Multiple Correspondence Analysis Using Burt Matrix: A Study of Bandung Institute of Technology student Characteristics

A W Mahdiyasa\textsuperscript{1} and U S Pasaribu\textsuperscript{1}

\textsuperscript{1}Statistics Research Division, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Indonesia

adilan@math.itb.ac.id

Abstract  Student characteristics can provide important information for universities development. However, it is not easy to analyze or make interpretation related to student’s characteristic, because we have to deal with qualitative observation and large data set. This article recommends a new technique for Multiple Correspondence Analysis (MCA) that is appropriate when the data is big. We propose Burt matrix that makes the calculation process simpler and take less time. The new technique will be applied to questionnaire data on the features of students at the Bandung Institute of Technology (ITB), which is collected by The Students Loop (TSL) Company in 2017.

1. Introduction

It is possible to divide the data into two types, qualitative data and quantitative data. Qualitative data is a categorical observation, not in terms of number, for instance type of disaster in a certain area, type of rock or performance services in certain hotels. Quantitative data is a numerical observation represented by a number, for instance depth of corrosion, the number of palm trees, or the earthquake intensity. Many statistical techniques for analysing quantitative information are available. However, the condition is very different for qualitative data. Correspondence analysis (CA), established by [1], is one of the most favourite technique for analysing qualitative data. In many research areas such as in psychology [2], contemporary style of life and behaviour [3][4], archaeology [5], and food quality [6] correspondence analysis was used.

Although CA has many advantages especially for uncover the relationship between categorical variables, this method cannot perform very well when the data is big. Researchers develop a new method called Multiple Correspondence Analysis (MCA) to solve this problem. MCA’s benefits to analyse large data set and the easiness to interpret the results make MCA very popular. MCA and principal component analysis have many similarities, except the type of data. MCA for qualitative data while principal component analysis for quantitative data [7][8][9][10]. Das et al. (2018) use MCA to investigate the wrong way driving crash patterns [11]. The findings showed that MCA helps in presenting a proximity map of the variable categories in a low dimensional space. MCA also has a significant contribution to the economics or financial fields, such as for measuring the circular economy [12] or assessing financial inclusion [13].
Knowing the student characteristics will give many benefits, not only for the university but also for the companies who want to expand their market to the university students. The university long term plan should be made in line with the student characteristics to achieve the best results. Besides that, it will reduce the misunderstanding between university and students because they have the same mind set. On the other hand, for the companies, knowing the characteristics of students will give information about what products will be sold, what marketing strategy that will be appropriate with them, and how to reduce the production cost.

This article uses MCA to analyse the characteristics of ITB students. ITB is a public university located in Bandung, West Java, Indonesia. Currently, ITB is a multi-campus with the first campus in Ganesha and the second campus in Jatinangor. ITB is the oldest engineering campus in Indonesia with 13 faculties. Almost all of the ITB study program has been accredited with A score from BAN-PT, Indonesia Accreditation Committee. Besides that, 12 study program has been accredited by Accreditation Board of Engineering and Technology (ABET) and 7 study program has been accredited by ASIIN. ITB has three types of a student organization that are student union, student activity unit and ITB student family. Student organizations are expected to provide soft skill and experience related to student background which is very important for their future. It also becomes a means to improve student hobby or interest.

The data for this research are collected by TSL Company, which is a start-up social enterprise in the field of education. They provide consulting services on the talent and personality interest of the students. TSL purpose is to make a connection between students with company so they can choose the right career.

