응용통계연구 제 7권 1호
1994년 3월 pp. 69 - 74

INFLUENCE FUNCTIONS IN MULTIPLE CORRESPONDENCE ANALYSIS[1)]

Honggie Kim[2)]

Abstract

Kim (1992) derived influence functions of rows and columns on the eigenvalues obtained in correspondence analysis (CA) of two-way contingency tables. As in principal component analysis, the eigenvalues are of great importance in CA. The goodness of a two dimensional correspondence plot is determined by the ratio of the sum of the two largest eigenvalues to the sum of all the eigenvalues. By investigating those rows and columns with high influence, a correspondence plot may be improved. In this paper, we extend the influence functions of CA to multiple correspondence analysis (MCA), which is a CA of multi-way contingency tables. An explicit formula of the influence function is given.

1. Introduction

Compared to multiple correspondence analysis, simple correspondence analysis primarily consists of techniques for displaying the rows and columns of a two-way contingency table. On the contrary, multiple correspondence analysis is used to investigate the categories of the multi-way contingency tables. These techniques provide flexible and powerful tools for statistical analysis. In many problems two- or three-dimensional displays of the table are highly informative. These graphical displays play an important role in providing insight and understanding of the data. The mathematical origins of correspondence analysis date back to the early 1940's, but it was developed and applied to a great variety of problems by Benzecri and other French statisticians who emphasize its geometric approach (Lebart et al., 1984). Interest in the method has been rekindled among English speaking statisticians by, for instance, Hill (1974), and Greenacre (1981).

Many of the computational techniques and mathematical structures of correspondence analysis are similar to those of principal component analysis. Jolliffe (1986) describes correspondence analysis as weighted principal component analysis. The techniques are briefly outlined in Kim (1992), and more details can be found in Greenacre (1984), and Lebart et al. (1984).

Critchley (1985) studied influence in principal component analysis, and Campbell (1978) obtained some interesting results on influence in discriminant analysis. Kim (1992) derived

---

1) This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1991.

2) Department of Statistics, Choongnam National University, Daejon, Republic of Korea

influence functions of rows in simple correspondence analysis of two-way contingency tables. As in principal component analysis, the eigenvalues are of great importance in correspondence analysis. By using similar techniques, influence in multiple correspondence analysis can be investigated. In this paper, the influence function (IF) of categories on the eigenvalues is obtained along with its sample version, the empirical influence function (EIF).

2. Correspondence Analysis

To briefly outline the theory of correspondence analysis of two-way contingency tables, consider a population consisting of subjects each of which can be classified according to two characteristics: A, with possible outcomes $A_1,..., A_I$, and B, with possible outcomes $B_1,..., B_J$. Let p_{ij} be the proportion of subjects in the population with properties (A_i, B_j). Then the population can be characterized by the probability matrix $P=(p_{ij},\ i=1,...,I,\ j=1,...,J)$, where $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}=1$, and $p_{ij}\geq 0 \quad \forall i,j$.

To briefly illustrate correspondence analysis, let

$$r=P1_{(J\times 1)}=(r_1,...,r_I)^t, \qquad c=P^t1_{(I\times 1)}=(c_1,...,c_J)^t,$$

where $r_i=\sum_{j=1}^{J} p_{ij}$ and $c_j=\sum_{i=1}^{I} p_{ij}$ are the marginal probabilities for the i^{th} row and the j^{th} column, respectively. And let

$$D_r=diag(r), \qquad D_c=diag(c).$$

The purpose of a correspondence analysis is to find a K(usually 2) dimensional subspace which best fits the row profile points in the row space, and to find a K dimensional subspace which best fits the column profile points in the column space. The 'best' fitting subspace of the row profile points is defined to be that for which the total squared distance of the row profile points to the fitted subspace is minimized, with the i^{th} row profile point having a weight r_i, the i^{th} row marginal probability. In the column space, the j^{th} column profile point has a weight c_j, the j^{th} column marginal probability.

The best fitting K dimensional subspace for rows can be obtained by an eigenanalysis of the matrix

$$\Omega_1=(P-r\ c^t)^tD_r^{-1}(P-r\ c^t)D_c^{-1}.$$

A correspondence analysis with $K=2$ is of special interest, since the result can be

plotted. The row and column profile points can be projected onto two separate planes, i.e. two separate plots. By overlaying one onto the other, a plot containing both row and column profiles can be constructed. This combined plot is called a *correspondence* plot, and constructing a correspondence plot is the final goal of a correspondence analysis.

A multi-way contingency table can be obtained from a survey consisting of more than two questions. Let Q denote the number of questions. A single question q consists of p_q of response categories. The total number of response categories, p, contained in the questionaire is

$$p = \sum_{q=1}^{Q} p_q.$$

Let n denote the number of individuals in the survey.

We denote by Z the matrix with n rows and p columns describing the response of the n individuals with binary coding. Regarding this binary matrix Z as a two-way contingency table, simple correspondence analysis can be performed. This procedure is called multiple correspondence analysis. Since n is usually large, the size of Z is large. Hence, we use the p by p square matrix

$$B = Z^t Z$$

called Burt's contingency table associated with Z. It can be shown that the correspondence analyses of Z and B are equivalent (Lebart et al (1984)).

3. Influence Functions

As in principal component analysis, the goodness-of-fit of a correspondence plot is measured by the ratio of the sum of the two largest eigenvalues to the sum of all the eigenvalues. That is,

$$\frac{\lambda_1+\lambda_2}{\sum_{k=1}^{K^*}\lambda_k},$$

where K^* is the rank of Ω_1. Hence, any row or column which has a high influence on the λ's will also have a high influence on the correspondence plot.

