Covariance and PCA for Categorical Variables

Hirotaka Niitsuma and Takashi Okada

Department of Informatics, Kwansei Gakuin University, 2-1 Gakuen-cho, Sanda 669-1323, Japan,
niitsuma@mub.biglobe.ne.jp,okada-office@ksc.kwansei.ac.jp

Abstract. Covariances from categorical variables are defined using a regular simplex expression for categories. The method follows the variance definition by Gini, and it gives the covariance as a solution of simultaneous equations using the Newton method. The calculated results give reasonable values for test data. A method of principal component analysis (RS-PCA) is also proposed using regular simplex expressions, which allows easy interpretation of the principal components.

1 Introduction

There are large collections of categorical data in many applications, such as information retrieval, web browsing, telecommunications, and market basket analysis. While the dimensionality of such data sets can be large, the variables (or attributes) are seldom completely independent. Rather, it is natural to assume that the attributes are organized into topics, which may overlap, i.e., collections of variables whose occurrences are somehow correlated to each other.

One method to find such relationships is to select appropriate variables and to view the data using a method like Principle Components Analysis (PCA) [4]. This approach gives us a clear picture of the data using KL-plot of the PCA. However, the method is not settled for the data including categorical data. Multinomial PCA [2] is analogues to PCA for handling discrete or categorical data. However, multinomial PCA is a method based on the parametric model and it is difficult to construct a KL-plot for the estimated result. Multiple Correspondence Analysis (MCA) [3] is analogous to PCA and can handle discrete categorical data. MCA is also known as homogeneity analysis, dual scaling, or reciprocal averaging. The basic premise of the technique is that complicated multivariate data can be made more accessible by displaying their main regularities and patterns as plots ("KL-plot") . MCA is not based on a parametric model and can give a "KL-plot" for the estimated result. In order to represent the structure of the data, sometimes we need to ignore meaningless variables. However, MCA does not give covariances or correlation coefficients between a pair of categorical variables. It is difficult to obtain criteria for selecting appropriate categorical variables using MCA.

In this paper, we introduce the covariance between a pair of categorical variables using the regular simplex expression of categorical data. This can give a criterion for selecting appropriate categorical variables. We also propose a new PCA method for categorical data.

Table 1. Fisher's data

		x_{hair}				
		fair	red	medium	dark	black
x_{eye}	blue	326	38	241	110	3
	light	688	116	584	188	4
	medium	343	84	909	412	26
	dark	98	48	403	681	85

2 Gini's Definition of Variance and its Extension

Let us consider the contingency table shown in Table 1, which is known as Fisher's data [5] on the colors of the eyes and hair of the inhabitants of Caithness, Scotland. The table represents the joint population distribution of the categorical variable for eye color x_{eye} and the categorical variable for hair color x_{hair}:

$$\begin{aligned} x_{hair} &\in \{\text{ fair red medium dark black}\} \\ x_{eye} &\in \{\text{ blue light medium dark}\}. \end{aligned} \tag{1}$$

Before defining the covariances among such categorical variables, $\sigma_{hair,eye}$, let us consider the variance of a categorical variable. Gini successfully defined the variance for categorical data [6].

$$\sigma_{ii} = \frac{1}{2N^2} \sum_{a=1}^{N} \sum_{b=1}^{N} (x_{ia} - x_{ib})^2 \tag{2}$$

where, σ_{ii} is the variance of the i-th variable, x_{ia} is the value of x_i for the a-th instance, and N is the number of instances. The distance of a categorical variable between instances is defined as $x_{ia} - x_{ib} = 0$ if their values are identical, and $= 1$ otherwise. A simple extension of this definition to the covariance σ_{ij} by replacing $(x_{ia} - x_{ib})^2$ to $(x_{ia} - x_{ib})(x_{ja} - x_{jb})$ does not give reasonable values for the covariance σ_{ij} [8]. In order to avoid this difficulty, we extended the definition based on scalar values, $x_{ia} - x_{ib}$, to a new definition using a vector expression [8]. The vector expression for a categorical variable with three categories $x_i \in \{r_1^i, r_2^i, r_3^i\}$ was defined by placing these three categories at the vertices of a regular triangle.

