Y. Escoufier and S. Junca

Unité de Biométrie, ENSA-INRA-USTL, 9, place Pierre Viala, 34060 Montpellier Cedex, France

Least-squares approximation of frequencies or their logarithms

The reflexions that are presented herewith about L.A. Goodman's paper are based on mathematical results which will be recalled without proof.

We shall first present certain aspects of correspondence analysis and then handle the log-bilinear model in the same context.

Goodman's notations are kept as far as possible to make comparisons easier.

Mathematical basis (Rao, 1980; Sabatier, Jan & Escoufier, 1984)

Let us consider two vector spaces E and F whose respective dimensions are J and I, and suppose that they both have scalar products defined respectively in E by a positive-definite matrix Q and in F by a positive-definite matrix D.

We will need the dual space of E, denoted by E^* and with scalar product defined by Q^{-1}. If A is the matrix associated to a linear transformation from E^* into F we will denote by A^{T} its transpose and by $A^* = QA^{\mathrm{T}}D$ its adjoint. Let us use the notation $N = \min(I, J)$.

It is well known that the best k-ranked approximation, with $k < N$, to A by a $I \times J$ matrix is provided by

$$A_k = \sum_{m=1}^{k} \lambda_m x_m y_m^{\mathrm{T}},$$

where the triplets (λ_m, x_m, y_m), for $m = 1, 2, \ldots, k$, are defined by

$$\begin{aligned} AA^*x &= AQA^{\mathrm{T}}Dx_m = \lambda_m^2 x_m \quad (x_m^{\mathrm{T}}Dx_m = 1), \\ A^*Ay &= A^{\mathrm{T}}DAQy_m = \lambda_m^2 y_m \quad (y_m^{\mathrm{T}}Qy_m = 1), \end{aligned} \tag{D4.1}$$

and the $\lambda_1, \lambda_2, \ldots, \lambda_m$ are ordered decreasingly.

It is easily checked that for $\lambda_m \neq 0$

$$x_m = AQy_m/\lambda_m, \quad y_m = A^{\mathrm{T}}Dx_m/\lambda_m, \tag{D4.2}$$

whereas, if $\lambda_m^2 \neq \lambda_{m'}^2$, then $x_m^{\mathrm{T}}Dx_{m'} = y_m^{\mathrm{T}}Qy_{m'} = 0$.

We have also

$$\|A - A_k\|^2 = \mathrm{Tr}\,[(A - A_k)Q(A - A_k)^{\mathrm{T}}D] = \sum_{m=k+1}^{N} \lambda_m^2, \tag{D4.3}$$

and no other k-ranked matrix provides a better approximation to A.

In particular

$$A = \sum_{m=1}^{N} \lambda_m x_m y_m^{\mathrm{T}}. \tag{D4.4}$$

Note as well that, if Δ_I is a diagonal matrix such that $\mathbf{1}_I^T \Delta_I A = 0$, then $\mathbf{1}_I^T \Delta_I x_m = 0$, for $m = 1, \ldots, N$; and, similarly, if Δ_J is a diagonal matrix for which $\mathbf{1}_J^T \Delta_J A^T = 0$, then $\mathbf{1}_J^T \Delta_J y_m = 0$, for $m = 1, \ldots, N$, where $\mathbf{1}_I$ is the R^I vector whose I components are equal to 1.

Correspondence analysis

Among the different ways of presenting this method we will choose, for brevity's sake, substitution of the contingency table P by the matrix $A = \Delta_I^{-1} P \Delta_J^{-1}$, where

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

We will apply the results just recalled with $Q = \Delta_J$ and $D = \Delta_I$.

Formula (D4.1) implies calculating the eigenvectors of

$$A \Delta_J A^T \Delta_I = \Delta_I^{-1} P \Delta_J^{-1} P^T, \quad A^T \Delta_I A \Delta_J = \Delta_J^{-1} P^T \Delta_I^{-1} P.$$

It is easy to check that $\mathbf{1}_I$ is the eigenvector of the first operator for the eigenvalue 1, as is $\mathbf{1}_J$ for the second operator. So let us decide to number the other eigenvalues and eigenvectors from 1 to $M = N - 1$.

