# A Two-Step Analysis with Common Quantification of Categorical Data

## カテゴリカルデータの共通数量化２段分析

OKAMOTO Yasuharu
岡本安晴

**[Abstract]** A data from a psychological research may contain categorical variables. This study proposes to use common quantification of categorical variables for various analyses, so that results from those analyses can be interpreted according to common quantification. The proposed method consists of two steps. In step 1, categorical variables are quantified independently of the subsequent analyses. After quantification, in step 2, various analyses are conducted based on common quantification. A hypothetical data set, including two categorical variables and one continuous variable, was prepared. Categorical variables were quantified in step 1; then in step 2, using the quantification in step 1, principal components, regression, and discriminant analyses were successfully conducted.

**[Key Words]** quantification, categorical variable, nominal variable, multivariate analysis, eigendecomposition

### Introduction

A data from a psychological research may contain categorical variables. This study proposes to use common quantification of categorical variables for various analyses so that results from those analyses can be understood according to common quantification. Usually, analysis of categorical data by quantification is done with criterion expression, set to make the quantification fit the analytical purpose. For example, homogeneity analysis uses the homogeneity function $\sigma_M(X, Y) = m^{-1} \sum_j SSQ(X - G_j Y_j)$ as a criterion, where $G_j$ is an indicator matrix (data), and $X$ and $Y_j$ are quantifications (Gifi, 1990). This quantification is meant to extract the homogeneity structure of the data $G_j$s. In general, optimal quantification is assigned according to a criterion set to reflect specific information among categorical variables, so, the quantification depends on the specified relations among them. However, when investigating various relationships in data, e.g., homogeneity, regression, discrimination, and so on, it would be informative to use common quantification for various analyses. This study proposes to use common quantification, based on which analyses can be conducted. Since common quantification is used for various subsequent analyses, it must not be influenced by them.

Categorical data, to which quantification is applied, might be gathered in various forms, and many quantification methods have been proposed. Explanations and discussions have been presented (Gifi, 1990; Greenacre, 2007; Nishisato, 2007). However, in this study, only one type of categorical data is employed. Data are assumed to be a matrix form, each row $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ comprising individual data. Furthermore, $x_{ij}$ might be categorical or continuous. Categorical variables can be analyzed with continuous variables after quantification. Furthermore, categories are assumed to be nominal, i.e., not ordered.

**Two-Step Analysis: Quantification and Analyses**

Let

$$X = (x_{ij}) = (X_1, \cdots, X_p, X_{p+1}, \cdots, X_{p+q})$$

denote the data, where

$$X_j = (x_{1j}, \cdots, x_{nj})'$$

and $n$ is the number of individuals. Assume that $X_1$, $\cdots$ , $X_p$ are continuous variables and $X_{p+1}$, $\cdots$ , $X_{p+q}$ are categorical variables. In the first step, each $X_{p+k}$ of categorical variables is quantified; then, in the second step, various multivariate analyses are applied to quantified categorical variables, possibly with continuous variables.

To quantify categorical variables independently of the specific subsequent analyses, each categorical variable is quantified on its own structure of information. Statistically, information of a categorical variable, which is nominal, is represented by frequency distribution of categories, so quantification is based on frequency of information.

**Step 1: Quantification**

Let $X_{p+k}$ have $C_k$ categories, say 1 to $C_k$. Set the indicator matrix $G_k = \left(g_{ij}^{(k)}\right)$ of $X_{p+k}$ as follows

$$g_{ij}^{(k)} = \begin{cases} 1 & \text{if } x_{i,p+k} = j \\ 0 & otherwise \end{cases} \qquad (1)$$

When category $j$ is quantified as $\omega_{kj}$, individual $i$’s quantified score of categorical variable $X_{p+k}$ is represented by

$$\sum_{j=1}^{C_k} g_{ij}^{(k)} \omega_{kj} \ . \qquad (2)$$

Since statistical structure about information on categorical variable $X_{p+k}$ is represented by frequency distribution of categories 1 to $C_k$ and we consider that information about frequency distribution is maximally reflected by quantification when variance of the quantified variable is maximized (Tenenhaus and Young, 1985), we choose quantification $\omega_{kj}$ so that variance of Equation 2 becomes (relative) maximum under the constraint on $\omega_{kj}$s.

Since

$$\mathbf{1}_n' G_k \boldsymbol{\omega}_k \ = \boldsymbol{f}_k \boldsymbol{\omega}_k \ ,$$

the variance is given by

$$\frac{1}{n}\left\{\left(G_k - \frac{1}{n}\mathbf{1}_n \boldsymbol{f}_k\right)\boldsymbol{\omega}_k\right\}'\left(G_k - \frac{1}{n}\mathbf{1}_n \boldsymbol{f}_k\right)\boldsymbol{\omega}_k$$

where $\boldsymbol{f}_k = \mathbf{1}_n' G_k$, $\boldsymbol{\omega}_k = \left(\omega_{k1} \cdots \omega_{kC_k}\right)$, and $\mathbf{1}_n = (1 \cdots 1)'$ is the $n$–dimensional column vector with all elements being 1.

