

Consideration on the variation of financial data of institutions for canonical correlation analysis

Masaaki Ida

National Institution for Academic Degrees and Quality Enhancement of Higher Education

Kodaira, Tokyo, Japan

ida@niad.ac.jp

Abstract— In these days, progress of e-government and spread of electrical data lead to the prevail of public open databases, and also lead to the large amount of data analysis applications applying to these official open data. With regard to data analysis method, Canonical Correlation Analysis, which is one of the basic data analysis method and also data visualization method, is becoming the requisite skill for data scientists in this Big Data era. This paper examines the open data of financial data of education institutions. Especially, we focus on the higher education institutions and their financial data. In addition, we examine the variation of data and its problem to the data analysis. We aim to apply this analysis method and the result of consideration for supporting the improvement of quality assurance of higher education institutions.

Keywords— data analysis, canonical correlation analysis, data variation, financial data, higher education institution

I. INTRODUCTION

In these days, progress of e-government and spread of electrical data lead to prevail of public open databases, and also to the development of various kinds of the data analysis applications applying to these official open database.

Education-related databases are important for college selections or various quality assurance activities, such as reporting and data analysis in higher education institution. Therefore, database analysis of higher education institutional database is desired to be developed.

“Japanese College and University Portraits” is an information system consisting of database with Web services for providing information concerning various activities undertaken by universities and junior colleges, covering national, prefectural, municipal, and private institutions [1]. System operation started in March 2015. The system is managed by “National Institution for Academic Degrees and Quality Enhancement of Higher Education, Japan (NIAD-QE)”, associated with Promotion and Mutual Aid Corporation for Private Schools of Japan.

NIAD-QE is also providing “Basic Information of Japanese National and Prefectural and Municipal Universities,” and “Financial Reports of Japanese National University Corporation” on its web site [2, 3].

With regard to data analysis method, “Canonical Correlation Analysis (CCA)” [4], which is one of the basic data analysis and visualization method, is becoming the requisite skill for data analysts in this Big Data era.

CCA is a generalized method of corresponding analysis that is useful for questionnaire analysis, and other and other multivariable analysis methods [5, 6].

This paper examines the open data of financial data. Especially, we focus on the higher education institutions and their financial data, and then on the variation of data. We aim to apply this analysis method for supporting the improvement of quality assurance of higher education institutions.

In the analysis process, we divide “revenue” and “cost and expenses” of the financial data and treat them respectively. We firstly examine the relationship between the revenue and expenses by correlation coefficients. Next, we examine the results of canonical correlation in the data item set and also in the data record set. In order to understand the whole tendency of the financial data, we utilize the simultaneous scatter plot of the result of revenue and expenses of canonical correlation analysis. It is important that the data visualization of results of the analysis method helps to understand the tendency of the data sets.

In addition to the analysis of data item, we examine the utilization of canonical correlation analysis for the variation of financial data. We examine the variation of the data, especially we consider on the problem from the viewpoint of numerical analysis. As for the characteristics of the data in this paper, we have to note a possibility of the occurrence of inappropriate situation of numerical calculation result. This is difficult problem for the beginner of data analysts. This paper examine the treatment for such a problem, i.e., the method of observing the situation of numerical analysis, and the method of improvement of the situation. We show numerical examples for illustrating the problem.

II. CANONICAL CORRELATION ANALYSIS

Canonical Correlation Analysis is a core analysis method in multivariate analysis field. Mathematical description is as follows.

Two multiple variable data matrices, X_R and Y_R , express $n \times p$ and $n \times q$ real data matrices. The number of data is n with p and q data items for X and Y . Values of the matrices are actual financial data. We define the following matrices for data average and deviation:

$$Q_n = I_n - (1/n)\mathbf{1}_n\mathbf{1}_n^T$$

$$X = Q_n X_R, \quad Y = Q_n Y_R$$

Where $\mathbf{1}_n$ means $(1, 1, \dots, 1)^T$, and Q_n means averaging.