Therefore it results from the orthogonalities of the vectors x_m and y_m noted at (D4.2) that

$$\mathbf{1}_I^T \Delta_I x_m = \sum_{i=1}^{I} x_{im} P_{i.} = 0, \quad \mathbf{1}_J^T \Delta_J y_m = \sum_{j=1}^{J} y_{jm} P_{.j} = 0. \tag{D4.5}$$

Formula (D4.4) is in this case written

$$A = \mathbf{1}_I \mathbf{1}_J^T + \sum_{m=1}^{M} \lambda_m x_m y_m^T$$

or, for all (i, j) in $I \times J$,

$$P_{ij} = P_{i.} P_{.j} \left(1 + \sum_{m=1}^{M} \lambda_m x_{im} y_{jm}\right). \tag{D4.6}$$

If we limit the sum to $k = M^* < M$ terms, (D4.3) provides

$$\begin{aligned} \sum_{m=k+1}^{M} \lambda_m^2 &= \mathrm{Tr}\left[(A - A_k) \Delta_J (A - A_k)^T \Delta_I\right] \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} \left[\frac{P_{ij}}{P_{i.} P_{.j}} - \left(1 + \sum_{m=1}^{M^*} \lambda_m x_{im} y_{jm}\right)\right]^2 P_{i.} P_{.j}. \end{aligned} \tag{D4.7}$$

Correspondence analysis thus appears as a least-squares approximation of the $P_{ij}/P_{i.}P_{.j}$, each (i, j)th element being weighted by the coefficient $P_{i.}P_{.j}$.

We remark that the notational problem that Goodman mentioned, that is λ_m or $\sqrt{\lambda_m}$, comes, in fact, from the eigenvalue notation in (D4.1), which is only conventional. Preference for the 'correlation' interpretation, as given by Goodman, leads to λ_m; in order to retrieve this notation the eigenvalues of AA^* are denoted by λ_m^2, this notation being in fact rarely used.

Passing from x_m to y_m and from y_m to x_m in the formula (D4.2) leads in this case to

$$x_m = \frac{\Delta_I^{-1} P y_m}{\lambda_m} \quad \text{or} \quad x_{im} = \frac{1}{\lambda_m}\left[\sum_{j=1}^{J} \frac{P_{ij}}{P_{i.}} y_{jm}\right],$$
$$y_m = \frac{\Delta_J^{-1} P^{\mathrm{T}} x_m}{\lambda_m} \quad \text{or} \quad y_{jm} = \frac{1}{\lambda_m}\left[\sum_{i=1}^{I} \frac{P_{ij}}{P_{.j}} x_{im}\right]. \tag{D4.8}$$

These equalities are known in France as the *transition formulae*. They are at the heart of the presentation of correspondence analysis as the 'reciprocal averaging' method. They lead to representations where the variable I's modalities have coordinates denoted by x'_m, where $x'_m = \lambda_m x_m$ and variable J's modalities have y as coordinates, or symmetrically $y'_m = \lambda_m y_m$ and x_m, thus exhibiting the feature known as the '*barycentric principle*'.

Figure 1 of Escoufier, (1982), cited by Goodman, is followed and developed by Fig. 2 which is an illustration of this technique. Benzécri et al. (1973) show this property in T II A, n° 26; another example of using this technique, written in English, is given by Greenacre (1984, pp. 76–77, 90). So the statement that the usual way of using correspondence analysis is with x' and y' is unfounded. The same is true of the statement that the use of this method systematically implies choice of two x' vectors and two y' vectors. Usually one chooses as many vectors as judged necessary for a good description of the table that is being analysed. The following section contains some remarks concerning this choice.

Log-bilinear model

Goodman considers the model

$$P_{ij} = \alpha_i \beta_j \exp\left(\sum_{m=1}^{M} \phi_m \mu_{im} \nu_{jm}\right),$$

where α_i and β_j must be positive numbers and the vectors μ_m and ν_m must satisfy, among other conditions, the equality

$$\sum_{i=1}^{I} \mu_{im} P_{i.} = \sum_{j=1}^{J} \nu_{jm} P_{.j} = 0.$$

If we develop this constraint we obtain

$$\log P_{ij} = \log \alpha_i + \log \beta_j + \sum_{m=1}^{M} \phi_m \mu_{im} \nu_{jm}, \tag{D4.9}$$

and therefore

$$\sum_{i=1}^{I} (\log P_{ij}) P_{i.} = \sum_{i=1}^{I} (\log \alpha_i) P_{i.} + \log \beta_j, \quad \sum_{j=1}^{J} (\log P_{ij}) P_{.j} = \sum_{j=1}^{J} (\log \beta_j) P_{.j} + \log \alpha_i.$$

It is evident that if the $I \times J$ couples $(\log \alpha_i, \log \beta_j)$ are solutions of (D4.9) then for every constant c the $I \times J$ couples $(\log \alpha_i + c, \log \beta_j - c)$ are solutions too. So it is possible to choose, for instance, a condition such as

$$\sum_{j=1}^{J} (\log \beta_j) P_{.j} = 0$$

to be imposed on the β_j. And thus we would have

$$\log \alpha_i = \sum_{j=1}^{J} (\log P_{ij}) P_{\cdot j},$$

$$\log \beta_j = \sum_{i=1}^{I} (\log P_{ij}) P_{i\cdot} - \sum_{i=1}^{I} \sum_{j=1}^{J} (\log P_{ij}) P_{i\cdot} P_{\cdot j}.$$