Set

$$H_k = G_k - \frac{1}{n}\mathbf{1}_n \boldsymbol{f}_k$$

Choose quantification $\boldsymbol{\omega}_k$, which optimally reflects information contained in categorical data $G_k$, so that $\boldsymbol{\omega}_k$ maximizes

$$(H_k\boldsymbol{\omega}_k)'(H_k\boldsymbol{\omega}_k)$$

under the constraint

$$\boldsymbol{\omega}_k'\boldsymbol{\omega}_k = 1.$$

Set a Lagrangian function as follows (Magnus & Neudecker, 1999):

$$Q = (H_k\boldsymbol{\omega}_k)'(H_k\boldsymbol{\omega}_k) - \lambda(\boldsymbol{\omega}_k'\boldsymbol{\omega}_k - 1).$$

Differentiating $Q$ with $\boldsymbol{\omega}_k$, we get

$$\frac{\partial Q}{\partial \boldsymbol{\omega}_k} = 2H_k'H_k\boldsymbol{\omega}_k - 2\lambda\boldsymbol{\omega}_k.$$

Hence, the optimal quantification $\boldsymbol{\omega}_k$ satisfies

$$H_k'H_k\boldsymbol{\omega}_k = \lambda_k\boldsymbol{\omega}_k,$$

where $\boldsymbol{\omega}_k$ is an eigenvector of $H_k'H_k$, and $\lambda_k$ is the associated eigenvalue.

The number of $\lambda_k$s is $C_k$. Denote these $\lambda_k$s as $\lambda_k^{(1)}$, $\cdots$, $\lambda_k^{(C_k)}$. We can assume that $\lambda_k^{(1)} \geq \cdots \geq \lambda_k^{(C_k)} \geq 0$. (See the Appendix for more technical details.)

Set

$$\boldsymbol{\omega} = \frac{1}{\sqrt{C_k}}\mathbf{1}_{C_k},$$

then

$$H_k'H_k\boldsymbol{\omega} = \frac{1}{\sqrt{C_k}}H_k'\left(G_k - \frac{1}{n}\mathbf{1}_n \boldsymbol{f}_k\right)\mathbf{1}_{C_k} = \mathbf{0}$$

Hence, we have

$$\lambda_k^{(C_k)} = 0,\ \boldsymbol{\omega}_k^{(C_k)} = \left(1/\sqrt{C_k}\right)\mathbf{1}_{C_k}.$$

The variance of the quantified $G_k$ by $\omega_k^{(h)}$ is given by

$$Var\left(G_k\boldsymbol{\omega}_k^{(h)}\right) = Var\left(H_k\boldsymbol{\omega}_k^{(h)}\right) = \frac{\lambda_k^{(h)}}{n}$$

Hence, for $\lambda_k^{(h)} > 0$, the standardized score $\boldsymbol{z}_k^{(h)}$ of $G_k\boldsymbol{\omega}_k^{(h)}$ is given by

$$\boldsymbol{z}_k^{(h)} = \left(\lambda_k^{(h)}/n\right)^{-1/2} H_k\boldsymbol{\omega}_k^{(h)}. \tag{3}$$

Set

$$G_k^z = \left(\boldsymbol{z}_k^{(1)} \quad \cdots \quad \boldsymbol{z}_k^{(K_k)}\right),$$

where $K_k$ denotes the number of $\lambda_k^{(h)}$s, which are not 0.

$G_k^z$ is the quantification of $X_{p+k}$, obtained independently of other $X_t$s and analyses planned after quantification. After quantification in step 1, various multivariate analyses can be applied to the data. For analyses, the same quantification $G_k^z$s of $X_{p+k}$ can be used.

**Step2: Analyses**

After quantifying categorical data $X_{p+k}$s, we have quantitative data

$$X_G = (X_1 \quad \cdots \quad X_p \quad G_1^z \quad \cdots \quad G_q^z).$$

To this data, we can apply various multivariate analyses such as principal components analysis (PCA), regression analysis, and discriminant analysis. As examples, these three analyses will be briefly discussed, considering the use of quantification $G_k^z$s.

**Principal Components Analysis.** First, select variables from $X_G$ to which PCA will be applied. If unstandardized variables are included, standardize them. Donate the standardized data to which PCA will be applied by $Z$. The PCA of $Z$ can be viewed as the orthogonal projection of $Z$, which maximizes variance of the projected $Z$ (Okamoto, 2006). Here, it should be emphasized that projection is performed in one step, although the classical explanation extracts principal components individually under orthogonal constraint (Morrison, 1976). Principal components are given in the subspace on which $Z$ is optimally projected, and mathematically, any coordinate system can be justified in the subspace. Rotation in PCA is associated with the selection of a coordinate system, so any rotation is mathematically justified. When eigenvectors of the data's correlation matrix are adopted as bases of the coordinate system and the coordinates are standardized, we obtain the so-called principal components.