A regular simplex can be used for a variable with more than four categories. This is a straightforward extension of a regular triangle when the dimension of space is greater than two. For example, a regular simplex in the 3-dimensional space is a regular tetrahedron. Using a regular simplex, we can extend and generalize the definition of covariance to

Definition 1 *The covariance between a categorical variable* $x_i \in \{r^i_1, r^i_2, ...r^i_{k_i}\}$ *with* k_i *categories and a categorical variable* $x_j \in \{r^j_1, r^j_2, ...r^j_{k_j}\}$ *with* k_j *categories is defined as*

$$\sigma_{ij} = \max_{L^{ij}}(\frac{1}{2N^2} \sum_{a=1...N} \sum_{b=1...N} (\mathbf{v}^{k_i}(x_{ia}) - \mathbf{v}^{k_i}(x_{ib}))L^{ij}(\mathbf{v}^{k_j}(x_{ja}) - \mathbf{v}^{k_j}(x_{jb}))^t), \quad (3)$$

where $\mathbf{v}^n(r_k)$ *is the position of the* k*-th vertex of a regular* $(n-1)$*-simplex* [1]. r^i_k *denotes the* k*-th element of the* i*-th categorical variable* x_i. L^{ij} *is a unitary matrix expressing the rotation between the regular simplexes for* x_i *and* x_j.

Definition 1 includes a procedure to maximize the covariance. Using Lagrange multipliers, this procedure can be converted into a simpler problem of simultaneous equations, which can be solved using the Newton method. The following theorem enables this problem transformation.

Theorem 1 *The covariance between categorical variable* x_i *with* k_i *categories and categorical variable* x_j *with* k_j *categories is expressed by*

$$\sigma_{ij} = trace(A^{ij}{L^{ij}}^t), \quad (4)$$

where A^{ij} *is* $(k_i - 1) \times (k_j - 1)$ *matrix :*

$$A^{ij} = \frac{1}{2N^2} \sum_a \sum_b (\mathbf{v}^{k_i}(x_{ia}) - \mathbf{v}^{k_i}(x_{ib}))^t(\mathbf{v}^{k_j}(x_{ja}) - \mathbf{v}^{k_j}(x_{jb})). \quad (5)$$

L^{ij} *is given by the solution of the following simultaneous equations.*

$$\begin{aligned} A^{ij}{L^{ij}}^t &= (A^{ij}{L^{ij}}^t)^t \\ L^{ij}{L^{ij}}^t &= \mathbf{E} \end{aligned} \quad (6)$$

Proof *Here, we consider the case where* $k_i = k_j$ *for the sake of simplicity. Definition 1 gives a conditional maximization problem :*

$$\begin{aligned} &\sigma_{ij} = \max_{L^{ij}} \frac{1}{2N^2} \sum_a \sum_b (\mathbf{v}^{k_i}(x_{ia}) - \mathbf{v}^{k_i}(x_{ib}))L^{ij}(\mathbf{v}^{k_j}(x_{ja}) - \mathbf{v}^{k_j}(x_{jb}))^t \\ &\textit{subject to} \quad L^{ij}{L^{ij}}^t = \mathbf{E} \end{aligned} \quad (7)$$

The introduction of Lagrange multipliers Λ for the constraint $L^{ij}{L^{ij}}^t = \mathbf{E}$ gives the Lagrangian function:

$$V = trace(A^{ij}{L^{ij}}^t) - trace(\Lambda^t L^{ij}{L^{ij}}^t - \mathbf{E}),$$

where Λ is $k_i \times k_i$ matrix. A stationary point of the Lagrangian function V is a solution of the simultaneous equations (6). $\square$

Instead of maximizing (3) with constraint $L^{ij}{L^{ij}}^t = \mathbf{E}$, we can get the covariance by solving the equations (6), which can be solved easily using the Newton method.

Application of this method to Table 1 gives

$$\sigma_{hair,hair} = 0.36409, \sigma_{eye,hair} = 0.081253, \sigma_{eye,eye} = 0.34985 \tag{8}$$

We can derive a correlation coefficient using the covariance and variance values of categorical variables in the usual way. The correlation coefficients for x_{eye}, x_{hair} for Table 1 is 0.2277.

