Proceedings of the International MultiConference of Engineers and Computer Scientists 2015 Vol II,
IMECS 2015, March 18 - 20, 2015, Hong Kong

First-Order Perturbation of Correspondence Analysis with Multiple Categories

Masaaki Ida

***Abstract*—Correspondence analysis is one of the requisite analysis skills for data analysts in this Big Data era. Especially, correspondence analysis with multiple categories is a popular research technique for analyzing questionnaire survey with mutually exclusive choices, and which is also applicable as a visualization technique for data mining and text mining. However, in practical situation, data perturbation is an important issue on corresponding analysis. We have so far considered perturbation of ordinal correspondence analysis. However, more generalized consideration on perturbation for multiple category case is required. This article presents mathematical results on first-order perturbation of correspondence analysis with multiple categories.**

***Index Terms*—correspondence analysis, perturbation, visualization, questionnaire analysis, text mining.**

I. Introduction

CORRESPONDENCE analysis is a technique to analyze corresponding relation between data in multiple categories expressed by cross table[1],[2]. This technique can analyze cross table or extended multiple cross table (Burt table) containing some measure of correspondence between the rows and columns. The results of correspondence analysis provide information similar to factor analysis. This method is closely related to homogeneity analysis, dual scaling and quantification methods. Correspondence analysis is frequently utilized for analyzing questionnaire survey and text mining visualization method. Various comprehensive considerations on overall accumulated data can be taken by executing the correspondence analysis. Those abilities will deepen the global understanding on the relations of accumulated multiple categorical information, and which have promising possibility leads to new knowledge discovery. We have so far focused on the educational text information. We conducted research on collecting and analyzing the information and text mining for analyzing and visualizing such educational information for grasping the global characteristics of higher education institutions[3],[4].

However, in practical situation, uncertain or perturbed information is necessarily included, such as variation of elements of cross table. In other case, the cross table will modified or updated by additional data. Then, it is necessary to examine the perturbations of cross table and its influence to the result of correspondence analysis.

This paper presents an extension of our former studies for sensitivity of ordinal correspondence analysis to correspondence analysis with multiple categories[5],[6]. Section two describes mathematical considerations on perturbation[7],[8] of correspondence analysis. Section three shows an application example of text mining data in education field. Obtained results in this paper are contributing for reducing the computational burden, and assisting interpretation and explanation of local influence by data perturbation.

Masaaki Ida is with the National Institution for Academic Degrees and University Evaluation, Tokyo, Japan, e-mail: ida@niad.ac.jp

II. Mathematical Formulation

A. Correspondence Analysis with Multiple Categories

Correspondence analysis with multiple categories is an extension of ordinal correspondence analysis which is essentially based on singular value decomposition[1],[2]. In this section, mathematical formulation of the case of multiple categories is discussed.

Table 1 illustrates a part of r categorical data table for questionnaire survey with mutually exclusive choices. Each row of this table corresponds to a binary coding of factors, also called dummy variables.

TABLE I
Example of categorical data table

	G_1^1	G_2^1	G_1^2	G_2^2	G_3^2	$\cdots$	G_1^r	G_2^r
1	1	0	1	0	0	$\cdots$	1	0
2	1	0	0	0	1	$\cdots$	1	0
3	0	1	1	0	0	$\cdots$	0	1
$\vdots$	$\vdots$		$\vdots$			$\vdots$	$\vdots$	

This table is expressed by an indicator matrix G as

$$G = (G^1, G^2, \cdots, G^r), \tag{1}$$

where G^i is a binary matrix that exactly one element in each row is equal to one.

