Communications for Statistical Applications and Methods
2021, Vol. 28, No. 3, 295–305

https://doi.org/10.29220/CSAM.2021.28.3.295
Print ISSN 2287-7843 / Online ISSN 2383-4757

Correlation plot for a contingency table

Chong Sun Hong[1,a], Tae Gyu Oh[a]

[a]Department of Statistics, Sungkyunkwan University, Korea

Abstract

Most graphical representation methods for two-dimensional contingency tables are based on the frequencies, probabilities, association measures, and goodness-of-fit statistics. In this work, a method is proposed to represent the correlation coefficients for each of the two selected levels of the row and column variables. Using the correlation coefficients, one can obtain the vector-matrix that represents the angle corresponding to each cell. Thus, these vectors are represented as a unit circle with angles. This is called a CC plot, which is a correlation plot for a contingency table. When the CC plot is used with other graphical methods as well as statistical models, more advanced analyses including the relationship among the cells of the row or column variables could be derived.

Keywords: angle, association, category, correlation, geometry

1. Introduction

There exist many graphical representation methods for a categorical data. Some methods express the frequencies and probabilities of each cell such as the bar chart, pie chart, and star chart for one categorical variable. For a 2×2 contingency table, Fienberg (1975) proposed the four-fold circular display. Other graphical representations that can be applied to a two-dimensional categorical data include the block chart, mosaic plot (Hartigan and Kleiner, 1981, 1984; Friendly, 1992, 1994), association plot (Cohen, 1980; Friendly, 1991), grouped bar graph (Tufte, 1985), grouped dot plot and framed rectangle chart (Cleveland and McGill, 1984), trellis display (Becker *et al.*, 1996), and the diamond graph (Li *et al.*, 2003), etc.

There are also other kinds of graphical methods to represent the relationships and fitting of the statistical models of their categorical variables. Fienberg (1968) and Fienberg and Gilbert (1970) proposed a method to geometrically represent the association measure using a tetrahedron for a 2×2 contingency table. Tukey (1977) suggested the two-way plot that represents the goodness-of-fit (GOF) for a two-dimensional contingency table. Darroch *et al.* (1980) developed graphical models that could describe an independent model and a conditional independent model for multidimensional contingency tables. There are two other methods that are based on the odds ratios and their confidence intervals for 2×2 contingency tables: the contour plot (Doi *et al.*, 2001; Yamamoto and Doi, 2001) and the raindrop plot (Barrowman and Myers, 2000, 2003). Moreover, Hong *et al.* (1999) proposed graphical methods to describe the relationship among the GOFs of the hierarchical log-linear models by constructing a right-angled triangle plot and a polyhedron plot.

There are other graphical methods to display the relations of the correlation coefficients. Corsten and Gabriel (1976) extended the biplot of Gabriel (1971) and proposed the h-plot to express the

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

Published 31 May 2021 / journal homepage: http://csam.or.kr

correlation coefficients as angles. Gower and Hand (1996) have extended and generalized the ideas of Gabriel. Trosset (2005) later proposed a correlation diagram using the cosine function of the angles. The correlation diagram represents a correlation coefficient matrix using a set of points on a unit circle. Pittelkow and Wilson (2005) developed the GE biplot using the biplot approaches to represent the relationships between the genes and samples. Park *et al.* (2008) compared the performance of the principal component analysis biplot, factor analysis biplot, multidimensional scaling biplot, and correspondence analysis biplot by analyzing various types of gene expression data. Hong and Lee (2006) suggested the G^2-plot that contains information about each log-linear model and all possible pairs of the hierarchical log-linear models using the ideas of the correlation diagram introduced by Trosset (2005).

For the two dimensional $I \times J$ contingency table, there are many graphical representation methods including the correspondence analysis. The correspondence analysis method explores the relationship of variables by simultaneously displaying the row and column categories of contingent table data based on the multidimensional reduction (scaling) method. Most of these methods are based on the frequencies, probabilities, and relationships of the statistical models. Even though there are some statistics to measure the association of categorical variables, it is not easy to find graphical methods to represent the relation of the measure of association. In this paper, a graphical method is proposed based on the correlation coefficients of each cell for a two-dimensional contingency table.

Section 2 defines the correlation coefficient for each of the two selected levels of the row and column variables of a $I \times J$ contingency table. One can then obtain a $I(I-1) \times J(J-1)$ correlation coefficient matrix. In Section 3, the $I \times J$ vector matrix can be found based on the correlation coefficient matrix by extending the ideas of the correlation diagram of Trosset (2005). Each element in the vector matrix can express an angle corresponding to each cell, so that a correlation plot for the correlation coefficient matrix is represented on a unit circle with angles. This is called the CC plot. The CC plots are explored for various 2×2 and 3×3 contingency tables. Some characteristics obtained from the CC plots are derived. An empirical 4×4 contingency table is discussed in Section 4. We explain that the CC plot can be expanded for a high-dimensional contingency table in Section 5. Section 6 summarizes the conclusions of this study.

