Journal of Multivariate Analysis 136 (2015) 175–189

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Diagnostics in a simple correspondence analysis model: An approach based on Cook's distance for log-linear models

Nirian Martín

Department of Statistics and Flores de Lemus Institute, Carlos III University of Madrid, 28903 Getafe (Madrid), Spain

ARTICLE INFO

Article history:
Received 17 December 2013
Available online 20 January 2015

AMS subject classifications:
62J20
62H17

Keywords:
Two-way contingency table
Multinomial sampling
Correspondence model
Log-linear model
Pearson's Chi-square residuals
Cook's distance

ABSTRACT

Diagnostics have not received much attention in the literature of simple correspondence analysis models. Since Cook's distance was defined to identify influential observations of the linear regression model, it has been extended to different models, in particular to log-linear models. In this paper we provide the asymptotic distribution of Cook's distance of any kind of log-linear models and also a method for diagnostics, based on it. By using Goodman's $RC(K)$ model as a log-linear model to approximate the ordinary simple correspondence analysis procedure, we follow a Cook's distance approach to identify influential cells and three examples illustrate the performance of this method.

1. From the ordinary simple correspondence analysis to Goodman's RC model: a log-linear approach

The original development of the simple correspondence analysis approach, popular from the publication of Benzécri [5], aims to describe association patterns between two categorical variables as well as using graphical procedures for that, once some departures from the model of independence are identified. A variety of extensions have been subsequently considered with respect to correspondence models, introduced by Goodman [11–13] for the first time, to make statistical inference when it is assumed that the data are generated under a specific sampling design. More thoroughly, when reducing the dimensions of the original correspondence analysis model, a new nested model is checked through a goodness of fit test in order to avoid losing representation of row and column profiles of the contingency table. In the following lines we shall describe the basic simple correspondence analysis model, but more details can be found in [3, Section 4]. It is worth mentioning that these models may also be useful for descriptive purposes and graphical displays.

Let X and Y be two categorical variables, with I and J categories respectively and n individuals cross classified in the two way contingency table with multinomial sampling. Without loss of generality, each possible outcome of (X, Y) may be identified by $(i, j) \in \{1, \ldots, I\} \times \{1, \ldots, J\}$. The probability of a specific outcome, denoted by $p_{ij} = \Pr(X = i, Y = j)$, is unknown and depends on $\boldsymbol{\theta}$ through a model $(p_{ij} = p_{ij}(\boldsymbol{\theta}))$ and it is assumed to be non null $(p_{ij} > 0)$. The contingency table of unknown probabilities is an $I \times J$ matrix $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{\theta}) = (p_{ij}(\boldsymbol{\theta}))_{i=1,\ldots,I;j=1,\ldots,J}$ which can be expressed in terms of the transposed row vectors as $\boldsymbol{P}^T(\boldsymbol{\theta}) = (\boldsymbol{p}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{p}_I(\boldsymbol{\theta}))$ and stacking these vectors we can express the whole contingency in a unique column vector through the vec operator (see [19, p. 343]) as

$$\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}) = \mathrm{vec}(\boldsymbol{P}^T(\boldsymbol{\theta})) = (p_{11}(\boldsymbol{\theta}), \ldots, p_{1J}(\boldsymbol{\theta}), \ldots, p_{I1}(\boldsymbol{\theta}), \ldots, p_{IJ}(\boldsymbol{\theta}))^T.$$

E-mail address: nirian.martin@uc3m.es.

http://dx.doi.org/10.1016/j.jmva.2015.01.008
0047-259X/

By following the same scheme of the previous contingency table, the observed frequencies may be denoted by a unique column vector, $\boldsymbol{n} = (n_{11}, \ldots, n_{1J}, \ldots, n_{I1}, \ldots, n_{IJ})^T$, and also its generator, $\boldsymbol{N} = (N_{11}, \ldots, N_{1J}, \ldots, N_{I1}, \ldots, N_{IJ})^T$, which is a multinomial IJ-dimensional random vector with parameters n and $\boldsymbol{p}(\boldsymbol{\theta})$, $\mathcal{M}(n, \boldsymbol{p}(\boldsymbol{\theta}))$.

The "ordinary or basic" simple correspondence analysis (CA) model establishes that

$$\begin{aligned} p_{ij} &= p_{i\bullet}p_{\bullet j}\left(1 + \sum_{k=1}^{K} \lambda_k u_{ki} v_{kj}\right) \\ &= p_{i\bullet}p_{\bullet j}\left(1 + \boldsymbol{u}_i^T \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{v}_j\right), \end{aligned} \tag{1}$$

where $p_{i\bullet} = \Pr(X = i) = \sum_{j=1}^{J} p_{ij}$, $p_{\bullet j} = \Pr(Y = j) = \sum_{i=1}^{I} p_{ij}$, $K = \min\{I - 1, J - 1\}$, u_{ik} is the k-th canonical score of category i of variable X, v_{jk} is the k-th canonical score of category j of variable Y, λ_k is the k-th canonical correlation such that $1 \geq \lambda_1 \geq \cdots \geq \lambda_K \geq 0$,

