Principal component analysis with successive interval in K-Means Cluster Analysis (Study case: Poverty data 2013 in East Nusa Tenggara)

A M Tamonob¹, A Saefuddin², A Wigena³

¹Faculty of Agricultural Technology, ArthaWacana Christian University, Kupang
²Department of Statistics, Bogor Agricultural University, Bogor, Indonesia

E-mail: arista_marlin@yahoo.com

Abstract. K-Means Cluster is a cluster analysis for continuous variables with the concept of distance used is a euclidean distance where that distance is used as observation variables, which are uncorrelated with each other. The case with the type of data that is correlated categorical can be solved by making categorical data into numerical data by the method called the successive interval and then used Principal Component Analysis. Applied this method in poverty data of East Nusa Tenggara Province in K-Means clusterobtained that Principal Component Analysis with Successive interval obtained variables that take effect to the cluster formation are toilet, fuel, and job.

1. Introduction

In the social and behavioral science, researchers are often confronted with a large number of variables. Often, variables used has a different measurement scale; there are numerical and categorical scale. Cluster analysis is a method in multivariate to grouping a set of objects in the same group (cluster) are more similar to each other than to those in other groups (clusters). In general, variables used as basic cluster consist of two types, there are categorical variables (nominal and ordinal) and numerical variables (interval and ratio). In general, there are two methods in cluster analysis, hierarchical and non-hierarchical method. The hierarchical method is used if the number of clusters is previously unknown while non-hierarchical method is used if the number of cluster is already known. One of non-hierarchical cluster analysis commonly used is the analysis of K-Means cluster used for numerical data and used Euclid distance as a measure to see similarity and dissimilarity measured. Euclidean distance is used if the observed variables are independent or uncorrelated with one another. One of the assumptions in cluster analysis is no multicollinearity between variables [1]. Euclid distance concept used for numerical data by transforming the variables using Principal Component Analysis [2]. Nominal and ordinal data can be an interval by using the method proposed by Hays [3] called successive interval. Principal Component Analysis can be used in the results of successive interval to reduce data and solve multicollinearity, so it can be analyzed by cluster analysis. Previous research related to cluster analysis for categorical data by Yuniato [4] was about the comparison between cluster analysis with nonlinear principal component analysis and two-step cluster in mix data and the result that the two-step cluster method is better and more specific than centroid linkage method with
principal component analysis transformation. This study focusses on principal component analysis with a successive interval for categorical data in K-Means cluster analysis using poverty data 2013 in East Nusa Tenggara Province. East Nusa Tenggara ranks fourth poorest province in Indonesia after Papua, West Papua and Maluku [5]. This cluster analysis aims to find out the significant factors of poverty in East Nusa Tenggara so that it can be an input for the government to reduce and even overcome poverty in East Nusa Tenggara.

1.1 Categorical data

Based on the type of data, it can be divided into numerical data (quantitative) and categorical data (qualitative). Numerical data is data expressed in numerical magnitude (numbers), for example, the data in per capita income, expending, price, etc., whereas categorical data is classified by category or a particular class. Categorical data consist of nominal data and ordinal data. Nominal data is data with an order or the value does not indicate the level merely as a label only, such as religion, gender, ethnicity or race. Ordinal data is data with the order of categories shows level or rank, such as education level, smoking habits. Categorical data can also be obtained by grouping continuous data, but having a risk of losing information. In practice, it is easier to record categorical data rather than continuous data; respondents are easier to answer sensitive questions; it is more practical significance. Presentation of categorical data can be either frequency, frequency tables and contingency tables [6].

In cases where the numerical variables are linked, correlation analysis is one option to see the relation between variables. However, if the two categorical variables are linked, then correlation analysis can not be used because the numbers on a category code aren't the only form of actual values. Another reason why the correlation analysis can not be used on categorical data is due to one type of categorical variable that can not be sorted as nominal category. Giving a different order will give different correlation values so that two people who calculate the correlation value likely give the same results. For this reason, the chi-squared analysis will be used to find relationships (associations) between the categorical variables. The analysis is based on the chi-square contingency table (often called cross-tabulations). A contingency table is a table whose cells contain the frequencies of the intersection of rows and columns. The general form of contingency tables with the first variable has \(m\) categories and the second variable has \(k\) categories. The hypotheses were:

\(H_0\): two variables are independent (associated)

\(H_1\): two variables are dependent (not associated)

The statistical test used was:

\[
\chi^2_{\text{hitung}} = \sum_{j=1}^{k} \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

(1)