2. Correlation coefficients for a contingency table

For a $I \times J$ contingency table, a partial 2×2 contingency table can be considered for the two selected levels of row and column variables, which for example are the i and $i^{'th}$ $(i \neq i^{'})$ levels of the row and the j and $j^{'th}$ $(j \neq j^{'})$ levels of the column. The correlation coefficient $\rho_{iji^{'}j^{'}}$ for the each of the two selected levels of the row and column can then be defined as

$$\rho_{iji^{'}j^{'}} = \frac{p_{i^{'}j^{'}} - p_{i^{'}+}p_{+j^{'}}}{\sqrt{p_{i+}p_{i^{'}+}p_{+j}p_{+j^{'}}}} = \frac{p_{ij}p_{i^{'}j^{'}} - p_{ij^{'}}p_{i^{'}j}}{\sqrt{p_{i+}p_{i^{'}+}p_{+j}p_{+j^{'}}}},$$

where $p_{i+} = p_{ij} + p_{ij^{'}}$ and $p_{+j} = p_{ij} + p_{i^{'}j}$. The correlation coefficient for each of the two selected levels of the row and column from a $I \times J$ contingency table exhibits the following properties:

(1) The value is invariant when both order of two levels of row and column variables are exchanged together, i.e., $\rho_{iji^{'}j^{'}} = \rho_{i^{'}j^{'}ij}$.

(2) The sign of the value is reversed when the order of the levels of either row or column is exchanged, i.e., $\rho_{iji^{'}j^{'}} = -\rho_{i^{'}jij^{'}} = -\rho_{ij^{'}i^{'}j}$.

This can be summarized as $\rho_{iji'j'} = \rho_{i'j'ij} = -\rho_{i'ji'j} = -\rho_{ij'i'j}$. For a $I \times J$ contingency table ($I \geq 3$, $J \geq 3$), one obtains a $I(I-1) \times J(J-1)$ correlation coefficient matrix, which is denoted as $\boldsymbol{P} = (\rho_{iji'j'})$. Nonetheless, it is enough to say that for a 2×2 contingency table, there exists one correlation coefficient since $\rho_{1122} = \rho_{2211} = -\rho_{1221} = -\rho_{2112}$. Hence, the correlation coefficient is represented as a scalar for a 2×2 table.

3. Correlation plot for the correlation coefficients matrix

The correlation diagram of Trosset (2005) is proposed to visualize a $p \times p$ correlation coefficient matrix $\boldsymbol{P} = (\rho_{ij})$ on a unit circle. The p vectors on a unit circle are represented with the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, whose element is the corresponding angle to satisfy the optimization problem for the following objective function.

$$\min 2 \sum_{i<j} \left[\cos(\theta_i - \theta_j) - \rho_{ij}\right]^2. \tag{3.1}$$

Trosset (2005) used the S-Plus function, nlminb, which is a quasi-Newtonian algorithm developed by Gay (1983, 1984), to minimize the optimization problem in (3.1).

In this work, using the $I(I-1) \times J(J-1)$ correlation coefficient matrix $\boldsymbol{P} = (\rho_{iji'j'})$ obtained in Section 2, one can find the IJ vector-matrix $\boldsymbol{\theta} = (\theta_{ij})$ that expresses the $I \times J$ angles corresponding to each correlation coefficient in $\boldsymbol{P}$ to solve the following objective function.

$$\min \sum_{i \neq i'}^{I} \sum_{j \neq j'}^{J} \left[\cos(\theta_{ij} - \theta_{i'j'}) - \rho_{iji'j'}\right]^2. \tag{3.2}$$

Since $\rho_{iji'j'}$ has a value from -1 to 1, the value of θ_{ij} belongs to $(0, 2\pi)$ so that the $I \times J$ elements in the vector-matrix can be represented on a unit circle with angles θ_{ij}. We call this correlation plot for the correlation coefficients matrix obtained from the contingency table as the CC plot.

If the correlation coefficient, $\rho_{iji'j'}$, has a positive and large value close to 1.0, this means that the difference between the two vectors, $\theta_{ij} - \theta_{i'j'}$, is close to 0.0 degree, so that two vectors θ_{ij} and $\theta_{i'j'}$ locate closely. On the other hand, when the correlation coefficient has a negative and large value close to –1.0, the difference between the two vectors is almost 180 degrees, and the two vectors are located opposite each other. Also, if the correlation coefficient, $\rho_{iji'j'}$, is close to 0.0, it means that the value of the difference between the two vectors, $\theta_{ij} - \theta_{i'j'}$, is close to 90 degrees, and the angle between θ_{ij} and $\theta_{i'j'}$ is close to a right angle (90 degrees).

