A unified approach for representing rows and columns in contingency tables

C.M. Cuadras[1], **D. Cuadras**[2]
[1]Universitat de Barcelona, Spain; *ccuadras@ub.edu*
[2]Servei d'Estadistica, UAB, Barcelona, Spain

Abstract

By using suitable parameters, we present a unified aproach for describing four methods for representing categorical data in a contingency table. These methods include: correspondence analysis (CA), the alternative approach using Hellinger distance (HD), the log-ratio (LR) alternative, which is appropriate for compositional data, and the so-called non-symmetrical correspondence analysis (NSCA). We then make an appropriate comparison among these four methods and some illustrative examples are given. Some approaches based on cumulative frequencies are also linked and studied using matrices.

Key words: Correspondence analysis, Hellinger distance, Non-symmetrical correspondence analysis, log-ratio analysis, Taguchi inertia.

1 Introduction

A common problem in multivariate analysis is the excesive number of methods available for doing practically the same. Often, several methods can be unified and made more flexible by using some parameters that linked them.

There are several methods for the graphical display of the rows (and columns) of a contingency table N. These methods aim to reinterpret N using a joint graphic for rows and columns. The representation depends on the distance used.

Correspondence analysis (CA) uses the so-called chi-square distance between the profiles of rows (and columns). This method is based in the decomposition of Pearson's contingency coefficent and is described in Greenacre (1984). As an alternative, Rao (1985) used canonical coordinates to represent the rows of a contingency table N, using the Hellinger distance (HD) between the row profiles.

For a contingency table, Goodman (1986) introduces the RC(M) association model. This model takes log-ratios and fits parameters by ML. Tsujitani (1982) uses the same model but assuming first independence, then fits a multiplicative model. Aitchison and Greenacre (2002) adapt biplot methodology by taking log-ratios (LR) to represent compositional data, which can be used for the same purpose.

Lauro and D'Ambra (1984) introduced non-symmetrical correspondence analysis (NSCA), a slight modification of CA, which is based on the decomposition of Goodman-Kruskal's tau.

In some way, all models can be considered as a particular case of the so-called generalized nonindependence analysis (GNA) introduced by Goodman (1993).

This paper presents a general approach for several methods of visulization of contingency tables by using a suitable parametrization.It extends the results by Cuadras et al. (2006), Cuadras and Cuadras (2006) and Greenacre (2007).

2 General parametrization

Let $N = (n_{ij})$ be an $I \times J$ contingency table and $P = n^{-1}N$ the correspondence matrix, where $n = \sum_{ij} n_{ij}$. Let $K = \min\{I, J\}$ and $r = P1, D_r$ =diag$(r), c = P'1, D_c$ =diag(c), the vectors and diagonal matrices with the marginal frequencies of P, where 1 is the column vector of ones of appropriate dimension.

In order to represent the rows and columns of N the so-called generalized nonindependence analysis (GNA), introduced by Goodman (1993), can be described as the SVD

$$D_r^{1/2}(I - 1r')(R[D_r^{-1}PD_c^{-1}])D_c^{1/2} = U\Lambda V'$$

where $R(x)$, with $x > 0$, is any monotonically increasing function. The coordinates for rows and columns are given by

$$A = D_r^{-1/2}U\Lambda, \qquad B = D_c^{-1/2}V\Lambda.$$

GNA reduces to CA when $R(x) = 1$.

A suitable choice of $R(x)$ is the Box-Cox transformation, i.e., the function

$$\begin{aligned} R(x) &= (x^\alpha - 1)/\alpha \quad \text{if} \quad x > 0, \\ &= \ln(x) \qquad\quad\ \text{if} \quad \alpha = 0. \end{aligned}$$

With this function, we consider the following SVD depending on three parameters:

$$D_r^{1/2}(I - \gamma 1r')(\frac{1}{\alpha}[D_r^{-1}PD_c^{-1}]^\alpha - 11'])D_c^\beta = U\Lambda V'.$$

Table 1: Four methods of representing rows and columns in a contingency table.