If \(\chi^2_{\text{hitung}}\) value is greater than \(\chi^2_{\text{table}(df,\alpha)}\) with the value of the degrees of freedom \((m - 1)(k - 1)\) then reject \(H_0\). Strong or weak associations can be seen from the association between the interval -1 to 1. Values association equals to 1 which means there is a strong association among variables, if the association equals to the value of zero, it means there is no association among variables.
Table 1. Correlation test or the chosen association

| Variable 1 | Variable 2 | Correlation test or the chosen association |
|------------|------------|---------------------------------------------|
| Nominal    | Nominal    | Contingency coefficient, Lambda, Phi, and Cramer |
| Nominal    | Ordinal    | Contingency coefficient, Lambda |
| Ordinal    | Ordinal    | Spearman, Gamma, Somers’d |
| Ordinal    | Numeric    | Spearman |
| Numeric    | Numeric    | Pearson |

1.2 K-Means cluster
K-Means cluster analysis has been developed by Mac Queen in 1967 and it is the most well-known non-hierarchical data grouping method that widely used in many fields because it is simple and easy to implement [7]. K-Means is a partitioning cluster method that separates data into different groups. The purpose of grouping the data is to minimize the variance within a group and maximize the variance between groups. Basic K-Means algorithm is as follows:
1. Determine the \( k \) value as the number of clusters to be formed
2. Generate the center point of the initial \( k \) randomly cluster
3. Calculate the distance of each data into each cluster center using Euclid distance
4. Group each data based on the distance between the nearest inter-data center
5. Determine new cluster center position by calculating the average value of the data existing on the same cluster center.

1.3 Principal component analysis
Principal component analysis is a multivariate method that aims to reduce the dimension of original variables to new variables (called principal component) that uncorrelated but have big information about original variables.

Principal component analysis is a weighted linear combination of the origin variables that are able to explain the data maximally.

The principal component \( j \) of \( p \) variables can be expressed as follows:

\[
Y_j = a_{j,1}x_1 + a_{j,2}x_2 + \ldots + a_{j,p}x_p = a_j^T x
\]

and the variance of principal components \( j \) is

\[
\text{Var}(Y_j) = \lambda_j; \ j = 1, 2, \ldots, p
\]

\( \lambda_1, \lambda_2, \ldots, \lambda_p \) is eigen values that obtained from the equation:

\[
|\Sigma - \lambda_j I| = 0
\]

Where \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \). Eigenvector \( a_j \) as the weight of transformation linear of origin variables obtained from the equation:

\[
(\Sigma - \lambda_j I) a_j = 0
\]

Variance total of principal component is

\[
\lambda_1 + \lambda_2 + \ldots + \lambda_p = \text{tr}(\Sigma)
\]

and the percentage of variance total data explained by the principal component \( j \) is:

\[
\left(\frac{\lambda_j}{\text{tr}(\Sigma)}\times 100\%\right)
\]

The percentage of total variance is considered sufficiently representative of variance if data is 75% or more.
1.4 Successive interval
Methods to convert data from nominal or ordinal data to interval data [8].
Steps that have to be carried out are:
a. Calculate the frequency of each data category
b. Calculate the proportion based on the frequency of each category
c. Calculate the cumulative proportion
d. Calculate the value of for each cumulative proportion
e. Calculate the value of the density function
f. Calculate the value scale (the average interval) for each category
g. Calculate caling

2. Material and Method

2.1 Data
The data used is the core data SUSENAS (National Socio-Economic Survey) 2013 of East Nusa Tenggara Province. Variables used consisted of 15 variables categorical: Education (X₁), Floor Area (X₂), Floor Type (X₃), Wall (X₄), Roof (X₅), Toilet (X₆), Water (X₇), Fuel (X₈), Assets (X₉), Literacy (X₁₀), Employment (X₁₁), Work Status (X₁₂), Credit (X₁₃), Poor Rice (X₁₄), Handphone (X₁₅). Groups in this analysis consist of poor and non-poor households.

2.2 Method
In this research, steps to analyze are:
1. Data exploration
2. Checked multicollinearity among variables
   Applied NLPCA to data
3. Carried out the Successive interval to data
4. Applied PCA to successive interval data
5. Applied K-Means Cluster Analysis on principal component scores of PCA with Successive interval

3. Results

3.1. Data description
Description SUSENAS core data of East Nusa Tenggara Province in 2013 to 10422 households. The following three variables were taken for describing the data:

Toilet

- Do not have
- Not his own
- own

![Pie Chart](image)

**Figure 1.** Percentage of toilet ownership
From Figure 1, we can see 64% of household heads in East Nusa Tenggara Province 2013 had their own toilet, 21% did not have toilet, and 15% not has their own. From Figure 2, we see that household fuel is still dominated by charcoal / briquettes / wood by 83.4%, electricity / gas by 0.9% and 15.7% kerosene. Figure 3 shows that 91% of the households had a job while 9% did not have job.

