ASSESSING THE NUMBER OF AXES THAT SHOULD BE CONSIDERED IN CORRESPONDENCE ANALYSIS

Yves Escoufier

ENSA-INRA-USTL
Unité de Biométrie
9, place Pierre Viala
34060 MONTPELLIER CEDEX
(France)

I. INTRODUCTION

When performing a correspondence analysis, statisticians most frequently have to cope with the following dilemma - choose either a simple interpretation based on the first two or three axes obtained, or a more subtle interpretation (obliging them to handle and view more information) based on a great number of axes.

One could wish to find in the inferential statistic approaches, efficient processes to objectivize the choice of the number of axes. The dramatic limitations of available tools are reviewed in Section IV.

The problem can then be solved to through more exploratory approaches based either on least squares approximation properties of the solutions, or on cross validation techniques.

Sections II and III of this article introduce those approaches.

ISBN 0-12-215485-1

II. CORRESPONDENCE ANALYSIS AS AN APPROXIMATION OF THE DATA MATRIX

A. Let P be a table of frequencies the size of which is IxJ, with $\sum_{i=1}^{I} \sum_{j=1}^{J} P_{ij} = 1$.

Δ_I and Δ_J are diagonal matrices defined by marginal weights of the rows and columns of P respectively.

$$\Delta_I = \begin{pmatrix} P_{1.} & & \\ & \ddots & \\ & & P_{I.} \end{pmatrix} \qquad \Delta_J = \begin{pmatrix} P_{.1} & & \\ & \ddots & \\ & & P_{.J} \end{pmatrix}$$

Consider the triplets $(\lambda_\alpha, \psi_\alpha, \phi_\alpha)$ defined for $\alpha = 1, \ldots, N = \min(I,J)$ by

$$\Delta_I^{-1} P \Delta_J^{-1} P' \psi_\alpha = \lambda_\alpha \psi_\alpha \quad \text{with } \psi'_\alpha \Delta_I \psi_\alpha = 1 \tag{1}$$

$$\Delta_J^{-1} P' \Delta_I^{-1} P \phi_\alpha = \lambda_\alpha \phi_\alpha \quad \text{with } \phi'_\alpha \Delta_J \phi_\alpha = 1 \tag{2}$$

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots > \lambda_N \tag{3}$$

For $\lambda_\alpha \neq 0$, the so called transition formulas are easily deduced from 1) and 2)

$$\phi_\alpha = \frac{\Delta_J^{-1} P' \psi_\alpha}{\sqrt{\lambda_\alpha}} \tag{4}$$

$$\psi_\alpha = \frac{\Delta_I^{-1} P \phi_\alpha}{\sqrt{\lambda_\alpha}} \tag{5}$$

Therefore $\sqrt{\lambda_\alpha}\, \phi_{\alpha j}$ is the barycenter of the $\psi_{\alpha i}$'s bearing the weights $\frac{P_{ij}}{P_{i.}}$, while $\sqrt{\lambda_\alpha}\, \psi_{\alpha i}$ is the barycenter of the $\phi_{\alpha j}$ bearing the weights $\frac{P_{ij}}{P_{.j}}$.

Geometric considerations (Lebart (1977), p. 58) lead then to the conclusion that for any α, $\lambda_\alpha \leqslant 1$.

Note $\underset{\sim}{1}_I$ as the vector of R^I, all the components of which are equal to the unit, and $\underset{\sim}{1}_J$ as the vector equivalent to R^J. It is easy to check that :

$$\Delta_I^{-1}\ P\ \Delta_J^{-1}\ P'\ \underset{\sim}{1}_I \ = \ 1 \times \underset{\sim}{1}_I \qquad (6)$$

$$\Delta_J^{-1}\ P'\ \Delta_I^{-1}\ P\ \underset{\sim}{1}_J \ = \ 1 \times \underset{\sim}{1}_J \qquad (7)$$