3.1. 2×2 contingency table

Consider a 2×2 contingency table. The objective function with $I = J = 2$ in (3.2) is then equal to

$$\begin{aligned}&\left[\cos(\theta_{11}-\theta_{22}) - \rho_{1122}\right]^2 + \left[\cos(\theta_{12}-\theta_{21}) - \rho_{1221}\right]^2 + \left[\cos(\theta_{21}-\theta_{12}) - \rho_{2112}\right]^2 + \left[\cos(\theta_{22}-\theta_{11}) - \rho_{2211}\right]^2\Big] \\ &= 2 \times \{\left[\cos(\theta_{11} - \theta_{22}) - \rho_{1122}\right]^2 + \left[\cos(\theta_{12} - \theta_{21}) - \rho_{1221}\right]^2\}.\end{aligned}$$

Setting $\theta_{11} = 0$ as an initial value, we then obtain $\theta_{22} = \cos^{-1}(\rho_{1122})$ and $\theta_{12} - \theta_{21} = \cos^{-1}(-\rho_{1122})$, which implies that $\theta_{12} = -\theta_{21} = \cos^{-1}(\rho_{1122})/2$ since $\rho_{1122} = -\rho_{1221} = -\rho_{2112}$. For example. if $\rho_{1122} = 0.2588$, then $\theta_{22} = 75°$ and $\theta_{12} = -\theta_{21} = 105°/2 = 52.5°$. If $\rho_{1122} = 0.9063$, then $\theta_{22} = 25°$ and $\theta_{12} = -\theta_{21} = 155°/2 = 77.5°$. And if $\rho_{1122} = -0.8660$, then $\theta_{22} = 150°$ and $\theta_{12} = -\theta_{21} = 30°/2 = 15°$. Three CC plots for this example are displayed in Figure 1.

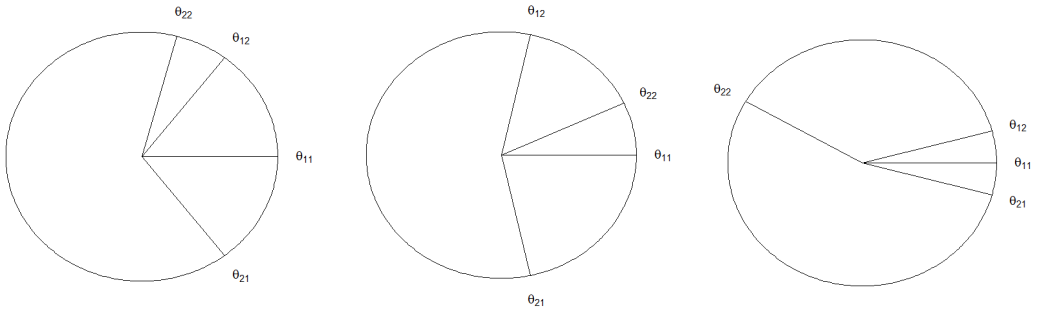

Figure 1: *CC plots for the* 2×2 *contingency tables.*

When a correlation coefficient ρ_{1122} has a positive and large value, one can find from Figure 1(b) that the vector θ_{22} is located close to the vector $\theta_{11} = 0$, and both the vectors $-\theta_{12}$ and θ_{21} are located between 0 and θ_{22}. When a correlation coefficient ρ_{1122} has a negative and large value, it is found from Figure 1(c) that both the vectors $-\theta_{12}$ and θ_{21} are located between 0 and θ_{22}, but the vector θ_{22} is located far away from $\theta_{11} = 0$. Moreover, we could say that the vector θ_{22} in Figure 1(a) is located between those in Figure 1(b) and Figure 1(c), since a correlation coefficient ρ_{1122} has a positive but small value.

3.2. 3×3 contingency table

Consider four 3×3 contingency tables in Table 1. The first and second tables show strong positive and negative relations and the third and fourth tables display two different odd relationships. It is then easy to obtain four correlation coefficient matrices

$$\boldsymbol{P} = \begin{pmatrix} \rho_{1122} & \rho_{1123} & \rho_{1221} & \rho_{1223} & \rho_{1321} & \rho_{1322} \\ \rho_{1132} & \rho_{1133} & \rho_{1231} & \rho_{1233} & \rho_{1331} & \rho_{1332} \\ \rho_{2112} & \rho_{2113} & \rho_{2211} & \rho_{2213} & \rho_{2311} & \rho_{2312} \\ \rho_{2132} & \rho_{2133} & \rho_{2231} & \rho_{2233} & \rho_{2331} & \rho_{2332} \\ \rho_{3112} & \rho_{3113} & \rho_{3211} & \rho_{3213} & \rho_{3311} & \rho_{3312} \\ \rho_{3122} & \rho_{3123} & \rho_{3221} & \rho_{3223} & \rho_{3321} & \rho_{3322} \end{pmatrix}.$$

(a)
$$\begin{pmatrix} 0.74 & 0.61 & -0.74 & -0.02 & -0.61 & 0.02 \\ 0.45 & 0.91 & -0.45 & 0.61 & -0.91 & -0.61 \\ -0.74 & -0.61 & 0.74 & 0.02 & 0.61 & -0.02 \\ -0.22 & 0.45 & 0.22 & 0.74 & -0.45 & -0.74 \\ -0.45 & -0.91 & 0.45 & -0.61 & 0.91 & 0.61 \\ 0.22 & -0.45 & -0.22 & -0.74 & 0.45 & 0.74 \end{pmatrix}$$

(b)
$$\begin{pmatrix} 0.02 & -0.61 & -0.02 & -0.74 & 0.61 & 0.74 \\ -0.61 & -0.91 & 0.61 & -0.45 & 0.91 & 0.45 \\ -0.02 & 0.61 & 0.02 & 0.74 & -0.61 & -0.74 \\ -0.74 & -0.45 & 0.74 & 0.22 & 0.45 & -0.22 \\ 0.61 & 0.91 & -0.61 & 0.45 & -0.91 & -0.45 \\ 0.74 & 0.45 & -0.74 & -0.22 & -0.45 & 0.22 \end{pmatrix}$$