Critchley (1985) studied influence of an observaion vector on the eigenvalues in principal component analysis, and Campbell (1978) obtained some interesting results on influence in discriminant analysis. Kim (1992) derived influence functions of row on the eigenvalues in simple correspondence analysis. The reason why we are more interested in the influence of rows rather than cells is as follows. To improve a correspondence plot, we are more concentrating on deleting influential rows. Deleting a cell (making a cell empty in the contingency table) will keep the row (with empty cell) in the correspondence plot,

which misrepresents the original frequencies of the row. When we delete a row, we can regard it as not being in the original caregory set, leaving other rows to represent their original frequencies. By extending these results, influence in multiple correspondence analysis can be investigated.

As mentioned in section 2, multiple correspondence analysis is an ordinary correspondence analysis of the n by p binary matrix Z. Hence, we start with the estimated probability matrix

$$\frac{Z}{nQ}$$

instead of the probability matrix P in correspondence analysis in section 2. Note that nQ is the total of the entries in Z since there are n rows in Z and each row of Z contains exactly Q many 1's.

Now correspondence analysis of Z is performed through an *eigen-analysis* of

$$\Omega_2 = A^t \left(\frac{I}{n}\right)^{-1} A \left[diag(c_1,...,c_p)\right]^{-1}$$

where

$$A = \frac{Z}{nQ} - \frac{1}{n} 1 (c_1,...,c_p)$$

and

$$1 = n \text{ dimensioal vector of } 1' \text{ s.}$$

Note that c_j is the j^{th} column sum of Z divided by nQ. Ω_2 can be simplified as

$$\Omega_2 = \frac{1}{nQ^2} (Z - Q1(c_1,...,c_p))^t (Z - Q1(c_1,...,c_p)) [diag(c_1,...,c_p)]^{-1}.$$

If we write

$$\Omega_2 = R \left[diag(c_1,...,c_p)\right]^{-1},$$

then it can be shown that

$$\Omega_3 = (diag(1/\sqrt{c_1},...,1/\sqrt{c_p})) R (diag(1/\sqrt{c_1},...,1/\sqrt{c_p}))$$

is a symmetric matrix having exactly the same set of eigenvalues as Ω_1. By applying the results of Kim (1992), the empirical influence function of i^{th} row of Z on the k^{th} largest eigenvalue is given by

$$EIF(\lambda_k, f_i) = Q\left(\widehat{z^2_{ki}}\, \hat{\lambda}_k \sum_{j=1}^{p} \frac{f_{ij}}{Q} \frac{\widehat{u^2_{kj}}}{c_j^2}\right),$$

where $\widehat{z_{ki}}$ is the coordinate of the i^{th} row projected on the k^{th} principal axis, $\widehat{u_{kj}}$ is the j^{th} component of the eigenvector corresponding to the k^{th} largest eigenvalue,

and f_{ij} is the ij^{th} element of matrix Z.

The empirical influence function given above measures the influence of rows, that is, individuals, on the eigenvalues. We are more intersted in the influence of the columns, which represent the response categories. By exchanging the roles of rows and columns of Z, the empirical influence function of j^{th} column of Z on the k^{th} largest eigenvalue is given by

$$EIF(\lambda_k, h_j) = f_{+j}(\widehat{z^2_{kj}} - n^2 \widehat{\lambda_k} \sum_{i=1}^{n} \frac{f_{ij}}{f_{+j}} \widehat{u^2_{ki}}),$$

where $\widehat{z_{kj}}$ is the coordinate of the j^{th} column projected on the k^{th} principal axis, $\widehat{u_{ki}}$ is the i^{th} component of the eigenvector corresponding to the k^{th} largest eigenvalue, and f_{+j} is the j^{th} column sum of matrix Z.

References

[1] Campbell, N. A. (1978), The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics,* Vol. 27, 251-258.

[2] Critchley, F. (1985), Influence in principal components analysis, *Biometrika,* Vol. 72, 627-636.

[3] Greenacre, M. J. (1981), Practical Correspondence Analysis, in : *Interpreting Multivariate Data,* Wiley, New York.

[4] ---- (1984), *Theory and Application of Correspondence Analysis,* Academic Press, London.

[5] Hill, M. O. (1974), Correspondence analysis: a neglected multivariate method, *Applied Statistics,* Vol. 23, 340-354.

[6] Jolliffe, I. T. (1986), *Principal Component Analysis,* Springer-Verlag, New York.

[7] Kim, H. (1992), Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation,* Vol. 40, 201-217.

[8] Lebart, L., Morineau, A., and Warwick, K. (1984), *Multivariate Descriptive Statistical Analysis,* Wiley, New York.

다중 대응 분석에서의 영향 함수[3]

김홍기[4]

요약

Kim(1992)은 이차원 분할표의 단순 대응 분석에서의 영향 함수를 유도하였다. 주성분 분석에서와 마찬가지로 특정 행렬의 고유치가 대응 분석에서도 중요한 역할을 한다. 이차원 대응 분석 그림의 정확도는 가장 큰 두개의 고유치 합의 전체 고유치 합에 대한 비율로 주어지게 된다. 고유치에 미치는 영향이 큰 행이나 열을 조사함으로써 대응 분석이 개선될 수 있다. 본 논문에서는 단순 대응 분석에서의 영향 함수를 다중 대응 분석으로 확장하였다.

3) 본 연구는 1991년도 학술진흥재단 자유공모과제 연구비에 의해서 수행되었음.
4) (305-764) 대전시 유성구 궁동 220 충남대학교 자연과학대학 통계학과 조교수