Let A and B be matrixes associated to the indicator matrix G,

$$A = G^{\mathrm{T}} G, \tag{2}$$

$$B = r\ \mathrm{diag}(G^{\mathrm{T}} G), \tag{3}$$

where $\mathrm{diag}(G^{\mathrm{T}}G)$ is a diagonal matrix of the elements of matrix $G^{\mathrm{T}}G$. Elements of the matrices are expressed as

$$A = \begin{pmatrix} {G^1}^{\mathrm{T}} G^1 & {G^1}^{\mathrm{T}} G^2 & \cdots & {G^1}^{\mathrm{T}} G^r \\ {G^2}^{\mathrm{T}} G^1 & {G^2}^{\mathrm{T}} G^2 & \cdots & {G^2}^{\mathrm{T}} G^r \\ \vdots & \vdots & & \vdots \\ {G^r}^{\mathrm{T}} G^r & {G^r}^{\mathrm{T}} G^2 & \cdots & {G^r}^{\mathrm{T}} G^r \end{pmatrix}, \tag{4}$$

$$B = r \begin{pmatrix} {G^1}^{\mathrm{T}} G^1 & & & 0 \\ & {G^2}^{\mathrm{T}} G^2 & & \\ & & \ddots & \\ 0 & & & {G^r}^{\mathrm{T}} G^r \end{pmatrix}. \tag{5}$$

The multiple correspondence analysis can be formulated as the following *Generalized eigenvalue problem* with symmetric matrixes[1],[2],

$$A\boldsymbol{x} = \lambda B\boldsymbol{x}. \tag{6}$$

ISBN: 978-988-19253-9-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Constraints on the eigenvectors are

$$\boldsymbol{x}^i B \boldsymbol{x}^i = \delta_{ij}, \tag{7}$$

where δ_{ij} is the *Kronecker delta*.

Each eigenvalue for this eigenvalue problem is denoted by λ^i $(i = 1, \cdots, \mathrm{rank}B)$, and it is noted that $\lambda^1 = 1$ and $\lambda^i \geq 0$. Corresponding eigenvector for the eigenvalue is denoted by $\boldsymbol{x}^i = (\boldsymbol{x}^i_1, \cdots, \boldsymbol{x}^i_r)$. Elements of the vector $\boldsymbol{x}^i$ are called *scores* of the correspondence analysis.

B. Perturbation of Correspondence Analysis

If $A \to A + \delta A$ and $B \to B + \delta B$ due to small change, then *first-order perturbation* of eigen system, $\lambda^i \to \lambda^i + \delta\lambda^i$ and $\boldsymbol{x}^i \to \boldsymbol{x}^i + \delta\boldsymbol{x}^i$ can be deduced as follows,

$$\delta\lambda^i = {\boldsymbol{x}^i}^{\mathrm{T}}(\delta A - \lambda^i \delta B)\boldsymbol{x}^i, \tag{8}$$

$$\delta\boldsymbol{x}^i = (-\frac{1}{2}{\boldsymbol{x}^i}^{\mathrm{T}}\delta B \boldsymbol{x}^i)\ \boldsymbol{x}^i + \sum_{j \neq i} \frac{{\boldsymbol{x}^j}^{\mathrm{T}}(\delta A - \lambda^j \delta B)\boldsymbol{x}^i}{\lambda^j - \lambda^i}\ \boldsymbol{x}^j. \tag{9}$$

We do not consider higher-order perturbation in this paper because our purpose is to grasp tendency of the effect of small change. We note that the special case of repeated eigenvalue problem ($\lambda^i = \lambda^j$) is not discussed which will be discussed in other papers.

Trivial case that the eigen vector of $\boldsymbol{x}^1 = \mathbf{1}$ with $\lambda^1 = 1$ is usually omitted as a score for correspondence analysis. In this paper, we consider the case that the eigenvalues, $0 < \lambda_i < 1$ $(i = 2, 3, \cdots)$, and $\lambda_i > \lambda_j$ for $(i < j)$. In actual situation the case of $i = 2$ or 3 is usually considered.

Here let a list g be as follows,

$$g = (g_1, g_2, \ldots, g_r). \tag{10}$$

For example, additional data to G is represented by

$$g = \big(\overset{i^1}{0, \ldots, 0, 1, 0, \ldots, 0} \ \big|\ \cdots \ \big|\ \overset{i^r}{0, \ldots, 0, 1, 0, \ldots, 0}\big).$$

This means that the additional data belongs to i^1-th group of g_1-th category, and to i^r-th group of g_r-th category. Additional data to G causes perturbation of cross table.