(c)
$$\begin{pmatrix} 0.46 & 0.08 & -0.46 & -0.27 & -0.08 & 0.27 \\ 0.43 & 0.48 & -0.43 & 0.08 & -0.48 & -0.08 \\ -0.46 & -0.08 & 0.46 & 0.27 & 0.08 & -0.27 \\ -0.04 & 0.43 & 0.04 & 0.46 & -0.43 & -0.46 \\ -0.43 & -0.48 & 0.43 & -0.08 & 0.48 & 0.08 \\ 0.04 & -0.43 & -0.04 & -0.46 & 0.43 & 0.46 \end{pmatrix}$$

(d)
$$\begin{pmatrix} 0.46 & 0.43 & -0.46 & -0.04 & -0.43 & 0.04 \\ 0.08 & 0.48 & -0.08 & 0.43 & -0.48 & -0.43 \\ -0.46 & -0.43 & 0.46 & 0.04 & 0.43 & -0.04 \\ -0.27 & 0.08 & 0.27 & 0.46 & -0.08 & -0.46 \\ -0.08 & -0.48 & 0.08 & -0.43 & 0.48 & 0.43 \\ 0.27 & -0.08 & -0.27 & -0.46 & 0.08 & 0.46 \end{pmatrix}$$

The 3×3 vector-matrices $\boldsymbol{\theta}$ could also be obtained to solve the optimization problem in (3.2) using

Table 1: Four 3×3 contingency tables

a	b	c	d
$\begin{pmatrix} 30 & 5 & 1 \\ 4 & 30 & 5 \\ 2 & 4 & 30 \end{pmatrix}$	$\begin{pmatrix} 1 & 5 & 30 \\ 5 & 30 & 4 \\ 30 & 4 & 2 \end{pmatrix}$	$\begin{pmatrix} 30 & 2 & 1 \\ 27 & 30 & 2 \\ 29 & 27 & 30 \end{pmatrix}$	$\begin{pmatrix} 30 & 27 & 29 \\ 2 & 30 & 27 \\ 1 & 2 & 30 \end{pmatrix}$

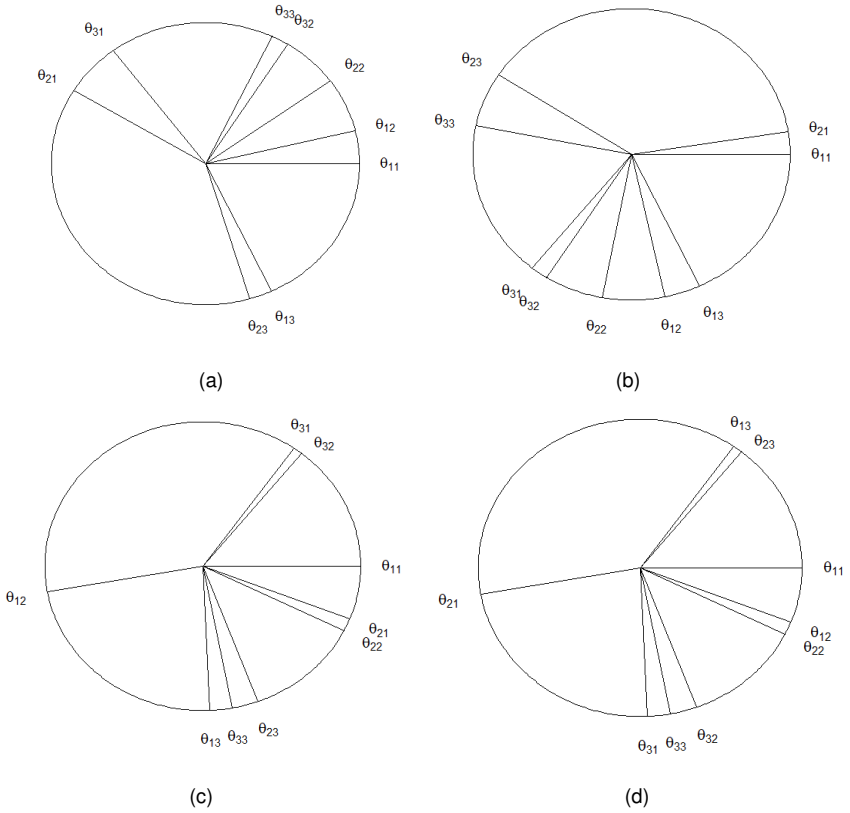

Figure 2: *CC plots for the* 3×3 *contingency tables.*

a quasi-Newtonian algorithm such as the nlminb. Setting $\theta_{11} = 0$ as an initial value.

