N° 8127

STATISTICAL SIGNIFICANCE CRITERIA
IN MULTIPLE-CHOICE DATA REDUCTION
AND VISUALIZATION

par

L. LEBART A. MORINEAU
STATISTICAL SIGNIFICANCE CRITERIA
IN MULTIPLE-CHOICE DATA REDUCTION AND VISUALIZATION

L. LEBART
C.N.R.S.
140, rue du Chevaleret - 75013 PARIS - FRANCE.

A. MORINEAU
CEPREMAP

Correspondence Analysis (also called Reciprocal Averaging, or Dual Scaling, or Canonical Analysis of Contingency Tables, according to various authors: GUTTMAN (1941), HAYASHI (1950), BENZECRI (1969), NISHISATO (1980), etc.), provides scales and graphical descriptions of the relations between two or more multiple-choice variables, as well as of similarities between subjects. We are concerned here with the validity of the results.

We shall focus on the laws of the eigenvalues produced by correspondence Analysis (in both case of contingency tables and multiple contingency tables) under the hypothesis of independence. We shall then briefly emphasize the inability of the percentages of variance to give a measurement of the quality of the results in terms of information, with the help of some counter-examples. The third paragraph is devoted to the confidence areas one may draw over the points of the scattering diagrams: the radii of such circular areas provide a valuable aid for the selection of the most relevant items in the processing of large binary data arrays.

I. THE INDEPENDENCE HYPOTHESIS (case of correspondence analysis)

The hypothesis of independence of the rows and columns of a table will generally be too restricted a hypothesis to be realistic. It is extremely
improbable that a table subjected to the analysis can be the analogue of of a random numbers table.

Although it is an extreme case with limited practical scope, the independence hypothesis will nevertheless allow us to define significance levels for the eigenvalues and the percentages of variance, which will act as a "brake" for the user.

As the eigenvalues follow nonparametric laws in the case of correspondence analysis of contingency tables, it has been possible to proceed to approximate tabulations, and the charts which summarize them have been plotted.

a) The law of eigenvalues

The law of eigenvalues produced by correspondence analysis has given rise to many an erroneous publication. Thus in the treatise on statistics of M.G. KENDALL and A. STUART (1961), the eigenvalues, like the total inertia, are assumed to follow $\chi^2$ laws.

H.D. LANCASTER (1963) has refuted this result by showing that the mathematical expectation of the first eigenvalue is always higher than the values derived from the assertions of M.G. KENDALL and A. STUART.

As a matter of fact, it can be shown (LEBART (1976), CORSTEN (1976), O'NEIL (1978)) that the law is related to the WISHART distribution in the following sense:

If $\lambda_\alpha$ is the $\alpha$th eigenvalue produced by the correspondence analysis of table $K$ of order $(n,p)$, with total sum $k$, then the distribution of $k\lambda_\alpha$ is approximately that of the $\alpha$th eigenvalue of a WISHART(*) matrix with parameters $(n-1)$ and $(p-1)$. ("FISHER-HSU" law).

(*) The probability density of the eigenvalues produced by a WISHART matrix has been explicated in 1939 by FISHER, GIRSHICK, HSU and ROY. The integration of this rather complex density has given rise to several publications; among the principal ones, those of K. PILLAI (1965), P.R. KRISHNAIAH and T.C. CHANG (1971), which are inspired by the works of the physicist M.L. MEHTA (1967). A table of the levels corresponding to two extreme eigenvalues has been published by CHOUDARY HANUMARA and THOMSON (1968) for matrices having their smallest side $p$ less than 10; by PILLAI and by CLEMM, KRISHNAIAH, and WALKER (1973) for $p \leq 20$. 
To show this result, we use an approximation analogous to that which is made in the establishment of the \( \chi^2 \) law relating to a contingency table: \( k \) will be assumed sufficiently large to allow the use of the normal approximation of the multinomial law.

The whole number \( k_{ij} \) being the general term of the contingency table \( K \) with \( n \) rows and \( p \) columns, we denote:

\[
    k = \sum_{i,j} k_{ij} \quad \text{and} \quad f_{ij} = \frac{k_{ij}}{k}.
\]

If \( p_{ij} \) denotes the probability corresponding to the cell \((i,j)\) and if the theoretical marginal totals are denoted by \( p_i \) and \( p_j \), the hypothesis of independence of the rows and columns is rendered by the relation:

\[
    p_{ij} = p_i p_j.
\]

We will assume besides that the observed marginal totals \( f_i \) and \( f_j \) can be substituted without consequence for the theoretical marginal totals \( p_i \) and \( p_j \) (without, however, disregarding the constraints implied by this substitution). Moreover these hypotheses allow the classical \( \chi^2 \) test to be found.