3.2 Checking multicollinearity
From the 15 independent variables of nominal and ordinal scale, the test multicollinearity assumptions were done by looking at the association between variables in the contingency tables. Test of the association among the variables was large enough association between the variables of literacy and education by 0.815, between jobs and education by 0.656, between the working status and education by 0.567, and between jobs and working status by 0.813.
3.3 Principal component analysis with successive interval

Results of the principal component analysis showed that when the two main components were taken to mean, only 35.2% could explain the variance of the initial data. The selection of the main components to be used in the K-Means cluster is based on the percentage of the cumulative variance of data about 75% or more [9]. To be able to represent the variance of the data in this study were taken 9 main components with a total variance of 79.1%. The final results of the principal component analysis area score of major components used for the analysis of K-Means cluster.

| Principal Component | Eigen Value | Variance Proportion | Cumulative Variance Proportion |
|---------------------|-------------|---------------------|--------------------------------|
| 1                   | 3.7140      | 0.248               | 24.8%                          |
| 2                   | 1.5729      | 0.105               | 35.2%                          |
| 3                   | 1.2984      | 0.087               | 43.9%                          |
| 4                   | 1.0467      | 0.070               | 50.9%                          |
| 5                   | 0.9811      | 0.065               | 57.4%                          |
| 6                   | 0.9432      | 0.062               | 63.6%                          |
| 7                   | 0.8481      | 0.057               | 69.3%                          |
| 8                   | 0.7543      | 0.050               | 74.3%                          |
| 9                   | 0.7160      | 0.048               | 79.1%                          |
| 10                  | 0.6941      | 0.046               | 83.7%                          |
| 11                  | 0.6606      | 0.044               | 88.1%                          |
| 12                  | 0.5827      | 0.039               | 92.0%                          |
| 13                  | 0.5545      | 0.037               | 95.7%                          |
| 14                  | 0.3616      | 0.024               | 98.1%                          |
| 15                  | 0.2808      | 0.019               | 100%                           |

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 | X14 | X15 |
|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|
| 0.372 | 0.201 | 0.229 | -0.302 | 0.284 | -0.244 | -0.229 | -0.328 | 0.163 | 0.236 | 0.308 | 0.221 | -0.000 | -0.312 |
| -0.040 | 0.417 | 0.085 | -0.174 | -0.287 | -0.296 | -0.065 | 0.213 | -0.328 | -0.426 | -0.430 | 0.145 | 0.170 | 0.081 |
| -0.378 | 0.375 | 0.206 | -0.208 | -0.069 | -0.070 | 0.076 | 0.058 | -0.228 | -0.319 | -0.277 | 0.358 | -0.409 | 0.081 |
| -0.201 | 0.063 | 0.175 | -0.214 | -0.082 | -0.115 | 0.234 | 0.058 | -0.228 | -0.319 | -0.277 | 0.358 | -0.409 | 0.081 |
| 0.169 | 0.069 | -0.154 | 0.037 | 0.104 | -0.150 | -0.264 | 0.070 | -0.228 | -0.319 | -0.277 | 0.358 | -0.409 | 0.081 |
| -0.121 | -0.088 | -0.638 | 0.132 | -0.079 | -0.302 | -0.264 | 0.070 | -0.228 | -0.319 | -0.277 | 0.358 | -0.409 | 0.081 |
| 0.134 | 0.128 | 0.040 | -0.099 | 0.354 | -0.007 | -0.264 | 0.070 | -0.228 | -0.319 | -0.277 | 0.358 | -0.409 | 0.081 |
| -0.073 | -0.076 | -0.288 | -0.065 | -0.099 | -0.578 | -0.578 | 0.578 | -0.076 | -0.288 | -0.099 | -0.578 | -0.578 | 0.578 |
| 0.012 | -0.312 | 0.237 | -0.019 | 0.519 | -0.336 | -0.336 | 0.578 | -0.336 | 0.237 | -0.336 | 0.578 | -0.336 | 0.578 |

Table 3. Principal component score of successive interval data

Table 2. Result of principal component analysis of successive interval data
From Table 3 above, variables X_1 value of the first principal component is 0.372 meaning that the principal component gives a contribution 0.372 but has a positive influence on the observation unit; the second component is 0.040 and has a negative influence on the observation unit. This also applies to principal components and other variables.

