The Reliability of Crash Car Protection Level Based on the Circle Confidence Region on the Correspondence Plot

K E Lestari¹, U S Pasaribu¹, S W Indratno¹ and H Garminia¹

¹Institut Teknologi Bandung
Jalan Ganesha 10, Bandung 40132, Indonesia
karunia.eka.lestari@s.itb.ac.id

Abstract. The important thing in vehicle engineering is how well the car protected a driver and passenger during the crash. In this case, there are many variables that might be associated with the car protection system. The association of these variables can be analyzed by multiple correspondence analysis via Burt matrix. By performing an eigen-decomposition on the transformed Burt matrix, we can determine the correspondence plot that visualize the association between variables in a reduced dimension space. While some researchers’ consideration on quantifying this association and visually depicting it, we interest to notice the reliability of the plot in representing the association. In this paper, we shall only consider the reliability by describing circle confidence regions for each point in a low-dimensional correspondence plot. The application on crash car protection data sets is given.

1. Introduction
Vehicle engineering improvements for safety achieved by modifying the vehicle to help the driver or passenger avoid a crash and provide protection against injury in the event of a crash for those inside and outside the vehicle. Some important things in vehicle engineering are how well the car protected a driver and passenger during the crash and how well the passenger compartment held up to force of a crash. Extensive literature has been found that investigated the impact of risk factors related to the driver, the vehicle, the road environment and crash characteristics on severity accidents. Yamamoto and Shankar [1] have studied severity of injuries based on crash characteristics. Kockelman and Kweon [2] investigated the injury severity based on type of crashes. Dissanayake and Lu [3] focused on the age of drivers affecting severity. Hoffman and De Leeuw [4] investigated the association between car body style, structural integrity and occupant protection.

Later on, Chang and Mannering [5] have identified the relationship between type of vehicles and severity. These studies mostly purpose to find out the relationship between risk factors and the injury severity of crashes and provide an idea to develop the car protection system to reduce the severity of accidents. Since there are many variables that might be associated with the car protection system, we can analyze it through multiple correspondence analysis (MCA). MCA provides useful data visualizations that highlight associations and patterns between several categorical variables [6]. While some researchers’ consideration on quantifying and depicting this association, we interest to notice the reliability of the plot in representing the association.
In some correspondence analysis, the focus has been on the reliability of a point proximity from another point and from the origin. In this paper, we shall only consider the reliability by describing circle confidence regions for each point in a low-dimensional correspondence plot. These regions have been used to identifying whether a category is statistically consistent with what is expected under the hypothesis of independence, which is based on the assumptions that underlie Person’s chi-squared statistics. As well as Lebart et al [7] and Beh [8] considered circle regions for nominal and ordinal categorical variables in simple correspondence analysis, respectively. We shows that the regions can be calculated in MCA.

This paper is organized as follows. Section 2 describes the theoretical method of multiple correspondence analysis and it’s circle confidence region. Data analysis and the result of crash car protection level based on the circle confidence region in Section 3. Summary is put forward as a conclusion in Section 4.

2. Theoretical Method
Suppose $X_1, X_2, \ldots, X_q$ be categorical variables on $n$ object, where variable $k$ has $j_k$ categories. Therefore, the total numbers of categories under consideration is $m = \sum_{k=1}^{q} j_k$. Let $X_k$ is the indicator matrix for $k$th variable. $X_k$ is a binary $n \times j_k$ matrix with exactly one nonzero element in each row $i$ that indicating in which category of variable $k$ observation $i$ falls. The $n \times m$ matrix

$$X = \begin{pmatrix} X_1 & X_2 & \cdots & X_q \end{pmatrix}$$

so that the Burt matrix $B$ form of this table has the following block structure [9]:

$$B = X^T X$$

$$B = \begin{bmatrix} D_{X_1} & N_{X_1X_2}^T & N_{X_1X_3}^T & N_{X_1X_4}^T \\ N_{X_1X_2} & D_{X_2} & N_{X_2X_3}^T & N_{X_2X_4}^T \\ N_{X_1X_3} & N_{X_2X_3} & D_{X_3} & N_{X_3X_4}^T \\ N_{X_1X_4} & N_{X_2X_4} & N_{X_3X_4} & D_{X_4} \end{bmatrix}$$

where $N_{X_iX_j}$ is the joint frequencies of the $i$th and $k$th variables, $i \neq k$. $N_{X_iX_j}^T$ be the transpose matrix of $N_{X_iX_j}$. Define $D_{X_k} = diag(n_{j_k})$, with $n_{j_k}$ is the column marginal frequencies of $k$th variables.