We have published approximate tables (Lebart, 1975) for contingency tables whose dimensions do not exceed \( 50 \times 100 \), and relating to the first five eigenvalues and corresponding rates of variance (estimates of the means, standard deviations, and unilateral 0.05 level for these quantities). Figure 1 summarizes a part of the results for the first eigenvalue. We read for instance on Figure 1 that, for a \((10,10)\) table, the first eigenvalue can attain 45% of the variance in 5% of the cases, in the hypothesis of independence of the rows and columns of the table.

b) Independence of the percentages of variance and of the trace

Let us note \( t \) the sum of the non-trivial eigenvalues: \( t = \sum_{\alpha=1}^{p-1} \lambda_\alpha \) and \( t_\alpha \) the percentage of variance: \( t_\alpha = \frac{\lambda_\alpha}{t} \).
Figure 1.
Upper percentage point (0.05) of the largest root.

Upper percentage point "percentages of variance"

\[ f \] = number of columns

Number of rows
If $k$ still denotes the sum of all the cells of the $(n,p)$ table $K$, it is well known that $kt$ is nothing but the classical chi-square with $(p-1)(n-1)$ df.

The following property holds, for the FISHER-HSU law: the percentages of variance $t_1, t_2, \ldots, t_{p-1}$ are independent of the trace $t$.

This property appears still valid in the case of correspondence analysis, for which the WISHART law is only an approximate law (the extensive simulations undertaken for constructing the charts have allowed this independence, which had been conjectured furthermore from empirical results) to be verified.

Thus, even if the trace does not allow the independence hypothesis to be rejected (the usual $\chi^2$ test), the first percentage of variance could nevertheless be significantly elevated: correspondence analysis could be used even for tables that the $\chi^2$ does not designate as being very rich in information.

Conversely, nonsignificant percentages of variance could correspond to a significantly elevated trace. Although the independence hypothesis is rejected, correspondence analysis is perhaps not the best tool in such a case to describe the dependence between the rows and columns of the table.

II. PERCENTAGES OF VARIANCE AND INFORMATION

Correspondence Analysis is used to describe the association between rows and columns of a contingency table as well as response pattern tables or other types of binary arrays. Apart from the case of contingency tables, the use of the percentages of variance is problematic. Some counter-examples will show that these coefficients are not appropriate to characterize satisfactorily the quality of a representation.

a) The case of disjunctive coding (response pattern table)

It is well known that, for the same representation, the analysis of two variables under disjunctive coding could yield considerably lower percentages
of variance than the analysis, nevertheless equivalent, of the contingency
table crossing the two variables. Therefore the percentages of variance
thus give a very pessimistic idea of the part of the information represented.

A transformation of the eigenvalues and of the trace is needed to make
them more meaningful (BENZECRI, 1979).

b) Case of analysis of the matrix associated with a symmetric graph
In several cases an exact analytical calculation can be made without
recourse to the computer: It is then interesting to study analytically
the variations of the representations as a function of the different
codings of the adjacency matrix. Let us denote by \( n \) the number of
vertices of the graph, and by \( M \) the adjacency matrix of the graph.
\( m_{ij} = 1 \) if there is an edge between the vertices \( i \) and \( j \); 
\( m_{ij} = 0 \) otherwise).

Let us examine for example the case of the analysis of a simple cycle. The
matrix \( M \) has only two nonzero elements (equal to 1) per row and per
column.

For \( n = 6 \), we have for example:

\[
M = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
\]

Since the sums of the elements of rows and columns are equal, the principal
axis \( x_\alpha \) is calculated by:

\[
\frac{1}{4} M^2 x_\alpha = \lambda x_\alpha
\]

It can be obtained directly from:

\[
\frac{1}{2} M x_\alpha = \epsilon \sqrt{\lambda} x_\alpha \quad (\epsilon = +1 \text{ or } -1)
\]

The preceding relation is again written, for \( 1 < j < n \):

\[
\frac{1}{2} (x_{\alpha,j-1} + x_{\alpha,j+1}) = \epsilon \sqrt{\lambda} x_{\alpha,j}
\]

The solution of this classical type of difference equations is, taking
into account the limit conditions:
and also
\[
\begin{align*}
    x_{\alpha,j} &= \cos \frac{2\alpha j \pi}{n} \\
y_{\alpha,j} &= \sin \frac{2\alpha j \pi}{n}
\end{align*}
\]  
(1)

These two vectors corresponding to the same eigenvalue:
\[\lambda_\alpha = \cos^2 \frac{2\alpha \pi}{n}\]

we obtain in the plan of the first two axes (\(\alpha = 1\)) the equation of a circle (Figure 2) and thus a satisfactory reconstitution of the structure, of which table \(M\) represents a particular coding.

