Comparing correspondence analysis and the log-ratio alternative for representing categorical data

C. M. Cuadras, D. Cuadras and M. J. Greenacre
Departament d'Estadística, Universitat de Barcelona.
Dep. d'Economia i Empresa, Universitat Pompeu Fabra.
Barcelona, Spain

Abstract

We compare correspondence analysis to the logratio approach based on compositional data. We also compare correspondence analysis and an alternative approach using Hellinger distance, for representing categorical data in a contingency table. We propose a coefficient which globally measures the similarity between these approaches. This coefficient can be decomposed into several components, one component for each principal dimension, indicating the contribution of the dimensions to the difference between the two representations. These three methods of representation can produce quite similar results. One illustrative example is given.

Key Words: Correspondence analysis; chi-square distance; Hellinger distance; compositional data; multidimensional scaling.

1 Introduction

Correspondence analysis (CA) is a multivariate method to visualize categorical data, typically presented as a two-way contingency table N. The distance used in the graphical display of the rows (and columns) of N is the so-called chi-square distance between the profiles of rows (and between the profiles

of columns). This method is described in Benzécri (1973) and Greenacre (1984).

In an early paper, Rao (1948) introduced the concept of canonical coordinates, also for graphical representation of multivariate data, especially quantitative multivariate data in several populations More recently, Rao (1995) also used canonical coordinates to represent the rows of a contingency table N, using the Hellinger distance decomposition (HD) between the profiles of rows. A third alternative for representing categorical data is based on compositional data and a distance between rows or columns, where the data are transformed to log-ratios (Aitchison, 1986).

The aim of this contributed paper is to compare these three approaches, supplying in particular a global measure of difference between CA, HD and LR (log-ratio) and decomposing this measure along the principal dimensions. One illustrative example is given.

2 Correspondence analysis

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 $r = P1$, D_r =diag(r), $c = P'1$, D_c =diag(c), the vectors and diagonal matrices with the marginal frequencies of P.

CA uses the SVD

$$D_r^{-1/2}(P - rc')D_c^{-1/2} = UD_\lambda V', \tag{1}$$

where D_λ is the diagonal matrix of singular values in descending order, U is an orthogonal matrix and the columns of V are orthonormal. To represent the I rows of N we may take as principal coordinates the rows of

$$A = D_r^{-1/2}UD_\lambda. \tag{2}$$

Then the Euclidean distance between rows i, i' of A equals the chi-square distance

$$\delta_{ii'}^2 = \sum_{j=1}^{J}(p_{ij}/r_i - p_{i'j}/r_{i'})^2/c_j. \tag{3}$$

Similarly, to represent the J columns of N we may use the principal coordinates contained in the rows of B or the standard coordinates B_0, where

$$B = D_c^{-1/2}VD_\lambda, \quad B_0 = D_c^{-1/2}V. \tag{4}$$

An advantage of CA is the possibility of a joint representation of rows and columns, called the symmetric representation, as a consequence of the transitive relations

$$A = D_r^{-1}PBD_\lambda^{-1}, \quad B = D_c^{-1}P'AD_\lambda^{-1}. \tag{5}$$

CA can also be approached by using the SVD

$$D_r^{-1/2}(P - rc')D_c^{-1/2} = UD_\lambda V', \tag{6}$$

and, from $(I - 1r')D_r^{-1}PD_c^{-1}(I - c1') = D_r^{-1}PD_c^{-1} - 11'$, a third approach uses the equivalent SVD

$$D_r^{1/2}(I - 1r')D_r^{-1}PD_c^{-1}(I - c1')D_c^{1/2} = UD_\lambda V'. \tag{7}$$