Multiple factor analysis (MFA) (cf. Bécue−Bertaut & Pagès, 2008) also employs a two-step PCA method for categorical data. However, MFA employs only the first eigenvalues of $G_k$s. It should be noted that quantification in this study uses all eigenvalues to construct $z_k^{(h)}$ (Equation 3).

**Regression Analysis.** To investigate dependence of a quantitative variable $X_s$ $(s = 1, \cdots, p)$ on categorical variables $X_{p+t}$s $(t = 1, \cdots, q)$, regression analysis, which has $X_s$ as a dependent variable and $G_t^z$s as predictor variables, can be employed. When some independent variables have strong correlations among them, the estimation of regression coefficient might become unstable (multicolinearity). We can use principal components (or rotated principal components) to avoid multicolinearity in this case.

**Discriminant Analysis.** Next, consider applying Fisher's linear discriminant function (Mardia, Kent, & Bibby, 1979) to $X_G$. Select $r$ variables to be used in discriminant analysis and denote them as

$$\boldsymbol{z}_{hi} = (z_{hi1} \quad \cdots \quad z_{hir})',$$

where $z_{hij}$ is the value of individual $i$ in group $h$ on the $j-$th variable selected. Assume that $\boldsymbol{z}_{hi}$s are centered, i.e.,

$$\sum_{h,i} z_{hij} = 0.$$

Set a linear combination of $z_{hij}$s as follows:

$$y_{hi} = \sum_{j} v_j z_{hij} = \boldsymbol{v}'\boldsymbol{z}_{hi},$$

where $\boldsymbol{v} = (v_1 \quad \cdots \quad v_r)'$ is a column vector. Determine $\boldsymbol{v}$, which maximizes the ratio of the between-group variance, $SS_{between}$, to the within-group variance, $SS_{within}$, under the constraint $\boldsymbol{v}'\boldsymbol{v} = 1$. $\boldsymbol{v}$, which gives the $k-$th relative maximum of $SS_{between}/SS_{within}$ under the constraint $\boldsymbol{v}'\boldsymbol{v} = 1$, defines the $k-$th canonical variate (Mardia, Kent, and Bibby, 1979) or the $k-$th discriminant factor (Cooley & Lohnes, 1971).

The prominent property of the two-step analysis is that the categorical data's common quantification can be used for various multivariate analyses. In the next section, a simple example is presented.

**Table 1** A Hypothetical Data Set

| ID | Group | Y | CV−1 | CV−2 |
|---|---|---|---|---|
| A | GA | 10 | Dog | Apple |
| A0 | GA | 12 | Dog | Bean |
| A1 | GA | 15 | Dog | Bean |
| A2 | GA | 17 | Fish | Apple |
| B | GB | 50 | Cat | Apple |
| B1 | GB | 47 | Cat | Bean |
| B2 | GB | 51 | Fish | Apple |
| C | GC | 49 | Dog | Pear |
| C1 | GC | 52 | Dog | Bean |
| C2 | GC | 47 | Fish | Pear |
| D | GD | 90 | Cat | Pear |
| D0 | GD | 88 | Fish | Pear |
| D1 | GD | 87 | Cat | Bean |
| D2 | GD | 89 | Fish | Pear |

**Example of an Application**

The proposed two-step analysis was applied to a hypothetical data set (Table 1), comprising five variables, i.e., ID (individual identification); Group (group identification); Y (a continuous variable); CV−1 (a categorical variable with categories, Dog, Cat, and Fish); and CV−2 (a categorical variable with categories, Apple, Pear, and Bean). Groups A, B, C, and D are characterized by (Dog, Apple); (Cat, Apple); (Dog, Pear); and (Cat, Pear); respectively. Categories Fish and Bean are irregular. Variable Y takes relatively lesser values in group A, intermediate ones in groups B and C, and greater ones in group D.

**Step 1: Quantification**

Indicator matrices, $G_1$ and $G_2$, of categorical variables CV−1 and CV−2 were constructed by assigning integers 1, 2, and 3 to categories Dog, Cat, and Fish of CV−1, and to categories Apple, Pear, and Bean of CV−2, respectively (Figure 1). From $G_k$s, $H_k$s were calculated. Eigenvalues and eigenvectors of $H_k'H_k$s are shown in Table 2. From these eigenvalues and eigenvectors, the standardized scores $z_k^{(h)}$s were calculated (Table 3). $z_k^{(h)}$s denote the $h$−th standardized scores of the categorical variable CV−k.

$$\begin{array}{cc}
\text{CV}-1 & \text{CV}-2 \\
G_1 = \begin{bmatrix} 1&0&0 \\ 1&0&0 \\ 1&0&0 \\ 0&0&1 \\ 0&1&0 \\ 0&1&0 \\ 0&0&1 \\ 1&0&0 \\ 1&0&0 \\ 0&0&1 \\ 0&1&0 \\ 0&0&1 \\ 0&1&0 \\ 0&0&1 \end{bmatrix} &
G_2 = \begin{bmatrix} 1&0&0 \\ 0&0&1 \\ 0&0&1 \\ 1&0&0 \\ 1&0&0 \\ 0&0&1 \\ 1&0&0 \\ 0&1&0 \\ 0&0&1 \\ 0&1&0 \\ 0&1&0 \\ 0&1&0 \\ 0&0&1 \\ 0&1&0 \end{bmatrix}
\end{array}$$

*Figure 1*. The indicator matrix $G_k$ for categorical variable CV−$k$. To make indicator matrices by Equation 1, categories in Table 1 are converted to integers as follows: Dog and Apple to 1, Cat and Pear to 2, and Fish and Bean to 3.

