

Correspondence Analysis and Perturbation of Chi-Squared Value

Masaaki Ida*

*National Institution for Academic Degrees and University Evaluation, Kodaira, Tokyo, Japan

ida@niad.ac.jp

Abstract— Correspondence analysis is a popular research technique in this Big Data era in order to analyze correspondence relation between data in multiple categories expressed by cross tabulation. Correspondence analysis is frequently utilized for text mining and for analyzing questionnaire survey with mutually exclusive choices, which is also applicable as a visualization technique for data mining. However, data addition or data fluctuation is one of the inevitable issues for correspondence analysis in practical situation. We have so far considered perturbation of ordinal and multiple correspondence analysis. These considerations has focused on perturbation of each element of cross tabulation. However, consideration on the whole influence of perturbation in cross tabulation due to data variation is required. This consideration is closely related to chi-squared value of test statistic. This article presents mathematical consideration and application example for perturbation of chi-squared value of cross tabulation.

Keywords— Correspondence analysis, Perturbation, Chi-squared value, Data mining, Questionnaire analysis.

I. INTRODUCTION

Correspondence analysis is a popular research technique to analyze correspondence relation between data in multiple categories expressed by cross tabulation (contingency table) or Burt table[1],[2]. This technique can analyze cross tabulation containing some measure of correspondence between the rows and columns. Burt table is a matrix composed of two-way cross tabulations between the categorical variables. Burt tabulation is used for generalization of correspondence analysis. 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 categorical information, and which have promising possibility leading to new knowledge discovery.

As an application field of corresponding analysis, we have so far focused on education-related data, and conducted research on collecting and analyzing such data for grasping global characteristics of higher education institutions[3],[4]. In such a practical application, data addition to cross tabulation, or data fluctuation and data modification of cross tabulation should be considered inevitably. In other words, the cross tabulation will be updated by additional data or modified by data replacement. Therefore, it is necessary to

examine the variation of cross tabulation and its influence to the result of correspondence analysis.

Previous paper[6] presented an extension of our former studies for sensitivity of ordinal correspondence analysis to correspondence analysis with multiple categories[5]. However, whole influence of perturbation in cross tabulation due to data variation is not considered.

In this paper, section two describes mathematical consideration on whole effect of cross tabulation by data variation. This consideration is closely related to chi-squared value of test statistic. 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 for total influence by data addition.

II. MATHEMATICAL FORMULATION

A. Formulation of Correspondence Analysis

In this section, mathematical formulation of correspondence analysis and χ^2 value are discussed. Data addition and its influence to the variation of χ^2 value is deduced.

Correspondence analysis is a kind of canonical correlation analysis and essentially based on singular value decomposition[1],[2]. Table 1 illustrates a part of categorical data table. This kind of data table is frequently used in the analysis process of text mining or questionnaire survey with mutually exclusive choices.

TABLE I
EXAMPLE OF CATEGORICAL DATA TABLE

	$G_{1,1}$	$G_{1,2}$	$G_{1,3}$	$\cdots$	$G_{2,1}$	$G_{2,2}$	$G_{2,3}$	$\cdots$
1	0	0	1	$\cdots$	1	0	0	$\cdots$
2	1	0	0	$\cdots$	0	0	1	$\cdots$
3	0	1	0	$\cdots$	1	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

This data table is expressed by an indicator matrix G as

$$G = (G_1, G_2), \quad (1)$$

where G_i is a binary matrix that exactly one element in each row is equal to one. Each row of this table corresponds to a binary coding of factors, also called dummy variables. For example, the first row data of this table belongs to category $G_{1,3}$ and category $G_{2,1}$.

Cross tabulation N_{12} for this data is expressed by

$$N_{12} = G_1^T G_2, \quad (2)$$

$$N_{21} = N_{12}^T. \quad (3)$$

Diagonal matrix D_1 and D_2 are defined by

$$D_1 = G_1^T G_1, \quad (4)$$

$$D_2 = G_2^T G_2. \quad (5)$$

Scores of categories of correspondence analysis are the values of elements of eigenvectors $\mathbf{p}$ and $\mathbf{q}$ of the following equations,

$$N_{12}\mathbf{q} = \mu D_1\mathbf{p}, \quad (6)$$

$$N_{21}\mathbf{p} = \mu D_2\mathbf{q}, \quad (7)$$

where $\mu(\mu_1, \mu_2, \dots)$ is the singular value of these equations.