2. Methods
The main advantage of MCA is the ability to calculate large data set with many categorical variables. The ordinary MCA uses indicator matrix \( X \) to record the presence of every respondent [14] with the element of \( X \) is a binary number. \( X \) can be very a big matrix if the data is big and it will be very time consuming. To overcome this problem, we define Burt matrix \( B \) which is cross tabulation from \( X \) with its categorical variables [15].

\[
B_{jk} = X^T X = \begin{bmatrix}
X_1^T X_1 & X_1^T X_2 & \cdots & X_1^T X_p \\
X_2^T X_1 & X_2^T X_2 & \cdots & X_2^T X_p \\
\vdots & \vdots & & \vdots \\
X_p^T X_1 & X_p^T X_2 & \cdots & X_p^T X_p
\end{bmatrix}
\]  

(1)

The matrix \( B \) has a different dimension with matrix \( X \), hence the transformation is needed to make sure that the dimension of \( B \) is suitable for the graphical interpretation. We define \( E \) as the transformation of \( B \), \( p \) as the number of variables and \( n \) the number of observations as follow

\[
E = \frac{1}{p^2 n} D^{-1} B
\]

(2)

where

\[
D_{jk} = \begin{bmatrix}
D_1 & 0 & \cdots & 0 \\
0 & D_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & D_p
\end{bmatrix}
\]

Suppose that \( B_k = \begin{bmatrix}
b_{1,1,k} & b_{1,2,k} & \cdots & b_{1,j,k} \\
b_{2,1,k} & b_{2,2,k} & \cdots & b_{2,j,k} \\
\vdots & \vdots & \ddots & \vdots \\
b_{n,1,k} & b_{n,2,k} & \cdots & b_{n,j,k}
\end{bmatrix} \), the marginal probability for the \( j_k^{th} \) column is

\[
p_{jk} = \frac{\sum_{i=1}^{n} b_{ij_k}}{n}
\]

(3)
Furthermore, the diagonal matrix for \( k \)th variable \( D_k \) can be written as:
\[
D_k = \begin{bmatrix}
p_{1k} & 0 & \cdots & 0 \\
0 & p_{2k} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & p_{jk}
\end{bmatrix}
\]

Eigen value decomposition (EVD) of matrix \( E \) is needed to make our calculation easier
\[
\text{EVD} (E) = BA_EB^T \tag{4}
\]
where
\[
\Lambda_E = \begin{bmatrix}
\lambda_1^E & 0 & \cdots & 0 \\
0 & \lambda_2^E & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_J^E
\end{bmatrix}
\]
\[
B = \begin{bmatrix}
\vec{v}_1^E \\
\vec{v}_2^E \\
\vdots \\
\vec{v}_J^E
\end{bmatrix}^T
\]

Finally, principal coordinate from MCA can be obtained from
\[
G = BA_X \tag{5}
\]
where
\[
\Lambda_X = \begin{bmatrix}
\lambda_1^X & 0 & \cdots & 0 \\
0 & \lambda_2^X & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_J^X
\end{bmatrix}
\]

Inertia of the data can be expressed in term of \( \lambda_i^E \) or \( G \) with equation as follow
\[
f_i = \frac{\lambda_i^E}{\sum_{i=1}^{J} \lambda_i^E} \text{ or } F = \text{trace} (G^T DG) \tag{6}
\]
The flow chart for calculating MCA is shown in Figure 1.

3. Data and Descriptive Statistics
The data are collected through questionnaire questions by TSL Company. The purpose of the questions is to find out the characteristics of ITB students. There are 74 questions that have to be fulfilled by the students via internet, however, there are many missing data or the students do not answer the question. Because of that, we decided to analyse just three questions that have the highest number of respondents. The questions and its answers are summarized in Table 1.

| Question | Choices of Answer |
|----------|-------------------|
| Gender \((Y)\) | 1. Female \((y_1)\)  
2. Male \((y_2)\) |
| Imagination about ITB \((Z)\) | 1. Good buildings \((z_1)\)  
2. Clever students \((z_2)\) |
| Favourite things about ITB \((U)\) | 1. Campus facility \((u_1)\)  
2. Academic facility \((u_2)\)  
3. Environment \((u_3)\)  
4. Reputation of campus \((u_4)\) |
The descriptive statistics of our data is quite different because it is qualitative data. We cannot calculate the mean, median, quartile, variance, or standard deviation of the data. Measures of location and measures of variability for quantitative data do not have any meanings for qualitative data. To overcome this problem, we make a histogram for each question which is displayed in Figure 2. The total number of respondents are 448; 198 female and 250 male students. The conditional probability for $Z$ given that $U = u$ or conditional probability for $U$ given that $Z = z$ are displayed in Table 2.