**TABLE 2** Eigenvalues and Eigenvectors of $H_k'H_k$

| | | $H_1'H_1$ | | | $H_2'H_2$ | |
|---|---|---|---|---|---|---|
| $\lambda_k^{(h)} > 0$ | | 5.000 | 4.286 | | 5.000 | 4.286 |
| $\omega_k^{(h)}$ | Dog | 0.707 | –0.408 | Apple | 0.000 | 0.816 |
| | Cat | 0.000 | 0.816 | Pear | 0.707 | –0.408 |
| | Fish | –0.707 | –0.408 | Bean | –0.707 | –0.408 |

**TABLE 3** The $h-$th Standardized Scores, $z_k^{(h)}$, of Categorical Variable CV−k

| ID | $z_1^{(1)}$ | $z_1^{(2)}$ | $z_2^{(1)}$ | $z_2^{(2)}$ |
|---|---|---|---|---|
| A | 1.183 | –0.632 | 0.000 | 1.581 |
| A0 | 1.183 | –0.632 | −1.183 | –0.632 |
| A1 | 1.183 | –0.632 | −1.183 | –0.632 |
| A2 | −1.183 | –0.632 | 0.000 | 1.581 |
| B | 0.000 | 1.581 | 0.000 | 1.581 |
| B1 | 0.000 | 1.581 | −1.183 | –0.632 |
| B2 | −1.183 | –0.632 | 0.000 | 1.581 |
| C | 1.183 | –0.632 | 1.183 | –0.632 |
| C1 | 1.183 | –0.632 | −1.183 | –0.632 |
| C2 | −1.183 | –0.632 | 1.183 | –0.632 |
| D | 0.000 | 1.581 | 1.183 | –0.632 |
| D0 | −1.183 | –0.632 | 1.183 | –0.632 |
| D1 | 0.000 | 1.581 | −1.183 | –0.632 |
| D2 | −1.183 | –0.632 | 1.183 | –0.632 |

**Step2: Multivariate Analyses**

**Principal Component Analysis.** PCA was applied to a set of variables $z_1^{(1)}$ through $z_2^{(2)}$. Singular values, derived from PCA, are shown in Table 4 (as to the relationship between singular value decomposition and PCA, c.f. Okamoto (2006)). The principal components are shown in Table 5. Correlation coefficients between the principal components and quantification $z_k^{(h)}$s are shown in Table 6. Figure 2 shows the configuration of individuals on the first and second principal components. Uniform random jitters centered at zero were added to the coordinates to avoid the complete overlapping of circles with same coordinate values. We see that the first principal component, denoted as Comp. 1, represents the irregular categories, i.e., Fish and Bean. Figure 3 shows the configuration of the second and third principal components, denoted as Comp. 2 and Comp. 3, respectively, with uniform random jitters added. We observe that the second principal component, i.e., Comp. 2, discriminates among groups B, A and D, and C, and the third one, i.e., Comp. 3, differentiates among groups A, B and C, and D.

**Regression Analysis.** Regression analysis was applied to the dependent variable Y with independent variables $z_1^{(1)}$, $z_1^{(2)}$, $z_2^{(1)}$, and $z_2^{(2)}$. That is, the following model 4 was set:

$$\mathrm{Y} = \beta_0 + \beta_1 z_1^{(1)} + \beta_2 z_1^{(2)} + \beta_3 z_2^{(1)} + \beta_4 z_2^{(2)} + Residual \tag{4}$$

**TABLE 4**　Singular Values from PCA

| | | | |
|---|---|---|---|
| 1.252 | 1.009 | 0.991 | 0.658 |

**TABLE 5**　Principal Components. Comp. k denotes the $k$−th principal component

| ID | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 |
|---|---|---|---|---|
| A | 0.204 | 0.343 | 1.772 | 1.543 |
| A0 | 1.257 | −0.834 | 0.574 | −0.462 |
| A1 | 1.257 | −0.834 | 0.574 | −0.462 |
| A2 | −1.053 | 0.907 | 1.198 | −0.850 |
| B | 0.000 | 2.085 | 0.000 | 1.156 |
| B1 | 1.053 | 0.907 | −1.198 | −0.850 |
| B2 | −1.053 | 0.907 | 1.198 | −0.850 |
| C | 0.000 | −1.398 | 0.000 | 1.931 |
| C1 | 1.257 | −0.834 | 0.574 | −0.462 |
| C2 | −1.257 | −0.834 | −0.574 | −0.462 |
| D | −0.204 | 0.343 | −1.772 | 1.543 |
| D0 | −1.257 | −0.834 | −0.574 | −0.462 |
| D1 | 1.053 | 0.907 | −1.198 | −0.850 |
| D2 | −1.257 | −0.834 | −0.574 | −0.462 |