Following equations can be obtained.

Proposition 1:

$${\boldsymbol{x}^i}^{\mathrm{T}}\ \delta A\ \boldsymbol{x}^j = \left(\sum_k \boldsymbol{x}^i_{k,g_k}\right)\ \left(\sum_k \boldsymbol{x}^j_{k,g_k}\right), \tag{11}$$

$${\boldsymbol{x}^i}^{\mathrm{T}}\ \delta B\ \boldsymbol{x}^j = r \sum_k \boldsymbol{x}^i_{k,g_k}\boldsymbol{x}^j_{k,g_k}. \tag{12}$$

These equations and (8), (9) directly deduce the following result of this paper.

Proposition 2:

$$\delta\lambda^i = \left(\sum_k \boldsymbol{x}^i_{k,g_k}\right)^2 - \lambda^i\ r \sum_k (\boldsymbol{x}^i_{k,g_k})^2, \tag{13}$$

$$\delta\boldsymbol{x}^i = -\frac{r}{2}\sum_k (\boldsymbol{x}^i_{k,g_k})^2\ \boldsymbol{x}^i + \sum_{j \neq i} \frac{\sum_k \boldsymbol{x}^i_{k,g_k} \sum_k \boldsymbol{x}^j_{k,g_k} - \lambda^i\ r \sum_k \boldsymbol{x}^i_{k,g_k}\boldsymbol{x}^j_{k,g_k}}{\lambda^i - \lambda^j}\ \boldsymbol{x}^j. \tag{14}$$

By using these equations, direction and quantity of *score perturbation* of correspondence analysis can be easily calculated compared to previous research results. Therefore, our results contribute to reducing the computational burden. Moreover, the results contribute to assisting explanation of local influences by data perturbation.

Difference from the results of our previous papers[5],[6] is that in this paper's method we directly utilize *Generalized eigenvalue problem* (equation (6)) and we do not use inverse matrixes and various transformation matrixes, so that we can reduce the computation of perturbation, and furthermore we can easily extend this method to other mathematical formulations.

III. Example

As an example, we show a simple two category case of $r = 2$ which was describe in reference[3],[4]. We developed the curriculum analysis system based on clustering and correspondence analysis for syllabus data, which was applied for Japanese universities. Procedure of the analysis system is described as follows,

1) selection (collection) of syllabuses of curricula applicable to analysis,
2) extraction of technical terms from syllabus contents,
3) calculation of similarity degrees between syllabuses,
4) clustering based on the degrees of similarity,
5) visualization of the distribution of syllabus clusters and curricula by *correspondence analysis*.

In the following analysis example, we choose 13 departments of 13 Japanese universities as shown in Table II (U# denotes each department). Department names are directly translated from Japanese[3],[4].

TABLE II
DEPARTMENTS OF 13 UNIVERSITIES (G_1)

U07	dept. of mechanical engineering
U09	dept. of computer and information sciences
U13	dept. of bio-system engineering
U15	dept. of systems engineering
U19	dept. of information and computer sciences
U32	dept. of systems engineering
U35	dept. of chemistry and chemical engineering
U37	dept. of information and systems engineering
U50	dept. of mechanical and system engineering
U61	dept. of systems engineering
U63	dept. of computer science and systems engineering
U73	dept. of mechanical systems engineering
U78	dept. of mechanical systems engineering

Totally 638 syllabuses are selected. Seven cluster summaries obtained in the step four of clustering procedure of the analysis system is shown in Table III (C# denotes each cluster).

Cluster summaries in Table III indicate as follows: Cluster C1 contains many terms in fields of the mechanics, the chemistry, the systems engineering, and the control engineering, C2 contains many technical terms in relation to computer software, C3 is related to communication engineering, C4 is

ISBN: 978-988-19253-9-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

related to electrical engineering, C5 is related to mathematics, C6 is related to image processing, C7 is related to terms of experimentation.

