The Method to Analyse the Association between Objects and Variables in the Form of $I \times 2$ Contingency Tables with Continuous Variables Additional Data

Irlandia Ginanjar¹, N Sunengsih¹ and Sudartianto¹

¹ Department of Statistics, Universitas Padjadjaran, Indonesia

Abstract. If the row and column categories of a contingency table sequentially seen, as objects and variables, then they are object data with discrete variables. Often from objects obtained additional data in the form of continuous variables. Based on these discrete and continuous variables, the more accurate method is necessary to analyse the associations between these variables. Treating discrete data as continuous is wrong, so this article aims to analyse the association of data in the form of $I \times 2$ contingency tables with additional data that is continuous. From the data in the form of $I \times 2$ contingency table, it converted using the simplification of correspondence analysis (SoCA), so that continuous principal coordinates obtained. Furthermore, the association between continuous variables was analysed using the cosine value of the angle between the two vectors. Case studies use poverty data in Indonesia, which published by the Central Statistics Agency (BPS-Statistics Indonesia). Data in the form of contingency tables are people population lived in poverty based on province and area of residence (urban or rural). Additional variables are poverty depth index, severity index, Gini ratio, food poverty line and non-food poverty line. The results of the analysis obtained information that in urban areas tend to have high Gini ratio, food poverty lines and non-food poverty lines, for rural areas tend to have a high poverty depth and severity index.

1. Introduction

In the current era of big data, many data with different types. If we intend to analyse objects that are sub-populations, then the data with qualitative types often converted in the form of cross-tabulations, so that they become contingency tables with row categories as those objects. The row and column categories of a contingency table sequentially, as objects and variables. Therefore, those are object data with discrete variables. Often from objects obtained additional data in the form of continuous variables. If the data is continuous type, the value for each object is the central tendency, which often uses the mean. Furthermore, researchers often want to identify associations objects and variables with different data types, namely discrete and continuous.

Verrusio et al. [1] identified differences between groups using one-way analysis of variance (ANOVA) for continuous variables, and Chi-square test ($\chi^2$) for discrete variables. Zilko et al. [2] investigates the effect of discrete variables on the Non-Parametric Bayesian Network (NPBN) method and presents the analysis process of a mixed model that is discrete and continuous. Heck [3] proposes a new approach that expands the multinomial processing tree (MPT) model, so that discrete and continuous variables can be used, which were previously limited to discrete variables. Hosaka [4]
introduced estimation method with using the functional connectivities of networks, with observations consisting of discrete and continuous variables. Mori and Mahalec [5] partitioned a large number of values into clusters and produced more robust parameter estimation, and produce a decision tree that could efficiently handle large domains of discrete and continuous variables. Based on that some of the articles shows the researchers are needed the simultaneous data analyses from a mixture of discrete and continuous variables.

The identification of associations using correlations will be misleading because treating discrete data as continuous data is wrong. Discrete data is in the form of natural numbers, while continuous is in the form of real numbers. A few studies have been carried out to analyse the association between variables from different data types simultaneously. Sokolova et al. [6] propose an efficient Bayesian information criterion (BIC) score calculation method for Bayesian hybrid networks so that data analysis with a combination of discrete and continuous variables can be handled, but the relationship between variables is assumed to be monotinous. Pleis [7] deriving the joint probability distribution of continuous and discrete random variables as a product of the conditional and marginal probability distribution with using the general location model (GLOM) method, by approximating the mixture with a non-mixture density from the same parametric family. Anderson-Bergman et al. [8] propose an extended PCA (XPCA) method that uses Gaussian copula and nonparametric marginal with calculations for discrete variables using the induced likelihood function, but these discrete variables are not contingency tables.

The study for analyses the simultaneous association between objects and mixed variables are discrete or continuous variables, with discrete variables constituting a contingency table have been introduced by Ginanjar [9], using hybrid correspondence analysis and correlation (HCAC). HCAC uses correspondence analysis (CA) to analyse discrete variables in the data. The calculation of correlation requires to mean, if the mean used is the normal distribution estimator that can be misled. Furthermore, the disadvantages of CA include the eigenvalues obtained by numerical processes and complex calculation procedures. Ginanjar et al. [10] also introduced a simpler method of CA calculation for \( I / J = 2 \) called the simplification of correspondence analysis (SoCA), with the principal coordinate directly calculated from frequency data in the cross-tabulation matrix, so that it does not use numerical processes and minimizes rounding values.