(a) $\begin{pmatrix} 0° & 13° & 295° \\ 149° & 36° & 287° \\ 127° & 58° & 64° \end{pmatrix}$ (b) $\begin{pmatrix} 0° & 282° & 295° \\ 9° & 259° & 147° \\ 231° & 238° & 169° \end{pmatrix}$ (c) $\begin{pmatrix} 0° & 190° & 272° \\ 339° & 334° & 290° \\ 55° & 51° & 281° \end{pmatrix}$ (d) $\begin{pmatrix} 0° & 339° & 55° \\ 190° & 334° & 51° \\ 272° & 290° & 281° \end{pmatrix}$

Figure 2 shows the four CC plots for the four correlation coefficient matrices. Now we explore the relations between Table 1 and Figure 2. Table 1(a) exhibits a strong and positive relation. The CC plot in Figure 2(a) tells that the diagonal vector set $(\theta_{11}, \theta_{22}, \theta_{33})$ have similar values but each of the three vector pairs $(\theta_{12}, \theta_{21})$, $(\theta_{13}, \theta_{31})$, and $(\theta_{23}, \theta_{32})$ is located opposite each other. In other words, the diagonal cells $(1,1)$, $(2,2)$, and $(3,3)$ have a positive relation but the three pairs of cells $((1,2),(2,1))$, $((1,3),(3,1))$, and $((2,3),(3,2))$ that are facing each other around the diagonal cells have negative relations.

Table 1(b) has a strong but negative relation. The CC plot in Figure 2(b) shows that another diagonal vector set $(\theta_{13}, \theta_{22}, \theta_{31})$ have similar values but each of the three vector pairs $(\theta_{11}, \theta_{33})$, $(\theta_{12}, \theta_{23})$, and $(\theta_{21}, \theta_{32})$ is located on the opposite side of each other. In other words, the diagonal cells $(1, 3)$, $(2, 2)$, and $(3, 1)$ have a positive relation but the three pairs of cells $((1, 1), (3, 3))$, $((1, 2), (2, 3))$, and $((2, 1), (3, 2))$ that are facing each other around the other diagonal cells have negative relations.

The diagonal cells of Table 1(c) and Table 1(d) have large and analog values. However, the cells under the diagonal of Table 1(c) and those over the diagonal of Table 1(d) have also large values. Hence, the three vectors $(\theta_{11}, \theta_{22}, \theta_{33})$ are observed to be located closely from the CC plots in Figure 2(c) and (d). One can say that the $(1, 1)$, $(2, 2)$, and $(3, 3)$ diagonal cells have positive but weak relations.

From the CC plot in Figure 2(c), the vector set $(\theta_{21}, \theta_{31}, \theta_{32})$ whose vectors correspond to under the diagonal has similar values and are located close to the vectors θ_{11} and θ_{22}. However, among θ_{12}, θ_{13} and θ_{23} vectors that belong to above the diagonal cells, both vectors θ_{13} and θ_{23} have similar values and are located close to vector θ_{33} but vector θ_{12} is located far away from θ_{11}. It is found that the $(2, 1)$, $(3, 1)$, and $(3, 2)$ cells that are under the diagonal have strong and positive relations with the $(1, 1)$ and $(2, 2)$ cells. On the other hand, among the $(1, 2)$, $(1, 3)$, and $(2, 3)$ cells that are over the diagonal, the $(1, 3)$ and $(2, 3)$ cells have similar relations, while the $(1, 2)$ cell has a strong but negative relation with $(1, 1)$ cell.

It is evident that the diagonal cells and above the diagonal cells in Table 1(d) have large frequencies, opposite that of Table 1(c). From the CC plot in Figure 2(d), the vector set $(\theta_{12}, \theta_{13}, \theta_{23})$ whose vectors are over the diagonal has similar values and are located close to the vectors θ_{11} and θ_{22}. However, among θ_{21}, θ_{31}, and θ_{32} vectors that belong to under the diagonal cells, both the vectors θ_{31} and θ_{32} have similar values and are located close to vector θ_{33}, but the vector θ_{21} is located far away from θ_{11}. Hence, it is found that the $(1, 2)$, $(1, 3)$, and $(2, 3)$ cells that are over the diagonal have strong and positive relations with the $(1, 1)$ and $(2, 2)$ cells. On the other hand, among the $(2, 1)$, $(3, 1)$, and $(3, 2)$ cells that are under the diagonal, the $(3, 1)$ and $(3, 2)$ cells have similar relations, while the $(2, 1)$ cell has a strong but negative relation with $(1, 1)$ cell.

Therefore, we might derive some characteristics from the CC plots in Figure 1. The (i, j) cells have positive relations when the corresponding vectors θ_{ij} have similar values. On the other hand, the (i, j) cells have negative relations with others when the corresponding vectors θ_{ij} are close to 180 degrees from the other vectors.

- Positive relations are found in
  1. The diagonal cells $(1, 1)$, $(2, 2)$, and $(3, 3)$ in Figure 2(a),
  2. The other diagonal cells $(1, 3)$, $(2, 2)$, and $(3, 1)$ in Figure 2(b),
  3. The diagonal cells $(1, 1)$, $(2, 2)$, and $(3, 3)$ in Figure 2(c) and (d) but weak relations,
  4. The cells under the diagonal with the $(1, 1)$ and $(2, 2)$ cells in Figure 2(c),
  5. The cells over the diagonal with the $(1, 1)$ and $(2, 2)$ cells in Figure 2(d).
- Negative relations are found in
  1. Three pairs of cells $((1, 2), (2, 1))$, $((1, 3), (3, 1))$, and $((2, 3), (3, 2))$ in Figure 2(a),
  2. Three pairs of cells $((1, 1), (3, 3))$, $((1, 2), (2, 3))$, and $((2, 1), (3, 2))$ in Figure 2(b),
  3. The $(1, 2)$ cell with $(1, 1)$ cell in Figure 2(c) with strong relation,
  4. The $(2, 1)$ cell with $(1, 1)$ cell in Figure 2(d) with strong relation.

