

Multivariable panel data cluster analysis and its application

Bingyun Zheng*, Sui Li

School of Management Science and Engineering, Anhui University of Finance and Economics, P.R. China, 233030

Received 18 December 2014, www.cmmt.lv

Abstract

Although the research on statistical analysis is very mature, the study on cluster analysis of multivariable panel data is little in China. This paper firstly analyses the numeric characteristic of multivariable panel data, and reconstructs a new the distance function of multivariable panel data and the function of sum of squares. On the base of clustering analysis basic thought, the paper explains the arithmetic and process of cluster analysis. At last, an experimental analysis is done on productive efficiency of the industrial enterprises in China, this example results shows good applicability.

Keywords: Multivariable; Panel data, Cluster analysis; Productive efficiency

1 Introduction

Panel data are comprised of section data and time serials, so they have the character in space and time dimensionality. Because the excellent character, panel data are attached importance in the research increasingly. In the recent years, correlate researches indicate that the modeling by panel data have a good effect, with the wide and in-depth development in the theory and application study. While, existing studies mostly research from the point of metric model [1], a few scholars think over how to analyze the problem by panel data in multivariable statistical theory.

Bonze D.C and Hermosilla A.Y introduce innovatively the panel data to multivariable statistical theory, employ optimization process to replace clustering process, use the random heuristic technology to optimize object function and modify the algorithms of cluster analysis with probability link function and genetic algorithms [2]. Henceforth, the relative research is infrequent overseas. Based on the Fisher ordinal clustering theory, Ren J. & Shi S L. put forwards a kind of multivariable panel data ordinal clustering method by using Frobenius norm to construct square deviation function [3-4]. But, these researches do not study in detail the dynamic categorized statistical character of multivariable panel data.

Michel Mouchart, Jeroen V.K. Rombouts proposes a clustering method of single-variable panel data by using the stepwise regression method, based on panel data with severe deficiencies, namely short times series and many missing data[5].Zhu J P. research clustering method of single variable penal data, and an experimental analysis appears in his article [6].Xiao Z L. et al reduce the dimension of panel data to one by the principal

component analysis, then use the traditional clustering analysis to the single variable penal data[7].

The cluster analysis of single-variable panel data is a simplifying to the problem, the cluster algorithms and process are same to the cluster analysis of section data, so it is simple and easy to deal with the data.

Based on time series and cross section's characteristics of panel data , Under comprehensive consideration of the panel data's characteristics with "absolute index", "incremental index" and "time fluctuation", Li Y.G., He X.Q. reconstructs the distance function of similarity measure and Ward clustering algorithm, presents a panel data clustering analysis method [8].

Apparently, natural and social problems are complicated, single variable includes a little information and cannot reflect the character of the phenomena. So, the cluster analysis of single-variable panel data is restricted badly in actual application. Multivariable can reflect the excellent character of the panel data, but its complexity baffles the relative research on panel data in the multivariable statistical theory. Interiorly, the research on statistical analysis of multivariable panel data is deficient [9]. The article will try to do the basal study and simple experimental analysis on cluster analysis of multivariable panel data.

2 The numeric format and characteristic of panel data

Panel data is a complex data construction format, so before in-depth analysis, it is necessary to do a pretreatment to panel data, which can help us to understand the data format and descriptive statistical character and to obtain some useful information. All the knowledge is a basement to the latter cluster analysis.

*Corresponding author's E-mail: engzh519@163.com

2.1 SINGLE-VARIABLE PANEL DATA

The format of the single-variable panel data can be shown through a bi-dimensionality table. Supposing that the number of the sample in the collectivity is N , the character of every sample is marked as X , the length is T , so $X_i(t)$ is the index value of the sample i at the time of t . As we known, the format of the section data is shown also as a bi-dimensionality table. Supposing that the number of the sample in the collectivity is N , the character of every sample is denoted by p indexes, so X_{ij} is the value of index j of sample i . It is easy to find by comparison that the statistical descriptive character of them is analogical, if the time dimensionality of the panel data is transferred into index dimensionality of the section data. The functions of their statistics are same, such as average value, variance and covariance, and so on. And in the cluster analysis, the arithmetic of the

sample distance and cluster process are also same. So, the cluster analysis of section data can be used for reference to single-variable panel data. The cluster result and dendrogram can be obtained though some sorts of software. So, it is easy to deal with single-variable panel data.