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}(K) = (\lambda_1, \ldots, \lambda_K)^T, \qquad \boldsymbol{u}_i = \boldsymbol{u}_i(K) = (u_{1i}, \ldots, u_{Ki})^T, \qquad \boldsymbol{v}_j = \boldsymbol{v}_j(K) = (v_{1j}, \ldots, v_{Kj})^T,$$

and $\boldsymbol{D}_*$ is the diagonal matrix of $*$. The value of K represents the dimension of the "ordinary or basic" simple correspondence analysis model, the values of the $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ vectors, the ith row and jth column canonical scores respectively, while the values of the matrices $\boldsymbol{U} = \boldsymbol{U}(K) = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_I)_{K\times I}$ and $\boldsymbol{V} = \boldsymbol{V}(K) = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J)_{K\times J}$, all the row and column canonical scores respectively and the values of the $\boldsymbol{\lambda}$ vector, all the canonical correlations. If the summation term $\sum_{k=1}^{K} \lambda_k u_{ki} v_{kj}$ is null, (1) is equivalent to the independence model, otherwise the value of $\sum_{k=1}^{K} \lambda_k u_{ki} v_{kj}$ indicates the deviation from independence of cell (i, j). For identification purposes, since (1) is overparameterized, several constraints have to be considered on parameters. One way to establish these constraints is based on maximizing the correlation between the k-th row of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, by following successively the order $k = 1, \ldots, K$, that is

$$\sum_{i=1}^{I}\sum_{j=1}^{J} u_{ki} v_{kj} p_{ij} = \lambda_k, \quad k = 1, \ldots, K \tag{2}$$

is maximized given that $\boldsymbol{U}$ and $\boldsymbol{V}$ are centered and standardized,

$$\sum_{i=1}^{I} u_{ki} p_{i\bullet} = \sum_{j=1}^{J} v_{kj} p_{\bullet j} = 0, \qquad \sum_{i=1}^{I} u_{ki}^2 p_{i\bullet} = \sum_{j=1}^{J} v_{kj}^2 p_{\bullet j} = 1, \quad k = 1, \ldots, K.$$

Maximizing (2) subject to the previous $4K$ restrictions, $\boldsymbol{U}$ and $\boldsymbol{V}$ are obtained, which are uncorrelated in different dimensions, i.e.

$$\sum_{i=1}^{I} u_{ki} u_{k'i} p_{i\bullet} = \sum_{j=1}^{J} v_{kj} v_{k'j} p_{\bullet j} = 0, \quad k, k' \in \{1, \ldots, K\},\ k \neq k'.$$

The previous expressions in matrix notation are given by

$$\boldsymbol{U}\boldsymbol{p}_X = \boldsymbol{V}\boldsymbol{p}_Y = \boldsymbol{0}_K, \qquad \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{p}_X}\boldsymbol{U}^T = \boldsymbol{V}\boldsymbol{D}_{\boldsymbol{p}_Y}\boldsymbol{V}^T = \boldsymbol{I}_K, \tag{3}$$

where $\boldsymbol{p}_X = \boldsymbol{P}\boldsymbol{1}_J = (p_{1\bullet}, \ldots, p_{I\bullet})^T$, $\boldsymbol{p}_Y = \boldsymbol{P}^T\boldsymbol{1}_I = (p_{\bullet 1}, \ldots, p_{\bullet J})^T$ are marginal probability vectors ($\boldsymbol{1}_*$ is the $*$-th dimensional vector of 1's). Note that $\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{p}_X}^{\frac{1}{2}}$ and $\boldsymbol{V}\boldsymbol{D}_{\boldsymbol{p}_Y}^{\frac{1}{2}}$ are orthogonal matrices. Hence, performing the singular value decomposition of the Pearson chi-square residuals for the independence model

$$\frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}}} = \sqrt{p_{i\bullet}p_{\bullet j}} \sum_{k=1}^{K} \lambda_k u_{ki} v_{kj}, \qquad \boldsymbol{D}_{\boldsymbol{p}_X}^{-\frac{1}{2}}\left(\boldsymbol{P} - \boldsymbol{p}_X\boldsymbol{p}_Y^T\right)\boldsymbol{D}_{\boldsymbol{p}_Y}^{-\frac{1}{2}} = \boldsymbol{D}_{\boldsymbol{p}_X}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{D}_{\boldsymbol{\lambda}}\boldsymbol{V}\boldsymbol{D}_{\boldsymbol{p}_Y}^{\frac{1}{2}},$$

we obtain the values of $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\lambda}$. This idea serves as linking between the correspondence analysis and the correlation analysis. It is worth noting that (1) is a saturated model, that is the adjusted probabilities coincide with the relative frequencies and so in practice the singular value decomposition is

$$\boldsymbol{D}_{\widehat{\boldsymbol{p}}_X}^{-\frac{1}{2}}\left(\widehat{\boldsymbol{P}} - \widehat{\boldsymbol{p}}_X\widehat{\boldsymbol{p}}_Y^T\right)\boldsymbol{D}_{\widehat{\boldsymbol{p}}_Y}^{-\frac{1}{2}} = \boldsymbol{D}_{\widehat{\boldsymbol{p}}_X}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{D}_{\widehat{\boldsymbol{\lambda}}}\boldsymbol{V}\boldsymbol{D}_{\widehat{\boldsymbol{p}}_Y}^{\frac{1}{2}}. \tag{4}$$

Taking into account that the chi-square statistic for testing independence is defined as n times the sum of the square of the left hand side of (4), i.e., $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{\widehat{p}_{i\bullet}\widehat{p}_{\bullet j}}(\widehat{p}_{ij} - \widehat{p}_{i\bullet}\widehat{p}_{\bullet j})^2$, the chi-square statistic for independence might be also expressed in terms of the square of the Euclidean norm of the vector of canonical correlations as

$$X^2 = n\left\|\widehat{\boldsymbol{\lambda}}\right\|^2 = n\sum_{k=1}^{K}\widehat{\lambda}_k^2. \tag{5}$$