Table 2: Job satisfaction data

Income($)	Job satisfaction			
	Very dissatisfied	Little dissatisfied	Moderately satisfied	Very satisfied
$< 6,000$	20 (14.2)	24 (24.7)	80 (72.9)	82 (94.2)
6,000–15,000	22 (19.9)	38 (34.6)	104 (102.3)	125 (132.5)
15,000–25,000	13 (16.2)	28 (28.2)	81 (83.2)	113 (107.5)
$> 25,000$	7 (11.8)	18 (20.5)	54 (60.5)	92 (78.2)

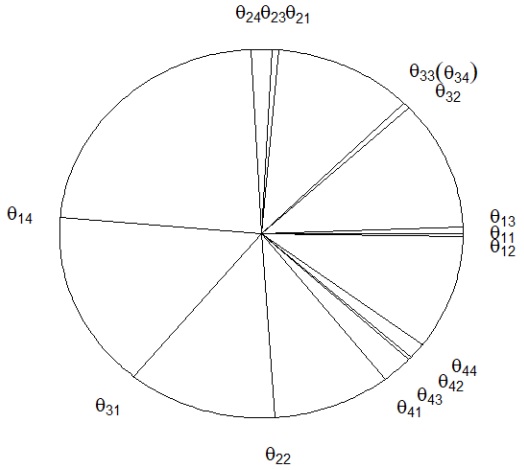

Figure 3: *CC plot for Table 2.*

4. Correlation plot for an illustrated example

Consider a 4×4 contingency table in Table 2 with the income and job satisfaction variables (Norušis, 1988). The independent model is accepted for this data so that the income variable is independent of the job satisfaction variable (Hong, 1995). However, the linear–by-linear uniform association model is fitted better than the independent model (Hong, 1995). Therefore, it is found that the two variables have a linear relation.

Based on the correlation coefficient matrix (the description of the matrix is omitted since it is a 12×12 matrix), the vector-matrix, $\boldsymbol{\theta}$, from Table 2 is obtained and the CC plot is represented in Figure 3 using the vector-matrix.

$$\boldsymbol{\theta} = \begin{pmatrix} 0^\circ & 359^\circ & 2^\circ & 175^\circ \\ 85^\circ & 274^\circ & 87^\circ & 93^\circ \\ 231^\circ & 43^\circ & 45^\circ & 45^\circ \\ 308^\circ & 318^\circ & 317^\circ & 323^\circ \end{pmatrix}$$

Let us take a look at the vectors corresponding to the row variable in Figure 3. The four vector sets $(\theta_{11}, \theta_{12}, \theta_{13})$, $(\theta_{21}, \theta_{23}, \theta_{24})$, $(\theta_{32}, \theta_{33}, \theta_{34})$, and $(\theta_{41}, \theta_{42}, \theta_{43}, \theta_{44})$ have similar values. The income

Table 3: Deceased person data

Year	Speed limit	Road condition	
		Main road	Secondary road
2010	Restrict	8	42
	Free	57	106
2011	Restrict	11	37
	Free	45	69

variable $I = 1$ (lowest level) has relations with the job satisfaction levels $J = 1, 2, 3$ (very dissatisfied, little dissatisfied, moderately satisfied), while the income variable $I = 4$ (highest level) has relations with all levels of the job satisfaction. On the other hand, the income variable $I = 2$ (low level) has relations with the job satisfaction levels $J = 3, 4$ (moderately satisfied, very satisfied), and the income variable $I = 3$ (high level) has relations with the job satisfaction level $J = 2, 3, 4$ (little dissatisfied, moderately satisfied, very satisfied). Hence, those whose incomes belong to either very low or very high levels are neither dissatisfied nor satisfied with their job. Nonetheless, for middle income levels (low or high levels), the related job satisfaction levels are moderately satisfied and very satisfied levels excluding the very dissatisfaction level.

The vectors corresponding to the column variable are then considered. The two vector sets $(\theta_{12}, \theta_{22}, \theta_{32}, \theta_{42})$ and $(\theta_{13}, \theta_{23}, \theta_{33}, \theta_{43})$ have similar values. This means that the job satisfaction $J = 1$ (very dissatisfied) and $J = 4$ (very satisfied) have no relation with the levels of the income variable. However, the job satisfaction $J = 2$ (little dissatisfied) and $J = 3$ (moderately satisfied) have relations with all levels of the income variable. Hence, when the job satisfaction is either very dissatisfied or very satisfied, the income levels do not have relations with the job satisfaction. However, those with middle job satisfaction levels (little dissatisfied and moderately satisfied) have relations with all levels of the job satisfaction variable. Therefore, we can conclude that the job satisfaction variable has a linear relation with the income variable, which exhibits a similar analysis result of the linear-by-linear uniform association model.