3 The Hellinger distance alternative

Following Rao (1995), HD is described as the SVD

$$D_r^{1/2}(\sqrt{D_r^{-1}P} - 1\sqrt{c'}) = \widetilde{U}\widetilde{D}_\lambda\widetilde{V}', \tag{8}$$

where if $M = (m_{ij})$ is a matrix then $\sqrt{M} = (\sqrt{m_{ij}})$. The I rows of N can be represented by taking as principal coordinates the rows of $\widetilde{A} = D_r^{-1/2}\widetilde{U}\widetilde{D}_\lambda$. The Euclidean distance between rows i, i' of $\widetilde{A}$ equals the Hellinger distance

$$\widetilde{\delta}_{ii'}^2 = \sum_{j=1}^{J}(\sqrt{p_{ij}/r_i} - \sqrt{p_{i'j}/r_{i'}})^2. \tag{9}$$

Equivalently, HD can be approached by using the SVD

$$D_r^{1/2}(D_r^{-1/2}\sqrt{P}D_c^{-1/2} - 11')D_c^{1/2} = \widetilde{U}\widetilde{D}_\lambda\widetilde{V}', \tag{10}$$

see (6), but the HD version of (7) does not hold.

Unfortunately there is no similar formula for representing columns, so the duality row-columns does not apply in HD. To overcome this deficiency, let us propose a procedure for representing columns in HD, similar to the way standard coordinates may be defined in CA. Clearly, the Hellinger distance is the Euclidean distance with coordinates the rows of $Q = D_r^{-1/2}\sqrt{P}$. Let $H_m = I - m^{-1}11'$ the centring matrix. After centring Q and $\widetilde{A}$, the principal

coordinates transformation is $H_I\widetilde{A} = H_IQT$, where T is orthogonal. To obtain T, we can use the Procrustean rotation (see Mardia *et al.*, 1979) from Q to $\widetilde{A}$ via the singular value decomposition $(H_IQ)'(H_I\widetilde{A}) = RDS'$, where D is diagonal. Then $T = RS'$.

To represent columns in the row space, we may interpret the $j-$th column of the contingency table as a set of J frequencies where only the $j-$th category is present. Thus we may identify this $j-$th column as the dummy profile $(0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the j-th entry, which assigns probability one to this column-variable. See, for example, Gower and Hand (1996).

Accordingly, the profiles for the J columns is the $J \times J$ identity matrix. Also centring this matrix, in order to represent the columns of N in the centred principal coordinates row space, the coordinates of the columns are given by

$$\widetilde{B}_0 = H_JT. \tag{11}$$

4 CA features

CA has the following advantages:

1. The joint representation of rows and columns, called the symmetric representation.

2. CA satisfies the "principle of distributional equivalence".

3. The Pearson chi-square statistic for testing independence is $\phi^2 = n(\lambda_1^2 + \cdots + \lambda_{K-1}^2)$, where $K = \min\{I, J\}$ and $\lambda_i, i = 1, \ldots, K-1$, are the singular values.

4. CA can be related to the log-linear models via the reconstitution formula $P = rc' + D_rA_0D_\lambda B_0D_c$, where $A_0 = D_r^{-1/2}U$, $B_0 = D_c^{-1/2}V$.

5. CA has equivalent approaches: canonical correlation, dual scaling, reciprocal averaging. This last approach (Hill, 1973) is useful for large sparse matrices, as typically occurs when relating and representing species and sites in ecological communities.

Indeed, some nice properties of CA stem from its probabilistic interpretation (see Cuadras et al., 2000; Cuadras, 2002). However, as stated by Rao (1995), CA has some drawbacks, the most important being the dependence

of the chi-square distance between row profiles on the column totals. Thus the chi-square distance between row profiles is not a function of the rows i, i' only. It also depends on the number of observations in the columns. This means that if we obtain a configuration and we add another set of rows, the CA on the extended set changes the distance between the original rows.

As an alternative to CA, HD has some advantages:

1. The Hellinger distance depends only on the profiles of the concerned rows. It is not altered when an extended set is considered.

2. The statistic $4n(\lambda_1^2 + \cdots + \lambda_K^2)$ is asymptotically distributed as chi-square, where $\lambda_1, \ldots, \lambda_K$ are the singular values in HD.

