Estimation of Dimension Based on Certain Information Criterion in Correspondence Analysis

Toru Ogura[1,3], and Yasunori Fujikoshi[2]

[1]Department of Industrial and Systems Engineering, Chuo University,
1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, JAPAN
[2]Department of Mathematics, Hiroshima University,
1-3-1, Kagamiyama, HigashiHiroshima-shi, Hiroshima, 739-8526, JAPAN
[3]Corresponding author: Toru Ogura, e-mail: ogura@indsys.chuo-u.ac.jp

Abstract

This paper deals with the problem of estimating the dimensionality in correspondence analysis for a two-way contingency table. We regard the estimation problem as a model selection problem. Then, using a close relationship between correspondence analysis and canonical correlation analysis, we propose an AIC-type criterion. Through simulation experiments, it is shown that our method works well.

Keywords: dimensionality, information criterion, model selection.

1. Introduction

This paper deals with the problem of estimating the dimensionality in correspondence analysis for a two-way contingency table. Corresponding analysis is considered as a method of presenting the row categories and the column categories of a two-way contingency table as the coordinates of points in a low-dimensional space. The row categories and the column categories are called the row variables and the column variables, respectively. Then, it is well known (see, e.g., Greenacre (1984), Siotani et al. (1985), Izenman (2008)) that there is a close relationship between correspondence analysis and canonical correlation analysis. For the coordinates of points in a low-dimensional space, the first coordinates are defined so that the correlation between the row variables and the column variables is maximum. The second coordinates are defined so that the correlation between the row variables and the column variables, after removing the effects of the first coordinates, is maximum, and so on. Similarly, we call these correlations the first canonical correlation, the second canonical correlation, and so on. The population canonical correlations in correspondence analysis are defined the cell relative frequencies with the cell probabilities.

In general, the dimensionality in correspondence analysis for a two-way contingency table may be defined as the number of nonzero population canonical correlations. The estimation method based on model selection approach is described as follows. Let M_k be the model such that the dimension is k, and let $\{M_0, M_1, \ldots, M_m\}$ be the set of possible candidate models. Then, we apply a model section criterion to the set of candidate models. If a model M_k is selected, then we estimate the dimensionality as k.

Akaike (1973) proposed a criterion for choice of models as follows. If k indexes the model, choose k to minimize

$$AIC = -2\log L(\hat{\theta}^{(k)}) + 2p_k, \tag{1}$$

where $L(\theta)$ is the likelihood function of observations, $\hat{\theta}^{(k)}$ is the maximum likelihood estimate of θ under the model k and p_k is the dimensionality of unknown parameters θ. Fujikoshi and Veitch (1979) obtained an AIC for dimensionality of canonical correlation analysis with two vector variables of p and q $(p \leq q)$ components. The criterion is equivalent to choosing the model M_k to minimize

$$A_k = -n\log\prod_{j=k+1}^{p}(1-r_j^2) - 2(p-k)(q-k), \tag{2}$$

where $A_p = 0$ and r_j is the jth sample canonical correlation coefficient.

In Section 2, we propose an information criterion for estimating the dimensionality in correspondence analysis, by using a close relationship between canonical correlation analysis and correspondence analysis. In order to assess the effectiveness of our method we give on a simulation result in Section 3. Our conclusions are presented in Section 4.

2. Information criterion for estimation of dimensionality

In correspondence analysis for a contingency table with two items A and B, let $A_1, \cdots, A_r$ be the categories of A, and $B_1, \cdots, B_c$ be the categories of B. Let p_{ij} be the population probability of the (i,j)-th cell, and let

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} p_{11} & \cdots & p_{1c} \\ \vdots & \ddots & \vdots \\ p_{r1} & \cdots & p_{rc} \end{pmatrix}. \tag{3}$$

Further, let $\boldsymbol{\Delta}_r$ be the $r \times r$ diagonal matrix with the i-th diagonal element $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$, and $\boldsymbol{\Delta}_c$ be the $c\times c$ diagonal matrix with the j-th diagonal element $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$. Then we may express the latent roots of $\boldsymbol{\Theta} = \boldsymbol{\Delta}_r^{-1/2}\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_c^{-1}\boldsymbol{\mathcal{P}}'\boldsymbol{\Delta}_r^{-1/2}$ as $1 = \rho_0^2 \geq \rho_1^2 \geq \cdots \geq \rho_{m-1}^2$, where $m = \min(r,c)$. The possible non-zero roots $\rho_1^2 \geq \cdots \geq \rho_{m-1}^2$ are called the canonical correlations between two items A and B. Further, the number of non-zero roots is called the dimensionality in correspondence analysis. Let M_k be the model such that thee dimension is k, i.e.,

$$M_k;\ \rho_1^2 \geq \cdots \geq \rho_k^2 > \rho_{k+1}^2 = \cdots = \rho_{m-1}^2 = 0. \tag{4}$$

The model M_k is equivalent to $\mathrm{rank}(\boldsymbol{\Theta}) = k+1$.