There exists the correspondence analysis method that represents a contingency table data. This explores not only the relationship between the row and column variables with an emphasis on correspondence but also the relationship between each variable's categories. The CC plot is proposed to be an alternative method that can also describe a contingency table data graphically, and this plot can explain the relationship between each variable's categories. Moreover, the correspondence analysis is based on chi-squared distance with an emphasis on correspondence, whereas the CC plot is based on the correlation coefficients between row and column variable's category levels. And the CC plot represents the correlation coefficients as the angles between two vectors in a unit circle geometrically, whereas the correspondence analysis method is shown in a rectangle.

5. Correlation plot for high-dimensional contingency tables

We consider a three-dimensional $I \times J \times K$ contingency table. For a given *kth* category of the third layer variable $(K = k)$, the correlation coefficient for the two selected rows and columns are denoted as $\rho^k_{iji'j'}$. Then, the K correlation coefficient matrices, $(\boldsymbol{P^1}, \ldots, \boldsymbol{P^K})$, where $\boldsymbol{P^k} = (\rho^k_{iji'j'})$, $k = 1, \ldots, K$ can be obtained. For each *kth* correlation coefficient matrix, the $I \times J$ vector-matrix, $\boldsymbol{\theta^k} = (\theta^k_{ij})$, are calculated. With each vector-matrix, the CC plot can be represented. Hence, we could discuss K CC plots for a three-dimensional contingency table and derive some relationships from the CC plots.

For an example of a $2 \times 2 \times 2$ contingency table in Table 3, two correlation coefficients, $\boldsymbol{P^1} = (\rho^1_{1122} = -0.1746)$ and $\boldsymbol{P^2} = (\rho^2_{1122} = -0.1590)$, and two vector matrices, $\boldsymbol{\theta^1} = (\theta^1_{ij} = 0°, 40°, -40°, 100°)$

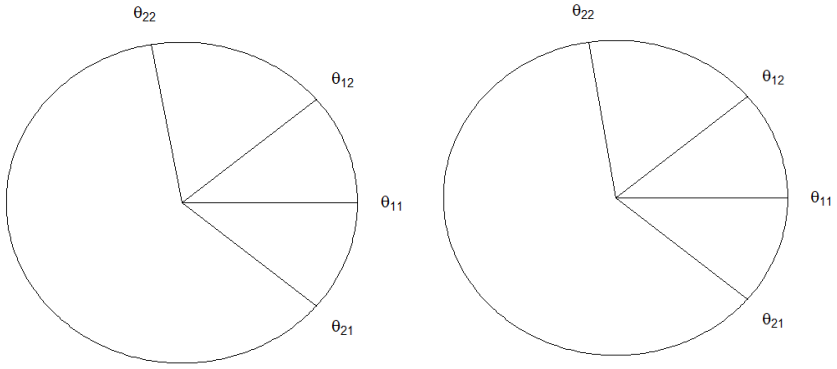

Figure 4: *The CC plots for the* $2 \times 2 \times 2$ *contingency table.*

and $\boldsymbol{\theta^2} = (\theta^2_{ij} = 0°, 40°, -40°, 99°)$, are calculated.

The two CC plots in Figure 4 have analog shapes since the correlation coefficient $\rho^1_{1122} = -0.1746$ is almost the same as $\rho^2_{1122} = -0.1590$. Also, these CC plots are similar to that in Figure 1(a) except for the larger angles of vector θ_{22} in Figure 4(a) and (b) compared to those in Figure 1(a). This is because these correlation coefficients in Figure 4 have signs that are opposite that of $\rho_{1122} = 0.2588$ in Figure 1(a) and the absolute values are not very different.

6. Conclusions

There are lots of graphical representation methods for the two-dimensional $I \times J$ contingency tables. Most methods are based on the frequencies, probabilities, association measures, and goodness-of-fit statistics. In this work, a graphical method is proposed using the correlation coefficient matrix, $\boldsymbol{P} = (\rho_{iji'j'})$, whose element is the correlation coefficient for the selected levels of the row and column variables from the $I \times J$ contingency table such as the i and $i^{'th}$ $(i \neq i^{'})$ levels of the row and the j and $j^{'th}$ $(j \neq j^{'})$ levels of the column.

Each value in the $I \times J$ vector-matrix, $\boldsymbol{\theta} = (\theta_{ij})$, is represented as the angle corresponding to each (i, j) cell. Therefore, the θ_{ij} vectors could be represented as a unit circle with angles. This plot is named as the CC plot, which is a correlation plot for the contingency table.

Some 2×2 and 3×3 contingency tables are implemented as the CC plots. From the CC plots, the relationships among the cells in a contingency table could be explained. It is found that the resulting relations are almost the same as those of the log-linear model analysis with an illustrated example.

The CC plot can also be extended to more than the two-dimensional contingency tables. The CC plots are explained for a three-dimensional contingency table and explored for a contingency $2 \times 2 \times 2$ table.

