VOL. 11, NO. 3, FEBRUARY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

SIMPLIFICATION OF CORRESPONDENCE ANALYSIS FOR MORE PRECISE CALCULATION WHICH ONE QUALITATIVE VARIABLES IS TWO CATEGORICAL DATA

I. Ginanjar[1], U. S. Pasaribu[2] and A. Barra[2]
[1]Departement of Statistics, Universitas Padjadjaran, Indonesia
[2]Departement of Mathematics, Institut Teknologi Bandung, Indonesia
E-Mail: irlandia@unpad.ac.id

ABSTRACT

The calculations of correspondence analysis (CA) are using the long stages matrix operations, so that through many times rounding process, and the eigenvalues obtained by numerical process. The CA is often using standard residual matrix to calculate the singular value decomposition (SVD), this paper proves that 0 is a singular value of standard residual matrix. Based on that, this paper introduce simplification of correspondence analysis (SoCA) of $2 \times J$ contingency table where $J = 2, 3, 4, \cdots$, where obtain the simpler and more precise calculation, because it managed to minimize rounding process, also does not use the numerical process, with use standardized residuals matrix as a matrix to calculate the SVD, it is very useful for data mining techniques.

Keywords: correspondence analysis, singular value decomposition, simplification, correspondence analysis, data mining.

1. INTRODUCTION

Tufféry [1] elucidate that "the most relevant fields of data mining are those where large volumes of data have to be analyzed, sometimes with the aim of rapid decision making". Data mining is a set of methods and techniques for exploring and analyzing data sets, by automatic or semi-automatic way, to find certain provisions of the unknown or hidden, and association or trends in the data set, the outputs specifically provide useful information while reducing the quantity of data. Data mining Data mining techniques was using the inferential statistics and 'conventional' data analysis including factor analysis, clustering analysis, discriminant analysis, correspondence analysis (CA), etc. Data mining analysis for two qualitative variables often use CA.

CA was first discovered and developed in the 1960s by Jean-Paul Benzécri and friends in France [2]. This analysis is defined as the mapping technique of a contingency table in an optimal small-dimensional vector space. This analysis is also used to grouping the categories of rows and columns from the contingency table.

CA has been applied in various fields of science, including Education, Economics, Safety, Medical, and others. Zhibo *et al.* [3] analyzed and compared the competitive power of steel industry of 30 provinces in China, with data containing 16 economic indicators to reflect each province's business conditions of steel industry. Lu *et al.* [4] investigated the associations between fatality levels and influence factors that involve place, cause, time of day, month, year and province. Zalewska *et al.* [5] described the relationship between asthma, region, and age, from Epidemiology of Allergy in Poland (ECAP) data survey in years 2006-2008.

The improvement study of CA performed by: Beh [6] introduced the Elliptical confidence regions for CA, so the quality of the correspondence plot configuration is better, because it involves the cumulative percentage of eigenvalues of the times more than the dimensions used. Takagi and Yadohisa [7] introduced the CA based on the interval algebra, so that the calculation method of CA performed contingency table with shaped cells intervals. Beh [8] introduced the CA using the adjusted residuals so that data map can show the cross-tabulation of data variability.

The calculations of CA are using matrix operations with the long stages. That through many times rounding process, it also the eigenvalues obtained by numerical process, so that the values obtained are less precise. The principal coordinates estimate are use standardized residuals matrix, which is Greenacre in 2011 [9] showed that this matrix can position an outlier in the CA map.

This paper proposed to perform mathematical analysis of CA with one's qualitative variables is two categorical data, to obtain the matrix model calculation method which is simpler and more precise (to minimize rounding process, also does not use numerical process and called simplification of correspondence analysis (SoCA). Ginanjar *et al* [10] has published the application of SoCA in 2014, while this paper writes more detailed theory of SoCA.

This paper is divided into five further sections. Section 2 will describe the methods of data analysis using CA. Section 3 will describe the detailed theory of SoCA from $2 \times J$ contingency table, to obtain the principal coordinates of rows and columns matrices. Section 4 will describe the example of SoCA used fraud consumer data with two qualitatively random variables are Card type and Countries based on IP Address. The paper is concluded

(Section 6) with a brief discussion of the procedure and some possible avenues for further research.

2. CORRESPONDENT ANALYSIS (CA)

First at all, we construct a contingency table. This table is a cross-tabulation of the two categories (variables). The first category variables is the row category and the second variables is the column category, see Table-1.

Table-1 gives a $I \times J$ contingency table, where the first variable has I row categories, and the second variable has J column categories. Suppose n_{ij} is the numbers of individuals in category i on the first variable and category j in the second variable, and the total of each row is $n_{i\bullet}$ and the total of column is $n_{\bullet j}$, where $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$. $n_{i\bullet}$ and $n_{\bullet j}$ are called marginal for first variable with the category i and second variable with the category j, respectively. The grand total is the total number of individuals which is denoted by n.

Table-1. Contingency table form.

