The comparative analysis of dependence for three-way contingency table using Burt matrix and Tucker3 in correspondence analysis

K E Lestari, U S Pasaribu, S W Indratno, H Garminia
Institut Teknologi Bandung, Indonesia
Email: karunia.eka.lestari@itb.ac.id

Abstract. In this paper, we confined our attention to compare two methods to obtain a graphical depiction of the association (dependency) between three categorical variables. We shall first describe how to recode a three-way contingency table by discussing the Burt matrix form of the data. This method is known as multiple correspondence analysis (MCA). Another method is to preserve a three-way contingency table form using Tucker3, it's known as a three-way correspondence analysis (CA3). As a case study, we pay attention to analyze the association between race and gender in occupation field that may have contributes to differences in employment opportunity and the continuing increases in women’s educational attainment. The results show that CA3 is more simple in computation and provide the graphical depiction of three-way association simultaneously, while MCA’s plot can’t. Consider to the cumulative inertia on the two-dimensional plot, the percentage inertia of CA3’s plot is better than MCA’s plot.

1. Introduction

Labor market analysis has been extensively studied and researchers have recently paid their attention to occupational race and gender segregation as well. Mintz and Krymkowski [1] investigate the intersection of race and gender in occupation segregation. Catanzarite [2] found much racial segregation among female workers. Kaufman [3] noticed that almost one-third of black or white workers would have to change occupations to achieve full integration. Moreover, Jacobs and Blair-Loy [4] found high levels of gender segregation among African American. These findings suggest that race and gender can interact in complex ways, but no information about how they associated with the occupation. Following these lead, we pay attention to analyze the three-way association between race and gender in occupation field. It may have contributes to differences in employment opportunity and the continuing increases in women’s educational attainment. This analysis is performed using correspondence analysis.

Correspondence analysis (CA) is a popular graphical method that is used to explore the association structure between categorical variables in contingency table [5]. However, there are many members of correspondence analysis family, see Greenacre and Blasius [6] and Beh and Lombardo [7]. In this paper, we interest to compare two member of the CA family for a three-way contingency table —i.e. multiple correspondence analysis (MCA) and three-way correspondence analysis (CA3). First, we shall a briefly explains how to recode a three-way contingency table by determining the Burt matrix form of the data.
Second, we shall use Tucker3 as a different way that preserve a three-way table form. Finally, we performed the comparison of the plots from the MCA by Burt matrix and CA3’s plot by Tucker3.

MCA differs fundamentally from CA3, especially in terms of the graphical display produced by both methods. CA3 provide the graphical depiction of three-way association simultaneously that MCA can’t, but it’s still powerful to use for depicting the association between more than two categorical variables. To verify the comparison, this paper is organised as follows. Section 2 provide description of the graphical depiction method for three-way contingency table. Data analysis and the result of comparative analysis of dependence for three-way contingency table using Burt matrix and Tucker3 is shown by a case study in Section 3. Summary and future works are put forward as a conclusion in Section 4.

2. Theoretical Methods
To assign the graphical depiction of three-way association, we performed a comparison between MCA and CA3 for three-way contingency table. By the MCA method, we obtain a two-dimensional graphical display of the information in the three-way contingency table. The method involves a recoding the three-way table into a two-way form that represented by the Burt matrix form. By doing so, no information about the three-way associations or interactions among variables [8]. On the other hand, CA3 analysed three-way contingency tables without collapsing them into a two-way form. Furthermore, the analysis of the association between the variables in CA3 involved a different decomposition model to the MCA. Therefore, the singular value decomposition that used in MCA is not appropriate. As a consequence, we consider the Tucker3 decomposition in CA3. Tucker3 involves the computations of principal components, which are derived for each of three categorical variables, and of the core array which is akin to generalized associations between these components [7].