![Figure 2](image)

The trace of the matrix to be diagonalized is written:
\[\text{tr} \frac{1}{4} M^2 = \frac{n}{2}\]

The percentage of variance corresponding to axis \(\alpha\) is thus:
\[t_\alpha = \frac{2}{n} \cos^2 \frac{2\alpha \pi}{n}\]

The result, seemingly paradoxical, is as follows: the percentage of variance of the subspace which restores the initial structure can be rendered as small as possible, provided that \(n\) is large enough, i.e. that a rather long cycle is chosen (if \(n = 10^3\), \(t_1 \leq 2 \times 10^{-3}\)). Other examples of problematic results could easily be found.

REMARK: We could see for example (LEBART et al., 1977) that the SHANNON-WIERNER information theory does not make easily visible the
percentage of variance as a measure of the information related to the corresponding axis. The JEFFREY's divergence allows us to express the distance between hypothesis of independence and general case as a function of the eigenvalues. Unfortunately, it involves small eigenvalues, whereas correspondence analysis (or related methods of scaling) retains only the large ones.

III. CONFIDENCE AREAS FOR POINTS ON GRAPHICAL DISPLAYS

The case of \((n,p)\) contingency tables

Correspondence analysis can be presented as a search for the principal axes of a multidimensional display of profile points. Each of the \(n\) rows of the contingency table can be represented as a point \(z_i\) with \(p\) coordinates \(z_i^j\) \((j \leq p)\).

Namely, (with the above notations): \(z_i^j = \frac{f_{ij}}{f_i}\). The row \(z_i\) is weighted by \(f_i\).

The mean-point \(g\) of the \(n\) profiles \(z_1, \ldots, z_n\) is a vector with coordinates \(\frac{f_{.j}}{f_{.}}\) \((j \leq p)\).

The so-called Chi-square distance between \(z_i\) and \(g\) is noted \(d(z_i, g)\)

\[
\begin{align*}
\text{with:} & & d^2(z_i, g) = \sum_{j=1}^{p} \frac{1}{f_{i,j}} (z_{i,j} - g_j)^2 \\
\text{that is:} & & d^2(z_i, g) = \sum_{j=1}^{p} \frac{1}{f_{i,j}} \left(\frac{f_{i,j}}{f_i} - f_{.j}\right)^2 \\
\end{align*}
\]

If \(k\) denotes the total sum of the elements of the contingency table, the quantity \(c_i^2 = k \frac{d^2(z_i, g)}{f_{i,.}}\) can be written:

\[
c_i^2 = k \frac{(f_{i,j} - f_{.j} f_{.,j})^2}{f_{i,.} f_{.,j}}
\]

The quantity \(c_i^2\) approximately follows a chi-square law with \((p-1)\) d.f., if the row "!" is supposed to be filled according to a multinomial law, the theoretical probabilities for each cell being defined by the margins. (In other words: \(z_i\) only differs from \(g\) on account of sampling fluctuation).
In projection onto any 2 dimensional subspace containing \( g \) (e.g.: the subspace of the first two principal axes of correspondence analysis), the squared distance will follow a chi-square law with two d.f. (the idempotent matrix of projection being of rank 2).

This leads to a simple procedure to test the significance of the position of certain points on the graphic displays.

We can draw a confidence circle centered at the origin with radius:

\[
    r = \sqrt{\frac{5.99}{k_f}}
\]

(5.99 is the value given by the tables for \( p = 0.05 \) and d.f. = 2).

The projection of \( z_i \) will fall outside this circle with the probability 0.05, if the \( i \)th row of the table is statistically equivalent to the margin.

In practice, instead of drawing concentric circles around the origin, it is clearer and easier to draw them around each point concerned, and look at the position of the origin (see example below).

The property of reduction of the d.f. from \((p-1)\) to 2 only holds if the two dimensional subspace is fixed in advance. However, it remains valid if the concerned points \( z_i \) do not participate much in the construction of the principal axes (cases of small weighted points, of points closed to the origin, or of supplementary points, i.e.: plotted afterwards).

IV. NUMERICAL APPLICATION

An example will be shown which deals with a \((12,8)\) contingency table crossing two items out of a large questionnaire survey: the first item (socio-economic categories) has 8 options, the second item (sources of information about environmental problems) has 12 options of response.