Method	Uncentered		Centered	
	$\gamma = 0$		$\gamma = 1$	
	α	β	α	β
CA	1	1/2	1	1/2
HD	1/2	1/2	1/2	1/2
LR	0	1/2	0	1/2
NSCA	1	1	1	1

Then the principal coordinates for the I rows and the standard coordinates for the J columns of N are given in A and B_0, respectively, where:

$$A = D_r^{-1/2} U \Lambda, \qquad B_0 = D_c^{\beta - 1} V \Lambda.$$

Implicit with this (row) representation is the distance between rows

$$\delta_{ii'}^2 = \sum_{j=1}^{J} [(\frac{p_{ij}}{r_i c_j})^\alpha - (\frac{p_{i'j}}{r_{i'} c_j})^\alpha]^2 c_j^{2\beta}$$

The first principal coordinates account for a relative high percentege of inertia. These four methods can be summarized in Table 1.

There are two overall measures of inertia, or dispersion between rows and columns.

2.1 Generalized Pearson contingency coefficient.

This parametric measure is given by:

$$\phi^2(\alpha, \beta) = \sum_{i=1}^{I} \sum_{j=1}^{J} [(\frac{p_{ij}}{r_i c_j})^\alpha - 1]^2 r_i c_i^{2\beta}.$$

2.2 Geometric variability

This is an average measure of the differences between rows:

$$\mathfrak{v} = \frac{1}{2} r' \Delta^{(2)} r,$$

where $\Delta^{(2)} = (\delta_{ii'}^2)$ is the $I \times I$ matrix of squared parametric distances. This measure was considered by Light and Margolin (1971) in categorical data analysis and by Cuadras et al. (1997) for discriminant analysis.

Note that $\mathfrak{v} = \phi^2(\alpha, \beta) = 0$ under statistical independence between rows and columns. Iin general $\mathfrak{v} \neq \phi^2(\alpha, \beta)$.

Table 2 summarizes these inertias.

3 Testing independence

The test of independence between rows and columns can be performed with

$$\phi^2(\alpha, 1) = \sum_{i=1}^{I} \sum_{j=1}^{J} ((\frac{p_{ij}}{r_i c_j})^\alpha - 1)^2 r_i c_i,$$

Table 2: Inertia expressions for four methods of representing rows in contingency tables. In CA and NSCA the geometric variability coincides with the contingency coefficient.

Method	Inertia (centered) $\mathfrak{v} = \sum \lambda_i^2$	Inertia (uncentered) $\phi^2 = \sum \lambda_i^2$
CA Pearson	$\mathfrak{v} = \sum_{i,j}(\frac{p_{ij}}{r_i c_j} - 1)^2 r_i c_j$	$\phi^2(1, 1/2) = \mathfrak{v}$
HD Matusita-Rao	$\mathfrak{v} = 1 - \sum_j (\sum_i \sqrt{p_{ij} r_i})^2$	$\phi^2(1/2, 1/2)$ $= 2(1 - \sum_{i,j} \sqrt{p_{ij} r_i c_j})$
LR Aitkinson	$\mathfrak{v} = \sum_{i,j} c_j r_i (\ln(p_{ij}/r_i))^2$ $- \sum_j c_j (\sum_{i=1}^I r_i \ln(p_{ij}/r_i))^2$	$\phi^2(0, 1/2)$ $= \sum_{i,j}(\ln \frac{p_{ij}}{r_i c_j})^2 r_i c_j.$
NSCA Goodma-Kruskal	$\mathfrak{v} = \sum_{i,j}(\frac{p_{ij}}{r_i} - c_j)^2 r_i$	$\phi^2(1, 1) = \mathfrak{v}$

Suppose $\alpha > 0$ a fix value. Under independence we have

$$\frac{n^2}{\alpha^2}\phi^2(\alpha, 1) \to \chi^2_{(I-1)(J-1)}$$

as $n \to \infty$, where the convergence is in law.