2.1 Burt matrix
Suppose we consider an \( I \times J \times K \) three-way contingency table \( \mathbf{N} \) that involves the cross-classification of three categorical variables. For the recoding of three-way contingency table, the Burt matrix consist of the concatenation of the row, column, and tube marginal frequencies. Let \( \mathbf{D}_I \) denotes the diagonal matrix of row marginal frequencies, \( \mathbf{D}_J \) is the diagonal matrix of column marginal frequencies, and \( \mathbf{D}_K \) is the diagonal matrix of column marginal frequencies. Let \( \mathbf{N}_{IJ}, \mathbf{N}_{IK}, \) and \( \mathbf{N}_{JK} \), respectively, be the matrix of bivariate marginal frequencies. The Burt matrix \( \mathbf{B} \) form of this table has the following block structure [9]:

\[
\mathbf{B} = \begin{bmatrix}
\mathbf{D}_I & \mathbf{N}_{IJ} & \mathbf{N}_{IK} \\
\mathbf{N}_{IJ}^T & \mathbf{D}_J & \mathbf{N}_{JK} \\
\mathbf{N}_{IK}^T & \mathbf{N}_{JK}^T & \mathbf{D}_K
\end{bmatrix}
\]  

(1)

The correspondence analysis of three-way contingency table using its Burt matrix may be undertaken by performing an eigen-decomposition of the \( M \times M \) Burt matrix \( \mathbf{B} \), where \( M = I + J + K \) such that

\[
\text{ED}(\mathbf{B}) = \mathbf{U}\Lambda_\mathbf{B}\mathbf{U}^T
\]  

(2)

Here, \( \Lambda_\mathbf{B} \) is diagonal matrix of the eigenvalues \( \lambda_m^B \), for \( m = 1, 2, \cdots, M \) such that \( \Lambda_\mathbf{B} = \text{diag}(\lambda_m^B) \). The matrix \( \mathbf{U} \) contain the eigenvector of \( \mathbf{B} \).

When performing the MCA to obtain a graphical representation of the association between three categorical variables, the nature of the association is represented by considering the proximity of the points that represent each category [10]. Based on the Burt matrix \( \mathbf{B} \), we define the set of variable coordinates (principal coordinates) as
\[ F = U \Lambda_B \] \hspace{1cm} (3)

The quality of the plot is identified based on the percentage of variance which is contained in each dimension of the plot. The percentage of variance on each principal coordinate is determined by the inertia [11]. The percentages of the principal inertia values are computed by dividing each eigenvalue by the sum of the eigenvalues. The total inertia of the data can be expressed as

\[ \text{Total inertia} = \text{trace}\ (\Lambda_B) \] \hspace{1cm} (4)

2.2 Tucker3

Suppose we consider an \( I \times J \times K \) three-way contingency table \( \mathbf{N} \) that involves the cross-classification of three categorical variables. In this section, we analyze the association between variables using Tucker3 by preserving a three-way table form that consists of \( I \) row, \( J \) column, and \( K \) tube categories. Using Tucker3, [7] model the total, marginal, and partial dependence such that

\[ S_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} u_{ip} v_{jq} w_{kr} + e_{ijk} \] \hspace{1cm} (5)

where \( P, Q, \) and \( R \) (\( p \leq I; q \leq J; r \leq K \)) are the fixed number of the components in the row, column, and tube matrices \( U, V, \) and \( W, \) respectively. The term \( e_{ijk} \) is the error of approximation. In matrix notation, we have

\[ \text{Tucker3} \ (\mathbf{S}) = \mathbf{UG} \left( \mathbf{V}^T \otimes \mathbf{W}^T \right) \] \hspace{1cm} (6)

where \( \mathbf{G} \) is the core array flattened into a two-way form size \( P \times QR. \) The core elements help explain the strength of the association between the variables in three-way contingency table.

To portray the association structure of the variables in a three-way contingency table, we define the row coordinates based on Tucker3 as follows

\[ F = \mathbf{UG}_{(P \times QR)} \left( = f_{l,qr} = \sum_{p=1}^{P} u_{ip} g_{pqr} \right) \] \hspace{1cm} (7)

On the other hand, the \((j,k)\)th pair of column-tube categories will be represented by a single point. Hence, the column-tube coordinates are defined as follows

\[ H = (\mathbf{V} \otimes \mathbf{W}) \mathbf{G}_{(QR \times P)} \left( = h_{jk,p} = \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} v_{jq} w_{kr} \right) \] \hspace{1cm} (8)

The total inertia of the three-way contingency table can be quantified by

\[ \text{Total inersia} = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{p=1}^{P} p \cdot j \cdot p \cdot k \cdot h_{jk,p}^2 = \sum_{i=1}^{I} \sum_{q=1}^{Q} \sum_{r=1}^{R} f_{l,qr}^2 \] \hspace{1cm} (9)