**TABLE 6**　Structure Matrix, i.e., Correlations between $z_k^{(h)}$ and Comp. j

| | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 |
|---|---|---|---|---|
| $z_1^{(1)}$ | 0.833 | −0.242 | 0.238 | 0.437 |
| $z_1^{(2)}$ | 0.301 | 0.671 | −0.659 | 0.158 |
| $z_2^{(1)}$ | −0.833 | −0.242 | −0.238 | 0.437 |
| $z_2^{(2)}$ | −0.301 | 0.671 | 0.659 | 0.158 |

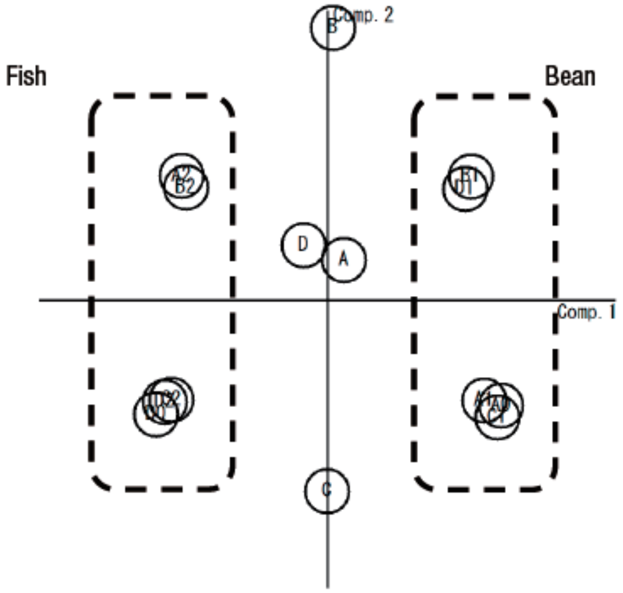

*Figure 2*. Configuration of individuals on the first and second components denoted by Comp. 1 and Comp. 2, respectively. To avoid complete overlapping, uniform random jitters centered at zero were added to the coordinates.

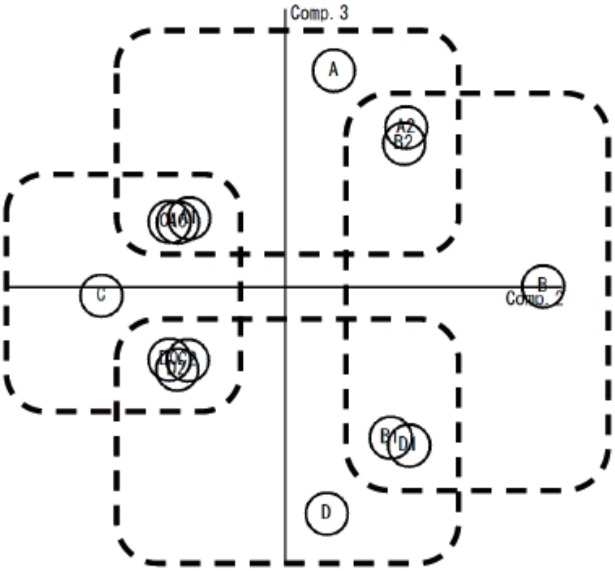

*Figure 3*. Configuration of individuals on the second and third components denoted by Comp. 2 and Comp. 3, respectively. To avoid complete overlapping, uniform random jitters centered at zero were added to the coordinates.

The estimates $\hat{\beta}_j$s of parameters $\beta_j$s by the least squares method are shown in Table 7, and the configuration of point $(\mathrm{Y}, \hat{\mathrm{Y}})$ in Figure 4, where $\hat{\mathrm{Y}}$ is an estimate by the model, i.e.,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 z_1^{(1)} + \hat{\beta}_2 z_1^{(2)} + \hat{\beta}_3 z_2^{(1)} + \hat{\beta}_4 z_2^{(2)}$$

**TABLE 7**　Estimates of Parameters of the Regression Model:
$\mathrm{Y} = \beta_0 + \beta_1 z_1^{(1)} + \beta_2 z_1^{(2)} + \beta_3 z_2^{(1)} + \beta_4 z_2^{(2)} + Residual$

| $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|---|---|---|---|---|
| 50.29 | −10.54 | 12.71 | 9.79 | −12.90 |

The configuration in Figure 4 shows that the global tendency of variation in Y is captured by the regression model (4). The figures in Table 7 (see also weights $\omega_k^{(h)}$s in Table 2) show that Ys tend to be greater for category Cat than for category Dog ($\hat{\beta}_1 = -10.54$, $\hat{\beta}_2 = 12.71$) and greater for category Pear than for category Apple ($\hat{\beta}_3 = 9.79$, $\hat{\beta}_4 = -12.90$).