To prove this asymptotic result, note that

$$(\frac{p_{ij}}{r_i c_j} - 1)^2 r_i c_i = \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

and

$$\frac{(r_i^{-\alpha} p_{ij}^{\alpha} c_j^{-\alpha} - r_i c_i)^2}{r_i c_i} = \frac{(r_i^{-\alpha} p_{ij}^{\alpha} c_j^{-\alpha} - 1)^2}{1/r_i c_i} = (x^\alpha - 1)^2 r_i c_i$$

with $x = p_{ij}/r_i c_j$. Then

$$(x^\alpha - 1)^2 r_i c_j = (\frac{x^\alpha - 1}{x - 1})^2 (x - 1)^2 r_i c_j$$

But $\lim_{x \to 1}[(x^\alpha - 1)/(x - 1)]^2 = \alpha^2$. Hence, under independence, $x \to 1$ as $n \to \infty$. Thus

$$\begin{aligned} \lim n^2 \sum_{i=1}^{I}\sum_{j=1}^{J}((\frac{p_{ij}}{r_i c_j})^\alpha - 1)^2 r_i c_i &= \alpha^2 \lim n^2 \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r c_j} \\ &= \alpha^2 \chi^2 \end{aligned}$$

When $\alpha \to 0$ then

$$\lim_{x \to 1, \alpha \to 0} \frac{1}{\alpha^2}(\frac{x^\alpha - 1}{x - 1})^2 = 1$$

and the asymptotic limit reduces to

$$n^2 \phi^2(0, 1) \to \chi^2_{(I-1)(J-1)}.$$

4 Analysis based on Taguchi inertia

Let $N = (n_{ij})$ the $I \times J$ contingency table, $n_{i\cdot}$ and $n_{\cdot j}$ the row and column marginals. Given a row i let us consider the cumulative frequencies

$$z_{i1} = n_{i1}, z_{i2} = n_{i1} + n_{i2}, \ldots, z_{iJ} = n_{i1} + \cdots + n_{iJ}$$

and cumulative column proportions

$$d_1 = \frac{n_{\cdot 1}}{n}, d_2 = \frac{n_{\cdot 1} + n_{\cdot 2}}{n}, \ldots, d_J = \frac{n_{\cdot 1} + \cdots + n_{\cdot J}}{n}$$

the so-called Taguchi statistic (Taguchi, 1974), is given by

$$T = \sum_{j=1}^{J-1} w_j (\sum_{i=1}^{I} n_{i\cdot} (\frac{z_{ik}}{n_{i\cdot}} - d_j)^2$$

where $w_1, \ldots, w_{J-1}$ are weights. Two choices are possible: $w_j = (d_j(1-d_j))^{-1}$ and $w_j = 1/J$. With the $(J-1) \times J$ matrix

$$A = \begin{pmatrix} 1-d_1 & -d_1 & \cdots & -d_1 & -d_1 \\ 1-d_2 & -d_2 & \cdots & -d_2 & -d_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-d_{J-1} & -d_{J-1} & \cdots & -d_{J-1} & -d_{J-1} \end{pmatrix}$$

then T can be expressed as

$$T = n\mathrm{trace}(D_r^{-1/2} N A' W A N' D_r^{-1/2}).$$

Using the $J \times J$ triangular matrix

$$M = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

and $d = (d_1, d_2, \cdots)$ then

$$d = Mc, \;\; Z = NM', \;\; A = M - d1', \;\; NA' = NM' - N1d' = NM' - nrC'.$$

Thus if $d = Mc$, Taguchi's T depends on $PM' - rd' = (P - rc')M'$, i. e., the differences between the cumulative frequencies of rows assuming P and independence rc', respectively

$$\begin{aligned} T &= n\mathrm{trace}(D_r^{-1/2}(NM' - nrd')W(NM' - nrd')'D_r^{-1/2} \\ &= n\mathrm{trace}(D_r^{-1/2}(nPM' - nrd')W(nPM' - nrd')'D_r^{-1/2} \\ &= n^3\mathrm{trace}(D_r^{-1/2}(P - rc')M'WM(P - rc')'D_r^{-1/2}. \end{aligned}$$

Now, taking into account the SVD in CA, with inertia

$$\mathrm{trace}(D_r^{-1/2}(P - rc')D_c^{-1}(P - rc')'D_r^{-1/2}),$$

Beh et al. (2007) consider the following SVD

$$D_r^{-1/2}(P - rc')M'W^{1/2} = U\Lambda V',$$

which is equivalent towhich is equivalent to

$$D_r^{1/2}(D_r^{-1}PM' - 1c'M')W^{1/2} = U\Lambda V'.$$

Then the principal and standard coordinates are

$$A = D_r^{-1/2}U\Lambda, \quad B = W^{-1/2}V\Lambda.$$

Table 3: Unified cumulative correspondence analysis.