Row category	Column category				
	Column 1	**Column 2**	**⋯**	**Column J**	**Total**
Row 1	n_{11}	n_{12}	⋯	n_{1J}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	⋯	n_{2J}	$n_{2\bullet}$
⋮	⋮	⋮	⋱	⋮	⋮
Row I	n_{I1}	n_{I2}	⋯	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	⋯	$n_{\bullet J}$	n

2.1 CA algorithm

Consider the following contingency table (cross-tabulation matrix):

$$\mathbf{N} = \left(n_{ij}\right). \tag{1}$$

Calculate the empirical joint distributions of row and column with the following formula:

$$\mathbf{P} = \left(p_{ij}\right) = \left(\frac{n_{ij}}{n}\right). \tag{2}$$

Calculate the vectors of row ($\vec{r}$) and column ($\vec{c}$) marginal distribution:

$$\vec{r} = \left(r_i\right) = \left(\sum_{j=1}^{J} p_{ij}\right) = \left(\frac{n_{i\bullet}}{n}\right), \tag{3}$$

and

$$\vec{c} = \left(c_j\right) = \left(\sum_{i=1}^{I} p_{ij}\right) = \left(\frac{n_{\bullet j}}{n}\right). \tag{4}$$

Let the diagonal matrices of the row and column marginal distribution is denoted by R and C.

The standardized residuals matrix [9], which is a matrix that represents the association in contingency table, is calculated as follows:

$$\mathbf{S} = \left(s_{ij}\right) = \left(\frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}\right) = \left(\frac{n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}}{\sqrt{n_{i\bullet} n_{\bullet j}}}\right). \tag{5}$$

The singular value decomposition (SVD) of the standardized residuals matrix is given as follows:

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}^t, \tag{6}$$

where $\mathbf{U}^t\mathbf{U} = \mathbf{I}$, $\mathbf{V}^t\mathbf{V} = \mathbf{I}$, and

$$\mathbf{D} = \mathrm{diag}\left(\sqrt{\vec{\lambda}_l}\right) = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_L} & \cdots & 0 \end{pmatrix}, \tag{7}$$

where $\sqrt{\vec{\lambda}_l}, l = 1,2,\cdots,L$ is square root of descending eigenvalues $\left(\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_L}\right)$ from $\mathbf{S}\mathbf{S}^t$ or $\mathbf{S}^t\mathbf{S}$, and L is a number of eigenvalues are obtained.

The matrix of principal coordinates of row and column, are calculated as follows:

$$\mathbf{Y} = \mathbf{R}^{-\frac{1}{2}}\mathbf{U}\mathbf{D}, \tag{7}$$

and column principal coordinates is:

$$\mathbf{Z} = \mathbf{C}^{-\frac{1}{2}}\mathbf{V}\mathbf{D}^{t}, \quad (8)$$

respectively.

The first two columns of the principal coordinates of row and column are the coordinates to build the two-dimensional map.

3. MATHEMATICAL ANALYSIS (CA)

3.1 Matrix for calculates the CA eigenvalues

Standard residual matrix can be calculated directly from the elements of cross-tabulation matrix (Equation (5)). SVD calculation begins by calculating the eigenvalues of the matrix product $\mathbf{S}\mathbf{S}^{t}$ or $\mathbf{S}^{t}\mathbf{S}$. Let $I < J$ then the size of the matrix $\mathbf{S}\mathbf{S}^{t} < \mathbf{S}^{t}\mathbf{S}$, thereby calculating eigenvalues of $\mathbf{S}\mathbf{S}^{t}$ will be simplified.

Lemma 1

If the size of a matrix $\mathbf{N}$ is $I \times J$ (Equation (1)), and $\mathbf{S}\mathbf{S}^{t} = \mathbf{A}$, then the elements of $\mathbf{A}$ is

$$a_{ik} = \frac{1}{\sqrt{n_{i\bullet}n_{k\bullet}}}\left(\sum_{j=1}^{J}\frac{n_{ij}n_{kj}}{n_{\bullet j}} - \frac{n_{i\bullet}n_{k\bullet}}{n}\right), \quad (9)$$

where i and $k = 1,2,\cdots,I$.

Proof:

$$\mathbf{A} = (a_{ik}) = \left(\sum_{j=1}^{J} s_{ij}s_{kj}\right)$$

with the result that

$$a_{ik} = \sum_{j=1}^{J}\frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}\frac{p_{kj} - r_k c_j}{\sqrt{r_k c_j}}$$

$$= \frac{1}{\sqrt{r_i r_k}}\left(\sum_{j=1}^{J}\frac{p_{ij}p_{kj}}{c_j}\right) - \sqrt{r_i r_k}$$

based equations (2) and (3), so,

$$a_{ik} = \frac{1}{\sqrt{\frac{n_{i\bullet}}{n}\frac{n_{k\bullet}}{n}}}\left(\sum_{j=1}^{J}\frac{\frac{n_{ij}}{n}\frac{n_{kj}}{n}}{\frac{n_{\bullet j}}{n}}\right) - \sqrt{\frac{n_{i\bullet}}{n}\frac{n_{k\bullet}}{n}}$$

$$= \frac{1}{\sqrt{n_{i\bullet}n_{k\bullet}}}\left(\sum_{j=1}^{J}\frac{n_{ij}n_{kj}}{n_{\bullet j}} - \frac{n_{i\bullet}n_{k\bullet}}{n}\right). \blacksquare$$

Lemma 1 is a formula to calculate $\mathbf{S}\mathbf{S}^{t}$ which is simpler and more precise, because each element of the matrix is calculated directly from the elements of the contingency Table. The variable often consists of two categories ($I = 2$), e.g. gender, yes or no, and others. Based on that, this paper formulates the lemma of $\mathbf{S}\mathbf{S}^{t}$ for $2 \times J$ contingency table.