Let n_{ij} be the frequency of the (i,j)-th cell, and let $\boldsymbol{N} = (n_{ij})$. A sample version of $\boldsymbol{\mathcal{P}}$ is $\boldsymbol{F} = \frac{1}{n}\boldsymbol{N} = (f_{ij})$, which is an estimator of $\boldsymbol{\mathcal{P}}$, where n is the total observation. Let $\boldsymbol{D}_r$ be the $r \times r$ diagonal matrix with the i-th diagonal element $f_{i\cdot} = \sum_{j=1}^{c} f_{ij}$, and $\boldsymbol{D}_c$ be the $c\times c$ diagonal matrix with the j-th diagonal element $f_{\cdot j} = \sum_{i=1}^{r} f_{ij}$. The sample latent roots $1 = \ell_0^2 \geq \ell_1^2 \geq \cdots \geq \ell_{m-1}^2$ corresponding to $1 = \rho_0^2 \geq \rho_1^2 \geq \cdots \geq \rho_{m-1}^2$ are the latent roots of $\boldsymbol{D}_r^{-1/2}\boldsymbol{F}\boldsymbol{D}_c^{-1}\boldsymbol{F}'\boldsymbol{D}_r^{-1/2}$.

The first term in (1) may be regarded as a measure of badness of fitness for M_k. Some modifications of the term have been considered in Ichikawa and

Konishi (1999), Fujikoshi and Kurata (2008), Fujikoshi et al. (2010), etc. Instead of $-2\log L(\hat{\theta}^{(k)})$ we use a quantity

$$\begin{aligned}&\min_{M_k} n\|\boldsymbol{D}_a^{-1/2}\boldsymbol{F}\boldsymbol{D}_b^{-1/2} - \boldsymbol{\Delta}_a^{-1/2}\boldsymbol{P}\boldsymbol{\Delta}_b^{-1/2}\|^2 \\ &\qquad = n(\ell_{k+1}^2 + \cdots + \ell_{m-1}^2) \\ &\qquad \approx -n\log(1-\ell_{k+1}^2)\cdots(1-\ell_{m-1}^2),\end{aligned} \tag{5}$$

where for a matrix $\boldsymbol{Q}$, $\|\boldsymbol{Q}\|^2 = \mathrm{tr}\boldsymbol{Q}'\boldsymbol{Q}$. The number of independent parameters under M_k is $d_k = (k+1)\{r+c-(k+1)\}-1$. Therefore, as a modification of AIC we propose an information criterion defined by

$$DIC_k = -n\log(1-\ell_{k+1}^2)\cdots(1-\ell_{m-1}^2) + 2d_k,$$

where $k = 0, 1, \ldots, m-1$. The method based on D_k is equivalent to the one based on D_k $(k = 0, 1, \cdots, m-1)$:

$$\begin{aligned} D_k &= DIC_k - DIC_{m-1} \\ &= -n\log(1-\ell_{k+1}^2)\cdots(1-\ell_{m-1}^2) - 2(r-k-1)(c-k-1),\end{aligned} \tag{6}$$

where $D_{m-1} = 0$. The criterion D_k may be regarded as a correspondence analysis version of A_k in (2).

3. Simulation study

In this section, we demonstrate the relative performance of the D_k. In our simulation $r = 5$ and $c = 5$. Let M_k be the model with dimension k defined by (4). We consider the following two cases: (1) The true model is M_k with $k = 1$; and (2) The true model is M_k with $k = 2$. More precisely, the elements of the true probability (covariance) structure are defined as follows:

Case 1: The population probability;

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} 0.168 & 0.008 & 0.008 & 0.008 & 0.008 \\ 0.008 & 0.048 & 0.048 & 0.048 & 0.048 \\ 0.008 & 0.048 & 0.048 & 0.048 & 0.048 \\ 0.008 & 0.048 & 0.048 & 0.048 & 0.048 \\ 0.008 & 0.048 & 0.048 & 0.048 & 0.048 \end{pmatrix}.$$

From this setting, we obtain the population latent roots of correspondence analysis, which are $(\rho_1^2, \rho_2^2, \rho_3^2, \rho_4^2) = (0.8^2, 0.0^2, 0.0^2, 0.0^2)$. We generate $\boldsymbol{N}$ of $n = 100, 150, 200$ from a multinomial distribution with probability $\boldsymbol{\mathcal{P}}$, and calculate each D_k from $\boldsymbol{N}$ in correspondence analysis. The ratios of selected models in 1,000,000 repetitions are provided in Table 1.