Based on that, this article will introduce the more accurate method to analyse the associations between discrete and continuous variables, for \( I \) objects as the column, the discrete variables are in the form of \( I \times 2 \) contingency tables, and the objects contains additional data in the type of continuous variables. In this article also uses the contribution statistical test dependencies between objects and two discrete variables, were that introduced by Ginanjar et al. [11]. This article aims to analyse the association of data in the form of \( I \times 2 \) contingency tables with additional data that is continuous. From the data in the form of \( I \times 2 \) contingency table, it converted using the simplification of correspondence analysis (SoCA), so that continuous principal coordinates obtained. Furthermore, the association between continuous variables analysed using the cosine value of the angle between the two vectors. Case studies use poverty data in Indonesia, which published by BPS-Statistics Indonesia [12]. Data in the form of contingency tables are people population lived in poverty based on province and area of residence (urban or rural). Additional variables are poverty depth index, severity index, Gini ratio, food poverty line and non-food poverty line. The calculations are using R.3.5.1. software.

2. Methods

Associations identification between objects and variables will be very necessary if researchers want to know the characteristics of each object based on the variables. Instance, based on the results of the cross-tabulation from two categorical variables, the row and column categories respectively represent the object and characteristics. Accordingly, CA is often used to identify associations between objects and their characteristics. If row and column respectively represent objects and continuous variables, then Biplot factor analysis is used to identify the association between the object and the variable. The method proposed in this paper is a hybrid of the two methods. However, the correlation calculation is using the cosine value. Based on that, this method is called Hybrid SoCA with cosine value association.
2.1. Simplification of Correspondence Analysis (SoCA)

If the size of object/individual is \( n \), with \( X_1 \) and \( X_2 \) are qualitative variables. The categories of variables \( X_1 \) and \( X_2 \) respectively are \( i \) and \( j \), with \( i = 1, 2, \ldots, I \) and \( j = 1, 2, \ldots, J \). if \( I = 2 \) or \( J = 2 \), for instance \( I = 2 \) so that Contingency Tables can be formulated as the following Table 1.

| \( X_1 \) | \( X_2 \) | Total |
|---|---|---|
| 1 | \( n_{11} \) | \( n_{1*} \) |
| 2 | \( n_{21} \) | \( n_{2*} \) |
| Total | \( n_i \) | \( n_j \) | \( n \) |

Ginajar et al. [10] wrote for \( 2 \times J \) contingency table the standard residual matrix can be directly calculated from the element of cross tabulation matrix, so calculating eigenvalue can be simplified to:

\[
\lambda_i = \frac{n}{n_i n_2} \left( \sum_{j=1}^{J} n_{ij}^2 - \frac{n_i^2}{n} \right) \quad \text{and} \quad \lambda_2 = 0
\]  

(1)

The objective of SoCA is to estimate principal coordinate of row and column categories. For \( I = 2 \) and the non-zero eigenvalue is only one, therefore the principal coordinate obtained is consist of one dimension. The principal coordinate of row categories are calculated by:

\[
y = (y_j) = \frac{n}{n_i} \sum_{j=1}^{J} \left( \frac{n_{ij}^2}{n_{ij}} - 1 - \frac{1}{n_2} \right)
\]

(2)

and the principal coordinate of column categories are calculated by:

\[
z = (z_j), \quad z_j = \frac{n_{2j} n_{1*} - n_{1j} n_{2*}}{n_{1*} n_{2*}}
\]

(3)

From Table 1 consider \( N = [n_{ij}] \) is the cross tabulation matrix, the calculation also applies to \( I \times 2 \) cross tabulation matrices, with \( I > 2 \), by looking at \( N^T = [n_{ji}] \) a \( 2 \times I \) cross tabulation matrix.