Calculating the correlation matrices, R_{XX} R_{YY} R_{XY} for X and Y , that is,

$$R_{XX} = X^T X, \quad R_{YY} = Y^T Y, \quad R_{XY} = X^T Y.$$

The result of canonical correlation analysis is regarded as the singular value μ and corresponding vectors a and b , which satisfy the following matrix equation:

$$\begin{pmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = (1 + \mu) \begin{pmatrix} R_{XX} & 0 \\ 0 & R_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

We call the maximum singular value μ^1 as the first canonical correlation coefficient, and μ^2 as second one, and μ^3 as third one and so on. Its corresponding vectors, a^i and b^i for i , are called score vectors.

Solving the equation is equivalent to obtain the eigenvalue and eigenvectors of the following matrices:

$$(R_{xx}^{-1} R_{xy}) (R_{yy}^{-1} R_{yx}) \quad \text{or} \quad (R_{yy}^{-1} R_{yx}) (R_{xx}^{-1} R_{xy})$$

In addition, we calculate the vectors f^i and g^i for data row level scores as follows;

$$f^i = X a^i$$

$$g^i = Y b^i$$

In this paper, we consider the vectors a^i , b^i , f^i and g^i for understanding the arrangement of each element and tendency of whole data set.

As an example of canonical correlation analysis, we show one numerical result of analysis for university financial data (work in progress).

In this example, the numerical data of items for X expresses revenues (e.g., management expenses grant, tuition, research grant, donation, and so on), and the numerical data of items for Y expresses the cost and expenses (e.g., general management expenses, research expenses, education expenses, and so on). Figure 1 shows an example of the result of canonical correlation analysis visualized in two dimensions (vectors a^i and b^i , $i = 1$ and 2), which are high accumulation contribution of eigenvalues. In this figure, blue points indicate revenues and red points indicate cost and expenses. We can grasp the global feature of financial situations of universities by this arrangement.

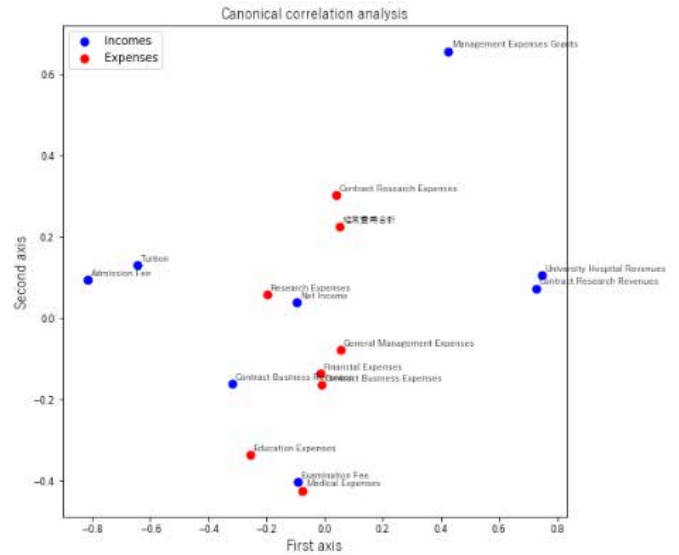


Figure 1. Example of CCA (1)

Next we examine the data in data record field (the vectors f^i and g^i), university set, Figure 2 show the arrangement of university revenues (blue) and expenses (red) in two dimensions ($i = 1$ and 2). This figure shows the relations among universities. It shows the proximity between universities and the tendency of whole university set. The numbers in this figure indicate the universities.

This figure shows that the “First axis” means the effect of scale of universities. Most universities located in the two major groups or located as the shattered points. This indicates that we can divide two typical university sets concerning revenue and expenses. However, this figure shows some large scale universities have the particular characteristics in contrast with other typical universities.

In this way, we can take a global view of the clustering. Various comprehensive considerations on overall of accumulated data can be taken by executing the analysis. Those abilities will deepen the global understanding on the relations of accumulated information, and which have promising possibility leads to new knowledge discovery in data analysis.