3. HD also satisfies the "principle of distributional equivalence".

But HD has the drawback of needing for the full representation $K = \min\{I, J\}$ dimensions, whereas CA only needs $K - 1$. Furthermore, it has no relation with log-linear models.

5 The log-ratio alternative

A third alternative to represent categorical data is inspired by the analysis on compositional data (Aitchison, 1986; Aitchison and Greenacre, 2000). Suppose now that the elements of P are positive values. Let us consider the weighted double-centering of $\log(D_r^{-1} P D_c^{-1})$, that is, the matrix with elements $[\log(p_{ij}/r_i c_j)]$:

$$D_r^{1/2}(I - 1r') \log(D_r^{-1} P D_c^{-1})(I - c1')D_c^{1/2} = U^* D_\lambda^* V^{*\prime}.$$

We may represent the rows of N using the principal coordinates $A^* = D_r^{-1/2} U^* D_\lambda^*$. This representation implicitly uses the following squared distance between rows

$$\delta_{ii'}^{*2} = \sum_{j=1}^{J} (\log \frac{p_{ij}}{(p_{i1} \cdots p_{ij})^{1/J}} - \log \frac{p_{i'j}}{(p_{i'1} \cdots p_{i'j})^{1/J}})^2.$$

We compare below this log-ratio (LR) approach to CA for a general P. Note that all entries in N must be strictly positive.

6 Comparing CA and HD

Let us compare (1) to (8). As D_c and D_r are diagonal, we can write

$$\sqrt{PD_c^{-1}} = \sqrt{P}D_c^{-1/2}, \quad 1'D_c^{1/2} = \sqrt{c'}, \quad D_r^{-1/2}r = \sqrt{r}.$$

Hence (8) is

$$D_r^{-1/2}(D_r^{1/2}\sqrt{P}D_c^{1/2} - rc')D_c^{-1/2} = \widetilde{U}\widetilde{D}_\lambda\widetilde{V}'. \tag{12}$$

If we compare (1) and (12), it is clear that CA and HD provide a similar representation of the rows of N when $P \approx D_r^{1/2}\sqrt{P}D_c^{1/2}$.

Next, an agreement measure between P and $D_r^{1/2}\sqrt{P}D_c^{1/2}$ is suggested. We define

$$\theta = \sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij} - h_{ij}) = 1 - \sum_{i=1}^{I}\sum_{j=1}^{J}h_{ij},$$

where $h_{ij} = \sqrt{r_i p_{ij} c_j}$. This measure satisfies $0 \leq \theta < 1$ with $\theta = 0$ if $P = rc'$. To see this, note that $\theta = 1 - \gamma$, where $\gamma = \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{r_i c_j}\sqrt{p_{ij}}$ is the affinity coefficient between p_{ij} and $r_i c_j$, which lies between 0 and 1.

7 Comparison along principal dimensions

The standard coordinates for representing columns in CA are $B_0 = D_c^{-1/2}V$. Then from (1) we have $P - rc' = D_r A B_0' D_c$. Thus the full representation using A, B_0 for rows and columns, called the asymmetric representation, represents graphically the whole frequency matrix N. In HD the joint representation uses $\widetilde{A}, \widetilde{B}_0$, although this representation does not reconstruct N.

Let us introduce the "standard coordinates" $\widetilde{B}_* = D_c^{-1/2}\widetilde{V}$ in HD, which have a similar interpretation to the standard coordinates in CA. Then

$$\begin{aligned} D_r^{1/2}\sqrt{P}D_c^{1/2} - rc' &= D_r\widetilde{A}\widetilde{B}_*'D_c, \\ P - D_r^{1/2}\sqrt{P}D_c^{1/2} &= D_r(AB_0' - \widetilde{A}\widetilde{B}_*')D_c. \end{aligned}$$

Theorem 1 *The measure θ satisfies the inequality $0 \leq \theta \leq 1 - 1/\sqrt{K}$, and can be expressed as*

$$\theta = -(\mu_1\nu_1 + \mu_2\nu_2 + \ldots + \mu_K\nu_K), \tag{13}$$

where μ_i, ν_i are the weighted means of the i-th principal and "standard coordinates" in HD, respectively.