**Figure 1.** The flow chart for calculating MCA using Burt Matrix

**Figure 2.** The Histogram from ITB students for each questions

**Table 2.** The conditional probability for each random variable

| Gender | Imagination About ITB | Preferred Things About ITB |
|--------|------------------------|---------------------------|
|        | Female | Male | Female | Male | Male | Male |
|        |        |      |        |      |      |      |
| $P(Z = z|U = u) = \begin{cases} \frac{55}{198}, z = z_1 \\ \frac{143}{198}, z = z_2 \end{cases}$ | $P(U = u|Z = z) = \begin{cases} 17/198, u = u_1 \\ 24/198, u = u_2 \\ 20/198, u = u_3 \\ 137/198, u = u_4 \end{cases}$ | $P(Z = z|U = u) = \begin{cases} \frac{98}{250}, z = z_1 \\ \frac{152}{250}, z = z_2 \end{cases}$ | $P(U = u|Z = z) = \begin{cases} 29/250, u = u_1 \\ 56/250, u = u_2 \\ 22/250, u = u_3 \\ 143/250, u = u_4 \end{cases}$ |

---

Start  
Data  
Create indicator matrix $X$ and Burt matrix $C$  
Create transformation Burt matrix $M$  
Calculate graphics coordinate  
Calculate inertia  
Graphics and interpretation  
Finish
4. Results and Analysis
We can make $\mathbf{X}$ based on our questionnaire data and it is a relatively simple step if we compared with the other process in our analysis. Table 3 gives the sample of $\mathbf{X}$ with 0 means the respondent does not give a response to the corresponding variable, while 1 means the respondent gives a response to the corresponding variable. We can interpret the results of $\mathbf{X}$ for the first row, the first respondent is male, think that ITB buildings are good and the favourite thing about ITB is its facility. The other respondents can be interpreted in a similar way.

Using Equation (1) we can make $\mathbf{B}$ and it is shown in Table 4. To get the interpretation from Table 4, we focus on second question $Z$ which has two choices of answer $z_1$ or $z_2$. There are 295 (65.84%) respondents choose $z_2$ and 153 (34.16%) respondents choose $z_1$. If we look more detail the number of female students and male students who choose $z_2$ are similar (143 and 152) however, this condition is different for $z_1$ (55 and 98). For third questions $U$, there are 280 (62.5%) respondents choose $u_4$. The lowest choice for these questions is $u_3$ with the number of respondents are 42 (9.375%).

| Table 3. Indicator Matrix for ITB Respondents |
|---|
| Respondent | $y_1$ | $y_2$ | $z_1$ | $z_2$ | $u_1$ | $u_2$ | $u_3$ | $u_4$ |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 446 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 447 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 448 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

| Table 4. Burt Matrix for ITB Respondent |
|---|
| $y_1$ | $y_2$ | $z_1$ | $z_2$ | $u_1$ | $u_2$ | $u_3$ | $u_4$ |
| $y_1$ | 198 | 0 | 55 | 143 | 17 | 24 | 20 | 137 |
| $y_2$ | 250 | 98 | 152 | 29 | 56 | 22 | 143 |
| $z_1$ | 153 | 0 | 21 | 28 | 15 | 89 |
| $z_2$ | 295 | 25 | 52 | 27 | 191 |
| $u_1$ | 46 | 0 | 0 | 0 | 
| $u_2$ | 80 | 0 | 0 |
| $u_3$ | 42 | 0 |
| $u_4$ | 280 |

Equation (2) can be used to make $\mathbf{E}$ with $p$ is equal to 3 and $n$ is equal to 448 and we also get $\lambda_1^E = 0.4087, \lambda_2^E = 0.34566, \lambda_3^E = 0.3333, \lambda_4^E = 0.30499, \lambda_5^E = 0.02746$. The other dimensional space has
zero eigenvalues because the maximum non zero eigenvalues of $E$ are $J - p$. Inertia can be calculated using Equation (6) with value $f_1 = 28.75\%, f_2 = 24.35\%, f_3 = 23.48\%, f_4 = 21.48\%$ and $f_5 = 1.93\%$.