The triplet $(1, \underset{\sim}{1}_I, \underset{\sim}{1}_J)$ is therefore one of the triplets $(\lambda_\alpha, \psi_\alpha, \phi_\alpha)$, which leads, when isolating it, to renumber the remaining triplets from 1 to N-1. These notations being implemented, any presentation of Correspondence Analysis gives the two following formulas :

Data reconstitution formula

$$\forall(i,j) \in IxJ \qquad P_{ij} = P_{i.}\,P_{.j}\,(1 + \sum_{\alpha=1}^{N-1} \sqrt{\lambda_\alpha}\ \psi_{\alpha i}\ \phi_{\alpha j}) \qquad (8)$$

Sum of latent root significance

$$\sum_{i=1}^{I}\ \sum_{j=1}^{J}\ \frac{(P_{ij} - P_{i.}\,P_{.j})^2}{P_{i.}\ P_{.j}} = \sum_{\alpha=1}^{N-1} \lambda_\alpha \qquad (9)$$

B. When seen in the context of an approximation of a matrix by a matrix of lower rank, these two formulas seen to be particular cases of more general formula.

In fact consider $A = \Delta_I^{-1}\ P\ \Delta_J^{-1}$ and define N matrices A_k by :

$$A_0 = \underset{\sim}{1}_I\ \underset{\sim}{1}'_J$$

$$\forall k = 1,\ldots, N-1 \quad A_k = \underset{\sim}{1}_I\ \underset{\sim}{1}'_J + \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\ \psi_\alpha\ \phi'_\alpha \qquad (10)$$

It was shown (Escoufier and Junca (1986)) that A_k is the best approximation of order k+1 for A and very accurately that :

$$\mathrm{Tr}((A-A_k)\ \Delta_J(A-A_k)'\ \Delta_I) = \sum_{\alpha=k+1}^{N-1} \lambda_\alpha \tag{11}$$

which also can be written for k ⩾ 1

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{P_{ij}}{P_{i.}\ P_{.j}} - (1 + \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\ \psi_{\alpha i}\ \phi_{\alpha j})\right)^2 P_{i.}\ P_{.j} = \sum_{\alpha=k+1}^{N-1} \lambda_\alpha \tag{12}$$

For A_0, the following is obtained

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{P_{ij}}{P_{i.}\ P_{.j}} - 1\right)^2 P_{i.}\ P_{.j} = \sum_{\alpha=1}^{N-1} \lambda_\alpha \tag{12'}$$

Noticing that $A - A_k = \sum_{\alpha=k+1}^{N-1} \sqrt{\lambda_\alpha}\ \psi_\alpha\ \phi'_\alpha$, formula 11 can be rewritten in order to show the factorizations of the total error made in the approximation $\sum_{\alpha=k+1}^{N-1} \lambda_\alpha$, according to rows, columns or even cells (i,j) of array P. The following is obtained :

$$\sum_{\alpha=k+1}^{N-1} \lambda_\alpha = \mathrm{Tr}((A-A_k)\ \Delta_J(A-A_k)'\ \Delta_I)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} P_{i.}\ P_{.j}\ \left(\sum_{\alpha=k+1}^{N-1} \sqrt{\lambda_\alpha}\psi_{\alpha i}\ \phi_{\alpha j}\right)^2$$

Noticing that for $\alpha \neq \beta$ $\sum_{i=1}^{I} P_{i.}\psi_{\alpha i}\psi_{\beta i} = \sum_{j=1}^{J} P_{.j}\phi_{\alpha j}\ \phi_{\beta j} = 0$ the following is obtained :

$$\sum_{\alpha=k+1}^{N-1} \lambda_\alpha = \sum_{\alpha=k+1}^{N-1} \sum_{i=1}^{I} \sum_{j=1}^{J} P_{i.} P_{.j}\ \lambda_\alpha (\psi_{\alpha i})^2\ (\phi_{\alpha j})^2 \tag{13}$$

It can be deduced that as long as the triplet $(\lambda_\alpha, \psi_\alpha, \phi_\alpha)$ is not used in the approximation, an error λ_α is made, which can be attributed to it.