Case 2: The population probability;

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} 0.168 & 0.008 & 0.008 & 0.008 & 0.008 \\ 0.008 & 0.138 & 0.018 & 0.018 & 0.018 \\ 0.008 & 0.018 & 0.058 & 0.058 & 0.058 \\ 0.008 & 0.018 & 0.058 & 0.058 & 0.058 \\ 0.008 & 0.018 & 0.058 & 0.058 & 0.058 \end{pmatrix}.$$

From this setting, we obtain the population latent roots of correspondence analysis, which are $(\rho_1^2, \rho_2^2, \rho_3^2, \rho_4^2) = (0.8^2, 0.6^2, 0.0^2, 0.0^2)$. Similarly, we generate $\boldsymbol{N}$

Table 1: The ratios selected by the D_k of case 1 (1,000,000 times)

$n\backslash k$	0	1	2	3	4
100	0.0%	69.3%	26.3%	3.8%	0.6%
150	0.0%	70.5%	25.4%	3.5%	0.5%
200	0.0%	71.1%	25.0%	3.4%	0.5%

of $n = 100, 150, 200$, and calculate each D_k from $\boldsymbol{N}$ in correspondence analysis. The ratios of selected models in 1,000,000 repetitions are provided in Table 2.

Table 2: The ratios selected by the D_k of case 2 (1,000,000 times)

$n\backslash k$	0	1	2	3	4
100	0.0%	0.1%	66.2%	27.9%	5.7%
150	0.0%	0.0%	66.4%	27.9%	5.7%
200	0.0%	0.0%	66.5%	27.8%	5.7%

From the results of simulation studies for two cases, it is seen that the true dimensions are selected with about 70 % successes for case 1 and about 66 % for case 2. Our estimation method will have a tendency of overestimating the dimensionality, but does not underestimate. These properties are not so effected for the sample sizes within the limits 100 $\sim$ 200.

4. Conclusions

We proposed an information criterion for estimating the dimensionality in correspondence analysis, by using a close relationship between correspondence analysis and canonical correlation analysis. In the estimation method, we used $D_k = DIC_k - DIC_{m-1}$, instead of DIC_k. This method is equivalent to that of estimating the dimensionality as k if $\min\{D_0, D_1, \ldots, D_{m-1}\} = D_k$, which corresponds to (2) in canonical correlation analysis. Through simulation experiments, we confirmed that the true dimension is estimated with a rather high probability. Therefore, it is anticipated that our information criterion would work considerably well. In general, if we know the true dimension in correspondence analysis, we can decide which coordinates in a low space are meaningful, and we can avoid an unnecessary interpretation in applications.

References

[1] Akaike, H. (1973) “Information theory and an extension of the maximum likelihood principle”, *2nd International Symposium on Information Theory*, Eds.B.N.Petrov and F.Csáki, 267–281, Budapest: Akadémia Kiado.

[2] Fujikoshi, Y. and Kurata, H. (2008) “Information criterion for some condition independence structures”, In *New Trends in Psychometrics* (Shigemasu,

K., Okada, A. Imaizumi, T. and Hoshino, T. eds.), Universal Academy Press, Tokyo, 69–79.

[3] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010) *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*, Wiley, Hoboken, New Jersey.

[4] Fujikoshi, Y. and Veitch, L. G. (1979) "Estimation of dimensionality in canonical correlation analysis", *Biometrika*, 66, 345–351.

[5] Greenacre, M. J. (1984) *Theory and Applications of Correspondence Analysis*, Academic Press, New York.

[6] Ichikawa, M. and Konishi, S. (1999) "Model evaluation and information criteria in covariance structure analysis", *British J. Math.* & *Statist.l Psych.*, **52**, 285–302.

[7] Izenman, A. J. (2008) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer, New York.

[8] Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985) *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Ohio.