−0.073	0.203	0.051	0.842	−0.129	0.775
x7	−0.242	0.095	−0.858	0.095	0.157	0.837	−0.048	−0.044	0.162	−0.017	−0.927	0.889
x8	0.046	0.117	−0.006	0.003	0.313	0.113	0.110	0.345	0.650	0.003	0.249	0.615
x9	0.091	0.319	0.111	−0.145	0.386	0.292	0.066	0.327	0.518	−0.191	0.014	0.416
x10	0.898	−0.054	0.286	−0.063	0.146	0.916	0.682	0.214	−0.080	−0.131	0.616	0.913
x11	0.433	−0.047	−0.138	0.638	−0.297	0.703	0.364	0.018	−0.489	−0.167	0.166	0.428
x12	0.696	−0.021	−0.031	−0.022	−0.029	0.487	0.089	−0.048	−0.540	−0.240	0.374	0.500
x13	0.884	−0.056	0.296	−0.097	0.156	0.907	0.511	0.197	−0.056	−0.169	0.728	0.862
x14	0.920	−0.033	0.107	0.319	0.014	0.960	0.944	0.155	−0.103	−0.001	0.146	0.948
x15	0.763	0.021	0.383	0.376	0.088	0.878	0.937	−0.162	−0.067	−0.020	0.077	0.914
x16	0.684	0.109	0.094	0.313	−0.347	0.707	0.358	0.486	−0.507	0.220	0.155	0.694
x17	0.943	0.047	0.138	−0.127	0.061	0.930	0.217	0.917	0.037	0.062	0.135	0.910
x18	0.387	−0.063	0.004	−0.115	−0.415	0.339	−0.029	0.256	−0.226	0.391	0.159	0.295
x19	0.944	0.046	0.138	−0.128	0.056	0.931	0.212	0.919	0.024	0.083	0.142	0.917
x20	0.942	0.039	0.135	−0.137	0.062	0.929	0.161	0.926	0.022	0.133	0.116	0.914
x21	−0.170	−0.198	−0.099	0.104	0.098	0.098	−0.063	−0.212	0.145	−0.323	−0.348	0.295
x22	−0.051	−0.100	−0.317	−0.027	0.785	0.730	0.101	−0.258	0.822	−0.101	−0.223	0.812
x23	−0.116	0.157	−0.569	−0.249	0.351	0.547	−0.444	0.140	0.429	0.062	−0.344	0.522
x24	0.911	0.044	0.047	0.341	0.024	0.952	0.739	0.440	0.320	0.143	0.032	0.863
x25	0.693	0.092	−0.063	0.531	−0.008	0.774	0.719	0.394	0.339	0.118	0.082	0.807
x26	0.374	0.113	0.814	−0.086	0.222	0.871	0.055	0.064	−0.085	−0.133	0.943	0.921
x27	0.250	−0.093	0.879	−0.108	−0.101	0.865	0.074	0.007	−0.129	0.022	0.936	0.899
x28	0.327	0.164	0.091	−0.170	0.764	0.755	0.082	−0.050	0.783	0.013	−0.186	0.658
Contribution	8.853	3.568	3.687	1.850	2.226		4.862	5.046	3.413	3.002	4.274	
Contribution rate	0.316	0.127	0.132	0.066	0.079		0.174	0.180	0.122	0.107	0.153	
Cumulative contribution rate	0.316	0.444	0.575	0.641	0.721		0.174	0.354	0.476	0.583	0.736	

main explanations of the corresponding variables. Aside from the red font elements in the factor loading matrix, the element with absolute value greater than or equal to 0.3 is displayed in green font. These green elements can partly explain the corresponding variables. From **Table 3**, we find that the cumulative contribution rate of the first five factors calculated by the classical estimator reaches 72.1%, and that by the robust Ogk estimator reaches 73.6%.