The error can be factorized by the rows

$$\lambda_\alpha = \sum_{i=1}^{I} P_{i.} \left[\sum_{j=1}^{J} P_{.j}\lambda_\alpha\ (\psi_{\alpha i})^2\ (\phi_{\alpha j})^2\right]$$

$$= \sum_{i=1}^{I} P_{i.} \lambda_\alpha (\psi_{\alpha i})^2$$

Similarly, it can be factorized by columns

$$\lambda_\alpha = \sum_{j=1}^{J} P_{.j} \lambda_\alpha (\phi_{\alpha j})^2$$

and also by cells :

$$\lambda_\alpha = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{i.} P_{.j} \lambda_\alpha (\psi_{\alpha i})^2 (\phi_{\alpha j})^2$$

C. These results can be used to assess the quality of the approximations made to the order k+1 in a finer way than for the simple consideration of the quantity $\sum_{\alpha=k+1}^{N-1} \lambda_\alpha$.

When constructing the matrix of the item $\sum_{\alpha=k+1}^{N-1} P_{i.} P_{.j} \lambda_\alpha (\psi_{\alpha i})^2 (\phi_{\alpha j})^2$ as a contribution from the cell (i,j) of P to the approximation error $\sum_{\alpha=k+1}^{N-1} \lambda_\alpha$, the sum of terms of the i^{th} row is equal to $\sum_{\alpha=k+1}^{N-1} \lambda_\alpha (\psi_{\alpha i})^2 P_{i.}$, as a contribution from the i^{th} row of P to the approximation error.

Similarly, the sum of the items of the j^{th} column is equal to $\sum_{\alpha=k+1}^{N-1} \lambda_\alpha (\phi_{\alpha j})^2 P_{.j}$, as a contribution from the j^{th} column of P to the approximation error.

A simple visualization of the matrix makes it possible to see whether factorizations of the total error show an identifiable structure or not. When this error is both small and unstructured, the number of axes used is therefore sufficient.

It should be noted that small $P_{i.} P_{.j}$ values can hide very high real residuals $\sum_{\alpha=k+1}^{N-1} \lambda_\alpha (\psi_{\alpha i})^2 (\phi_{\alpha j})^2$. These residuals are negligible from a methodological point of view since they are linked to very low weightings.

III. DETERMINING THE NUMBER OF AXES TO BE RETAINED BY CROSS VALIDATION

A. The calculation of item $(A_k)_{i,j}$ as carried out in formula (10) requires all items A, and more especially A_{ij}. It may seem highly favorable to use item A_{ij} in order to estimate this item itself. The principle of cross validation processes consists precisely in trying to avoid that situation. Approximation $(A_k)_{i,j}$ shall be computed without considering A_{ij}. A poor reconstruction of A_{ij}'s by $(A_k)_{i,j}$'s will show that the model underlying the approximation of A by A_k lets too great a part of the data variability escape : the specificity of a cell A_{ij} escapes the approximation $(A_k)_{i,j}$ that can be deduced from other items of A.

B. Omitting item A_{ij} could be considered with references to the processes used when data are missing (Greenacre (1984)).

In order to allow for the objective of cross validation this algorithm shall be used for each cell of matrix A, i.e., IxJ times. Assuming that r iterations of the algorithm are necessary (or deliberately chosen) for each cell of matrix A, r x I x J analyses should then be performed.

A less costly process is preferred. It was introduced within a Principal Component Analysis of a triplet (X, Q, D) by Holmes-Junca (1985). Adapting it to the Correspondence Analysis must allow for the fact that in Correspondence Analysis, matrices Q and D depend on data X.

Consider table $P^{(i)}$ obtained when removing row i from P. The size of this table is (I-1) x J. Correspondence Analysis of this table will produce $N^{(i)} = \min((I-1), J)$ triplets $(\lambda_\alpha^{(i)}, \psi_\alpha^{(i)}, \phi_\alpha^{(i)})$.