3. Data analysis and Result

Our data comes secondary from Bureau of Labor Statistics website, www.bls.gov. This data containing information about the employed person by occupation, race, and sex. We defined the row variable is given by the occupation consisting ten categories: management (R1), professional (R2), service (R3), sales (R4), office (R5), natural resources (R6), construction (R7), installation (R8), production (R9), and transportation (R10). While the column and tube variables defined by race, with three categories: white (C1), black (C2), and Asian (C3), and sex, with two categories: men (T1) and women (T2). Our aim here is to graphically display the associations (dependencies) that exist among the three categorical variables: occupation, race, and sex.
Table 1. Employed persons by occupation, race and sex.

| Category          | Men (T1) |          |          | Women (T2) |          |          |
|-------------------|----------|----------|----------|------------|----------|----------|
|                   | White (C1) | Black (C2) | Asian (C3) | White (C1) | Black (C2) | Asian (C3) |
| Management (R1)   | 12,025  | 944      | 898      | 8,993 | 1,127 | 727 |
| Professional (R2) | 11,830  | 1,276    | 1,801    | 15,890 | 2,372 | 1,493 |
| Service (R3)      | 8,580   | 1,844    | 644      | 10,869 | 2,658 | 941 |
| Sales (R4)        | 6,630   | 716      | 454      | 5,959 | 1,033 | 428 |
| Office (R5)       | 3,705   | 778      | 274      | 10,042 | 1,821 | 530 |
| Natural resources (R6) | 780 | 52 | 10 | 225 | 9 | 9 |
| Construction (R7) | 6,890   | 533      | 120      | 220 | 19 | 9 |
| Installation (R8) | 4,095   | 402      | 150      | 165 | 29 | 13 |
| Production (R9)   | 4,810   | 720      | 289      | 1,710 | 413 | 232 |
| Transportation (R10) | 5,655  | 1,477    | 349      | 1,103 | 364 | 76 |

3.1 The principal coordinates from Burt matrix

Consider the three-way contingency table that summarizes the employed persons data in Table 1. Its Burt matrix form from the data is given in Table 2.

Table 2. Burt table of Table 1

|       | R1     | R2     | R3     | R4     | R5     | R6     | R7     | R8     | R9     | R10    |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| R1    | 24,714 | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 21,018 |
| R2    | 0      | 34,662 | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 27,720 |
| R3    | 0      | 0      | 25,536 | 0      | 0      | 0      | 0      | 0      | 0      | 19,449 |
| R4    | 0      | 0      | 0      | 15,220 | 0      | 0      | 0      | 0      | 0      | 12,589 |
| R5    | 0      | 0      | 0      | 0      | 17,150 | 0      | 0      | 0      | 0      | 13,747 |
| R6    | 0      | 0      | 0      | 0      | 0      | 1,085 | 0      | 0      | 0      | 1,005 |
| R7    | 0      | 0      | 0      | 0      | 0      | 0      | 7,791 | 0      | 0      | 7,110 |
| R8    | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 4,854 | 0      | 4,260 |
| R9    | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 8,156 | 6,520 |
| R10   | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 6,758 |

In cases where the variables consist of more categories (big data), the Burt matrix can consist of hundreds or even many more cells. In addition, if we have three-way contingency table of size $I \times J \times K$, then we have to arrange a Burt matrix of size $(I + J + K) \times (I + J + K)$. As a consequence, the application of Burt matrix for the big data is unfeasible since it requires a longer computational time. Furthermore, by using eigen-decomposition and Equation (2), the principal coordinates of the data are summarized in Table 3.
more simple in computation. It is because we still treat core array. Determination of the principal coordinates using CA3 through Tucker3 dekomposition is given below.

\[ \begin{array}{ccc}
0.0315 & 0.0400 & -0.0108 \\
0.0385 & 0.0433 & -0.0080 \\
0.0364 & -0.0461 & 0.0051 \\
0.0033 & 0.0072 & -0.0159 \\
0.0073 & -0.0152 & -0.0534 \\
0.0185 & 0.0080 & -0.0192 \\
0.0826 & 0.0073 & -0.0314 \\
0.0622 & 0.0010 & -0.0057 \\
0.0318 & -0.0130 & 0.0203 \\
0.0524 & -0.0599 & 0.0303 \\
0.0162 & 0.0249 & -0.0290 \\
0.0286 & -0.0865 & 0.0165 \\
0.0177 & 0.0324 & 0.0804 \\
-0.0108 & -0.0051 & 0.0091 \\
0.1084 & 0.0054 & -0.0097
\end{array} \]