There exists the correspondence analysis method which represents a contingency table data. This method explore not only the relationship between the row and column variables with an emphasis on correspondence but also the relationship between each variable's categories on a rectangle. The CC plot is proposed to be an alternative graphical method for a contingency table. Moreover, the correspondence analysis is based on chi-squared distance with an emphasis on correspondence, whereas the CC plot is based on the correlation coefficients between row and column variable's category lev-

els. And the CC plot represents the correlation coefficients as the angles between two vectors in a unit circle geometrically, whereas the correspondence analysis method is shown in a rectangle.

Since the CC plot has some advantages that it is easy to use the algorithm for obtaining the angles between two vectors, and simple to interpret the CC plot represented in a unit circle, the CC plot could be used with other graphical methods as an alternative method for a contingency table. Therefore, the CC plot proposed in this work can be a good and worthwhile graphical representation method for categorical data.

References

Barrowman NJ and Myers RA (2000). Still more spawner-recruitment curves: the hockey stick and its generalizations, *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 665–676.

Barrowman NJ and Myers RA (2003). Raindrop plots: a new way to display collections of likelihoods and distributions, *The American Statistician*, **57**, 268–274.

Becker RA, Cleveland WS, and Shyu MJ (1996). The visual design and control of trellis display, *Journal of computational and Graphical Statistics*, **5**, 123–155.

Cleveland WS and McGill R (1984). Graphical perception: theory, experimentation, and application to the development of graphical methods, *Journal of the American statistical association*, **79**, 531–554.

Cohen A (1980). On the graphical display of the significant components in two-way contingency tables, *Communications in Statistics-Theory and Methods*, **9**, 1025–1041.

Corsten LCA and Gabriel KR (1976). Graphical exploration in comparing variance matrices, *Biometrics*, **9**,851–863.

Darroch JN, Lauritzen SL, and Speed TP (1980). Markov fields and log-linear interaction models for contingency tables, *The Annals of Statistics*, **8**,522–539.

Doi M, Nakamura T, and Yamamoto E (2001). Conservative tendency of the crude odds ratio, *Journal of the Japan Statistical Society*, **31**, 53–65.

Fienberg SE (1968). *The estimation of cell probabilities in two-way contingency tables* (Doctoral dissertation), Harvard University, USA

Fienberg SE and Gilbert JP (1970). The geometry of a two by two contingency table, *Journal of the American Statistical Association*, **65**, 694–701.

Fienberg SE (1975). Perspective Canada as a social report, *Social Indicators Research*, **2**, 153–174.

Friendly M (1991). *SAS System for Statistical Graphics* (1st ed), SAS Publishing, USA.

Friendly M (1992). Mosaic displays for loglinear models. In *Proceedings of the Statistical Graphics Section*, American Statistical Association, 61–68.

Friendly M (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association*, **89**, 190–200.

Gabriel KR (1971). The biplot graphical display of matrices with applications to principal component analysis, *Biometrika*, **58**, 453–467.

Gay DM (1983). Algorithm 611: subroutines for unconstrained minimization using a model / trust-region approach, *ACM Transactions on Mathematical Software*, **9**, 503–524.

Gay DM (1984). A trust region approach to linearly constrained optimization. In *Proceedings of the Numerical Analysis*(Dundee 1983, pp. 171–189, ed. F.A. Lootsma), Springer, Berlin.

Gower J and Hand D (1996). *Biplots*, Chapman and Hall, London.

Hartigan JA and Kleiner B (1981). Mosaic for contingency tables, Computer Science and Statistics. In *Proceedings of the 13th Symposium on the Interface* (Eddy DWF ed, pp. 268–273), Springer-

Verlag, New York.
Hartigan JA and Kleiner B (1984). A mosaic of the television ratings, *American Statisticians*, **38**, 32–35.
Hong CS (1995). *Loglinear Model*, Freedom Academy, Seoul.
Hong CS, Choi HJ, and Oh MG (1999). Geometric descriptions for hierarchical log-linear models, *InterStat on the Internet*.
Hong, CS and Lee UK (2006). Graphical methods for hierarchical log-linear models, *Communications for Statistical Applications and Methods*, **13**, 755–764.
Li X, Buechner JM, Tarwater PM, and Munoz A (2003). A diamond-shaped equiponderant graphical display of the effects of two categorical predictors on continuous outcomes, *American Statisticians*, **57**, 193–199.
Norusis MJ (1988). *SPSS-X Advanced Statistics Guide* (2nd ed), SPSS, Chicago.
Park M, Lee JW, Lee JB, and Song SH (2008). Several biplot methods applied to gene expression data, *Journal of Statistical Planning and Inference*, **138**, 500-–515.
Pittelkow Y and Wilson S (2005). Use of principal component analysis and of the GE-biplot for the graphical exploration of gene expression data, *Biometrics*, **61**, 630-–632.
Trosset MW (2005). Visualizing correlation, *Journal of Computational and Graphical Statistics*, **14**, 1–19.
Tufte ER (1985). The visual display of quantitative information, *The Journal for Healthcare Quality (JHQ)*, **7**, 15.
Tukey JW (1977). *Exploratory Data Analysis*, Addison-Wesley Publishing Company, California.
Yamamoto E and Doi M (2001). Noncollapsibility of common odds ratios without/with confounding. In *Bulletin of The 53rd Session of the International Statistical Institute*, Seoul, Korea, 39–40.

Received January 29, 2021; Revised March 10, 2021; Accepted March 18, 2021