On the other hand, correspondence analysis can be formulated in more generalized form as following generalized eigenvalue problem[7]. Let A and B be matrixes associated to the indicator matrix G ,

$$A = G^T G, \quad (8)$$

$$B = 2 \text{diag}(G^T G), \quad (9)$$

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

$$A = \begin{pmatrix} D_1 & N_{12} \\ N_{21} & D_2 \end{pmatrix}, \quad (10)$$

$$B = \begin{pmatrix} 2D_1 & 0 \\ 0 & 2D_2 \end{pmatrix}. \quad (11)$$

The ordinal correspondence analysis can be formulated as the following generalized eigenvalue problem with symmetric matrixes[1],[2],

$$A\mathbf{x} = \lambda B\mathbf{x}, \quad (12)$$

where $\lambda(\lambda_1, \lambda_2, \dots; \lambda_i \geq \lambda_j \text{ for } i < j)$ is the eigenvalue for this eigenvalue problem. It is noted that $\lambda_1 = 1$ and $\lambda_i \geq 0$. Eigenvector corresponding to eigenvalue λ_i is denoted by $\mathbf{x}_i$. Elements of the vector $\mathbf{x}_i$ are corresponded to the scores of correspondence analysis. Constraints for normalization on the eigenvectors are

$$\begin{aligned} \mathbf{x}_i B \mathbf{x}_i &= 1, \\ \mathbf{x}_i B \mathbf{x}_j &= 0 \quad (i \neq j) \end{aligned} \quad (13)$$

for i and j .

We have the following relation in the above two eigenvalue formulations:

$$\lambda_i = \frac{1 + \mu_i}{2} \quad (14)$$

B. Perturbation of the Result of Correspondence Analysis

Due to data addition (or small data change) to cross tabulation, if $A \rightarrow A + \delta A$ and $B \rightarrow B + \delta B$, then first-order perturbation of eigensystem, $\lambda_i \rightarrow \lambda_i + \delta \lambda_i$ and $\mathbf{x}_i \rightarrow \mathbf{x}_i + \delta \mathbf{x}_i$ is deduced as follows[7],

$$\delta \lambda_i = \mathbf{x}_i^T (\delta A - \lambda_i \delta B) \mathbf{x}_i, \quad (15)$$

$$\begin{aligned} \delta \mathbf{x}_i &= \left(-\frac{1}{2} \mathbf{x}_i^T \delta B \mathbf{x}_i \right) \mathbf{x}_i \\ &+ \sum_{j \neq i} \frac{\mathbf{x}_j^T (\delta A - \lambda_j \delta B) \mathbf{x}_i}{\lambda_j - \lambda_i} \mathbf{x}_j. \end{aligned} \quad (16)$$

We do not consider higher-order perturbation in this paper. It should be noted that these results are based on first order approximation method, so that calculated data contains numerical errors to some extent. Therefore, we should not consider the accuracy of data value, but consider the tendency of data values. In addition to this, we do not consider the special case of repeated eigenvalue problem ($\lambda_i = \lambda$ for $i \neq j$).

Trivial case that the eigenvector of $\mathbf{x}_1 = \mathbf{1}$ with $\lambda_1 = 1$ is usually omitted as a score for correspondence analysis. Usually we consider the case that the eigenvalues, $0 < \lambda_i < 1$ ($i = 2, 3, \dots$), and $\lambda_i > \lambda_j$ for ($i < j$). In practical situation we usually consider the two dimensional case of $i = 2$ and 3.

We have so far considered perturbation of each element in cross tabulation. However, consideration on whole perturbation of cross tabulation by small data variation is required. This consideration is related to chi-squared test for independence and homogeneity of cross tabulation. Perturbation of chi-squared value $\delta \chi^2$ can be regarded as the estimation of whole influence of the result of correspondence analysis by data variation.

Main result of this paper is as follows,

Proposition 1:

$$\begin{aligned} \chi^2 &\rightarrow \chi^2 + \delta \chi^2, \\ \delta \chi^2 &= n \sum_i (4(2\lambda_i - 1) \mathbf{x}_i^T (\delta A - \lambda_i \delta B) \mathbf{x}_i) \end{aligned} \quad (17)$$

where n denotes total number of data.

By using the result, overall estimation of tendency for χ^2 value variation (of cross tabulation of correspondence analysis) can be easily calculated.

III. APPLICATION EXAMPLE

A. Example of Correspondence Analysis

We show a numerical example which is a modified version of the example in references[3],[4].