Lemma 2

If the size of a matrix $\mathbf{N}$ is $2 \times J$ (Equation (1)), and $\mathbf{S}\mathbf{S}^{t} = \mathbf{A}$, then the elements of $\mathbf{A}$ is obtained

$$a_{11} = \frac{n_{2\bullet}}{n_{1\bullet}}\, a_{22} = -\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}\, a_{12} = -\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}\, a_{21}. \quad (10)$$

Proof

Let $n_{1\bullet} = \frac{n_{1\bullet}}{n_{2\bullet}}n_{2\bullet}$

$$\sum_{j=1}^{J}\left(n_{1j} - n_{2j}\right) = \frac{n_{1\bullet}}{n_{2\bullet}}n_{2\bullet} - n_{2\bullet}$$

$$\prod_{j=1}^{J} n_{\bullet j}\left(\sum_{j=1}^{J}\left(n_{1j} - n_{2j}\right)\right) = \prod_{j=1}^{J} n_{\bullet j}\left(\frac{n_{1\bullet}}{n_{2\bullet}}n_{2\bullet} - n_{2\bullet}\right)$$

$$\frac{1}{n_{1\bullet}}\left(\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}} - \frac{n_{1\bullet}^2}{n}\right) = \frac{n_{2\bullet}}{n_{1\bullet}}\left(\frac{1}{n_{2\bullet}}\left(\sum_{j=1}^{J}\frac{n_{2j}^2}{n_{\bullet j}} - \frac{n_{2\bullet}^2}{n}\right)\right),$$

by Lemma 1 is obtained $a_{11} = \frac{n_{2\bullet}}{n_{1\bullet}}a_{22}$.

Let $n_{1\bullet} = \frac{n_{1\bullet}}{n_{2\bullet}}n_{2\bullet}$

$$\frac{\sum_{j=1}^{J} n_{1j}}{n_{2\bullet}} = \frac{n_{1\bullet}}{n_{2\bullet}} \times \frac{n_{1\bullet} + n_{2\bullet}}{n}$$

$$\frac{1}{n_{2\bullet}}\sum_{j=1}^{J} n_{1j}\frac{n_{1j} + n_{2j}}{n_{1j} + n_{2j}} = \frac{n_{1\bullet}\left(n_{1\bullet} + n_{2\bullet}\right)}{n_{2\bullet}n}$$

$$\frac{1}{\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}n_{1\bullet}n_{2\bullet}}}\sum_{j=1}^{J}\frac{n_{1j} + n_{2j}}{n_{\bullet j}} - \frac{n_{1\bullet}\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}n_{1\bullet}n_{2\bullet}}}{n_{2\bullet}n}$$

$$= -\left(\frac{1}{n_{2\bullet}}\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}} - \frac{n_{1\bullet}^2}{n_{2\bullet}n}\right)$$

$$\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}}\left(\frac{1}{\sqrt{n_{1\bullet}n_{2\bullet}}}\sum_{j=1}^{J}\frac{n_{1j}+n_{2j}}{n_{\bullet j}}-\frac{\sqrt{n_{1\bullet}n_{2\bullet}}}{n}\right)$$

$$=-\frac{n_{1\bullet}}{n_{2\bullet}}\left(\frac{1}{n_{1\bullet}}\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{n_{1\bullet}^2}{n}\right)$$

by Lemma 1 is obtained, $\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}}a_{12}=\frac{n_{1\bullet}}{n_{2\bullet}}a_{11}$

$$a_{11}=-\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}a_{12}=-\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}a_{21}$$

then,

$$a_{11}=\frac{n_{2\bullet}}{n_{1\bullet}}a_{22}=-\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}a_{12}=-\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}a_{21}\,.\ \blacksquare$$

3.2 Eigenvalues and orthonormal eigenvector

The equations for calculating eigenvalues are

$$\det(\lambda\mathbf{I}-\mathbf{SS}^t)=0\,. \tag{11}$$

The calculation of $\mathbf{SS}^t$ eigenvalues has uniqueness. $\mathbf{SS}^t$ is a real symmetric matrix that has the following properties: 1) is positive semidefinite, 2) is always diagonalizable, 3) has orthogonal eigenvectors, and 4) has only real eigenvalues. Other than that 0 is eigenvalue of $\mathbf{SS}^t$, it is proved in Theorem 1.

Theorem 1

If the size of a matrix $\mathbf{N}$ is $I \times J$ (Equation (1)), and $\mathbf{SS}^t=\mathbf{A}$, then 0 is eigenvalue of $\mathbf{A}$.

Proof

Let $\vec{w}=\left(\sqrt{r_1},\sqrt{r_2},\cdots\sqrt{r_I}\right)$ with $\sum_{i=1}^{I}r_1=1$.

Then

$$\mathbf{S}^t\vec{w}=\begin{pmatrix}s_{11}&\cdots&s_{I1}\\ \vdots&\ddots&\vdots\\ s_{1J}&\cdots&s_{IJ}\end{pmatrix}\begin{pmatrix}\sqrt{r_1}\\ \vdots\\ \sqrt{r_I}\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{I}\sqrt{r_i}s_{i1}\\ \vdots\\ \sum_{i=1}^{I}\sqrt{r_i}s_{iJ}\end{pmatrix}$$

$$=\begin{pmatrix}\sum_{i=1}^{I}\frac{p_{i1}-r_ic_1}{\sqrt{c_1}}\\ \vdots\\ \sum_{i=1}^{I}\frac{p_{iJ}-r_ic_J}{\sqrt{c_J}}\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{c_1}}\left(\sum_{i=1}^{I}p_{i1}-\sum_{i=1}^{I}r_ic_1\right)\\ \vdots\\ \frac{1}{\sqrt{c_J}}\left(\sum_{i=1}^{I}p_{iJ}-\sum_{i=1}^{I}r_ic_J\right)\end{pmatrix}$$

$$=\begin{pmatrix}\frac{1}{\sqrt{c_1}}\left(c_1-c_1\sum_{i=1}^{I}r_i\right)\\ \vdots\\ \frac{1}{\sqrt{c_J}}\left(c_J-c_J\sum_{i=1}^{I}r_i\right)\end{pmatrix}=\vec{0}$$

so, we obtained

$\mathbf{A}\vec{w}=\left(\mathbf{SS}^t\right)\vec{w}=\mathbf{S}\left(\mathbf{S}^t\vec{w}\right)=\mathbf{S}\vec{0}=\vec{0}$, has a non-trivial solution. Then 0 is eigenvalue of $\mathbf{A}$. ■