From an eigenvalue-decomposition on the transformed Burt matrix, we can determine the correspondence plot that visualize the association between variables in a reduced dimension space. The transformed Burt matrix form is [10]

$$B^* = \frac{1}{n^2} D B^{-1}$$

Here $D = diag(p_{j_k})$, where $p_{j_k} = \frac{n_{j_k}}{n}$, such that

$$ED(B^*) = U \Lambda_B U^T$$

$\Lambda_B = diag(\lambda_B^B)$ is diagonal matrix where $\lambda_B^B$ are eigenvalues, for $\ell = 1, 2, \cdots, L$. The eigenvector of $B$ that corresponds to $\lambda_B^B$ represented by matrix $U$. According to the Burt matrix $B$, Beh and Lombardo [10] define the set of variable coordinates as
According to Lebart et al. [7], circle confidence regions is a method for observing of a variable coordinate in correspondence plot. It explain the reliability of a variable coordinate from another coordinate and the origin proximately. Particularly, if the origin lies in the confidence circle for any category, then that category has small contribution to the association among variables. Referring to Lebart et al [7], Beh and Lombardo [11] calculated the radii length of confidence circle on CA for jth column coordinates, that is

\[ r = \sqrt{\frac{\chi^2_{L-1}}{np_{jk}}} \]

where \( \chi^2_{L-1} \) is chi-squared value with degree of freedom \( L - 1 \). Mostly, the CA’s plot depict on two-dimensional plot, but it can include three or more, see [12],[13],[14],[15].

Consider indicator supermatrix \( \mathbf{X} \) consist of \( n \) rows, and \( m \) columns, where \( m = \sum_{k=1}^{q} j_k \). Notice that \( j_k \) categories is a column containing \( n_{jk} \) value of 1 and \( n - n_{jk} \) value of 0. The statistical hypothesis of the independency of rows and columns is generally too strict to be realistic. It allows us to define thresholds of significance for the eigenvalue. Kendall and Stuart [16] said that the distribution of eigenvalue extracted from a correspondence analysis under the hypothesis of independence is follow the chi-squared distribution, as the total variance. Under these condition, the expected value of the coordinate of \( j_k \) on an axis is zero. Therefore, its variance is \( \frac{(n-n_{jk})}{n_{jk}(n-1)} \) and the covariance of the coordinates of the point in two distinct axes is zero.

By applying the theorem of central limit to the mean variable, the coordinate distribution on an axis is approximated by the normal distribution. Also, the coordinates of a point on distinct axes are independent. The square of the distance to the origin in a \( \ell \)-dimensional subspace thus follows a chi-squared distribution with \( \ell \) degree of freedom. Therefore, the radii length of confidence circle for \( j_k \)th column coordinates in a two-dimensional correspondence plot can be calculated by

\[ r = \sqrt{\frac{\chi^2_{2}}{np_{jk}}} \cdot \frac{(n-n_{jk})}{n_{jk}(n-1)} \]

If the circle region of \( j_k \) coordinates does not contain the origin, then this point has a position that is different from the origin. It indicate that the \( j_k \) category play a main role for representing the association among variables. Conversely, the coordinates are located close to the origin indicate that categories do not play main role in this association. Confidence circles are able to be calculated with the origin as a centre. Best and Rayner [17] considered confidence circle with the position of the profile coordinate. The disadvantages of drawing a circle with the origin as a centre, it assumes that coordinate close to the origin has no contribution to the association among variables.

3. Result and Discussion
Table 1 presents the data obtained from [4] that recorded 24 small cars that Consumers Union judged by crash protection level. There are four categorical variables observed, i.e. the two occupant protection variables indicates how well the car protected a driver dummy and passenger dummy during crash tests.
Structural integrity indicates how well the passenger compartment held up to the forces of a crash. The remaining categorical variable indicates car body style.

We code the variables using indicator supermatrix to allow for easy expression in matrix notation. The purpose of MCA via Burt matrix is to construct a joint map of cars and variable categories in such a way that a car is relatively close to a category it is in, and relatively far from the categories it is not in. The Burt matrix that constructed from this indicator supermatrix is put forward as a Figure 1.

### Table 1. Indicator matrix constructed from Consumer Union’s judgment on crash protection.

| Car make and model | Body style (B) | Driver protection (D) | Passenger protection (P) | Structural integrity (S) |
|--------------------|----------------|-----------------------|--------------------------|--------------------------|
|                    | 1   | 2   | 3   | 4   | 5   | 1   | 2   | 3   | 4   | 5   | 1   | 2   | 3   | 4   | 5   |
| Acura Integra (C1) | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Daihatsu Charade* (C2) | 1   | 0   | 0   | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Doge Colt (C3)     | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Eagle Summit (C4)  | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Ford Escort* (C5)  | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Ford Festiva* (C6) | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Honda Civic* (C7)  | 1   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Hyundai Excel* (C8) | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Hyundai Excel (C9) | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Isuzu I-Mark (C10) | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Mazda 323* (C11)   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Mazda RX-7* (C12)  | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Mitsubishi Mirage (C13) | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Mitsubishi Starion* (C14) | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   |
| Nissan Pulsar NX* (C15) | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   |
| Nissan Sentra (C16) | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   |
| Nissan Sentra* (C17) | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Plymouth Colt (C18) | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   |
| Pontiac LeMans (C19) | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   | 0   |
| Subaru Justy* (C20) | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 1   | 0   |
| Toyota Celica (C21) | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Toyota Tercel* (C22) | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   |
| Volkswagen Golf* (C23) | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   |
| Yuga GV* (C24)      | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 0   |

B: 1 = 2-door, 2 = 4-door, 3 = wagon; asterisk indicates a hatchback.
D&P: 1 = no injury or minor injury, 2 = possible moderate injury, 3 = certain injury-possibly severe, 4 = high likelihood of severe or fatal injury, 5 = severe or fatal injury virtually certain.
S: 1 = much better than average, 2 = better than average, 3 = average, 4 = worse than average, 5 = much worse than average.