Proof. Write $\theta = r'(AB_0' - \widetilde{A}\widetilde{B}_*')c$ and use some properties of concave and convex functions. □

Thus, a global agreement measure is given by $\tau = \theta/(1 - 1/\sqrt{K})$ and a partial agreement measure for a specific dimension is $\tau_i = |\mu_i \nu_i/\tau|$. Note that $\tau_i = 0$ if the coordinates are identical. Hence CA and HD give the same two-dimensional representation of the rows of N provided that τ_1 and τ_2 are close to 0.

8 Joint comparison

The three approaches for representing categorical data can be jointly compared. Let us write $D_r^{-1}PD_c^{-1} - 11'$ as $[p_{ij}/r_i c_j - 1]$ and $D_r^{-1/2}\sqrt{P}D_c^{-1/2} - 11'$ as $[\sqrt{p_{ij}/r_i c_j} - 1]$. Then CA, HD and LR uses the SVD (canonical coordinates or weighted double-centering, see above) of the expressions contained in Table 1, showing that CA is formally similar to both HD and LR, but HD is not similar to LR.

Table 1. Matrices which are decomposed in CA, HD and LR.

	Canonical coordinates	Weighted double-centering
CA	$D_r^{1/2}[p_{ij}/r_i c_j - 1]D_c^{1/2}$	$D_r^{1/2}(I - 1r')[p_{ij}/r_i c_j](I - c1')D_c^{1/2}$
HD	$D_r^{1/2}[\sqrt{p_{ij}/r_i c_j} - 1]D_c^{1/2}$	-
LR	-	$D_r^{1/2}(I - 1r')[\log(p_{ij}/r_i c_j)](I - c1')D_c^{1/2}$

Theorem 2 *CA can be understood as the first-order approximation to the alternatives HD and LR.*

Proof. Use Taylor expansions on $\sqrt{p_{ij}/r_i c_j} - 1$ and $\log(p_{ij}/r_i c_j)$, see Table 1. Then CA is the first term of this expansion.□

As a consequence, when rows and columns are almost independent, i.e., $p_{ij} \approx r_i c_j$, then

$$\begin{aligned} \sqrt{p_{ij}/r_i c_j} - 1 &\approx \tfrac{1}{2}(p_{ij}/r_i c_j - 1), \\ \log(p_{ij}/r_i c_j) &\approx (p_{ij}/r_i c_j - 1), \end{aligned} \tag{14}$$

and these three methods may provide quite similar graphic displays.

These approximations justify θ and suggest a measure φ for comparing CA and LR. The measure θ can also be expressed as:

$$\begin{aligned} \theta &= 1'(\tfrac{1}{2}(P - rc') - (D_r^{1/2}\sqrt{P}D_c^{1/2} - rc'))1 \\ &= -r'(D_r^{-1/2}\sqrt{P}D_c^{-1/2} - 11')c. \end{aligned}$$

A global measure φ for comparing CA and LR can be defined as follows:

$$\begin{aligned}\varphi &= -1'(D_r \log(D_r^{-1} P D_c^{-1}) D_c - P - rc')1 \\ &= -r' \log(D_r^{-1} P D_c^{-1}) c.\end{aligned}$$

As $\log(x)$ is concave, $\varphi \geq 0$ being 0 in case of independence. No upper bound exists for φ.

9 A classical example

Table 2 reports the hair colour and eye colour of 5,383 individuals.

Table 2. Hair colour combined with eye colour.

	Hair colour					
Eye colour	Fair	Red	Medium	Dark	Black	Total
LIGHT	688	116	584	188	4	1,580
BLUE	326	38	241	110	3	718
MEDIUM	343	84	909	412	26	1,774
DARK	98	48	403	681	81	1,311
Total	1,455	286	2,137	1,391	114	5,383

There is a significant dependence between eye and hair colour ($\chi^2 = 1,230$ for 12 d.f.). However, the association is moderate, as the Cramér coefficient is $\sqrt{1,230/(5,383 \times 3)} = 0.276$.