III. DETERMINING THE NUMBER OF AXES TO BE RETAINED BY CROSS VALIDATION

A. The calculation of item $(A_k)_{i,j}$ as carried out in formula (10) requires all items A, and more especially A_{ij}. It may seem highly favorable to use item A_{ij} in order to estimate this item itself. The principle of cross validation processes consists precisely in trying to avoid that situation. Approximation $(A_k)_{i,j}$ shall be computed without considering A_{ij}. A poor reconstruction of A_{ij}'s by $(A_k)_{i,j}$'s will show that the model underlying the approximation of A by A_k lets too great a part of the data variability escape : the specificity of a cell A_{ij} escapes the approximation $(A_k)_{i,j}$ that can be deduced from other items of A.

B. Omitting item A_{ij} could be considered with references to the processes used when data are missing (Greenacre (1984)).

In order to allow for the objective of cross validation this algorithm shall be used for each cell of matrix A, i.e., IxJ times. Assuming that r iterations of the algorithm are necessary (or deliberately chosen) for each cell of matrix A, r x I x J analyses should then be performed.

A less costly process is preferred. It was introduced within a Principal Component Analysis of a triplet (X, Q, D) by Holmes-Junca (1985). Adapting it to the Correspondence Analysis must allow for the fact that in Correspondence Analysis, matrices Q and D depend on data X.

Consider table $P^{(i)}$ obtained when removing row i from P. The size of this table is (I-1) x J. Correspondence Analysis of this table will produce $N^{(i)} = \min((I-1), J)$ triplets $(\lambda_\alpha^{(i)}, \psi_\alpha^{(i)}, \phi_\alpha^{(i)})$.

Similarly column j can be removed from table P in order to obtain a table $P^{(j)}$. The correspondence analysis of this table will produce $N^{(j)} = \min\,(I, (J-1))$ triplets $(\lambda_\alpha^{(j)}, \psi_\alpha^{(j)}, \phi_\alpha^{(j)})$.

Let $N^{-} = \min(N^{(i)}, N^{(j)})$. For $k = 1, \ldots, N^{-}-1$ the following is then defined

$$(A_k)_{ij} = 1 + \sum_{\alpha=1}^{k} \sqrt{\sqrt{\lambda_\alpha^{(i)} \; \lambda_\alpha^{(j)}}}\; \psi_{\alpha i}^{(j)} \; \phi_{\alpha j}^{(i)}$$

The quality of the reconstruction can be determined in calculating the matrix of the following terms

$$(A_{ij} - (A_k)_{ij})^2 \; P_{i.} \; P_{.j}$$

IV. TESTS ON THE SUM OF THE NON-RETAINED LATENT ROOTS

A. Basing the determination of the number of axes to be retained on a testing process, first implies considering that the observations performed consist in a sample representative of a greater population. Let E be the table, I x J, of numbers from which table P of frequencies was calculated. $E_{i.}$, $E_{.j}$ and $E_{..}$ will be the total numbers for the line i, the column j and the entire table, respectively. It is usual practice to rewrite formula (9) as

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (E_{ij} - E_{i.} E_{.j})^2 \; / \; (E_{i.} E_{.j} / E_{..}) = \left(\sum_{\alpha=1}^{N-1} \lambda_\alpha \right) \times E_{..}$$

The following conventional statement is thus obtained. Under the Independence Hypothesis of rows and columns of E, $\left(\sum_{\alpha=1}^{N-1} \lambda_\alpha \right) \times E_{..}$ respects the law of chi-square with (I-1) x (J-1) degrees of freedom.

In fact, this result comes down to checking that none of the latent roots observed is significantly different from zero.

Used strictly, this result should lead not to perform a correspondence analysis when the independence hypothesis is not rejected.