2.2 MULTIVARIABLE PANEL DATA

The structure of multivariable panel data is more complex, and cannot be marked by a simple bi-dimensionality table, which is different from the above data. Strictly speaking, it should be marked though a bi-dimensionality table, we can transfer it into a bi-dimensionality bearing two grades as shown in table 1. Supposing that the number of the sample in the collectivity is N , the character of every sample is denoted by p indexes ($X_1, X_2, \dots, X_j, \dots, X_p$), the length is T . So $X_{ij}(t)$ is the value of index j of sample i at the time t .

TABLE 1 The format of multivariable panel data

time index sample	1	...	t	...	T
	X1 ... Xj ... Xp	...	X1 ... Xj ... Xp	...	X1 ... Xj ... Xp
1	X ₁₁ (1) ... X _{1j} (1)	X ₁₁ (t) ... X _{1j} (t) ... X _{1p} (t)	...	X ₁₁ (T) ... X _{1j} (T) ... X _{1p} (T)
2	X ₂₁ (1) ... X _{2j} (1)	X ₂₁ (t) ... X _{2j} (t) ... X _{2p} (t)	...	X ₂₁ (T) ... X _{2j} (T) ... X _{2p} (T)
...
i	X _{i1} (1) ... X _{ij} (1) ... X _{ip} (1)	...	X _{i1} (t) ... X _{ij} (t) ... X _{ip} (t)	...	X _{i1} (T) ... X _{ij} (T) ... X _{ip} (T)
...
N	X _{N1} (1) ... X _{Nj} (1) ... X _{Np} (1)	...	X _{N1} (t) ... X _{Nj} (t) ... X _{Np} (t)	...	X _{N1} (T) ... X _{Nj} (T) ... X _{Np} (T)

Now, we will give some statistics of multivariable panel data. Thereinto, $i \in [1, N]$; $j \in [1, p]$; $t \in [1, T]$. These statistics will be used frequently in the cluster analysis.

The average of index j at the time i .

$$\bar{X}_j(t) = \frac{1}{N} \sum_{i=1}^N X_{ij}(t) \tag{1}$$

The average of index j

$$\bar{X}_j = \frac{1}{T} \frac{1}{N} \sum_{t=1}^T \sum_{i=1}^N X_{ij}(t) \tag{2}$$

The variance of index j at the time i .

$$VAR_{X_j(t)} = \frac{1}{N-1} \sum_{i=1}^N [X_{ij}(t) - \bar{X}_j(t)]^2 \tag{3}$$

The variance of index j

$$VAR_{X_j} = \frac{1}{T} \frac{1}{N-1} \sum_{i=1}^N [X_{ij}(t) - \bar{X}_j(t)]^2 \tag{4}$$

Then, we can do the analysis of metric model and multivariable statistics of multivariable panel data. In this paper, we just discuss cluster analysis.

3 The cluster analysis of multivariable panel data

The cluster analysis of multivariable panel data is comparatively complex. Presently, there is no relevant software to use directly, which is an important reason why panel data seldom be used in the multivariable

statistics research. If the request is not very strict to the problem, we can adopt an idea of retrogression. Every index will be averaged in the time dimensionality, and abstracted as a case in some special time. So the time dimensionality is obliterated, and the panel data will become section data as retrogression. But this disposal of retrogression has two points of default at least. First, it will lose a lot of information, because average value only show the average change of the object, and not show the other distributing character, such as degree of deviation. Second, a recessive hypothesis exists, that is every same index of all samples changes along same direction. Otherwise, inaccurate or wrong conclusion will appear.

But if the request is strict to the problem, we need to reconstructs the distance function and the clustering algorithm. Now we will presents a panel data clustering analysis method.

3.1 THE BASIC THOUGHT OF CLUSTER ANALYSIS

On the base of data known, the proximity of the samples or variables will be observed. According some rules, the samples or variables that have bigger proximity than others will be aggregate into one kind, and others that have bigger proximity will be converged into another kind. In this process, it is necessary to assure that the difference in the same kind is small, and the difference

among the kinds is bigger. This process will circulate down in the same way. At last, all the samples or variables will be divided into some kinds.

There are two core questions to deal with, one is what statistic will be used as a token to the proximity of the samples, the other one is which method will be adopted from many material hierarchical cluster methods, or which rule is appropriate to judge the proximity of the kinds.

3.2 THE PROXIMITY INDEX

To bring a simple structure of kinds from a group of complex data, it is necessary to measure the proximity. When these samples are aggregated, ‘approach’ can be depicted by some distance. The distance between sample r and sample k in the collectivity can be marked as d_{rk} , and d_{rk} should satisfy some conditions as follows.