3 Principal Component Analysis

3.1 Principal Component Analysis of Categorical Data using Regular Simplex (RS-PCA)

Let us consider categorical variables $x_1, x_2 ... x_J$. For the a-th instance, x_i takes value x_{ia}. Here, we represent x_{ia} by the vector of vertex coordinates $\mathbf{v}^{k_i}(x_{ia})$. Then, the values of all the categorical variables $x_1, x_2 ... x_J$ for the a-th instance can be represented by the concatenation of the vertex coordinate vectors of all the categorical variables:

$$\mathbf{x}(a) = (\mathbf{v}^{k_1}(x_{1a}), \mathbf{v}^{k_2}(x_{2a}), ..., \mathbf{v}^{k_J}(x_{Ja})). \tag{9}$$

Let us call this concatenated vector the *List of Regular Simplex Vertices* (LRSV). The covariance matrix of LRSV can be written as

$$\mathcal{A} = \frac{1}{N}\sum_{a=1}^{N}(\mathbf{x}(a) - \bar{\mathbf{x}})^t(\mathbf{x}(a) - \bar{\mathbf{x}}) = \left[\begin{array}{c|c|c|c} A^{11} & A^{12} & ... & A^{1J} \\ \hline A^{21} & A^{22} & ... & A^{2J} \\ \hline ... & ... & ... & ... \\ \hline A^{J1} & A^{J2} & ... & A^{JJ} \end{array}\right]. \tag{10}$$

where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{a=1}^{N}\mathbf{x}(a)$ is an averege of the LRSV. The equation (10) shows the covariance matrix of LRSV. Since its eigenvalue decomposition can be regarded as a kind of Principal Component Analysis (PCA) on LRSV, we call it the *Principal Component Analysis using the Regular Simplex for categorical data* (RS-PCA).

When we need to interpret an eigenvector from RS-PCA, it is useful to express the eigenvector as a linear combination of the following vectors. The first basis set, d, shows vectors from one vertex to another vertex in the regular simplex. The other basis set, c, show vectors from the center of the regular simplex to one of the veritices.

$$\mathbf{d}^{k_j}(a \to b) = \mathbf{v}^{k_j}(b) - \mathbf{v}^{k_j}(a) \qquad a, b = 1, 2...k_j \tag{11}$$

$$\mathbf{c}^{k_j}(a) = \mathbf{v}^{k_j}(a) - \frac{\sum_{b=1}^{k_j} \mathbf{v}^{k_j}(b)}{k_j} \qquad a = 1, 2...k_j \tag{12}$$

Eigenvectors defined in this way change their basis set depending on its direction to the regular simplex, but this has the advantage of allowing us to grasp its meaning easily. For example, the first two principal component vectors from the data in Table 1 are expressed using the following linear combination.

$$\begin{aligned} \mathbf{v}_1^{rs-pca} &= -0.63 \cdot \mathbf{d}^{eye}(medium \to light) - 0.09 \cdot \mathbf{c}^{eye}(blue) - 0.03 \cdot \mathbf{c}^{eye}(dark) \\ &\quad -0.76 \cdot \mathbf{d}^{hair}(medium \to fair) + 0.07 \cdot \mathbf{d}^{hair}(dark \to medium) \qquad (13) \\ \mathbf{v}_2^{rs-pca} &= 0.64 \cdot \mathbf{d}^{eye}(dark \to light) - 0.13 \cdot \mathbf{d}^{eye}(medium \to light) \\ &\quad -0.68 \cdot \mathbf{d}^{hair}(dark \to medium) + 0.30 \cdot \mathbf{c}^{hair}(fair) \qquad (14) \end{aligned}$$

This expression shows that the axis is mostly characterized by the difference between $x^{eye} = light$ and $x^{eye} = medium$ values, and the difference between $x^{hair} = medium$ and $x^{hair} = fair$ values. The KL-plot using these components is shown in Figure 1 for Fisher's data. In this figure, the lower side is mainly occupied by data with values: $x^{eye} = medium$ or $x^{hair} = medium$. The upper side is mainly occupied by data with values $x^{eye} = light$ or $x^{hair} = fair$. Therefore, we can confirm that $(\mathbf{d}^{eye}(medium \to light) + \mathbf{d}^{hair}(medium \to fair))$ is the first principal component. In this way, we can easily interpret the data distribution on the KL-plot when we use the RS-PCA method.