The eigenvalues calculation involves the calculation of polynomial roots, which has been obtained through numerical process. The eigenvalues calculation of the matrix size 2×2 can be performed using mathematical analysis, so the eigenvalues are obtained more precise. If the size of a matrix $\mathbf{SS}^t$ is 2×2, with Lemma1, Lemma 2, and Theorem 1, then the eigenvalues can be calculated directly from the elements of the contingency table, which is described in Lemma 3.

Lemma 3

If the size of a matrix $\mathbf{N}$ is $2 \times J$ (Equation (1)), and $\mathbf{SS}^t=\mathbf{A}$, then the first eigenvalues (λ_1) corresponding to $\mathbf{A}$ is $\frac{n}{n_{1\bullet}n_{2\bullet}}\left(\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{n_{1\bullet}^2}{n}\right)$, and the second eigenvalues (λ_2) corresponding to $\mathbf{A}$ is 0.

Proof

From Theorem 1 than the result that $\lambda_2=0$.

From Lemma 2 be discovered $a_{22}=\frac{n_{1\bullet}}{n_{2\bullet}}a_{11}$, than the result that:

$$\lambda_1=\mathrm{tr}(\mathbf{A})=a_{11}+a_{22}=a_{11}+\frac{n_{1\bullet}}{n_{2\bullet}}a_{11}=a_{11}\left(1+\frac{n_{1\bullet}}{n_{2\bullet}}\right)$$

with Lemma 1 than the result that:

$$\lambda_1=\frac{1}{n_{1\bullet}}\left(\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{n_{1\bullet}^2}{n}\right)\left(1+\frac{n_{1\bullet}}{n_{2\bullet}}\right)$$

$$\lambda_1=\frac{n}{n_{1\bullet}n_{2\bullet}}\left(\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{n_{1\bullet}^2}{n}\right).\ \blacksquare$$

As a result of the Lemma 3 it can be seen that for the size of a matrix $\mathbf{N}$ is $2 \times J$, will be obtained 1-dimensional principal coordinate, with 100% contribution ratio.

The equations for calculating eigenvector are

$$\left(\lambda\mathbf{I}-\mathbf{SS}^t\right)\vec{u}^*=\vec{0} \tag{12}$$

If the size of a matrix $\mathbf{SS}^t$ is 2×2, and λ is the eigenvalues of $\mathbf{SS}^t$, then with Lemma 2, and Theorem 1 the orthonormal eigenvectors can be calculated. The orthonormal eigenvectors are the columns for the matrix $\mathbf{U}$ (Equation (6)), which can be calculated directly from the elements of the contingency table, and written on Lemma 4.

Lemma 4

If the size of a matrix $\mathbf{N}$ is $2 \times J$ (Equation (1)), and $\mathbf{SS}^t = \mathbf{A}$, then the orthonormal eigenvectors from $\mathbf{A}$ are

$$\begin{pmatrix} \vec{u}_1 & \vec{u}_2 \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} -\sqrt{n_{2\bullet}} & \sqrt{n_{1\bullet}} \\ \sqrt{n_{1\bullet}} & \sqrt{n_{2\bullet}} \end{pmatrix}.$$

Proof

With Lemma 1, Theorem 1, and the Equation (12) are obtained

$$\begin{pmatrix} a_{11} - \lambda & -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} a_{11} \\ -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} a_{11} & \frac{n_{1\bullet}}{n_{2\bullet}} a_{11} - \lambda \end{pmatrix} \begin{pmatrix} u_{11}^* \\ u_{12}^* \end{pmatrix} = \vec{0},$$

let the eigenvector for $\lambda_1 = a_{11} + \frac{n_{1\bullet}}{n_{2\bullet}} a_{11}$ is:

$$\begin{pmatrix} -\frac{n_{1\bullet}}{n_{2\bullet}} a_{11} & -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} a_{11} \\ -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} a_{11} & -a_{11} \end{pmatrix} \begin{pmatrix} u_{11}^* \\ u_{12}^* \end{pmatrix} = \vec{0}$$

by using elementary row operations, so obtained

$$\begin{pmatrix} 1 & \sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_{11}^* \\ u_{12}^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow u_{11}^* = -u_{12}^* \sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}}.$$

let $u_{12}^* = t$, so the eigenvector is

$$\begin{pmatrix} u_{11}^* \\ u_{12}^* \end{pmatrix} = \begin{pmatrix} -t\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}} \\ t \end{pmatrix}$$