Curriculum analysis system was developed based on clustering and correspondence analysis for syllabus data of higher education institutions, which was applied for Japanese universities[3],[4]. 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*. We use the data of 13 Japanese universities (departments) as shown in Table II. Department names are directly translated from Japanese[3],[4].

Totally 637 syllabuses are selected. Cluster summaries obtained in the step four of clustering procedure of the analysis system is shown in Table III. Cluster summaries indicate that cluster C1 contains many technical terms in fields of the mechanical engineering (chemistry engineering, systems engineering and the control engineering), C2 contains many technical terms in relation to computer science, C3 is related to communication engineering, C4 is to electrical engineering, C5 is to mathematics, C6 is to cognitive science.

Cross tabulation for 13 departments and 6 clusters is shown in Table IV, and the result of correspondence analysis

TABLE II
DEPARTMENTS OF 13 UNIVERSITIES (G_1)

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

TABLE III
CLUSTER SUMMARIES ANALYZING CURRICULA (G_2)

C1	(Mechanical Engineering) reaction, design, stress, movement, cycle, control, energy, flow, structure,
C2	(Computer Science) language, programming, program, computer, algorithm, C language, procedure,
C3	(Communication Engineering) information, system, probability, management, signal, technology,
C4	(Electrical Engineering) circuit, electric, electronic, light, alternating current, semiconductor, electric,
C5	(Mathematics) integral, differential, expansion, matrix, differential equation, vector, function,
C6	(Cognitive Science) image, recognition, voice, pattern, discrimination, pattern recognition,

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

	C1	C2	C3	C4	C5	C6
U07	10	1	3	0	0	2
U09	7	10	6	2	10	2
U13	17	14	13	29	3	2
U15	10	5	8	8	6	0
U19	5	8	6	4	19	2
U32	4	1	11	1	1	0
U35	58	0	2	5	3	0
U37	2	10	24	16	5	3
U50	46	2	8	2	23	1
U61	25	7	16	6	7	1
U63	7	9	18	9	8	2
U73	31	1	4	5	0	0
U78	32	1	3	1	4	0

is shown in Figure 1. Interpretation of the result of correspondence analysis is performed for understanding the cross tabulation of Table IV. Points of Figure 1 shows visually summarized information in two dimensions, which are high accumulation contribution of eigenvalues. Each point shows university or cluster. Right side of Figure 1 seems to be related to Computer Science (C2, C3, C4, C6). On the other hand, left side is related to Mechanical Engineering (C1). In this way, we can take a global view of the clustering, and it is possible for understanding the global feature of tendency[3],[4].

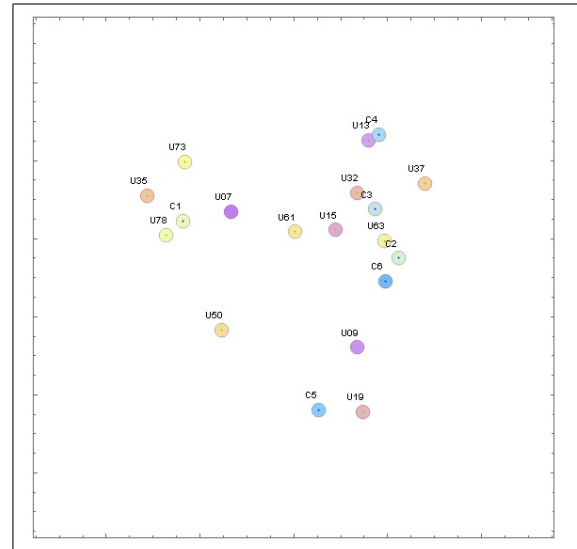


Figure 1. Result of Correspondence Analysis (x_2, x_3)

B. χ^2 value and Data Addition to the Cross tabulation in Correspondence Analysis

Table IV shows cross tabulation for 13 departments and 6 clusters. χ^2 value of this cross tabulation is 358.797 and the degree of freedom is 60. Apparently this tabulation has some tendency to express the dependency in multiple categories. Figure 1, which is a result of correspondence analysis, shows the tendency in graphical manner. Therefore, we can obtain global view of clusters in this figure.

However, elements of the cross tabulation, namely, numbers of data will be changed by data addition or data fluctuation. Then it is necessary to examine the effects on correspondence analysis result by data addition.

Utilizing the result of section II in this paper, Figure 2 illustrates a Heat Map (Temperature Map) of the effects of data addition to each element of cross tabulation. Yellow or warm colors mean the increase of χ^2 value, and blue or cool colors mean the decrease of χ^2 value by data addition. However, the result of section II is based on first order approximation method as mentioned, so that the calculated data contains numerical errors to some extent. Therefore, we should carefully consider this result.