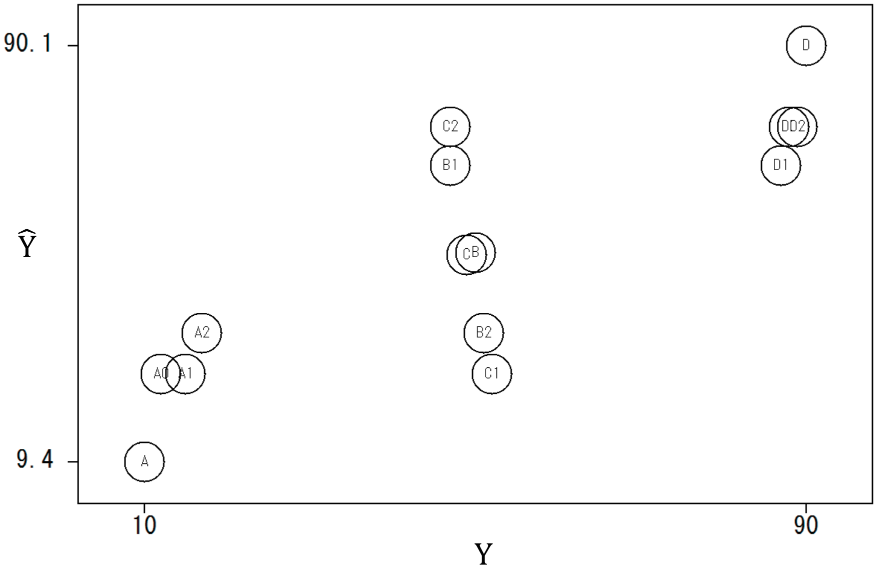

*Figure 4*. Configuration of $(\mathrm{Y}, \hat{\mathrm{Y}})$, where $\hat{\mathrm{Y}}$ is an estimate by the regression model.

When the principal components were used as independent variables, i.e., the following model

$$\mathrm{Y} = \beta_0 + \beta_1 Comp.1 + \beta_2 Comp.2 + \beta_3 Comp.3 + \beta_4 Comp.4 + Residual$$

was used, the results shown in Table 8 were obtained. The value −21.72 of $\hat{\beta}_3$ indicates the overall tendency of Y, which increases from group A through groups B and C to group D (Figure 3). Coefficient $\hat{\beta}_1 =$ −9.21 means that Y tends to be less for category Bean than for category Fish (Figures 2 and 4).

**TABLE 8** Estimates of Parameters of the Regression Model:

$\mathrm{Y} = \beta_0 + \beta_1 Comp.1 + \beta_2 Comp.2 + \beta_3 Comp.3 + \beta_4 Comp.4 + Residual$

| $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|---|---|---|---|---|
| 50.29 | –9.21 | 0.05 | –21.72 | –0.36 |

**Discriminant Analysis.** Discriminant analysis was applied to the data set that has a group identification variable Group (Table 1) and independent variables $z_1^{(1)}$, $z_1^{(2)}$, $z_2^{(1)}$ and $z_2^{(2)}$ (Table 3). Set a linear combination of $z_1^{(1)}$ to $z_2^{(2)}$ as follows:

$$y = v_1 z_1^{(1)} + v_2 z_1^{(2)} + v_3 z_2^{(1)} + v_4 z_2^{(2)} . \tag{5}$$

Note that means of $z_k^{(h)}$s are all zero. The $k-$th canonical variate or discriminant factor is given as Equation 5 that gives the $k-$th largest relative maximum of the ratio of the between-groups sum of squares to the within-groups sum of squares (Mardia, Kent, & Bibby, 1979).

The configuration of individuals in the discriminant plane is shown in Figure 5. Small circles represent individuals, which are plotted at their positions (the first and the second discriminant factor). Uniform random jitters centered at zero were added to avoid complete overlapping of individuals having the same coordinates. Members of each group gather around the squares, each of which represents the group mean's position. Figure 5 shows that the first discriminant factor separates individuals into groups A, B and C, and D, and that the second discriminant factor separates individuals into groups B, A and D, and C.

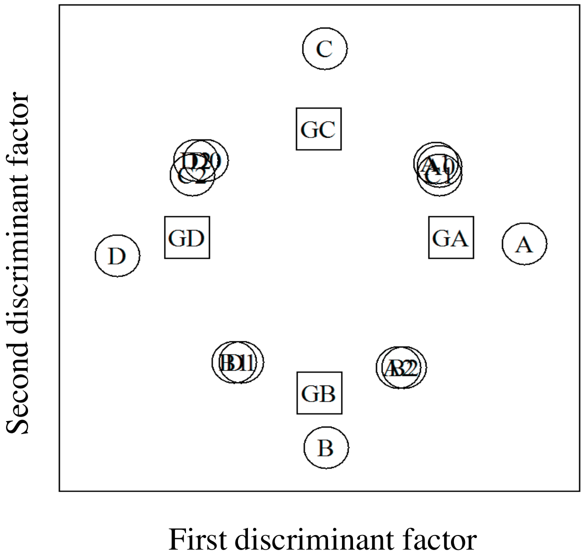

*Figure 5*. Configuration of individuals (circles) and group means (squares) in the discriminant plane. To avoid complete overlapping, uniform random jitters centered at zero were added to the coordinates.