2.2. Cosine Value Association

Pearson correlation is not suitable to be used to identify associations between two vectors if the data are not normally distributed. Based on that, in writing the association is calculated based on cosine from the angle between the two vectors. Let the continuous variables additional data are \( V_m \) with the number of additional variables is \( q \), so \( m = 1, 2, \ldots, q \) and the principal coordinate of column categories are \( z_j \) as written in equation (3). So the association between the principal coordinate of column categories with the \( m^{th} \) continuous variables is as follows:

\[
\rho_m = \cos(\theta_m) = \frac{\sum_{j=1}^{J} z_j v_{jm}}{\sqrt{\sum_{j=1}^{J} z_j^2} \sqrt{\sum_{j=1}^{J} v_{jm}^2}}
\]

(4)

2.3. Hybrid SoCA with Cosine Value Association

Based on the same theory with the Biplot factor analysis method, then the cosine association values are the continuous variables coordinate vectors.
To produce the vector coordinates that correspond to the principal coordinates of the SoCA results, then \( \alpha_1 \) becomes a multiplier for the continuous variables coordinate vectors, where \( 0 \leq \alpha \leq 1 \). Furthermore, the coordinates of continuous variables were calculated using the following equation:

\[
\Psi = \lambda_1^\alpha \rho = (\rho_m) \tag{5}
\]

Based on the principal coordinate, and the coordinates of continuous variables, one plot can be obtained which can display three types of coordinates simultaneously. The three coordinates are objects obtained from equation (3), discrete variables from equation (2), and continuous variables from equation (6), so the plot is called Triplot.

3. Data Analysis

Case studies use poverty data in Indonesia, which published by the Central Statistics Agency (BPS-Statistics Indonesia). Data in the form of contingency tables are people population lived in poverty based on province and area of residence (urban or rural). Additional variables are poverty depth index, severity index, Gini ratio, with the other two variables being the food poverty line and the non-food poverty line in units of Rupiah/capita/month. Data standardization is done because the units in continuous variables vary. The data of people population lived in poverty with the form of contingency tables and additional variables on standardization results are presented in Table 2.

The principal coordinate and the contribution statistical test of row categories are presented in Table 3. Based on principal coordinates (PCo), negative signs for the urban category and positive in the rural category in Table 3, serve as a benchmark for the characterization of the province's association with the area of residence. The principal coordinate and the contribution statistical test of column categories are presented in Table 4. Provinces with negative principal coordinates are those that are positively associated with urban areas, including DKI Jakarta, West Java, DI Yogyakarta, and others. Thus, poverty alleviation programs in these provinces would be better implemented in urban areas. In contrast, provinces with positive principal coordinates are positively associated with rural areas, including Papua, West Papua, Gorontalo, and others. Thus, poverty alleviation programs in these provinces will be better implemented in rural areas.

The coordinates of continuous variables with \( \alpha = 0 \) are presented in Table 5. The negative and positive signs of the coordinate of continuous variables reflect the characteristics of the provincial categories. Negative values on the coordinates of the Gini ratio, the food poverty line and the non-food poverty line. Thus, it shows that, even though the poverty line for provinces in the left-hand (negative principal coordinate) is high, it has a high level of population expenditure disparity. While the positive value on the coordinates of poverty depth and severity index shows that the provinces that are in the right-hand (positive principal coordinate), must be the main concern for poverty alleviation. The Triplot is presented in Figure 1. The results of the analysis based on Figure 1 obtained information that in urban areas tend to have high Gini ratio, food poverty lines and non-food poverty lines, for rural areas tend to have a high poverty depth and severity index.
Table 2. *The Data of People Population Lived in Poverty and Standardized Additional Variables*