The Burt matrix that constructed from this indicator supermatrix is put forward as a Figure 1. Notice that the diagonal sub-matrices consists of the column marginal frequencies of each category for each variable. For example, there are 15 small car that have 2-door, 6 cars have 4-door, and 3 cars is wagon or hatchback. The off-diagonal sub-matrices consists of joint frequencies for each pair categories or its transpose. For example, there are 4 cars that have 2-door with no injury or minor injury of driver protection (B1D1).
Figure 1. The Burt matrix corresponding to the data of Table 1.

The association among variables depicted in Figure 2. It shows that the most important contributors is high likelihood of fatal injury of driver protection (D4) which is relatively far from the origin. Thus, if D4 is omitted, the different results will be obtained. Overall, the figure shows that the car protection system with the good quality (D1, D2, D3, P1, P2, P3, S1, S2, S3) mainly associated with 2-door body style (B1).

Figure 2. Correspondence plot of the categories of car protection system (left).

Next, the Figure 3 shows that each circle confidence regions does not contain the origin. This indicates that each category of variables is reliable in describing associations between variables. Since the origin, which is associated with zero predictability of the response variable given the explanatory variables (i.e. independence), does not lie within any of the circle, all of the category of car body style, occupant protection, and structural integrity are statistically influential in crash car protection.
Figure 3. Circle confidence regions coordinates from decomposing Burt matrix (right).

Furthermore, we can observe the association between these categorical variables based on the type of car, as illustrated in Figure 4.

Figure 4. Correspondence plot of the category of the crash car protection data using multiple correspondence analysis via Burt matrix.

The horizontal direction in map separates car on the basis of occupant protection. Car on the top are associated with no injuries to driver and passenger, while car on the bottom are associated with injuries to driver and passenger. Meanwhile, the car protection system of Dogde colt (C3), Isuzu I-mark (C10), Nissan sentra (C16), and Plymouth colt (C18) are relatively different from other cars, which can be classified as having a worst protection system. From the plot above we also can find out that the passenger protection that high likelihood of severe or fatal injury (P4) mainly associated with structural integrity that much worse than average (S5).
4. Conclusion
This paper discusses the reliability of points in representing associations between variables by describing their circle confidence regions. A point said to be reliable if its region does not contain the origin. As a case study, we use the data that recorded 24 small cars that Consumers Union judged based on crash protection level. The correspondence plot of the data show that the origin does not lie within any of the circle. It indicate that all of the category of car body style, occupant protection, and structural integrity are statistically influential in crash car protection.

5. References
[1] Yamamoto T and Shankar V N 2002 Transportation Research Board Annual Meeting CD-ROM
[2] Kockelman K M and Kweon J Y 2002 Accident Analysis and Prevention 34 313-321
[3] Dissanayake S and Lu J 2002 Accident Analysis and Prevention 34 609-618
[4] Hoffman D L and De Leeuw J 1992 Marketing Letter 3 259-272
[5] Chang L Y and Mannering F 1998 Transportation Research Record 1635 93-103
[6] Lebart L 2006 J (eds) Chapman and Hall 179-196
[7] Lebart L, Morineau A, and Warwick K M 1984 Multivariate Descriptive Statistical Analysis (New York: Wiley) 231
[8] Beh E J 2001 Journal of Applied Mathematics and Decision Sciences (JMDS) 5 35-45
[9] Burt C 1950 Journal of Statistical Psychology 3 166-185
[10] Beh E J and Lombardo R 2014 Correspondence analysis theory, practice, and new strategies (West Sussex:John Wiley & Sons, Ltd) 593
[11] Beh E J and Lombardo R 2015 Communications in Statistics-Theory and Methods 44 95-114
[12] Andrews D F 1972 Plot of high dimensional data Biometrics 28 125-136
[13] Chernoff H 1973 Journal of American Statistical Associations (JASA) 68 361-368
[14] Tukey P A and Tukey J W 1981 Interpreting Multivariate Data 189-213
[15] Tukey P A and Tukey J W 1981 Interpreting Multivariate Data 245-275
[16] Best D J and Rayner J C W 1997 The Statistician 46 347-354
[17] Kendall M G and Stuart A 1961 The advanced theory of statistics (London: Griffin) 237