### 3.4 Clustering results

#### Table 4. Cluster variance

| Method                          | Variance |          |
|---------------------------------|----------|----------|
| PCA with successive interval    | 9.548    | 24287.65 |
| Ratio                           | 0.00039  |          |

From Table 4, it can be seen that the ratio between the variance in the cluster of method showed that the method of Principal Component Analysis method with successive interval has a ratio of 0.00039. We can see variance within a cluster smaller than between clusters. Ratio 0.00039 can be used if we want to compare two methods of clustering, the smaller ratio shows that the method better than the other method.

#### Table 5. ANOVA of K-Means cluster for principal component analysis with successive interval

| Variables | Deviation standard within a cluster | Deviation standard between cluster | F    | p-value |
|-----------|-------------------------------------|-----------------------------------|------|---------|
| X_1       | 0.2527                              | 0.0320                            | 0.06 | 0.8083  |
| X_2       | 0.2125                              | 0.0409                            | 0.15 | 0.7134  |
| X_3       | 0.3213                              | 0.0395                            | 0.06 | 0.8140  |
| X_4       | 0.1547                              | 0.0064                            | 0.01 | 0.9364  |
| X_5       | 0.2048                              | 0.1125                            | 1.21 | 0.3142  |
| X_6       | 0.1628                              | 0.2004                            | 6.06 | 0.0489  |
| X_7       | 0.1420                              | 0.0478                            | 0.45 | 0.5260  |
| X_8       | 0.2129                              | 0.3993                            | 14.06| 0.0095  |
| X_9       | 0.3599                              | 0.0745                            | 0.17 | 0.6932  |
| X_10      | 0.2658                              | 0.2270                            | 2.92 | 0.1385  |
| X_11      | 0.2062                              | 0.2312                            | 5.03 | 0.0662  |
| X_12      | 0.4184                              | 0.0629                            | 0.09 | 0.7739  |
| X_13      | 0.2207                              | 0.0489                            | 0.20 | 0.6733  |
| X_14      | 0.2351                              | 0.2200                            | 3.50 | 0.1104  |
| X_15      | 0.1839                              | 0.0398                            | 0.19 | 0.6802  |

ANOVA results show that the variables that differentiate in cluster formation are variable with p-value less than 0.01. In this case, the significant variables are X_6 (toilet), X_8 (fuel), X_11 (job). Therefore, the factors need to be considered by the East Nusa Tenggara government to reduce problems. Toilet is related to public health and cleanliness, so every household must have it. The result of SUSENAS shows that about 36% of households did not have a toilet. From fuel, there are still many households that use charcoal/briquettes/woods about 83.4%, the government must solve these problems, one way is to distribute electricity to the village. From SUSENAS about 91% households had a job, but mostly the jobs not a proper job with a good salary, so there are many households classified to poverty households.
4. Conclusions

In this study, it was concluded that PCA with successive interval produced 9 principal components with variance total about 79%. In K-Means cluster analysis with PCA successive interval had variance ratio within a cluster and between cluster about 47.26725. Variables that have influenced the household clustering were $X_6$ (toilet), $X_8$ (fuel), $X_{11}$ (job). From this research we can see that this method can be used for categorical data, in PCA we can transform it back way and from the successive interval we can get the numerical data, so we can use them in another analysis using numerical data.

References

[1] Hair J F, Black W C, Babin B J and Anderson R E 2010 *Multivariate Data Analysis* 7th Edition (New Jersey: Prentice Hall International Inc.)

[2] Kaufma L and Peter J R 2005 *Finding Groups in Data: An Introduction to Cluster Analysis* (New York: John Wiley & Sons Inc.)

[3] Hays W L 1976 *Quantification in Psychology* (New Delhi: Prentice Hall)

[4] Yunianto Y 2011 Perbandingan Metode Penggerombolan dengan Komponen Utama Nonlinier dan Gerombol Dua Langkah pada Data Campuran *Skripsi* (Bogor: Institut Pertanian Bogor)

[5] Badan Pusat Statistik 2013 *Nusa Tenggara Timur dalam Angka 2013* (Kupang: BPS Nusa Tenggara Timur)

[6] Agresti A 2002 *An Introduction to Categorical Data Analysis* (New York, US: John Wiley and Sons)

[7] Wichern D W and Johnson R A 2002 *Applied Multivariate Statistical Analysis* (New Jersey: Prentice Hall)

[8] Waryanto B and Millafati Y A 2006 Transformasi Data Skala Ordinal ke Interval menggunakan Makro Minitab *Informatika Pertanian* 15 881-95

[9] Morrisson D F 1990 *Multivariate Statistical Methods Third Edition* (New York: McGraw-Hill Inc.)