The orthonormal eigenvector is eigenvector vector with length is 1, so the orthonormal eigenvector for λ_1, is eigenvector for λ_1 which each element is divided by

$$\left\| \vec{u}_1^* \right\| = \sqrt{\left(-t\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}} \right)^2 + t^2} = \frac{t\sqrt{n}}{\sqrt{n_{1\bullet}}},$$

so the orthonormal eigenvector for λ_1 is:

$$\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = \begin{pmatrix} -t\sqrt{\frac{n_{2\bullet}}{n_{1\bullet}}} \\ t \end{pmatrix} \frac{\sqrt{n_{1\bullet}}}{t\sqrt{n}} = \frac{1}{\sqrt{n}} \begin{pmatrix} -\sqrt{n_{2\bullet}} \\ \sqrt{n_{1\bullet}} \end{pmatrix},$$

and the eigenvector for $\lambda_2 = 0$ is:

$$\begin{pmatrix} a_{11} & -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} a_{11} \\ -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} a_{11} & \frac{n_{1\bullet}}{n_{2\bullet}} a_{11} \end{pmatrix} \begin{pmatrix} u_{21}^* \\ u_{22}^* \end{pmatrix} = \vec{0}$$

by using elementary row operations, so obtained

$$\begin{pmatrix} 1 & -\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_{21}^* \\ u_{22}^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow u_{21}^* = u_{22}^* \sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}}.$$

Let $u_{22}^* = t$, so the eigenvector is

$$\begin{pmatrix} u_{21}^* \\ u_{22}^* \end{pmatrix} = \begin{pmatrix} t\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} \\ t \end{pmatrix},$$

the orthonormal eigenvector for λ_2, is eigenvector for λ_2 which each element is divided by

$$\left\| \vec{u}_2^* \right\| = \sqrt{\left(t\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} \right)^2 + t^2} = \frac{t\sqrt{n}}{\sqrt{n_{2\bullet}}}$$

so the orthonormal eigenvector for λ_2 is:

$$\begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} = \begin{pmatrix} t\sqrt{\frac{n_{1\bullet}}{n_{2\bullet}}} \\ t \end{pmatrix} \frac{\sqrt{n_{2\bullet}}}{t\sqrt{n}} = \frac{1}{\sqrt{n}} \begin{pmatrix} \sqrt{n_{1\bullet}} \\ \sqrt{n_{2\bullet}} \end{pmatrix},$$

then, the orthonormal eigenvectors from $\mathbf{A}$ are:

$$\begin{pmatrix} \vec{u}_1 & \vec{u}_2 \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} -\sqrt{n_{2\bullet}} & \sqrt{n_{1\bullet}} \\ \sqrt{n_{1\bullet}} & \sqrt{n_{2\bullet}} \end{pmatrix}. \blacksquare$$

With using Equation (5), Equation (6), and Lemma 4, than the elements of the orthonormal eigenvectors $(\vec{v}_1)$ of the first eigenvalues λ_1 corresponding to $\mathbf{S}^t\mathbf{S}$, this can be calculated directly from the elements in contingency table, which is described in Lemma 5.

Lemma 5

If the size of a matrix $\mathbf{N}$ is $2 \times J$ (Equation (1)), and $\mathbf{SS}^t = \mathbf{A}$, then the elements of orthonormal eigenvectors $(\vec{v}_1)$ from the first eigenvalues (λ_1) corresponding to $\mathbf{S}^t\mathbf{S}$ is $v_{j1} = \frac{n_{2j}n_{1\bullet} - n_{1j}n_{2\bullet}}{\sqrt{\lambda_1 n_{1\bullet}n_{2\bullet}n_{\bullet j}n}}$ for $j = 1, 2, \cdots, J$.

Proof

Based on the SVD (Equation (6)) are obtained:

$$s_{1j} = u_{11}v_{j1}\sqrt{\lambda_1} + u_{21}v_{j2}\sqrt{\lambda_2}$$

and

$$s_{2j} = u_{12}v_{j1}\sqrt{\lambda_1} + u_{22}v_{j2}\sqrt{\lambda_2} .$$

Calculate v_{j2} from s_{2j},

$$s_{2j} - u_{12}v_{j1}\sqrt{\lambda_1} = u_{22}v_{j2}\sqrt{\lambda_2}$$

$$\frac{s_{2j} - u_{12}v_{j1}\sqrt{\lambda_1}}{u_{22}\sqrt{\lambda_2}} = v_{j2},$$

substitution v_{j2} into s_{1j},

$$s_{1j} = u_{11}v_{j1}\sqrt{\lambda_1} + u_{21}\sqrt{\lambda_2}\frac{s_{2j} - u_{12}v_{j1}\sqrt{\lambda_1}}{u_{22}\sqrt{\lambda_2}}$$

$$v_{j1} = \frac{s_{1j}u_{22} - u_{21}s_{2j}}{(u_{11}u_{22} - u_{21}u_{12})\sqrt{\lambda_1}}$$

with Lemma 4 than the result that:

$$v_{j1} = \frac{1}{\sqrt{n}}\frac{s_{1j}\sqrt{n_{2\bullet}} - \sqrt{n_{1\bullet}}s_{2j}}{\left(\sqrt{n_{2\bullet}}\sqrt{n_{2\bullet}} - \sqrt{n_{1\bullet}}\sqrt{n_{1\bullet}}\right)\sqrt{\lambda_1}}$$

$$= \frac{s_{1j}\sqrt{n_{2\bullet}} - \sqrt{n_{1\bullet}}s_{2j}}{-\sqrt{\lambda_1}\sqrt{n}}$$