| Province          | Urban   | Rural   | Depth Index | Severity Index | Gini Ratio | Food Poverty Line | Non-food Poverty Line |
|-------------------|---------|---------|-------------|----------------|------------|-------------------|-----------------------|
| Aceh              | 168110  | 651330  | 0.5610      | 0.3448         | -0.8899    | 0.4538            | -0.0631               |
| Sumatera Utara    | 675740  | 606300  | -0.3322     | -0.3414        | -0.9465    | 0.1583            | -0.1198               |
| Sumatera Barat    | 121350  | 226870  | -0.6346     | -0.5374        | -1.2579    | 0.6453            | 0.0682                |
| Riau              | 175930  | 314790  | -0.5010     | -0.4198        | -0.4654    | 0.3821            | 0.4723                |
| Jambi             | 115080  | 159240  | -0.4307     | -0.3610        | -0.8333    | 0.0420            | -0.4184               |
| Sumatera Selatan  | 384530  | 689220  | 0.1953      | 0.0507         | -0.5503    | -0.5479           | -0.4266               |
| Bengkulu          | 96520   | 205780  | 0.4484      | 0.1880         | -0.2955    | 0.4400            | 0.3360                |
| Lampung           | 231860  | 831800  | 0.0616      | -0.0865        | -0.6069    | -0.4300           | -0.4195               |
| Kep. Bangka Belitung | 23310  | 45070   | -0.9370     | -0.7531        | -2.3051    | 2.3724            | 1.9366                |
| Kep. Riau         | 104210  | 24250   | -0.8104     | -0.7139        | -0.2672    | 0.8593            | 2.3187                |
| Dki Jakarta       | 365550  | 0       | 0           | -0.9440        | -0.7335    | 1.2329            | 1.3900                |
| Jawa Barat        | 2268750 | 1130410 | -0.5291     | -0.4786        | 1.4593     | -0.9297           | -0.4249               |
| Jawa Tengah       | 1633960 | 2109260 | -0.2197     | -0.3610        | 0.2989     | -1.0801           | -0.6361               |
| Di Yogyakarta     | 304660  | 143810  | -0.0720     | -0.2041        | 2.0537     | -0.4719           | 0.0612                |
| Jawa Timur        | 1449270 | 2662980 | -0.0298     | -0.0669        | 0.5536     | -0.6691           | -0.5762               |
| Banten            | 378730  | 275730  | -0.7612     | -0.5963        | 0.4121     | -0.1486           | 0.3559                |
| Bali              | 97980   | 65870   | -0.9229     | -0.7335        | 0.4404     | -1.0214           | 0.1700                |
| Nusa Tenggara Barat | 384650 | 351310  | 0.3430      | -0.0001        | 0.8083     | -0.8486           | -0.6154               |
| Nusa Tenggara Timur | 114120 | 1032200 | 1.6229      | 1.2663         | 0.1573     | -0.7636           | -1.0962               |
| Kalimantan Barat  | 81640   | 296770  | -0.4939     | -0.4394        | -0.6635    | -0.0049           | -0.6223               |
| Kalimantan Tengah | 48560   | 86040   | -0.8034     | -0.6747        | -0.4087    | 0.0838            | -0.7977               |
| Kalimantan Selatan| 70520   | 121970  | -0.8034     | -0.6551        | -0.4654    | -0.2427           | 0.3707                |
| Kalimantan Timur  | 107670  | 112250  | -0.6557     | -0.5374        | -0.5786    | 1.3782            | 1.7890                |
| Kalimantan Utara  | 22880   | 25900   | -0.4799     | -0.3414        | -1.5692    | 2.1469            | 1.5879                |
| Sulawesi Utara    | 65499   | 126200  | -0.4658     | -0.4590        | 0.4687     | -0.8589           | -0.9950               |
| Sulawesi Tengah   | 84740   | 325620  | 0.3430      | 0.2860         | -0.6635    | -0.0682           | -0.4310               |
| Sulawesi Selatan  | 170100  | 597690  | -0.2759     | -0.2826        | 1.0914     | -1.4616           | -1.0798               |
| Sulawesi Tenggara | 71820   | 230760  | 0.2726      | 0.2076         | 1.3744     | -1.4920           | -1.0954               |
| Gorontalo         | 21270   | 164760  | 0.5610      | 0.2860         | 1.6008     | -1.2980           | -1.2928               |
| Sulawesi Barat    | 31280   | 120120  | -0.0649     | -0.1257        | 0.4121     | -1.3550           | -1.3305               |
| Maluku            | 45600   | 272090  | 1.0392      | 0.7762         | -0.7484    | 0.6918            | 0.1315                |
| Maluku Utara      | 15320   | 69280   | -0.6698     | -0.5571        | -1.0880    | 0.0779            | -0.5988               |
| Papua Barat       | 22610   | 189880  | 2.6427      | 2.9133         | 1.0065     | 1.5760            | 0.3671                |
| Papua             | 41010   | 885350  | 3.7468      | 4.1485         | 1.2329     | 0.9943            | 0.4832                |