Figure 3(a), 3(b), and 3(c) are the results of our analysis in the one-dimensional, two-dimensional and three-dimensional space respectively. Those figures become the advantages of MCA because it is easier for us to make an interpretation based on graphic rather than just a group of number. Furthermore, it is answers the question why MCA is very popular in the social science fields like sociology, psychology or economy. The interpretation of MCA quite straight forward because it just depends on the distance of the variable. If we want to know about the characteristics of ITB students based on their gender, then we have to find variables with the smallest distance to $y_1$ or $y_2$. It can be seen that $z_2$ and $u_4$ are variable with the smallest distance to $y_1$, while $z_1$ and $u_2$ are variable with the smallest distance to $y_2$. We can say that female student thinks ITB students are clever and the most favourite things about ITB is its reputation. On the other hand, male student think that ITB buildings are good and they like the academic facility of ITB.

![Figure 3(a)](image1)
![Figure 3(b)](image2)
![Figure 3(c)](image3)

**Figure 3.** Correspondence graphics for (a) 1-D, (b) 2-D, and (c) 3-D.

5. **Conclusion**

We have developed a new statistical technique that combines MCA with Burt matrix. This technique has two advantages that are easy to be interpreted and very suitable for analysing a large data set. We use this technique to analyse the characteristics of ITB student based on their gender. We analyse the data from 448 students with proportion 44% female and 56% male. Data is collected using questionnaire
question by TSL Company in 2017. We get the conclusion that female students think ITB students are clever and they like the reputation of ITB. Male students have different characteristics, they think ITB buildings are good and they like the academic facility of ITB.

References
[1] Benzecry J P 1973 L’Analyse Des Donnees, Tome 1: La Taxinomie (Paris: Dunod).
[2] Mauleman R and Savage M A 2013 *Cul. Sociol.* 7 230-56.
[3] Silva E, Alan W and David W 2009 *Cul. Sociol.* 3 299-316.
[4] Stroud R A, Papandonatos G D, Borba K, Kehoe T and Scott-Sheldon L A J 2019 *Addic. Behav.* 91 21-9.
[5] Beh E J, Lombardo R and Alberti G 2018 *Comput. Stat. Data Anal.* 128 73-86.
[6] Cariou V and Qannari E M 2018 *Food Qual. Prefer.* 68 1-11.
[7] Beh E J 2014 *Correspondence Analysis Theory, Practice, and New Strategies* (United Kingdom: West Sussex).
[8] Tenenhaus M and Young F 1985 *Psychometrika* 50 91-119.
[9] Greenacre M J 1988 *J. Appl. Stat.* 20 251-69.
[10] Tuffery S 2011 *Data Mining and Statistics for Decision Making* (United Kingdom: John Wiley and Sons).
[11] Das S, Avelar R, Dixon K and Sun X *Accident Anal. Prev.* 111 43-55.
[12] Parchomenko A, Nelen D, Gillabel J and Rechberger H J. *Cleaner Prod.* 210 200-16.
[13] Dungey M, Tchatoka D F and Yanotti M B 2018 *Int. Rev. Finance. Anal.* 59 212-22.
[14] Greenacre M 2006 *Multiple Correspondence Analysis in Practice 3rd ed* (Florida: CRC Press)
[15] Burt C 1961 *Brit. J. Statist. Psych.* 14 3-24.