Based on Equation (5), than the result that:

$$v_{j1} = \frac{\frac{n_{1j} - \frac{n_{1\bullet}n_{\bullet j}}{n}}{\sqrt{n_{1\bullet}n_{\bullet j}}}\sqrt{n_{2\bullet}} - \frac{n_{2j} - \frac{n_{2\bullet}n_{\bullet j}}{n}}{\sqrt{n_{2\bullet}n_{\bullet j}}}\sqrt{n_{1\bullet}}}{-\sqrt{\lambda_1}\sqrt{n}}$$

$$= \frac{\frac{n_{1j}\sqrt{n_{2\bullet}}}{\sqrt{n_{1\bullet}n_{\bullet j}}} - \frac{n_{1\bullet}n_{\bullet j}\sqrt{n_{2\bullet}}}{n\sqrt{n_{1\bullet}n_{\bullet j}}} - \frac{n_{2j}\sqrt{n_{1\bullet}}}{\sqrt{n_{2\bullet}n_{\bullet j}}} + \frac{n_{2\bullet}n_{\bullet j}\sqrt{n_{1\bullet}}}{n\sqrt{n_{2\bullet}n_{\bullet j}}}}{-\sqrt{\lambda_1}\sqrt{n}}$$

$$= \frac{\frac{n_{1j}\sqrt{n_{2\bullet}}}{\sqrt{n_{1\bullet}n_{\bullet j}}} - \frac{n_{2j}\sqrt{n_{1\bullet}}}{\sqrt{n_{2\bullet}n_{\bullet j}}} + \frac{\sqrt{n_{2\bullet}n_{\bullet j}}\sqrt{n_{1\bullet}}}{n} - \frac{\sqrt{n_{1\bullet}n_{\bullet j}}\sqrt{n_{2\bullet}}}{n}}{-\sqrt{\lambda_1}\sqrt{n}}$$

$$= \frac{n_{2j}n_{1\bullet} - n_{1j}n_{2\bullet}}{\sqrt{\lambda_1 n_{1\bullet}n_{2\bullet}n_{\bullet j}n}} . \blacksquare$$

Because of $v_{j2} = \frac{s_{2j} - u_{12}v_{j1}\sqrt{\lambda_1}}{u_{22}\sqrt{\lambda_2}}$ and $\lambda_2 = 0$, then the value of v_{j2} is undefined.

3.3 Principal coordinates of rows and columns

The main objective of the CA is to estimate the principal coordinates, for mapping the row and column categories of a contingency table. The principal coordinates are the linear combinations vectors from each row or column category.

Based on Lemma 3, Lemma 4, and Lemma 5, then the author can make an equation to estimate the principal coordinates, which is simpler and more precise, which is calculated directly from the elements in contingency table. It was described in Theorem 2.

Theorem 2

If the size of a matrix $\mathbf{N}$ is $2 \times J$ (Equation (1)), then the row principal coordinates $\mathbf{Y}$ is

$$\mathbf{Y} = \sqrt{\frac{n}{n_{1\bullet}^2}\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}} - 1}\begin{pmatrix} -1 & 0 & \cdots & 0 \\ \frac{n_{1\bullet}}{n_{2\bullet}} & 0 & \cdots & 0 \end{pmatrix}$$

and the column principal coordinates $\mathbf{Z}$ is

$$\mathbf{Z} = \begin{pmatrix} \vec{z}_{j1} & \vec{0} \end{pmatrix} \text{ where } z_{j1} = \frac{n_{2j}n_{1\bullet} - n_{1j}n_{2\bullet}}{n_{\bullet j}\sqrt{n_{1\bullet}n_{2\bullet}}} .$$

Proof

The row principal coordinates (Equation (6)) is:

$$\mathbf{Y}=\begin{pmatrix}\frac{1}{\sqrt{r_1}} & 0\\ 0 & \frac{1}{\sqrt{r_2}}\end{pmatrix}\frac{1}{\sqrt{n}}\begin{pmatrix}-\sqrt{n_{2\bullet}} & \sqrt{n_{1\bullet}}\\ \sqrt{n_{1\bullet}} & \sqrt{n_{2\bullet}}\end{pmatrix}\begin{pmatrix}\sqrt{\lambda_1} & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\end{pmatrix}$$

$$=\frac{\sqrt{\lambda_1}}{\sqrt{n}}\begin{pmatrix}-\frac{\sqrt{n_{2\bullet}}}{\sqrt{r_1}} & 0 & \cdots & 0\\ \frac{\sqrt{n_{1\bullet}}}{\sqrt{r_2}} & 0 & \cdots & 0\end{pmatrix}=\sqrt{\lambda_1}\begin{pmatrix}-\frac{\sqrt{n_{2\bullet}}}{\sqrt{n_{1\bullet}}} & 0 & \cdots & 0\\ \frac{\sqrt{n_{1\bullet}}}{\sqrt{n_{2\bullet}}} & 0 & \cdots & 0\end{pmatrix}$$

with Lemma 4 than the result that:

$$\mathbf{Y}=\sqrt{\frac{n}{n_{1\bullet}n_{2\bullet}}\left(\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{n_{1\bullet}^2}{n}\right)}\begin{pmatrix}-\frac{\sqrt{n_{2\bullet}}}{\sqrt{n_{1\bullet}}} & 0 & \cdots & 0\\ \frac{\sqrt{n_{1\bullet}}}{\sqrt{n_{2\bullet}}} & 0 & \cdots & 0\end{pmatrix}$$

$$=\sqrt{\frac{1}{n_{1\bullet}}\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{n_{1\bullet}}{n}}\begin{pmatrix}-\sqrt{\frac{n}{n_{1\bullet}}} & 0 & \cdots & 0\\ \frac{\sqrt{n_{1\bullet}n}}{n_{2\bullet}} & 0 & \cdots & 0\end{pmatrix}$$

$$=\sqrt{\frac{1}{n_{1\bullet}^2}\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-\frac{1}{n}}\begin{pmatrix}-\sqrt{n} & 0 & \cdots & 0\\ \frac{n_{1\bullet}\sqrt{n}}{n_{2\bullet}} & 0 & \cdots & 0\end{pmatrix}$$

$$=\sqrt{\frac{n}{n_{1\bullet}^2}\sum_{j=1}^{J}\frac{n_{1j}^2}{n_{\bullet j}}-1}\begin{pmatrix}-1 & 0 & \cdots & 0\\ \frac{n_{1\bullet}}{n_{2\bullet}} & 0 & \cdots & 0\end{pmatrix}$$