In a comparison of the results from the discriminant and principal components analyses, correlation coefficients were calculated between the discriminant factors and principal components. Table 9 shows that the first, second, and third discriminant factors are almost identical to the third, second, and fourth principal components, respectively, although the directions of the second discriminant factor and the second principal component are opposite. Absolute values of the correlation coefficients of these pairs are greater than 0.9. The configurations of individuals in Figures 3 and 5 reflect this correspondence between the discriminant factors and the principal components.

**TABLE 9** Correlation Coefficients between Principal Components and Discriminant Factors: Disc. $i$ denotes the $i-$th discriminant factor, and Comp. j denotes the $j-$th principal component

| | Disc. 1 | Disc. 2 | Disc. 3 |
|---|---|---|---|
| Comp. 1 | 0.375 | 0.000 | 0.000 |
| Comp. 2 | 0.000 | −0.979 | 0.202 |
| Comp. 3 | 0.927 | 0.000 | 0.000 |
| Comp. 4 | 0.000 | 0.202 | 0.979 |

**Discussion**

In the example of application of the two-step analysis to the hypothetical data set (Table 1), common quantification of categorical data was used for the three analyses: principal components, regression, and discriminant analyses. The principal components were derived from the quantification $z_1^{(1)}$ to $z_2^{(2)}$ and used in the subsequent regression and discrimination analyses. By these principal components, derived from the common quantification in step 1, and used in various analyses in step 2, we obtain integrative interpretation of the results from the three analyses, i.e., principal components, regression, and discriminant analyses.

Although various methods for the different forms of categorical data have been proposed and discussed, in this study the data set is assumed to be a case-by-variable format, i.e., $X = (x_{ij})$, where $X_j = (x_{1j}, \cdots, x_{nj})'$ is a column vector of values of $n$ individuals on variable $j$. A two-step method of PCA for categorical data is also used in MFA (Bécue−Bertau & Pagès, 2008). However, MFA uses only the first eigenvalue for scaling values. The two-step method in this study uses all eigenvalues to standardize all quantifications. Furthermore, in the two-step analysis proposed in this study, quantified variables might not be used just for PCA, but also for other analyses as common quantification. While some quantification methods, e.g., analysis by meet loss, $\sigma_M(X, Y)$, use iterative algorithm, the quantification in step 1 of the two-step analysis is obtained using eigenvalues and eigenvectors. Iterative algorithm might not converge, but eigenvalues and eigenvectors can be obtained from eigendecomposition of a real symmetric matrix by effective algorithm (Burden & Faires, 1997; Press, Teukolsky, Vetterling, & Flannery, 2007).

The two-step analysis in this study can analyze a data that includes both continuous and categorical items, and the quantification in step 1 can be commonly used in the subsequent analyses (step 2). By this common quantification, interpretation of results from analyses can be easily compared. The two-step analysis in this study can be expected to broaden the possibility of various analyses' applicability to psychological data with continuous and categorical items, and to make it easier to compare results from various analyses.

## References

Bécue-Bertaut, M., & Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, 52(6), 3255−3268.

Burden, R. L., & Faires, J. D. (1997). *Numerical analysis, sixth edition*. Pacific Grove: Brooks/Cole Publishing company.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis*. New York: John Wiley & Sons, Inc.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: John Wiley & Sons, Inc.

Greenacre, M. (2007). *Correspondence analysis in practice, second edition*. Boca Raton: Chapman & Hall/CRC.

Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics, Revised edition*. Chichester: John Wiley & Sons.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Amsterdam: Academic Press.

Morrison, D. F. (1976). *Multivariate statistical methods, Second edition*. New York: McGraw−Hill Book Company.

Nishisato, S. (2007). *Multidimensional nonlinear descriptive analysis*. Boca Raton: Chapman & Hall/CRC.

Okamoto, Y. (2006). A Justification of Rotation in Principal Component Analysis : Projective viewpoint of PCA. *Faculty of Integrated Arts and Social Sciences journal, Japan Women's University*, 17, 59−71. Retrieved from http://ci.nii.ac.jp/naid/110006223025

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing, third edition*. Cambridge: Cambridge University Press.

Tenenhaus, M., & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50(1), 91-119.

## Appendix

More explicit but rather complicated explanation of the quantification in Step 1 is as follows.