\[ \begin{array}{cccccc}
-6.61E-34 & 4.43E-34 & 7.27E-36 \\
-7.83E-34 & 5.25E-34 & 8.62E-36 \\
-6.72E-34 & 4.51E-34 & 7.39E-36 \\
-5.19E-34 & 3.48E-34 & 5.71E-36 \\
-5.51E-34 & 3.69E-34 & 6.06E-36 \\
-1.38E-34 & 9.29E-35 & 1.52E-36 \\
-3.71E-34 & 2.49E-36 & 4.08E-36 \\
-2.93E-34 & 1.97E-34 & 3.22E-36 \\
-3.80E-34 & 2.55E-34 & 4.18E-36 \\
-3.99E-34 & 2.68E-34 & 4.40E-36 \\
-3.13E-33 & -6.12E-34 & 1.93E-35 \\
-1.23E-33 & -2.41E-34 & 7.57E-36 \\
-8.75E-34 & -1.72E-34 & 5.40E-36 \\
-2.62E-33 & -1.22E-34 & 2.09E-35 \\
-1.09E-34 & -1.15E-34 & 1.97E-35
\end{array} \]

Table 3. The principal coordinates from eigen-decomposition of Burt matrix on $p$th axis.

### 3.2 The principal coordinates from Tucker3

From the Tucker3 decomposition, we can derive marginal of row, column, and tube categories, also core array. Determination of the principal coordinates using CA3 through Tucker3 dekomposition is more simple in computation. It is because we still treat three-way contingency table of size $I \times J \times K$ as a cube matrix of size $I \times J \times K$. Additionally, we can obtain the principal coordinates of the data as given below.

\[ \begin{array}{cccccc}
0.0519 & -0.0370 & -0.1219 & 0.2825 & 0.0924 & 0.4076 \\
-0.5640 & 0.1799 & -0.0808 & 0.0472 & -0.6199 & 0.4427 \\
-0.6302 & -0.1984 & -0.0710 & -0.2712 & -0.1009 & -0.4336 \\
-0.1959 & -0.0498 & -0.1530 & 0.1545 & 0.1447 & 0.0655 \\
-1.3732 & 0.0678 & -0.0219 & 0.0360 & 0.2255 & -0.3946 \\
0.6915 & -0.7087 & 0.2227 & 0.3928 & 0.7416 & 0.3177 \\
1.4725 & -0.4390 & -0.4225 & 0.5399 & 1.0171 & -0.9944 \\
1.7292 & -0.1594 & -0.2999 & 0.4103 & 0.6747 & 0.0668 \\
0.6175 & -0.2606 & -0.0745 & -0.1142 & -0.0169 & -0.0917 \\
1.5323 & 0.1667 & 0.0477 & -0.5265 & 0.0558 & -0.8733 \\
-1.3634 & -0.8285 & -0.4613 & -0.1422 & -0.2824 & 0.0413 \\
0.9114 & -1.2718 & 0.2414 & 0.3283 & 0.0672 & 0.0048 \\
-2.1915 & 0.2183 & 0.7341 & 0.2666 & -0.1000 & -0.0003 \\
2.1461 & 0.4324 & 0.2753 & 0.1638 & -0.2459 & 0.0320 \\
0.0420 & 0.4019 & -0.6110 & 0.3736 & -0.1557 & -0.0481
\end{array} \]

Table 4. The principal coordinates from Tucker3 on $p$th axis.
3.3 The comparative analysis

In this section, we confine our attention to the simple comparison of the graphical display the associations between the principal coordinates from Burt matrix and Tucker3. Figure 1 gives a visual comparison of the principal coordinates on first two axis by considering Table 3 and Table 4.

These methods provide different graphical results. The correspondence plot obtained from the Burt matrix only allow two-way interaction of each category. On the other hand, the plot obtained from Tucker3 assign three-way interaction of each category simultaneously. For example, from the MCA's plot we can find out that white race (C1) has a strong association with management (R1) and natural resources occupation (R6), also men workers (T1) has a strong association with construction occupation (R7). It shows two-way interactions between row-column categories and row-tube categories. Unfortunately, the three-way interaction between row-column-tube categories cannot be determined through this plot. It's different from the CA3’s plot, we obtain information about the three-way interactions between each category. For the above case, we know that Asian women (C3T2) has a strong association with production (R9) and natural resources occupation (R6).