- (1) $d_{rk} \geq 0$, iff $X_r = X_k$, then $d_{rk} = 0$
- (2) $d_{irk} = d_{kr}$, to all X_r, X_k
- (3) $d_{rk} \leq d_{rj} + d_{kj}$, to all X_r, X_k, X_j

Familiar distance function includes Block distance, Euclidean distance, Minkowski distance, Chebychev distance, Mahalanobis distance, and so on. This paper chooses Euclidean distance to describe the proximity degree between samples. Certainly, the distance function of multivariable panel data added time dimensionality is different from the distance function of section data. Here, it is new called ‘Euclidean distance timed and spaced’, the ‘Euclidean distance timed and spaced’ between sample r and sample k is marked as d_{rk} , which is defined as follows.

$$d_{rk} = \left\{ \sum_{t=1}^T \sum_{j=1}^p [X_{rj}(t) - X_{kj}(t)]^2 \right\}^{1/2} \tag{5}$$

Then, the distance of all samples between two will form into a distance matrix, apparently, this matrix is symmetrical, and its diagonal elements are all zero. Here, it is marked as a lower triangular matrix shown as follows.

$$\begin{bmatrix} 0 & & & & & \\ d_{21} & 0 & & & & \\ d_{31} & d_{32} & \ddots & & & \\ \dots & \dots & \dots & \ddots & & \\ d_{N1} & d_{N2} & \dots & d_{NN-1} & 0 & \end{bmatrix} \tag{6}$$

3.3 THE METHOD OF CLUSTER

Cluster analysis provides rich classification methods, which approximately may be induced to some sorts, such as hierarchical cluster method, k-means cluster method, fuzzy cluster method, two step cluster method, and so on. And every kind includes some material methods. For the second question of cluster analysis, that is which rule is appropriate to judge the proximity of the kind, so two

kinds can be aggregated into n new kind. There are many familiar methods in the hierarchical cluster method kind, such as nearest distance method, furthest distance method, centroid cluster method, median cluster method, and sum of squares deviation method (Ward method). We choose Ward method to explain the cluster analysis process of multivariable panel data. Be the same as distance function, the function of sum of squares deviation of multivariable panel data is different from the function of section data, the sum of squares of the kind g is marked S_g , the function of S_g that structured in this paper is follow.

$$S_g = \sum_{t=1}^T \sum_{j=1}^p \sum_{i \in i^g} [X_{ij}(t) - \bar{X}_j^g(t)]^2 \tag{7}$$

Thereinto, i^g expresses the set of all sequence number in the kind g , $\bar{X}_j^g(t)$ expresses the average of the index j at the time t of all the samples in the kind g , the arithmetic can be seen in the formula 1. And the meanings of the other sign are same to the above.

Ward thinks that the added sum of squares can measure the distance from kind r to kind k , when the two kinds incorporate into one kind. And the two kinds that the added sum of squares is least will be incorporated. If the case appears that the added sums of squares are equal, relative kinds ought to respectively combine at the same time. By illation, when the kind r and kind k combine, the distance function between them is follow (supposing that there are two samples in the kind k . If there are many samples, we can think the kind k is composed of two sub-kinds, and recursion can continue).

$$D_{rk} = \Delta S_{rk} = \frac{n_r + n_p}{n_r + n_k} S_{rp} + \frac{n_r + n_q}{n_r + n_k} S_{rq} - \frac{n_r}{n_r + n_k} S_{pq} \tag{8}$$

Thereinto, n_r expresses the number of the samples in the kind r , S_{rp} expresses the sum of squares when the kind r and sub-kind p combine into a new kind. The meanings of other signs are similar to them. The Ward method takes on the monotonicity of the distance and the expansibility of the space.

Now, an experimental instance is done on productive efficiency of the industrial enterprises in China. This case will concretely explain the process of the cluster analysis of panel data.

4 An experimental analysis

Industry plays a dominant ruler and brings great guide effect, and it is important power to optimize and enhance industry. Along with the in-depth reformation, our industrial enterprises carry out a good many measures to step up the productive efficiency. Some domestic scholars have done a series of researches on industry efficiency, and found some important conclusion. But, these researches emphasize particularly on different type

and scale of enterprises. The difference of the gift is great among regions, so from the angel of region, it is meaning to compare the enterprise productive efficiency of different regions, while the study along this direction is less. This paper will do cluster analysis on industrial enterprise efficiency of 31 regions in China from 2000 to 2013. We will find out the grade type and imbalance of the industry efficiency, at last, the possible reasons will be discussed that affect the industrial enterprise efficiency simply.