The column principal coordinates (Equation (7)) is:

$$\mathbf{Z}=\begin{pmatrix}\frac{1}{\sqrt{c_1}} & 0 & \cdots & 0\\ 0 & \frac{1}{\sqrt{c_2}} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sqrt{c_J}}\end{pmatrix}\begin{pmatrix}v_{11} & v_{12} & \cdots & v_{1J}\\ v_{21} & v_{22} & \cdots & v_{2J}\\ \vdots & \vdots & \ddots & \vdots\\ v_{J1} & v_{J2} & \cdots & v_{JJ}\end{pmatrix}\begin{pmatrix}\sqrt{\lambda_1} & 0\\ 0 & 0\\ \vdots & \vdots\\ 0 & 0\end{pmatrix}$$

$$=\begin{pmatrix}\sqrt{\lambda_1}\frac{v_{11}}{\sqrt{c_1}} & 0\\ \sqrt{\lambda_1}\frac{v_{21}}{\sqrt{c_2}} & 0\\ \vdots & \vdots\\ \sqrt{\lambda_1}\frac{v_{J1}}{\sqrt{c_J}} & 0\end{pmatrix}$$

with Lemma 5 than the result that:

$$\mathbf{Z}=\begin{pmatrix}\vec{z}_{j1} & \vec{0}\end{pmatrix}=\begin{pmatrix}\sqrt{\lambda_1}\frac{v_{11}}{\sqrt{c_1}} & 0\\ \sqrt{\lambda_1}\frac{v_{21}}{\sqrt{c_2}} & 0\\ \vdots & \vdots\\ \sqrt{\lambda_1}\frac{v_{J1}}{\sqrt{c_J}} & 0\end{pmatrix}\text{ where }z_{j1}=\sqrt{\lambda_1}\frac{v_{j1}}{\sqrt{c_j}}$$

$$z_{j1}=\sqrt{\lambda_1}\frac{n_{2j}n_{1\bullet}-n_{1j}n_{2\bullet}}{\sqrt{\lambda_1 n_{1\bullet}n_{2\bullet}n_{\bullet j}nc_j}}=\frac{n_{2j}n_{1\bullet}-n_{1j}n_{2\bullet}}{n_{\bullet j}\sqrt{n_{1\bullet}n_{2\bullet}}}.\ \blacksquare$$

4. EXAMPLE

This section is presents an example to describe the steps of SoCA. The example is used fraud consumer data with two qualitatively random variables. The variables are Card type and Countries based on IP Address. The data are obtained from an online payment gateway company in Indonesia.

4.1 Transform two qualitative variables into a contingency table

Transform two qualitative variables (card type and IPID Country) into contingency table, so we got the data shown in Table-2.

Table-2. Contingency table: card type and IPID Country.

Card Type	IPID Country										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Sum**
Visa	1	1	2	2	54	0	5	3	3	2	73
MasterCard	0	8	0	2	27	1	8	1	0	0	47
Sum	1	9	2	4	81	1	13	4	3	2	120

4.2 Compute the principal coordinates of rows and column

Compute the principal coordinates of rows using Theorem 2, where

$$\mathbf{Y} = \sqrt{\frac{120}{73^2}\left(\frac{1^2}{1}+\frac{1^2}{9}+\frac{2^2}{2}+\frac{2^2}{4}+\frac{54^2}{81}+\frac{0^2}{1}+\frac{5^2}{13}+\frac{3^2}{4}+\frac{3^2}{3}+\frac{2^2}{2}\right)-1}$$

$$\times\begin{pmatrix}-1 & 0 & \cdots & 0\\ \frac{73}{47} & 0 & \cdots & 0\end{pmatrix}$$

$$=\sqrt{0.1098}\begin{pmatrix}-1 & 0 & \cdots & 0\\ \frac{73}{47} & 0 & \cdots & 0\end{pmatrix}=\begin{pmatrix}-0.3314 & 0 & \cdots & 0\\ 0.5147 & 0 & \cdots & 0\end{pmatrix}$$

Compute the principal coordinates of columns using Theorem 2, where

$$z_{11}=\frac{n_{21}n_{1\bullet}-n_{11}n_{2\bullet}}{n_{\bullet 1}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{0\times 73-1\times 47}{1\sqrt{73\times 47}}=-0.8024,$$

$$z_{21}=\frac{n_{22}n_{1\bullet}-n_{12}n_{2\bullet}}{n_{\bullet 2}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{8\times 73-1\times 47}{9\sqrt{73\times 47}}=1.0186,$$

$$z_{31}=\frac{n_{23}n_{1\bullet}-n_{13}n_{2\bullet}}{n_{\bullet 3}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{0\times 73-2\times 47}{2\sqrt{73\times 47}}=-0.8024,$$

$$z_{41}=\frac{n_{24}n_{1\bullet}-n_{14}n_{2\bullet}}{n_{\bullet 4}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{2\times 73-2\times 47}{4\sqrt{73\times 47}}=0.2219,$$