TABLE III
CLUSTER SUMMARIES ANALYZING CURRICULA (G_2)

C1	(Mechanical) reaction, design, stress, movement, cycle, control, energy, flow, structure,
C2	(Computer) language, programming, program, computer, algorithm, C language, procedure,
C3	(Communication) information, system, probability, management, signal, technology,
C4	(Electrical) circuit, electric, electronic, light, alternating current, semiconductor, electric,
C5	(Mathematical) integral, differential, expansion, matrix, differential equation, vector, function,
C6	(Image processing) image, recognition, voice, pattern, discrimination, pattern recognition,
C7	(Experiment) applied physics, report, measurement, equipment, microphone,

Cross table for 13 departments and 7 clusters is shown in Table IV.

TABLE IV
DEPARTMENT-CLUSTER CROSS TABULATION ($G_1 \times G_2$)

	C1	C2	C3	C4	C5	C6	C7	Sum
U07	10	1	3	0	0	2	0	16
U09	7	10	6	2	10	2	0	37
U13	17	14	13	29	3	2	0	78
U15	10	5	8	8	6	0	0	37
U19	5	8	6	4	19	2	0	44
U32	4	1	11	1	1	0	0	18
U35	58	0	2	5	3	0	0	68
U37	2	10	24	16	5	3	0	60
U50	46	2	8	2	23	1	1	83
U61	25	7	16	6	7	1	0	62
U63	7	9	18	9	8	2	0	53
U73	31	1	4	5	0	0	0	41
U78	32	1	3	1	4	0	0	41
Sum	254	69	122	88	89	15	1	638

Correspondence analysis is performed in order to understand the cross table of Table IV.

Points of Figure 1 shows visually summarized information in two dimensions, which are high accumulation contribution of eigenvalues. We can grasp the global feature of Table IV by this figure. Each point shows university or cluster.

Right side of Fig. 1 seems to be related to "Information science" (cluster C2, C3, C4, C6). On the other hand, left side is related to "Mechanical engineering" (cluster C1). In this way, we can take a global view of the clustering.

In this way, as discussed in the previous section, various comprehensive considerations on overall of accumulated data can be taken by executing correspondence analysis. Generally, results depend on the quality of syllabuses, however, we can assert that it is possible for understandings of global feature of tendency in curricula with many syllabus documents.

However, elements of the cross table, numbers of syllabuses, relates perturbation. Therefore, it is necessary to examine the effects of perturbation on correspondence analysis result. When perturbation of the cross table would make more changes in the result of correspondence analysis, the interpretation on the data should be careful. Sometimes we have to reexamine and correct the interpretation if necessary. On the other hand, result of correspondence analysis might not be fluctuated by the change of cross tab. In such robust case, the interpretation of the result of correspondence analysis is regarded as robust and important features. Therefore, it is important to examine data variation mathematically.

As an example, we consider following data addition that means perturbation of cross table,

$$\boldsymbol{g}_1 = (0,0,0,0,0,0,0,0,0,0,0,0,1),$$
$$\boldsymbol{g}_2 = (0,0,0,0,1,0,0).$$

We calculate the perturbation of the correspondence analysis as

$\delta\boldsymbol{x}^2 =$
(-0.000128504, 0.000205204, -0.000240658, 0, 0.000443406, -0.000199935, -0.0000231272, -0.000196163, 0.000316652, -0.0000228726, -0.00003866, -0.000144594, -0.00014737, -0.0000838663, -0.0000782423, -0.000158081, -0.000235003, 0.000746641, -0.0000265884, 0.000614631),

and $\delta\boldsymbol{x}^3 =$
(-0.000243794, 0.000297203, 0.0000676268, 0.0000132407, 0.000171452, 0.000494789, -0.000919747, 0.000410613, -0.000582341, -0.000182137, 0.000243881, -0.000562423, 0.00191112, -0.000375932, 0.000444631, 0.00033986, 0.0000299719, 0.00019226, 0.000324164, -0.00126688).