In fact, it is often interesting to explore deviations from independence model even if those deviations are not important enough to be judged significant by the test. Symmetrically the chi-square may be significant without the correspondence analysis being the method adapted for an explanation of the dependence structure. These results follow from the independence of latent roots of a Wishart matrix with regard to their sum. This point has ben detailed by Lebart (1976) and Lebart et al. (1977).

B. For the problem under discussion, the necessary result would be to check that, from a certain index k, latent roots are not significantly different from zero. In this respect several solutions were proposed. Relying on O'Neill's results (1978), Greenacre's (1984) and Goodman's (1986) articles have shown that all these proposals were erroneous and produced too optimistic results.

Let Π be the probability table in the population under consideration, and state

$$\Pi_{ij} = \Pi_{i.}\ \Pi_{.j}\ (1 + \sum_{\alpha=1}^{N-1} \sqrt{\mu_\alpha}\ \xi_{\alpha i}\ \eta_{\alpha j})$$

where triplets $(\mu_\alpha, \xi_\alpha, \eta_\alpha)$ are defined for Π by analogy with triplets $(\lambda_\alpha, \psi_\alpha, \phi_\alpha)$ linked to P through formulas (1), (2) and (3).

Let $k < N-1$

For any $s > k$, any $t > k$ and any $\alpha \leqslant k$, set down

$$E_c(s, t, \alpha) = \sum_{i=1}^{I} \xi_{si}\ \xi_{ti}\ \xi_{\alpha i}\ \Pi_{i.}$$

Similarly, for any $u > k$, any $v > k$ and any $\alpha \leqslant k$ set down

$$E_e(u, v, \alpha) = \sum_{j=1}^{J} \eta_{uj}\, \eta_{vj}\, \eta_{\alpha j}\, \Pi_{.j}$$

O'Neill has shown that with the hypothesis

$\mu_\alpha = 0$ for any $\alpha > k$

quantify $\left(\sum_{\alpha=k+1}^{N-1} \lambda_\alpha \right) \times E_{..}$ is distributed as the sum of N-k squares of normal dependant variables and that a chi-square is obtained with (I-1-k) x (J-1-k) degrees of freedom if, and only if

for any u,v,s,t > k $\quad \sum_{\alpha=1}^{k} \sqrt{\mu_\alpha}\, E_e(u,v,\alpha)\, E_c(s,t,\alpha) = 0$

Most frequently, this purely theoretical condition shows above all that any testing process dealing with the last latent roots will be questionable.

C. Conclusion

At the time being, Correspondence Analysis users are not provided with very efficient testing tools for retaining a number of axes to be studied. This lack should give rise to works on that subject. Experiments were made to use Bootstrap's and Jackknife's techniques which have not brought yet a precise and easy-to-implement tool to users.

REFERENCES

1. Escoufier, Y. and Junca S. (1986). Least squares approximation of frequencies or their logarithms. Inst. Stat. Rev. Vol. 54, n° 3, pp. 279-283.

2. Goodman, L. (1986). Some Useful Extensions of the Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables. Int. Stat. Rev. Vol. 54, n° 3, pp. 243-270.
3. Greenacre, M.J. (1984). Theory and Applications of Correspondence Analysis. Academic Press, Inc.
4. Holmes-Junca, S. (1985). Outils Informatiques pour l'Evaluation de la Pertinence d'un Résultat en Analyse des Données. Thèse de 3ème cycle. U.S.T.L. Montpellier.
5. Lebart, L. (1976). The Significance of Eigenvalues Issued from Correspondence Analysis. In "Proceedings in Computational Statistics" (Compstat), pp. 38-45. Physica-Verlag, Vienna.
6. Lebart, L., Morineau, A., Tabard, N. (1977). Techniques de la Description Statistique : Méthodes et Logiciels pour l'Analyse des Grands Tableaux. Dunod. Paris.
7. O'Neill, M.E. (1978). Distributional Expansions for Canonical Correlations from Contingency Tables. J.R.S.S. B. 40, n° 3, pp. 303-312.