Since the results of classical factor analysis are greatly affected by the outliers, we only analyze the results by the robust Ogk estimator. From **Table 3**, we see that factor 1 mainly explains x5 (provident fund per share), x10 (total assets), x14 (shareholders' equity), x15 (capital accumulation fund), x23 (PB, *i.e.*, price/book value ratio), x24 (total market value), x25 (circulation market value), partially explains x3 (return on equity), x11 (fixed assets), x13 (total liabilities), x16 (main business income), and thus it is called the **provident fund market value factor**; factor 2 mainly explains x3 (return on equity), x4 (operating cash per share), x17 (operating profit), x19 (total profit), x20 (net profit), partially explains x1 (earnings per share), x5 (provident fund per share), x8 (net profit on year-on-year basis), x9 (main business revenue on year-on-year basis), x16 (main business income), x24 (total market value), x25 (circulation market value), and thus it is called the **profit factor**; factor 3 mainly explains x8 (net profit on year-on-year basis), x9 (main business revenue on year-on-year basis), x11 (fixed assets), x12 (intangible assets), x16 (main business income), x22 (price-to-sales ratio), x28 (net profit margin of main business), partially explains x23 (PB, *i.e.*, price/book value ratio), x24 (total market value), x25 (circulation market value), and thus it is called the **market value profit rate factor**; factor 4 mainly explains x1 (earnings per share), x2 (net assets per share), x6 (unallocated per share), partially explains x4 (operating cash per share), x5 (provident fund per share), x18 (non-operating income and expenses), x21 (p/e ratio), and thus it is called the **value per share factor**; factor 5 mainly explains x7 (shareholders' equity ratio), x13 (total liabilities), x26 (equity ratio), x27 (asset-liability ratio), partially explains x10 (total assets), x12 (intangible assets), x21 (p/e ratio), x23 (PB, *i.e.*, price/book value ratio), and thus it is called the **asset liability factor**.

We can sort each of the 5 factor scores computed by the robust factor analysis from high to low in tables. Because of the limited space, the tables are omitted. By calculation, in the top 10 and the last 10 stocks (a total of 100 stocks), there are 81 stocks of the robust Ogk estimator that are outliers. This is a normal phenomenon, because the factor scores of maximum/minimum values of the sample points are likely to be outliers.

The ranges of 5 factors scores computed by the robust Ogk estimator are: $-1.5 \le \text{Factor1} \le 64.6$, $-6.1 \le \text{Factor2} \le 66.0$, $-25.8 \le \text{Factor3} \le 10.9$, $-51.6 \le \text{Factor4} \le 10.5$, $-11.4 \le \text{Factor5} \le 54.5$. Although the maximum and minimum values of each factor score is asymmetric about zero, even very asymmetric, but the average factor scores of the regular points are zero. If the score is close to zero, then the stock corresponding to the factor score is close to the average level.

The larger the value of Factor 1 is, the higher the provident fund market value is; conversely, the lower. Industrial and commercial bank of China, Construction bank, The bank of China, Agricultural bank of China, China's oil's values of Factor1 are larger, their provident fund market values are higher; Guodian south, Yanghe shares's values of Factor1 are smaller, their provident fund market values are lower.

The larger the value of Factor 2 is, the more strong the profitability is; conversely, the less. Industrial and commercial bank of China, Construction bank, Agricultural bank of China's value of Factor 2 are larger, their profitability are strong; Aluminum corporation of China, *ST Anshan Iron & Steel's values of Factor 2 are small, their profitability are weak.

The larger the value of Factor 3 is, the higher the market value profit rate value is; conversely, the lower. The bank of China, Construction bank's values of Factor 3 are larger, their market value profit rate values are higher; Aluminum corporation of China, Industrial and commercial bank of China, China petroleum & chemical corporation's values of Factor3 are smaller, their market value profit rate values are lower.

The larger the value of Factor 4 is, the higher the value per share is; conversely, the lower. Guizhou maotai, Ping an bank, China's Ping An's values of Factor 4 are larger, their values per share are higher; Industrial and commercial bank of China, Construction bank, Agricultural bank of China, The bank of China's value of Factor 4 are smaller, their values per share are lower.

The larger the value of Factor 5 is, the higher the asset liability value is; conversely, the lower. Industrial and commercial bank of China, The bank of China, Agricultural bank of China, Construction bank, Bank of Communications's values of Factor 5 are larger, their asset liability values are higher; Aluminum corporation of China, China's oil, China Shenhua energy's values of Factor 4 are smaller, their asset liability values are lower.