For this contingency table (13) gives

$$\theta = -(0.0001 + 0.0019 - 0.0309 - 0.0003) = 0.0292.$$

These values indicate a quite small difference between both representations. Indeed, the rank order for the distances along the first and the second axis is exactly the same (Figure 1).

We also obtain $\varphi = 0.1251$ and the LR alternative gives a slightly different display, as the differences between the quantities in (14) are not negligible owing to the moderate row-column association.

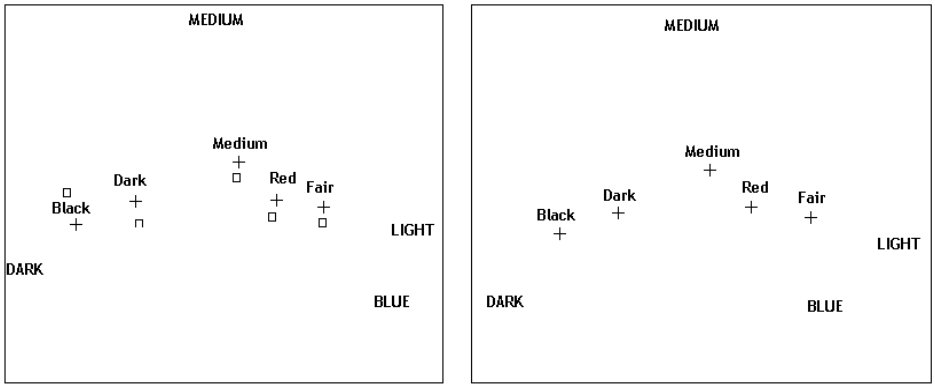

Figure 1: Asymmetric correspondence analysis of eye colour and hair colour (left, crosses, quality 99%), log-ratio representation (left, squares) and representation using Hellinger distance (right, quality 97%).

10 Conclusions

Correspondence analysis (CA) and the Hellinger distance approach (HD) for representing categorical data may provide similar results under some circumstances, for example, when rows and columns are almost independent. CA is the best for several reasons (joint representation, probabilistic interpretation, relation to log-linear models, reciprocal averaging approach), but may have some drawbacks when the rows are multinomial populations and the solution should not depend on the column frequencies, for which the HD approach may be preferable. Except in the case of almost independence, the log-ratio approach (LR) may provide quite different results, but cannot be applied with zero frequencies.

11 Acknowledgments

Work supported in part by grants Ministerio de Educacion y Ciencia MTM 2004-00440 and Generalitat de Catalunya CUR 2001 SGR 00067. An extended version of this contributed paper has been submitted for publication to a refereed journal.

12 References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Aitchison, J. and M. J. Greenacre (2000) Biplots of compositional data. *Applied Statistics*, **51**, 375-392.

Benzécri, J. P. (1973) *L'Analyse des Données. I. La Taxinomie. II. L'Analyse des Correspondances*. Dunod, Paris.

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

Cuadras, C. M., Fortiana, J. and M. J. Greenacre (2000) Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions. In: R.D.H. Heijmans, D.S.G. Pollock and A. Satorra, (Eds.), *Innovations in Multivariate Statistical Analysis*, pp. 101-116. Kluwer Ac. Publ., Dordrecht.

Gower, J. C. and D. J. Hand (1996) *Biplots*. Chapman and Hall, London.

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

Hill, M. O. (1973) Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology*, **61**, 237-249.

Mardia, K.V., Kent, J.T. and J.M. Bibby (1979) *Multivariate Analysis*. Academic Press, London.

Rao, C. R. (1948) The utilization of multiple measurements in problems of biological classification (with discussion) *Journal of Royal Statistical Society, Series B*, **10**, 159-193.

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