Forced Classification in Principal Component Analysis

Yasumasa Baba

Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku
Tokyo Japan 106-8569
Baba@ism.ac.jp

1. Forced classification in categorical data analysis

Let

$$\mathbf{F} = \left[f_{ij} \right] \quad (i = 1,..., N;\ j = 1,..., M)$$

be the data matrix obtained from item-category type questions or an frequency table of response patterns. Here let us consider the former case without generality. Let x_j, y_i be the quantities assigned to item-category j and subject i.

$$\mathbf{y} = [y_i]^t, \quad \mathbf{x} = [x_i]^t, \quad \mathbf{y} = \mathbf{F}\mathbf{x}.$$

In Hayashi's quantification method type III item-categories and subjects are described on a plane spanned by $\mathbf{x}$ or $\mathbf{y}$ which is the solution of an eigen equation derived from maximization of the correlation between $\mathbf{x}$ and $\mathbf{y}$.

Here let us consider the item-category type data matrix from N subjects and indicate it as

$$\mathbf{F} = [\mathbf{F}_1, \mathrm{K}\ , \mathbf{F}_s, \mathrm{K}\ , \mathbf{F}_p]$$

where $\mathbf{F}_j$ is an $N \times m_j$ matrix of 1's and 0's of item j and m_j is the number of the category j. Let us define two matrices

$$\mathbf{F}^* = [\mathbf{F}_1, \mathrm{K}\ , \mathbf{F}_s, \mathrm{K}\ , \mathbf{F}_s, \mathrm{K}\ , \mathbf{F}_p],$$
$$\mathbf{G} = [\mathbf{F}_1, \mathrm{K}\ k\mathbf{F}_s, \mathrm{K}\ , \mathbf{F}_p]$$

In $\mathbf{F}^*$ $\mathbf{F}_s$ is repeated k times. As the ranks of three matrices $\mathbf{F}$, $\mathbf{F}^*$ and $\mathbf{G}$ are same the properties of the eigen equations are not changed if we use $\mathbf{G}$ for $\mathbf{F}$. From principle of equivalent partitioning (Nishisato,1984) we can obtain same results about $\mathbf{x}$ and $\mathbf{y}$ form $\mathbf{F}^*$ and $\mathbf{G}$. However the weight of item j increases and other items are classified by item j as k increases. This is known as forced classification (Nishisato, 1986.)

2. Forced classification in principal component analysis

Forced classification can be extended to principal component analysis easily(Baba,1997.) Let us define a data matrix as

$$\mathbf{X} = [x_{ij}] \qquad (i = 1, \mathrm{K}\ , N;\ j = 1, \mathrm{K}\ , p)$$

where x_{ij} denotes the value of variable j of individual i. Using variable vectors

$$\mathbf{x}_j = [x_1, x, \mathrm{K}\ , x_n]^t$$

let us indicate the data matrix as

$$\mathbf{X} = [\mathbf{x}_1, \mathrm{K}\ , \mathbf{x}_s, \mathrm{K}\ , \mathbf{x}_p].$$

Here let us define a data matrix with parameter k as

$$\mathbf{H} = [\mathbf{x}_1, \mathrm{K}\ , k\mathbf{x}_s, \mathrm{K}\ , \mathbf{x}_p].$$

According to k-value the weight of the s-th variable in **H** changes. Here k is not limited in integers but it is rather regarded as continuous parameter. We can define it as

$$k \geq 0.$$

In principal component analysis based on the covariance matrix made from **H** the solution depends on parameter k. Large k corresponds to forced classification and small k corresponds to sensitivity analysis. We can illustrate the structure of variable space by changing k.

In this presentation an example of application of forced classification to family income and expenditure survey in Japan will be shown.

REFERENCE

Baba,Y. (1996). Application of forced classification to principal component analysis(in Japanese), *The Institute of Statistical Mathematics Cooperative Research Report* **No.100**, 41-49.

Nishisato,S.(1984). Forced classification: A simple application of a quantification technique, *Psychometirika*, **49**,25-36.

Nishisato, S.(1986). Generalized forced classification for quantifying categorical data, *Data Anaalysis and Information* (Eds. By Diday,E. et al), North-Holland, pp351-362.

Nishisato,S. and Baba,Y.(1999). On contingency, projection and forced classification of dual scaling, *Behaviormetrika*, **Vol.26,No.2**, 207-219.

RESUME

Forced classification was developed as discriminant analysis for categorical data. The procedure in the method is simple one that by multiplying a variable by a large constant the variable becomes dominant in quantification and variables are classified into the variables dependent on it and others. This procedure will be extended to principal component analysis. In this presentation an example by applying forced classification to family income and expenditure survey in Japan will be shown.