There are some special stocks with extreme factors values. For the factor score matrix, there are a total of 5 factors, the top 10 stocks whose factor values are "big", the last 10 stocks whose factor values are "small". The number of stock categories of the special factor scores from 2 factors is $\binom{5}{2}\times 2^2 = 10\times 4 = 40$. Of course, not all of the 40 categories have stocks. The number of stock categories of the special factor scores from 2, 3, 4, 5 factors is $40+80+80+32=232$. After calculation, there are 67 categories that have special factor value stocks in 232 categories. To save space, we only report the special stocks with 5 special factor scores: Construction bank, The bank of China, Agricultural bank of China, and China merchants bank have 5 factor scores (big, big, big, small, big); Industrial and commercial bank of China has 5 factor scores (big, big, small, small, big); China's oil and China petroleum & chemical corporation have 5 factor scores (big, big, small, small, small).

3. Summary

We apply the object-oriented robust factor analysis R package robustfa to the 28 financial indicators of the 100 listed companies in China's Chinese Securities Index (CSI) 100 index in the first quarter of 2013. First of all, according to the size of the data, we automatically choose a robust estimator, the robust Ogk estimator. By the Mahalanobis distances which are computed by the robust Ogk estimator, greater than the critical value, we find a total of 47 abnormal points. This paper discovers that the results of the sample correlation matrix, the rotated factor loading matrix, the contribution of the factors to the original variables, the contribution rate, the cumulative contribution rate, the screeplot of the eigenvalues of the sample correlation matrix, the scatter plot of the first two factor scores, factor scores, and the sorted scores according to factor scores etc. computed by the classical estimator and the robust Ogk estimator are quite different. Finally, we condense the 28 financial indicators to 5 factors by combining the principal component analysis method and the robust Ogk estimator: Provident fund market value factor, profit factor, market value profit rate factor, value per share factor, and asset liability factor. Finally, we sort the 5 factor scores from high to low of each factor, and also get some special stocks according to the factor scores, most of them are outliers. The number of stock categories of the special factor scores from 2, 3, 4, 5 factors is 232. There are 67 categories that have special factor value stocks in 232 categories. The robust factor analysis results provide a good basis for investors to choose the stocks.

Acknowledgements

Yingying Zhang thanks Natural Science Foundation Project of CQ CSTC CSTC2011BB0058.

References

[1] Yang, H. (2013) Multivariate Statistical Analysis. Chongqing University Press, Chongqing.

[2] Xue, Y. and Chen, L.P. (2009) Statistical Modeling and R Software. Tsinghua University Press, Beijing.

[3] Wang, X.M. (2009) Applied Multivariate Analysis. 3rd Edition, Shanghai University of Finance and Economics Press, Shanghai.

[4] Zhang, T.J., Yang, A.M. and Zhang, C.H. (2008) An Empirical Study of Operational Risk Control Model of State-Owned Commercial Banks—Based on Exploratory Factor Analysis and Confirmatory Factor Analysis Point Inspection. *Journal of Chongqing University* (*Social Science Edition*), **14**, 36-43.

[5] Pison, G., Rousseeuw, P.J., Filzmoser, P. and Croux, C. (2003) Robust Factor Analysis. *Journal of Multivariate Analysis*, **84**, 145-172. http://dx.doi.org/10.1016/S0047-259X(02)00007-6

[6] Maronna, R.A., Martin, D. and Yohai, V. (2006) Robust Statistics: Theory and Methods. John Wiley & Son, New York. http://dx.doi.org/10.1002/0470010940

[7] Todorov, V. and Filzmoser, P. (2009) An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, **32**, 1-47. http://www.jstatsoft.org/v32/i03/

[8] Rousseeuw, P.J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T. and Maechler, M. (2013) **Robustbase**: Basic Robust Statistics. R Package Version 0.9-10. http://CRAN.R-project.org/package=robustbase

[9] Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D. and Konis, K. (2013) **Robust**: Insightful Robust Library. R Package Version 0.4-15. http://CRAN.R-project.org/package=robust

[10] Todorov, V. (2013) **Rrcov**: Scalable Robust Estimators with High Breakdown Point. R Package Version 1.3-4. http://CRAN.R-project.org/package=rrcov

[11] Zhang, Y.Y. (2013) **Robustfa**: An Object Oriented Solution for Robust Factor Analysis. R Package Version 1.0-5. http://CRAN.R-project.org/package=robustfa