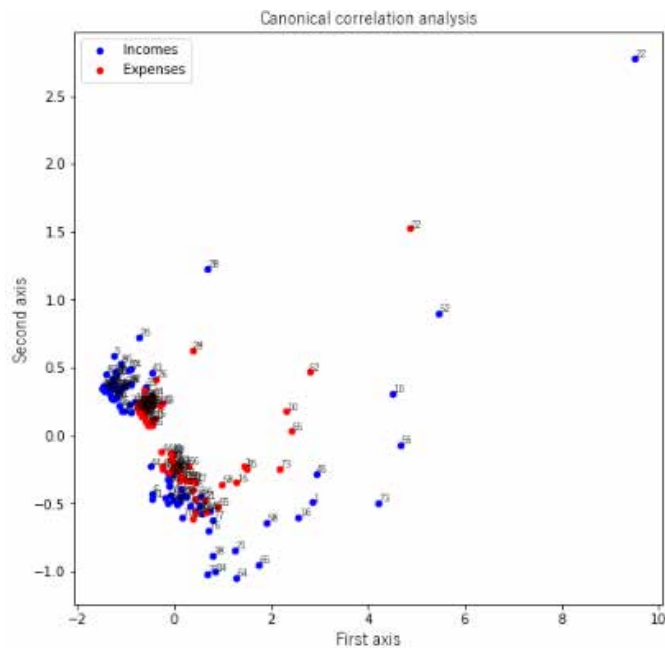


Figure 2. Example of CCA (2)

III. VARIATION OF DATA

In the consideration of ordinary data analysis process, we sometimes change the selections of data items and/or vary the values of the some data in order to understand the effect of variation of data. However, in such cases we sometimes encounter the difficulty of the data calculation.

The characteristics of the financial data here is that the elements of the correlation coefficient matrices R_{XX} , R_{YY} and R_{XY} may sometimes take very high values as shown in Table 1. It takes a possibility that the high correlation values occur. Therefore, the calculation of inverse matrices of R_{XX} and R_{YY} , which are needed for canonical correlation analysis, may become very difficult.

Table 1. Part of the correlation coefficient matrix

	Management Expenses Grants	Tuition	Admission Fee	Examination Fee	University Hospital Revenues	Contract Research Revenues	Contract Business Revenues	Net Income	Expenses	Education Expenses
Management Expenses Grants	1.000000	0.904880	0.911250	0.862543	0.792679	0.907159	0.923712	0.886704	0.985573	0.956425
Tuition	0.904880	1.000000	0.982126	0.947755	0.779438	0.745373	0.870235	0.776914	0.914894	0.924594
Admission Fee	0.911250	0.982126	1.000000	0.966454	0.761273	0.735622	0.871812	0.759650	0.910702	0.926344
Examination Fee	0.862543	0.947755	0.966454	1.000000	0.714552	0.699604	0.820096	0.730677	0.859296	0.901973
University Hospital Revenues	0.792679	0.779438	0.761273	0.714552	1.000000	0.577253	0.727956	0.731615	0.873718	0.735466
Contract Research Revenues	0.907159	0.745373	0.735622	0.699604	0.577253	1.000000	0.830832	0.835930	0.876551	0.836881
Contract Business Revenues	0.923712	0.870235	0.871812	0.820096	0.727956	0.830832	1.000000	0.775019	0.915627	0.885076
Net Income	0.886704	0.776914	0.759650	0.730677	0.731615	0.835930	0.775019	1.000000	0.878633	0.836806
Expenses	0.985573	0.914894	0.910702	0.859296	0.873718	0.876551	0.915627	0.878633	1.000000	0.933292
Education Expenses	0.956425	0.924594	0.926344	0.901973	0.735466	0.836881	0.885076	0.836806	0.933292	1.000000
Research Expenses	0.959049	0.835447	0.841639	0.771615	0.694951	0.939664	0.900970	0.851508	0.942790	0.876877
Medical Expenses	0.792836	0.775428	0.756090	0.706152	0.996065	0.582974	0.732373	0.711764	0.874245	0.731125

For example, Figure 3 shows the inappropriate numerical calculation result of CCA. In such case, we do not have the appropriate scatter plot of “cost and expenses”, so we may deduce inappropriate considerations from this result. Sometimes without noticing, this type of result of CCA may occurs in data analysis.

In order to avoid such inappropriate consideration, we have to check the condition of numerical calculation of CCA. We carefully observe the variation of some eigenvalues of the matrices in CCA, and check the condition numbers in numerical analysis. The condition number, $\text{cond}(A)$, is defined by the norm of matrix as follows:

$$\text{cond}(A) = \|A\| \|A^{-1}\|$$

where “A” is the matrix and $\|A\|$ is the norm of matrix. In the calculation of CCA, we check the condition number of R_{XX} and other matrices and variations of the values of eigenvalues.