4.1 THE SOURCE OF DATA AND METHOD TO DEAL WITH DATA

The industrial enterprises above designated size possess big proportion in total industrial enterprises and have detailed data, so the article chooses industrial enterprises above designated size by region as object, and chooses three indexes to analyses the status of industrial enterprises. They are overall labor productivity, productive efficiency of working capitals and productive

efficiency of fixed assets. The data used come from China statistical yearbooks.

By observation to the panel data on the three indexes of 31 regions in China from 2000 to 2013, it is able to find approximately that the over labor productivity of our industrial enterprises enhance continuously, but productive efficiency of working capitals and productive efficiency of fixed assets take on a descending trendy. The accurate opinion and difference can't come forth through this outcome. But cluster analysis is helpful to distinguish the grade type and imbalance of the efficiency from different regions.

For the data processing to cluster analysis of panel data, no ready-made software can deal with the problem directly, the statistical software such as SPSS and EVIEWS can't, which baffles the appliance research on panel data in the multivariable statistical analysis. This paper employs the CLUSTER module provided by R statistical software, on this base, and by a program to solve the problem.

TABLE 2. The statistic of the panel data

variable	description	N	mean	max	Min	Stand. dev
X1	overall labor productivity	434	71634.72	121696.5	47710.54	20677.86
X2	productive efficiency of working capitals	434	0.536554	0.848434	0.234619	0.153007
X3	productive efficiency of fixed assets	434	0.564626	0.838853	0.373396	0.115838

4.2 THE RESULT OF CLUSTER ANALYSIS

Through the arithmetic shown above, the dendrogram of the industrial enterprises efficiency of regions in China is shown in Figure 1 as follows.

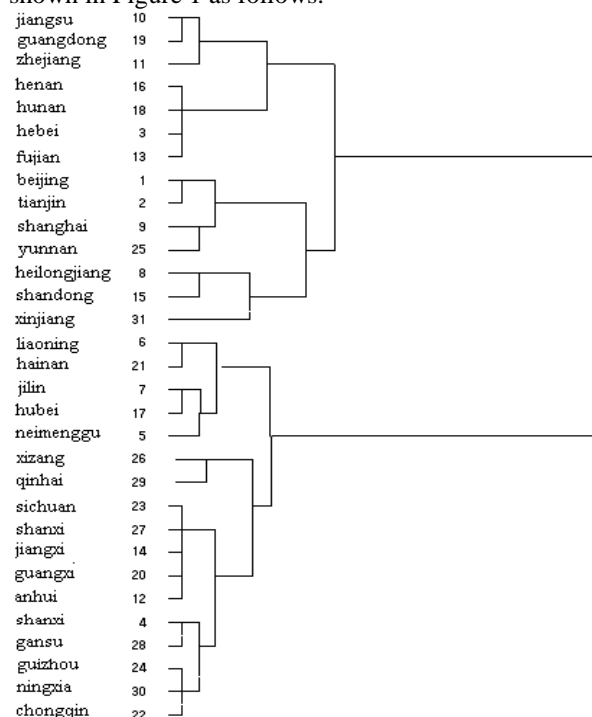


FIGURE 1 The dendrogram of the industrial enterprises efficiency of regions in China

From the dendrogram and fact, the productive efficiency of our industrial enterprises should be divided into three types. The first type includes Beijing, Tianjin, Shanghai, Yunnan, Heilongjiang, Shandong, and Xinjiang. The efficiency of industrial enterprise in these regions is highest. The second type includes Jiangsu, Guangdong, Zhejiang, Henan, Hunan, Hebei and Fujian. The efficiency of industrial enterprise in these regions is in the middling level. The third type includes Liaoning, Hainan, Jilin, Hubei, Neimenggu, Xizang, Qinhai, Sichuan, Shanxi, Jiangxi, Guangxi, Anhui, Shanxi, Gansu, Guizhou, Ningxia and Chongqing. The efficiency of industrial enterprise in these regions is least.

Through the result, we find that efficiency of industrial enterprise is higher in these regions whose economic level is higher. But Yunnan and Xinjiang are exception. Although the two regions belong to underdevelopment region, their productive efficiency belongs to the highest kind. The main cause is that behaviors of indexes are good of manufacture of tobacco in Yunnan and manufacture of foods in Xinjiang. Furthermore, they occupy a big proportion in the whole industrial enterprise.