$$z_{51}=\frac{n_{25}n_{1\bullet}-n_{15}n_{2\bullet}}{n_{\bullet 5}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{27\times 73-54\times 47}{81\sqrt{73\times 47}}=-0.1195,$$

$$z_{61}=\frac{n_{26}n_{1\bullet}-n_{16}n_{2\bullet}}{n_{\bullet 6}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{1\times 73-0\times 47}{1\sqrt{73\times 47}}=1.2463,$$

$$z_{71}=\frac{n_{27}n_{1\bullet}-n_{17}n_{2\bullet}}{n_{\bullet 7}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{8\times 73-5\times 47}{13\sqrt{73\times 47}}=0.4583,$$

$$z_{81}=\frac{n_{28}n_{1\bullet}-n_{18}n_{2\bullet}}{n_{\bullet 8}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{1\times 73-3\times 47}{4\sqrt{73\times 47}}=-0.2902,$$

$$z_{91}=\frac{n_{29}n_{1\bullet}-n_{19}n_{2\bullet}}{n_{\bullet 9}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{0\times 73-3\times 47}{3\sqrt{73\times 47}}=-0.8024,$$

$$z_{10\,1}=\frac{n_{2\,10}n_{1\bullet}-n_{1\,10}n_{2\bullet}}{n_{\bullet\,10}\sqrt{n_{1\bullet}n_{2\bullet}}}=\frac{0\times 73-2\times 47}{2\sqrt{73\times 47}}=-0.8024.$$

4.3 Plot the projections of data

The first two columns of the principal coordinates of rows and columns are the coordinates to make the map of SoCA where presented in Figure-1.

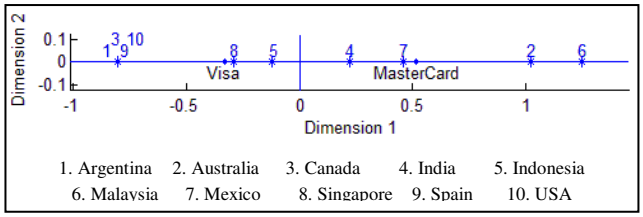

Figure-1. SoCA map of card type and IPID Country.

Figure-1 show that most of the fraud customer from Argentina, Canada, Spain, and the USA using Visa, and most of the fraud customers from Australia and Malaysia using MasterCard. The fraud customers from Singapore and Indonesia predominant use Visa and MasterCard partly use. The fraud customer from Mexico and India predominant use MasterCard and Visa partly use.

5. DISCUSSION AND CONCLUSIONS

This paper has shown that SoCA obtains the simpler and more precise calculation method, because it managed to minimize rounding process, also does not use numerical process. The standardized residuals matrix is used for calculate the SVD, which rare objects are often positioned as outliers in CA map, which gives the impression that they are highly influential, but their low weight offsets their distant positions and reduces their effect on the results. For each $\mathbf{N}_{(2\times J)}$ is a cross tabulation matrix (Equation 1) where $j = 1, 2, \cdots, J$, will be obtained one-dimensional visualization, with the contribution ratio of 100%. The idea of SoCA of $2 \times J$ contingency tables can be highly enlightening as to the properties of these methods. In future work we will generalize SoCA to be used for $I \times J$ contingency tables where $I = 3, 4, 5$ and $J = 2, 3, \cdots$.

ACKNOWLEDGEMENT

The authors thanks to Dr. Sapto W. Indartno, who have provided direction, correction and perfection solution for this paper, we also thank the editor. Work supported by grants BPP-DN from Indonesian directorate general of higher education.

REFERENCES

[1] S. Tufféry. 2011. Data Mining and Statistics for Decision Making. John Wiley and Sons, Ltd, United Kingdom.

[2] J.P. Benzécri. 1992. Correspondence Analysis Handbook. Marcel Dekker Inc., New York, USA.

[3] R. Zhibo, L. Kai and W. Wei. 2012. The Comparative Study of the Competitive Power of the Steel Industry

of Every Province in China Based on Correspondence Analysis Method. Physics Procedia. 25: 1671-1674.

[4] S. Lu, P. Mei, J. Wang and H. Zhang. 2012. Fatality and influence factors in high-casualty fires: A correspondence analysis. Safety Science. 50: 1019-1033.

[5] M. Zalewska, K. Furmańczyk, S. Jaworski, W. Niemiro and B. Samoliński. 2013. The Prevalence of Asthma and Declared Asthma in Poland on the Basis of ECAP Survey Using Correspondence analysis. Computational and Mathematical Methods in Medicine. 2013: 1-8.

[6] E.J. Beh. 2010. Elliptical confidence regions for simple correspondence analysis. Journal of Statistical Planning and Inference. 140: 2582-2588.

[7] I. Takagi and H. Yadohisa. 2011. Correspondence analysis for symbolic contingency tables based on interval algebra. Procedia Computer Science. 6: 352-357.

[8] J.E. Beh. 2012. Simple correspondence analysis using adjusted residuals. Journal of Statistical Planning and Inference. 142: 965-973.

[9] M. Greenacre. 2011. The Contributions of Rare Objects in Correspondence Analysis, Barcelona GSE Working Paper Series. No. 571.

[10] I. Ginanjar, U.S. Pasaribu and S.W. Indratno. 2014. Identification the characteristics of Indonesian credit card frauds by trough correspondence analysis (An application of simplification quantitative analysis). Proceedings of 2014 2nd International Conference on Technology, Informatics, Management, Engineering and Environment, TIME-E 2014. 246-251.