When $s-1$ quantifications $\boldsymbol{\omega}_k^{(1)}$, $\cdots$, $\boldsymbol{\omega}_k^{(s-1)}$ have been calculated for $H_k$, the next quantification $\boldsymbol{\omega}_k^{(s)}$ can be calculated by setting the following equation

$$Q_k^{(s)} = \left(H_k\boldsymbol{\omega}_k^{(s)}\right)'\left(H_k\boldsymbol{\omega}_k^{(s)}\right) - \lambda_k^{(s)}\left(\boldsymbol{\omega}_k^{(s)\prime}\boldsymbol{\omega}_k^{(s)} - 1\right)$$

$$-2\mu_{k,s}^{(1)}\left(\boldsymbol{\omega}_k^{(1)\prime}\boldsymbol{\omega}_k^{(s)} - 0\right) - \cdots - 2\mu_{k,s}^{(s-1)}\left(\boldsymbol{\omega}_k^{(s-1)\prime}\boldsymbol{\omega}_k^{(s)} - 0\right)$$

We have

$$\frac{\partial Q_k^{(s)}}{\partial \boldsymbol{\omega}_k^{(s)}} = 2H_k'H_k\boldsymbol{\omega}_k^{(s)} - 2\lambda_k^{(s)}\boldsymbol{\omega}_k^{(s)} - 2\mu_{k,s}^{(1)}\boldsymbol{\omega}_k^{(1)} - \cdots - 2\mu_{k,s}^{(s-1)}\boldsymbol{\omega}_k^{(s-1)}$$

Hence, $\boldsymbol{\omega}_k^{(s)}$, at that $\left(H_k\boldsymbol{\omega}_k^{(s)}\right)'\left(H_k\boldsymbol{\omega}_k^{(s)}\right)$ has a relative maximum under the constraint

$$\boldsymbol{\omega}_k^{(s)'}\boldsymbol{\omega}_k^{(s)} = 1,\ \boldsymbol{\omega}_k^{(1)'}\boldsymbol{\omega}_k^{(s)} = 0,\ \cdots\ ,\ \boldsymbol{\omega}_k^{(s-1)'}\boldsymbol{\omega}_k^{(s)} = 0,$$

satisfies the equation

$$H_k'H_k\boldsymbol{\omega}_k^{(s)} = \lambda_k^{(s)}\boldsymbol{\omega}_k^{(s)} + \mu_{k,s}^{(1)}\boldsymbol{\omega}_k^{(1)} + \cdots + \mu_{k,s}^{(s-1)}\boldsymbol{\omega}_k^{(s-1)}$$

Multiplying $\boldsymbol{\omega}_k^{(t)}$; $1 \leq t < s$, to the above equation, we have

$$\boldsymbol{\omega}_k^{(t)\prime}H_k'H_k\boldsymbol{\omega}_k^{(s)} = \lambda_k^{(s)}\boldsymbol{\omega}_k^{(t)'}\boldsymbol{\omega}_k^{(s)} + \mu_{k,s}^{(1)}\boldsymbol{\omega}_k^{(t)'}\boldsymbol{\omega}_k^{(1)} + \cdots + \mu_{k,s}^{(s-1)}\boldsymbol{\omega}_k^{(t)\prime}\boldsymbol{\omega}_k^{(s-1)}$$

$$= \mu_{k,s}^{(t)}$$

Noticing

$$\boldsymbol{\omega}_k^{(t)\prime}H_k'H_k\boldsymbol{\omega}_k^{(s)} = \left(H_k'H_k\boldsymbol{\omega}_k^{(t)}\right)'\boldsymbol{\omega}_k^{(s)} = \left(\lambda_k^{(t)}\boldsymbol{\omega}_k^{(t)}\right)'\boldsymbol{\omega}_k^{(s)} = 0,$$

we have–

$$\mu_{k,s}^{(t)} = 0.$$

Hence, we have

$$H_k'H_k\boldsymbol{\omega}_k^{(s)} = \lambda_k^{(s)}\boldsymbol{\omega}_k^{(s)}. \tag{A1}$$

**In the case where some eigenvalues are equal to each other**

When

$$\lambda_k^{(t)} = \lambda_k^{(t+1)} = \cdots = \lambda_k^{(t+u)} = \lambda, \tag{A2}$$

we have

$$\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)\begin{pmatrix}\lambda & & 0\\ & \ddots & \\ 0 & & \lambda\end{pmatrix}\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)' = \lambda\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)'.$$

Let $T$ be any orthogonal matrix of order $u$ by $u$, we have

$$TT' = \begin{pmatrix}1 & & 0\\ & \ddots & \\ 0 & & 1\end{pmatrix}.$$

Hence, we have

$$\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)\begin{pmatrix}\lambda & & 0\\ & \ddots & \\ 0 & & \lambda\end{pmatrix}\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)'$$

$$= \lambda\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)TT'\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)'$$

$$= \left(\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)T\right)\begin{pmatrix}\lambda & & 0\\ & \ddots & \\ 0 & & \lambda\end{pmatrix}\left(\left(\boldsymbol{\omega}_k^{(t)}\cdots\boldsymbol{\omega}_k^{(t+u)}\right)T\right)'. \qquad (A3)$$

Equation A3 shows that for $\lambda_k^{(s)}$s in Equation A1, which are the same value $\lambda$, i.e. which satisfy Equation A2, $\boldsymbol{\omega}_k^{(s)}$s can be determined under the restriction of orthogonal rotation in the space spanned by $\boldsymbol{\omega}_k^{(t)}, \cdots, \boldsymbol{\omega}_k^{(t+u)}$.