The developed regions commonly hold abundant and high quality human capital, and distributing mainly are in the east-south coast regions and partly central regions. Human capital can enhance the individual labor efficiency on the one hand, and take on positive external effect to the whole economics on the other hand [10]. The west-north regions and other central regions can't attract

abundant and good human capital because of lagged economic level and development environment.

In addition, the open degree also affects the productive efficiency. Generally, the open degree in coast regions and developed regions is high, and these regions import strong competition mechanism. Specially, the inflow of foreign direct investment improves obviously the economic benefit and productive efficiency. Niu shu-hai thinks that the difference of productive efficiency of industrial enterprise in the three big belt is an very important factor, which make the area gap expand in China[11]. From the result in the paper, on the whole, the productive efficiency of industrial enterprise is higher in these regions whose economic level is higher. So, to reduce the area economic gap by improving the productive efficiency of industrial enterprise in the underdevelopment regions is an important way to realize the sustainable and coordinated development, specially, on the condition of transformation of economic growth.

References



- [1]Cheng Hsiao 1986 Analysis of Panel Data *Cambridge University Pres: Cambridge* 25-76
- [2]Bonzo D. C, Hermosilla A.Y 2002 Cluster Panel Data Via Perturbed Adaptive Simulated Annealing and Genetic Algorithms *Advances in Complex Systems* 5(4):339-360
- [3] Ren J, Shi Sh. L 2009 Multivariable Panel Data Ordinal Clustering and Its Application in Competitive Strategy Identification of Appliance-wiring Listed Companies *Proc. ICMSE International Conference on Management Science & Engineering (16th)*, Moscow,Russia, 253-258
- [4]Ren J 2013Fusion Clustering Analysis of Multivariable Panel Data *Application of Statistical and Management in China* 32(1):57-67
- [5]Michel Mouchart, Jeroen V.K. 2005 Rombouts. Clustered panel data Clustered Panel Data Models: An Efficient Approach for Nowcasting from Poor Data *International Journal of Forecasting* 5:577-594
- [6]Zhu Jianping & Chen Minken 2007 The Cluster Analysis of Panel Data and Its Application *Statistical Research* 24(4):11-14
- [7]Xiao Z L., Li B. Y., Liu S. F 2009 The Discussion on The Clustering Way Based on The Multivariable Panel Data and Empirical Analysis *Application of Statistical and Management in China* 28(5):831-838
- [8]Li Y.G., He X.Q. 2010 Panel Data Clustering Method and Application *Statistical Research in China* 27(9):73-79
- [9]He xiaoqun 2012 The Multivariable Analysis *People's University of China Press:Beijing* 77-89
- [10]NIE Xiudong, WANG Zhi-gang 2006 The Regional Technical Efficiency Difference in China:1978-1999 *Journal of Shanxi Finance and Economics University* 28(2):27-32
- [11]NIU Shuhai 2006 A Research of the Regional Difference in the Production Efficiency of Different Types of Industrial Enterprises in Our Country *Economic Survey in China* 6 :47-49

5 Conclusion

The cluster analysis of panel data is only one aspect of research that panel data may be used in multivariable statistical analysis. The paper just does the basal study and simple experimental analysis on cluster analysis of multivariable panel data. And the relative study will still exist wide and large, such as Non-equal-time-span panel data and imbalance panel data, etc. The relative study will be more complex and needs scholars to discuss farther.

Acknowledgements

The authors are very grateful to the people that supplied suggestion during the research. This research has been partially supported by Anhui Provincial Natural Science Foundation (Grant NO. 1308085MG112) and Anhui Provincial colleges and universities Natural Science Foundation (Grant NO. KJ2013A004).

Authors	
	<p>Bingyun Zheng , 1977.7-,Xinyang County, Henan Province, P.R. China Current position, grades: Associate Professor of Anhui University of Finance and Economics, China. University studies: Graduated from Nanjing University of Aeronautics and Astronautics, China in 2011, received a doctor's degree in Management Science and Engineering. Scientific interest: Statistical method, et al. Publications: more than 10 papers published in various journals. Experience: teaching experience of 10 years, has completed 3 scientific research projects.</p>
	<p>Sui Li , 1980.3, Qinhuangdao County, Hebei Province, P.R. China Current position, grades: Associate Professor of Anhui University of Finance and Economics, China. University studies: Graduated from Nanjing University of Aeronautics and Astronautics, China in 2010, received a doctor's degree in Management Science and Engineering. Scientific interest: Statistical method, et al. Publications: more than 10 papers published in various journals. Experience: teaching experience of 9 years, has completed 2 scientific research projects.</p>