Multiple Correspondence Analysis (MCA) [7] provides a similar PCA methodology to that of RS-PCA. It uses the representation of categorical values as an indicator matrix (also known as a dummy matrix). MCA gives a similar KL-plot. However, the explanation of its principal components is difficult, because their basis vectors contain one redundant dimension compared to the regular simplex expression. Therefore, a conclusion from MCA can only be drawn after making a great effort to inspect the KL-plot of the data.

4 Conclusion

We studied the covariances between a pair of categorical variables based on Gini's definition of the variance for categorical data. The introduction of the regular simplex expression for categorical values enabled a reasonable definition of covariances, and an algorithm for computing the covariance was proposed. The regular simplex expression was also shown to be useful in the PCA analysis. We showed these merits through numerical experiments using Fisher's data.

The proposed RS-PCA method is mathematically similar to the MCA method, but it is much easier to interpret the KL-plot in RS-PCA than in MCA.

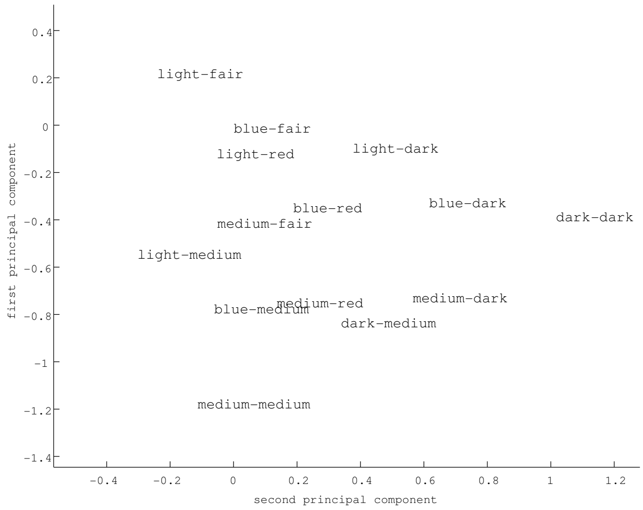

Fig. 1. KL-plot of Fisher's data calculated using RS-PCA. A point is expressed by a pair of eye and hair categories: $x^{eye} - x^{hair}$.

Acknowledgments This research was partially supported by the Ministry of Education, Culture, Sport, Science and Technology, of Japan, with a Grant-in-Aid for Scientific Research on Priority Areas, 13131210 and a Grant-in-Aid for Scientific Research (A) 14208032.

References

1. F. Buekenhout and M. Parker. The number of nets of the regular convex polytopes in dimension ≤ 4. *Disc. Math.*, 186:69–94, 1998.
2. W. Buntine. Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Machine Learning: ECML 2002. LNAI 2430*, pages 23–34. Springer-Verlag, 2002.
3. S.-E. Clausen. *Applied correspondence analysis: an introduction*. Thousand Oaks: Sage Publ, 1998.
4. K. Diamantaras and S. Kung. *Principal Component Neural Networks*. Wiley, New York, 1996.
5. R. A. Fisher. The precision of discriminant functions. *Annals of Eugenics (London)*, 10:422–429, 1940.
6. C. W. Gini. Variability and Mutability, contribution to the study of statistical distributions and relations. Studi Economico-Giuridici della R. Universita de Cagliari (1912). Reviewed in: R. J. Light and B. H. Margolin: An Analysis of Variance for Categorical Data. *J. American Statistical Association*, 66:534–544, 1971.
7. J. C. Gower and D. J. Hand. *Biplot*. Chapman and Hall, London, 1996.
8. T. Okada. A note on covariances for categorical data. In K. S. Leung, L. W. Chan, and H. Meng, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2000 LNCS 1983*, pages 150–157, 2000.