Let us then consider the array A defined by its elements

$$A_{ij} = \log P_{ij} - \log \alpha_i - \log \beta_j,$$

which satisfy

$$\sum_{i=1}^{I} A_{ij} P_{i\cdot} = \sum_{j=1}^{J} A_{ij} P_{\cdot j} = 0.$$

By applying the results discussed in the section 'Mathematical basis' above to A, to $Q = \Delta_J$ and to $D = \Delta_I$ we can find the coefficients ϕ_m and the vectors μ_m and ν_m by solving the equations corresponding to the formula (D4.1).

Formula (D4.2) leads to

$$\mu_m = A\Delta_J \nu_m / \phi_m, \quad \nu_m = A^{\mathrm{T}} \Delta_I \mu_m / \phi_m \tag{D4.10}$$

and finally

$$A_k = \sum_{m=1}^{k} \phi_m \mu_m \nu_m^{\mathrm{T}}$$

is the best k-ranked approximation to A. To be precise, as a consequence of (D4.3), we

Table D4.1

Results of analysis of Goodman's Table 1 using method described above; approximations made with three decimals for $m = 1$ and $m = 2$

	$m = 1$	$m = 2$
Eigenvalues	0·02930	0·00183
Percentages	93·16%	5·83%
Row coordinates (standardized for the P_i weights)	1·679 0·139 −0·133 −1·413	0·680 0·215 −1·840 0·832
Column coordinates (standardized for the $P_{\cdot j}$ weights)	1·076 1·148 0·366 −0·024 −0·967 −1·854	−0·421 −0·736 −0·414 1·272 1·008 −1·587

have

$$\sum_{m=k+1}^{M} \phi_m^2 = \mathrm{Tr}\,[(A - A_k)\Delta_J(A - A_k)^{\mathrm{T}}\Delta_I]$$
$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\log P_{ij} - \log \alpha_i - \log \beta_j - \sum_{m=1}^{k} \phi_m \mu_{im} \nu_{jm}\right]^2 P_{i.} P_{.j}.$$

In other words, *this approach gives a least-squares approximation to*

$$\log P_{ij} - \log \alpha_i - \log \beta_j,$$

where each couple (i, j) *bears the weight* $P_{i.} P_{.j}$. The results obtained by using this method on the table that Goodman has analysed are shown in Table D4.1.

Conclusions and perspectives

Correspondence analysis users do not usually speak of it in terms of a model as Goodman does. Perhaps they ought to admit that even if a method such as correspondence analysis is used for exploring data, it is based on an algebraic approach which defines the form of the solutions that one will be able to obtain.

There is in fact a model which can be explained by the following: the general form of possible solutions is known before the study, but the model is only a potential one since it is the results of the analysis that will provide the choice of the value of k.

Goodman insists on the practical interest of being able to impose constraints on the solutions, for instance $x_{i1} = x_{01} + i\Delta^*$. The presentation given above leads to solutions to the least-squares problems solved without constraints. However one could easily take constraints into account; the numerical problem of computing eigenvectors and eigenvalues would then be replaced by one of minimizing a function under certain constraints. It seems that this is probably a field to be explored.

Goodman concludes with the problem of choosing the right number of axes for a correspondence analysis study. In France the method most currently used consists in computing the proportion.

$$\sum_{i=1}^{k} \lambda_m^2 \Big/ \sum_{i=1}^{M} \lambda_m^2.$$

This proportion can also be read in terms of inertia, and as a consequence of formula (D4.7) it is an index of the quality of the approximation to $P_{ij}/P_{i.}P_{.j}$ by

$$\left(1 + \sum_{m=1}^{k} \lambda_m x_{im} y_{jm}\right).$$

Nishisato (1980, p. 42) proposes and uses Bock's adaptation of Bartlett's χ^2 approximation.

It might be preferable to use a bootstrap-inspired approach which would mean looking for the number of axes that keep stable through a number of new studies done by resampling in the original table. This seems quite a natural way of handling the choice because correspondence analysis is seen as a method for exploring data.

Additional references

Rao, C.R. (1980). Matrix Approximation and Reduction of Dimensionality in Multivariate Statistical Analysis. In *Multivariate Analysis,* **5,** Ed. P.R.Krishnaiah, pp. 3–22. Amsterdam: North Holland.

Sabatier, R., Jan, Y. & Escoufier, Y. (1984). Approximations d'Applications Linéaires et Analyse en Composantes Principales. In *Data Analysis and Informatics,* **3,** Ed. E. Diday et al., pp. 569–580. Amsterdam: Elsevier.

[*Received March* 1986, *revised June* 1986]