SVD	$D_r^{-1/2}L(P-rc')M'W^{1/2} = U\Lambda V'$
DA	L and M triangular, W weight
CA	$L = I$, $M = I$, $W = D_c^{-1}$
TA	$L = I$, M triang, W weight

The initial matrix of coordinates and the distance between rows are:

$$Q = D_r^{-1}(P - 1c')M'W^{1/2}, \quad \delta^2_{ii'} = \sum_{j=1}^{J} w_j(\frac{P_{ij}}{r_i} - \frac{P_{i'j}}{r_{i'}})^2,$$

where $P_{ij} = p_{i1}+\cdots+p_{ij}$ is the cumulative sum for row i. Matrix Q is centered. The decomposition of inertia is

$$\begin{aligned}\mathrm{trace}(D_r^{-1/2}(P-rc')M'WM(P-rc')'D_r^{-1/2} &= \textstyle\sum_{i,j=1}^{I,J} w_j(P_{ij}-r_iC_j)^2/r_i \\ &= \textstyle\sum_{i=1}^{K}\lambda_i^2,\end{aligned}$$

Correspondence analysis can be approached by using cumulative frequencies for rows and columns, Cuadras (2002).In this way, a more general approach based on double acumulative (DA) frequencies is

$$D_r^{-1/2}L(P-rc')M'W^{1/2} = D_r^{-1/2}(H-RC')W^{1/2} = U\Lambda V',$$

where L is triangular with ones. Then $H = LPM', R = Lr, C = Mc$ are the cumulative frequencies. Therefore DA, CA and Taguchi analysis can also be unified, see Table 3.

Acknowledgements

Work supported in part by grant MEC MTM-2004-00440.

References

Aitchison, J. and Greenacre, M. J. (2002). Biplots of compositional data. *Applied Statistics 51*, pp. 375–392.

Beh, E., D'Ambra, L., and Simonetti, B. (2007). Ordinal correspondence analysis based on cumulative chi-squared test. *Correspondence Analysis and Related Methods*, Rotterdam: CARME 2007.

Cuadras, C. M. (2002). Correspondence analysis and diagonal expansions in terms of distribution functions. *J. of Statistical Planning and Inference 103*, pp. 137–150.

Cuadras. C. M., Cuadras, D. (2006). A parametric approach to correspondence analysis. *Linear Algebra and its Applications 417*, pp. 64–74.

Cuadras, C. M., Fortiana, J. and Oliva, F. (1997). The proximity of an individual to a population with applications in discriminant analysis. *J. of Classification 14*, pp. 117–136.

Cuadras, C. M., Cuadras, D., Greenacre, M. (2006). A comparison of different methods for representing categorical data. *Communications in Statistics-Simul. and Comp. 35*(2), pp. 447–459.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review 54*, pp. 243–309.

Goodman, L. A. (1993). Correspondence analysis, association analysis, and generalized nonindependence analysis of contingency tables: Saturated and unsaturated models, and appropriate graphical displays. In C. M. Cuadras and C. R. Rao. (Eds.), *Multivariate Analysis: Future Directions 2*, Amsterdam: Elsevier.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press.

Greenacre, M. J. (2007). Power transformations in correspondence analysis. *Presented at Carme2007, Rotterdam, 2007.*

Lauro, N. and D'Ambra, L. (1984). L'analyse non symetrique des correspondances. In E. Diday et et al. (Eds.), *Data analysis and informatics III*, North Holland, Amsterdam, pp. 433–446. *rzneim. Forsch. (Drug Res.)* 26, pp. 1295–1300.

Light, R. J., Margolin, B. H. (1971). An analysis of variance for categoricala data. *J. of the American Statistical Association 66*, pp. 534–544.

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió 19*, pp. 23–63.

Taguchi, G. (1974). A new statistical analysis for clinical data, the accumulating analysis in contrast with the chi-square test. *Saishin Igaku (The New Medicine) 20*, pp. 806–813.