The result may have contributes to differences in employment opportunity and the continuing increases in women’s educational attainment. For example, CA3’s plot show that Asian men (C3T1) and Asian women (C3T2) have an equal-opportunity in management (R1) and sales occupation (R4). In addition, from the MCA’s plot we suggest the women (T2) to continue increases their education attainment in construction (R7) and installation occupation (R8). Other differences can be viewed based on the percentage of inertia in each dimension. Table 5 presents a comparison the principal inertias of the first two axes of a correspondence plot as well as the total inertia of the contingency table.

|          | Based on Burt matrix | Based on Tucker3 |
|----------|----------------------|------------------|
|          | Axis 1 | Axis 2 | Axis 1 | Axis 2 |
| Inertia values | 0.1562 | 0.1005 | 0.6097 | 0.0525 |

Figure 1. Comparison of the plots from a multiple correspondence analysis by Burt matrix (right) and a three-way correspondence analysis by Tucker3 (left) of the data in Table 1.
| Cumulatif  | 0.1562 | 0.2567 | 0.6097 | 0.6622 |
|------------|--------|--------|--------|--------|
| Percentage of inertia | 15.6222 | 10.0493 | 60.9677 | 5.2461 |
| Total inertia | 1.3701 | 17.0839 | |

The cumulative inertia values on the two-dimensional plot for each methods are 0.2567 and 0.6622, each representing 25.67% and 66.22% of the total association between the categorical variables. Therefore, the Tucker3 provide the graphical depiction of three-way association or interaction with a better of inertia.

4. Conclusion
We have studied the comparison of Burt matrix and Tucker3 in analyzing dependencies between three categorical variables graphically. The result shows that Tucker3 provide the graphical depiction of three-way association (interaction) with a more precise of inertia. In case where the sample size is very large (big data), MCA via Burt matrix treated the three-way contingency table of size $I \times J \times K$ as a matrix of size $(I + J + K) \times (I + J + K)$. Whereas in CA3 via Tucker3 we preserve it as a cube matrix of size $I \times J \times K$. Therefore, CA3 is more simple in computation to use for big data. In the future work, we can be extended this study to the $n$-way contingency table. From the data, the result may have contributes to differences in employment opportunity and the continuing increases in women’s educational attainment.

Acknowledgments
The authors thanks to editors who have provided correction. Work supported by P3MI ITB.

References
[1] Mintz B and Krymkowski D H 2011 The intersection of race/ethnicity and gender in occupational segregation International Journal of Sociology 4 31-58
[2] Catanzarite 2003 The factorial analysis of qualitative data British Journal of Statistical Psychology 3 166-185
[3] Kaufman R 2002 Assessing alternative perspective on race and sex employment segregation American Sociological Review 67 547-720
[4] Jacobs J and Blair-Loy M 1996 Gender, race, local labor markets and occupational devaluation Sociological Focus 29 209-300
[5] Beh E J, Lombardo R, and Simonetti B 2010 A European perception of food using two methods of correspondence analysis Food Quality and Preference 22 226-231
[6] Greenacre M and Blasius J 2006 Multiple correspondence analysis and related methods (London: Chapman & Hall/CRC) 607
[7] Beh E J and Lombardo R 2014 Correspondence analysis theory, practice, and new strategies (West Sussex: John Wiley & Sons, Ltd) 593
[8] Bartlett M S 1935 Contingency table interaction Journal of the Royal Statistical Society 2 148-259
[9] Burt C 1950 The factorial analysis of qualitative data British Journal of Statistical Psychology 3 166-185
[10] Le Roux B and Rouanet H 1988 Interpreting axes in multiple correspondence analysis: method of the contributions of points and deviations in visualization of categorical data (eds J Blasidus and M Greenacre) Academic Press 197-220
[11] Ginajar I, Pasaribu U S, and Indratno S W 2014 Identification the characteristic of Indonesian credit card frauds by trough correspondence analysis Proc. Int. Conf. on Technology, Informatics, Management, Engineering & Environment vol 2 (Bandung: Indonesia/IEEE) 246-251