*Table 1* shows that certain cells contain small numbers, likely to induce nonsignificant positions on the principal axes display.
### TABLE 1

**TABLE CROSSING**

12 SOCIO-ECONOMIC CATEGORIES  
8 SOURCES OF INFORMATION  
ABOUT ECOLOGICAL PROBLEMS  
(2328 SUBJECTS)

<table>
<thead>
<tr>
<th></th>
<th>T.V.</th>
<th>NEWS PAPER</th>
<th>RADIO</th>
<th>BOOKS</th>
<th>FRIENDS</th>
<th>CITY HALL</th>
<th>ASSOCIATIONS</th>
<th>MANY SOURCES</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FARMER</td>
<td>26</td>
<td>18</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>38</td>
<td>114</td>
</tr>
<tr>
<td>MANAGER</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>CRAFTSMAN</td>
<td>36</td>
<td>26</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>13</td>
<td>99</td>
</tr>
<tr>
<td>HIGH, EXECUTIVE</td>
<td>15</td>
<td>44</td>
<td>4</td>
<td>13</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>21</td>
<td>108</td>
</tr>
<tr>
<td>PROFESSIONAL</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>EXECUTIVE</td>
<td>44</td>
<td>87</td>
<td>4</td>
<td>39</td>
<td>13</td>
<td>3</td>
<td>14</td>
<td>38</td>
<td>240</td>
</tr>
<tr>
<td>WHITE COLLAR</td>
<td>83</td>
<td>87</td>
<td>13</td>
<td>24</td>
<td>19</td>
<td>1</td>
<td>5</td>
<td>43</td>
<td>275</td>
</tr>
<tr>
<td>BLUE COLLAR</td>
<td>181</td>
<td>107</td>
<td>16</td>
<td>31</td>
<td>41</td>
<td>7</td>
<td>7</td>
<td>70</td>
<td>460</td>
</tr>
<tr>
<td>STUDENT</td>
<td>14</td>
<td>43</td>
<td>1</td>
<td>17</td>
<td>15</td>
<td>0</td>
<td>5</td>
<td>21</td>
<td>116</td>
</tr>
<tr>
<td>HOUSEWIFE</td>
<td>158</td>
<td>87</td>
<td>12</td>
<td>33</td>
<td>20</td>
<td>5</td>
<td>17</td>
<td>72</td>
<td>404</td>
</tr>
<tr>
<td>RETIRED</td>
<td>167</td>
<td>95</td>
<td>29</td>
<td>15</td>
<td>15</td>
<td>7</td>
<td>7</td>
<td>88</td>
<td>424</td>
</tr>
<tr>
<td>UNEMPLOYED</td>
<td>27</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>51</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td>756</td>
<td>817</td>
<td>100</td>
<td>191</td>
<td>142</td>
<td>36</td>
<td>70</td>
<td>416</td>
<td>2328</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>EIGENVALUES</th>
<th>PERCENTAGES</th>
<th>CUMULATED PERCENTAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.069241</td>
<td>57.32</td>
<td>57.32</td>
</tr>
<tr>
<td>2</td>
<td>0.022702</td>
<td>18.80</td>
<td>76.12</td>
</tr>
<tr>
<td>3</td>
<td>0.011555</td>
<td>9.57</td>
<td>85.69</td>
</tr>
<tr>
<td>4</td>
<td>0.009719</td>
<td>8.05</td>
<td>93.73</td>
</tr>
<tr>
<td>5</td>
<td>0.005081</td>
<td>4.21</td>
<td>97.94</td>
</tr>
<tr>
<td>6</td>
<td>0.002013</td>
<td>1.67</td>
<td>99.61</td>
</tr>
<tr>
<td>7</td>
<td>0.000475</td>
<td>0.39</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td>0.120767</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3 is the scattering diagram produced by correspondence analysis.

a) Trace, percentage of variance
The total sum $k$ is equal to 2328 (number of interviewed subjects). The trace is equal to $t = 0.1208$.
Thus, $kt = 281$.
This value is highly significant, for a chi-square variable with 77 d.f.
The hypothesis of independence is rejected, and correspondence analysis allows us to understand why it is rejected, by pointing out the network of relationship between the rows and the columns of the table.
For the first axis, the percentage of variance is 57.3%.
The statistical table summarized in Figure 1 gives the value 45.8%, corresponding to the confidence level 0.05.
Therefore, this axis is highly significant. Without knowledge about the conditional law of the percentage corresponding to the second axis, it is not possible to easily appreciate its significance.

b) Confidence areas for points
Using the formula of the previous paragraph, four confidence circles have been drawn.
The source of information "RADIO" occupies a significant position (so do all the sources, in this example).
Despite its large distance from the origin, the point "MANAGER" (like the "WHITE-COLLAR" point) does not differ significantly from it (confidence level 0.05).
FIGURE 3
CORRESPONDENCE ANALYSIS

12 SOCIO-ECONOMIC CATEGORIES
8 SOURCES OF INFORMATION
2328 SUBJECTS
REFERENCES