In order to treat such difficult situation, one possible solution for the case is that we check the correlation coefficient matrix as shown in Table 1, and find the combination of data items with high values, and avoid to select such combination of items. We check the result of recalculation of CCA and some condition numbers in CCA calculation process.

For a concrete example, as for Figure 3, the value of condition number $\text{cond}(R_{YY})$ is extremely high. Then, we avoid select the combination of data items with high correlation coefficients. We do not select the data items with the correlation value is large. Then we eliminate the one item of them. And we observe the condition number fall in the appropriate range. This kind of phenomenon is related to the similar items.

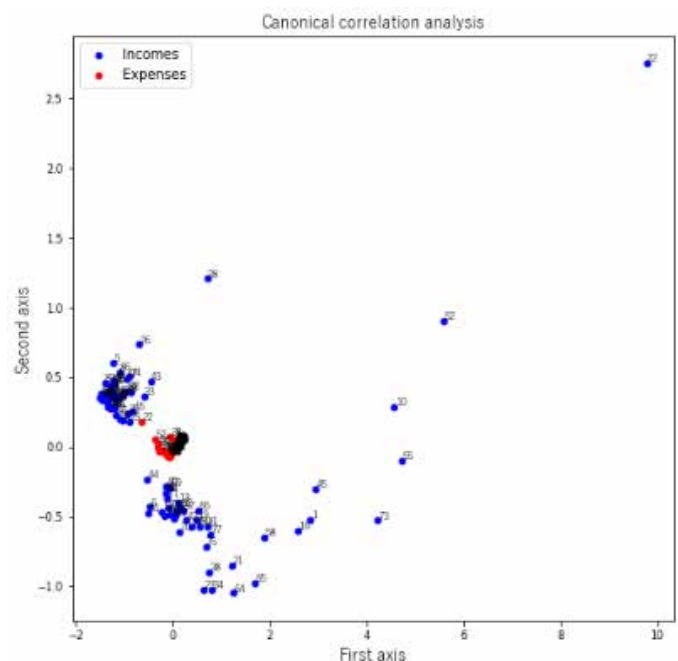


Figure 3. Inappropriate result of CAA and its visualization

IV. CONCLUSIONS

This paper describes the consideration on the variation of financial data of institutions for canonical correlation analysis.

We remark the development of data base concerning university institutional data. Especially, we focus on financial data of Japanese universities. We show the Canonical Correlation Analysis (CCA) for these financial data. In order to understand the whole tendency of university financial situation, CCA is adequate to be applied to these kind of data.

Moreover, we consider the variation of data and the difficulty of CCA calculation for these data. Sometimes occurrence of inappropriate numerical results are observed, then it has a possibility to deduce the inappropriate understanding of data tendency. Therefore, in this paper, we indicated the possible solution that how to observe such situation and how to find the possible solution of the problem by utilizing the correlation coefficient matrix and condition numbers.

We hope that our consideration and attempt to financial data of institutions will play an important role as an infrastructure for data utilization and data analysis in higher education quality assurance.

REFERENCES

- [1] Japanese College and University Portraits, URL; <https://top.univ-info.niad.ac.jp> [accessed: 2019-01-04]
- [2] Basic Information of Japanese National and Prefectural and Municipal Universities, URL; <http://portal.niad.ac.jp/ptrt/table.html> [accessed: 2019-01-04]
- [3] Financial reports of Japanese National University Corporation, URL; https://www.niad.ac.jp/support/university_finance/ and https://www.niad.ac.jp/support/university_finance/summary/heise_29.html [accessed: 2019-01-04]
- [4] B. Thompson, *Canonical Correlation Analysis: Uses and Interpretation*, Sage Publications, 1985.
- [5] J. P. Benzecri, *Correspondence Analysis Handbook*, Marcel Dekker, 1992.
- [6] M. Greenacre, *Correspondence Analysis in Practice, Second Edition*, Chapman and Hall/CRC, 2007.

Masaaki Ida was received the B.E., M.E, and Ph.D. degrees in engineering from Kyoto University, Kyoto, Japan. He is currently a professor of National Institution for Academic Degrees and Quality Enhancement of Higher Education. His current research interests are in computer and systems science, especially in advanced database system and communication technology in higher education.