The perturbation is graphically shown as the set of arrows in Figure 1. These arrows indicate directions and quantities of perturbation. Green arrows are corresponding to the variation of first category (university departments), and blue arrows are corresponding to the variation of second category (clusters). Two red arrows in the figure correspond to the perturbation element.

IV. CONCLUSION

In this paper, we considered first-order perturbation of correspondence analysis with multiple categories as one of the generalized eigenvalue problem. Results on the information of direction and quantity of scores due to small changes can be calculated. Moreover, these results are contributed for reducing the computational burden, and assisting explanation for local influences by data perturbation.

REFERENCES

[1] J. P. Benzecri, *Correspondence Analysis Handbook*, Marcel Dekker, 1992.
[2] M. Greenacre, *Correspondence Analysis in Practice, Second Edition*, Chapman and Hall/CRC, 2007.

ISBN: 978-988-19253-9-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

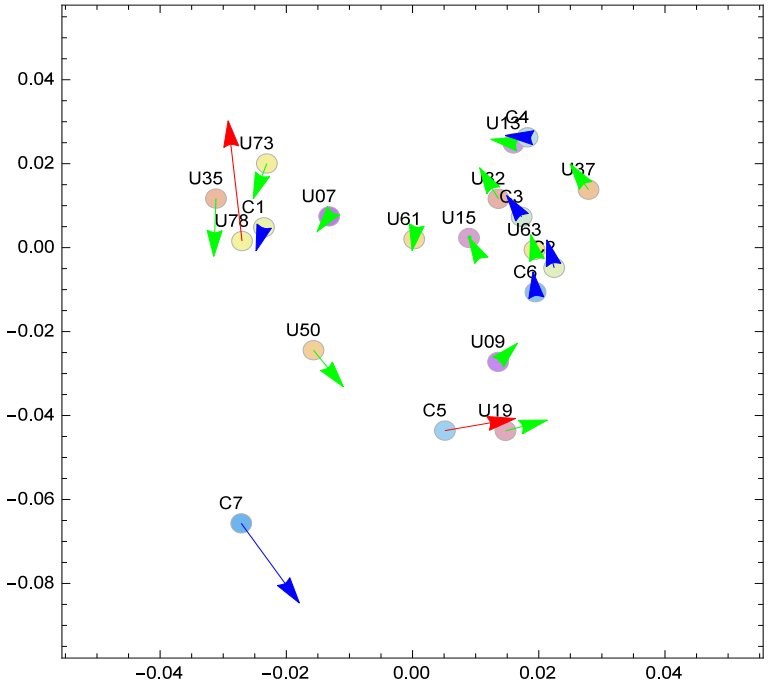

Fig. 1. Example of Correspondence Analysis and Data Perturbation $(\boldsymbol{x}^2, \boldsymbol{x}^3)$

[3] M. Ida, T. Nozawa, F. Yoshikane, K. Miyazaki, and H. Kita, Development of Syllabus Database and its Application to Comparative Analysis of Curricula among Majors in Undergraduate Education, *Research on Academic Degrees and University Evaluation*, 2, pp.85-97, 2005.
[4] T. Nozawa, M. Ida, F. Yoshikane, K. Miyazaki, and H. Kita, Construction of Curriculum Analyzing System based on Document Clustering of Syllabus Data, *Journal of Information Processing Society of Japan*, 46, 1, pp.289–300, 2005.
[5] M. Ida, Sensitivity Analysis for Correspondence Analysis and Visualization, *ICROS-SICE International Joint Conference 2009*, pp.735–740, 2009.
[6] M. Ida, Consideration on Sensitivity for Multiple Correspondence Analysis, *the International Multiconference of Engineers and Computer Scientists 2010*, pp.560–565, 2010.
[7] L.I. Schiff, *Quantum Mechanics*, McGraw-Hill, 1968.
[8] T. Kato, *Perturbation Theory for Linear Operators, 2nd ed.*, Springer, 1980.

ISBN: 978-988-19253-9-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)