We explain some interpretations to the result of correspondence analysis. For example, in case that a data addition to the cross tabulation would make more changes (or serious increase of χ^2 value) in the result of correspondence analysis, the interpretation on the result should be careful.

Sometimes we have to reexamine and correct the interpretation if necessary. On the other hand, whole perturbation of correspondence analysis might not be effected by the data addition. In such robust case, the interpretation of the result of correspondence analysis is regarded as stable and important features. In this way, this paper’s result contributes to estimate the effect of data variation on cross tabulation.

We consider the following cases of data addition which cause perturbations of the result of correspondence analysis (see Table IV and Figure 2).

Data addition to (U07, C6) or (1, 6) of cross tabulation: Perturbation of scores is graphically shown as the set of arrows in Figure 3. These arrows indicate directions and quantities of perturbation. Green arrows are corresponding to the variation of first category (universities, departments), and blue arrows are corresponding to the variation of second category (clusters). As shown in Figure 2, Figure 3 indicates the increase of χ^2 value, which means enlargement of differences of clusters.

Other influences by data addition to cross tabulation can be estimated by the colors in Figure 2.

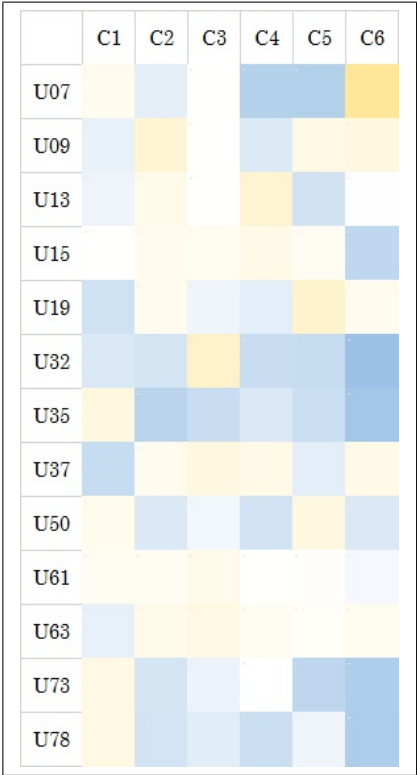


Figure 2. Heat Map of $\delta\chi^2$ Value of Cross Tabulation ($G_1 \times G_2$) due to Data Addition

IV. CONCLUSION

In this paper we showed mathematical consideration and application example of variation of chi-squared value with data addition to cross tabulation of correspondence analysis. Mathematical consideration is based on generalized eigen-value problem. Obtained Heat Map due to the data variation could be contributed for assisting our deep understanding and easy explanation on the influence by data addition.

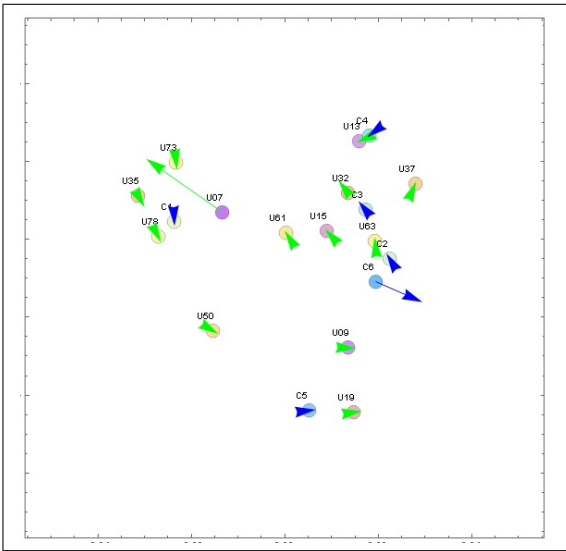


Figure 3. Perturbation of Scores of (x_2, x_3) by Data Addition to (U07, C6) or (1, 6).

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.
- [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] M. Ida, First-Order Perturbation of Correspondence Analysis with Multiple Categories, *the International Multiconference of Engineers and Computer Scientists 2015*, pp.991–994, 2015.
- [8] L.I. Schiff, *Quantum Mechanics*, McGraw-Hill, 1968.
- [9] T. Kato, *Perturbation Theory for Linear Operators, 2nd ed.*, Springer, 1980.

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 University Evaluation. His current research interests are in computer and systems science, especially in advanced database system and communication technology in higher education.