Table 3. *Row Categories Principal Coordinate and Contribution Statistical Test*

|               | PCo  | Confidence Interval | P-Values |
|---------------|------|---------------------|----------|
| Urban         | -0.4826279 | 0.0055553            | 0        |
| Rural         | 0.3184033  | 0.0045122            | 0        |
### Table 4. Column Categories Principal Coordinate and Contribution Statistical Test

| Category                  | PCo       | Confidence Interval | P-Values |
|---------------------------|-----------|---------------------|----------|
| Aceh                      | 0.3930275 | 0.0194015           | 0        |
| Sumatera Utara            | -0.2648049| 0.0155111           | 0        |
| Sumatera Barat            | 0.1001375 | 0.0297623           | 0        |
| Riau                      | 0.0796474 | 0.0250713           | 0        |
| Jambi                     | -0.0449922| 0.0335324           | 0        |
| Sumatera Selatan          | 0.0804551 | 0.0169489           | 0        |
| Bengkulu                  | 0.1598072 | 0.0319429           | 0        |
| Lampung                   | 0.3668088 | 0.0170291           | 0        |
| Kep. Bangka Belitung      | 0.1156621 | 0.0671628           | 0        |
| Kep. Riau                 | -0.845425 | 0.0490015           | 0        |
| DKI Jakarta               | -1.2311682| 0.0290483           | 0        |
| Jawa Barat                | -0.5516223| 0.0095259           | 0        |
| Jawa Tengah               | -0.0797338| 0.0090776           | 0        |
| DI Yogyakarta             | -0.5759136| 0.0262257           | 0        |
| Jawa Timur                | 0.0920846 | 0.0086607           | 0        |
| Banten                    | -0.370263 | 0.021796            | 0        |
| Bali                      | -0.4096908| 0.0433881           | 0        |
| Nusa Tenggara Barat      | -0.2557503| 0.0204723           | 0        |
| Nusa Tenggara Timur       | 0.6088899 | 0.0164037           | 0        |
| Kalimantan Barat          | 0.3713827 | 0.0285504           | 0        |
| Kalimantan Tengah         | 0.0750321 | 0.0478708           | 0        |
| Kalimantan Selatan        | 0.0636216 | 0.0400304           | 0        |
| Kalimantan Timur          | -0.188188 | 0.0374508           | 0        |
| Kalimantan Utara          | -0.1462115| 0.0795193           | 0        |
| Sulawesi Utara            | 0.1141169 | 0.0401138           | 0        |
| Sulawesi Tengah           | 0.3902703 | 0.0274164           | 0        |
| Sulawesi Selatan          | 0.3595307 | 0.0200434           | 0        |
| Sulawesi Tenggara         | 0.3272167 | 0.0319281           | 0        |
| Gorontalo                 | 0.5786011 | 0.0407195           | 0        |
| Sulawesi Barat            | 0.390059  | 0.0451368           | 0        |
| Maluku                    | 0.5189342 | 0.0311596           | 0        |
| Maluku Utara              | 0.4422017 | 0.0603821           | 0        |
| Papua Barat               | 0.5937801 | 0.0381899           | 0        |
| Papua                     | 0.7217751 | 0.0182475           | 0        |

### Table 5. The Coordinates of Continuous Variables

| Coordinates               | Coordinates |
|---------------------------|-------------|
| Depth Index               | 0.5908011   |
| Severity Index            | 0.552144    |
| Gini Ratio                | -0.5570903  |
| Food Poverty Line         | -0.1338218  |
| Non-food Poverty Line     | -0.1434457  |
4. Conclusion
The principal coordinate from Hybrid SoCA with cosine value association directly calculated from frequency data in the cross-tabulation matrix, so that it does not use numerical processes and minimizes rounding values. Furthermore, the calculation can identify the contribution of each category and construct the confidence interval in one dimension. Moreover, the calculation of associations between continuous variables does not depend on the distribution of the variables. The results from case study are obtained information that in urban areas tend to have high Gini ratio, food poverty lines and non-food poverty lines, for rural areas tend to have a high poverty depth and severity index.

The futures studies are developing Hybrid SoCA with cosine value for $I \times 3$ or $3 \times J$ cross tabulation matrices. In addition, Hybrid SoCA with cosine value will also be used for other case studies, so that the information obtained is more reliable for decision makers.

5. Acknowledgement(s)
The authors acknowledge the financial support provided by the Acceleration of Associate Professor Research Universitas Padjadjaran (RPLK) 2020 Number 3370/UN6.D/LT/2020. We also thank for all reviewers.

References
[1] W. Verrusio, A. Renzi, V. M. Magro, M. Musumeci, P. Andreozzi, and M. Cacciafesta, “Quality of life in mental health services with a focus on psychiatric rehabilitation practice,” Ann Ist Super Sanità, vol. 47, no. 4, pp. 363–372, 2011.
[2] A. A. Zilko, D. Kurowicka, A. M. Hanea, and R. M. P. Goverde, “The Copula Bayesian Network with Mixed Discrete and Continuous Nodes to Forecast Railway Disruption Lengths,” in 6th International Conference on Railway Operations Modelling and Analysis, 2015, no. 2014, pp. 020-1-020-18.
[3] D. W. Heck, “Processing Tree Models for Discrete and Continuous Variables,” University of Mannheim, 2017.
[4] R. Hosaka, “Inferring functional connectivities of networks with discrete and continuous observables,” in 2014 International Symposium on Nonlinear Theory and its Applications, 2014, pp. 446–449.
[5] J. Mori and V. Mahalec, “Inference in hybrid Bayesian networks with large discrete and continuous domains,” Expert Syst. Appl., vol. 49, pp. 1–19, 2016.
[6] E. Sokolova, P. Groot, T. Claassen, and T. Heskes, “Causal discovery from databases with discrete and continuous variables,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8754, pp. 442–457, 2014.
[7] J. R. Pleis, “MIXTURES OF DISCRETE AND CONTINUOUS VARIABLES: CONSIDERATIONS FOR DIMENSION REDUCTION by John R. Pleis BS, University of Maryland - University College, 1998 BS, Peru State College, 1990 Submitted to the Graduate Faculty of the Department of Biostatics,” University of Pittsburgh, 2018.

[8] C. Anderson-Bergman, T. G. Kolda, and K. Kincher-Winoto, “XPCA: Extending PCA for a Combination of Discrete and Continuous Variables,” in Machine Learning & Deep Learning Workshop, 2018, pp. 1–30.

[9] I. Ginanjar, “Hybrid correspondence analysis and correlation to analyze the market position from data with two qualitative and p-2 quantitative variables,” in AIP Conference Proceedings, 2015, vol. 1692, p. 020019.

[10] I. Ginanjar, U. S. Pasaribu, and A. Barra, “Simplification of Correspondence Analysis for More Precise Calculation which One Qualitative Variables is Two Categorical Data,” ARPN J. Eng. Appl. Sci., vol. 11, no. 3, pp. 1983–1991, 2016.

[11] I. Ginanjar, I. Nurhuda, N. Sunengsih, and Sudartianto, “Contribution of a categorical statistical test in examining dependencies among qualitative variables by means simplification of correspondence analysis,” J. Phys. Conf. Ser., vol. 1265, no. 1, 2019.

[12] Statistics Indonesia (BPS), “Indonesia Poverty and Inequality,” Statistics Indonesia, 2019.

[13] I. Ginanjar, U. S. Pasaribu, and A. Barra, “Simplification of correspondence analysis for more precise calculation which one qualitative variables is two categorical data,” ARPN J. Eng. Appl. Sci., vol. 11, no. 3, pp. 1983–1991, 2016.

[14] E. J. Beh, “Elliptical confidence regions for simple correspondence analysis,” J. Stat. Plan. Inference, vol. 140, no. 9, pp. 2582–2588, 2010.