A SAS PROGRAM FOR CORRESPONDENCE ANALYSIS USING PROC MATRIX

Bernard Clément, Ecole Polytechnique de Montréal
Daniel Lévesque, Ecole Polytechnique de Montréal

ABSTRACT

Correspondence analysis is primarily a technique for the simultaneous display of the rows and columns of a two-way contingency table in a low-dimensional vector space. Apart from contingency tables, the analysis can be applied to a table of positive numbers or to a suitably prepared table derived from one. As such, it is an alternative well worth considering to the more conventional methods of principal components and factor analysis. Noteworthy features of correspondence analysis are the well-developed aids for interpretation.

Since no correspondence analysis procedure existed in SAS, we undertook to write a correspondence analysis program for SAS users. The program is run under PROC MATRIX. It is easy to use and has a wide choice of options available, especially for the graphical output.

In this paper, an overview of correspondence analysis is presented and the general features of our program are given followed by an example.

CORRESPONDENCE ANALYSIS: AN OVERVIEW

The French statisticians have stressed the importance of a geometric approach to exploratory multidimensional analysis over the more classical approach of inferential statistics by hypothesis testing. One of the most popular methods they use is correspondence analysis and its strongest advocate is J.P. Benzecri (1969, 1979, 1980). The technique is theoretically equivalent to a number of statistical techniques: canonical correlation (a special case) (Hotelling 1936), contingency table analysis (Hirschfeld 1935), reciprocal averaging (Hill 1975), and dual scaling (Nishisato 1980).

Correspondence analysis should be thought of as a special case of factor analysis or as a technique for the simultaneous display of the rows and columns of a two-way table in a low-dimensional vector space. One presentation of the method uses a terminology drawn from geometry, mechanics and statistics: profile, mass, inertia, center of gravity and chi-square distance.

Let $N = \{n_{ij}\}$ be an $r \times c$ contingency table i.e. $n_{ij}$ is the joint frequency count of sampling units having characteristics $i$ and $j$ on a pair of qualitative variables with values identified by the cells $(i,j)$ of a two-way contingency table:

\[
\begin{array}{c|ccc}
| & j & & c \text{ total} \\
\hline
1 & \vdots & & \\
\vdots & \ddots & \vdots & \vdots \\
i & n_{ij} & \cdots & n_i. \\
\hline
r & & & \\
\hline
\text{total} & n_j & & n \\
\end{array}
\]

Let

\[
\begin{align*}
n_i & = \sum_j n_{ij} \\
n_j & = \sum_i n_{ij} \\
n & = \sum_i \sum_j n_{ij} \\
f_{ij} & = \frac{n_{ij}}{n} \\
f_i & = \frac{n_i}{n} \\
f_j & = \frac{n_j}{n} \\
f & = \frac{n}{n}
\end{align*}
\]

Each row of $N$ or $F$ is represented by its row-profile $f_i^c$ (conditional frequencies) in $R^c$ space:

\[
f_i^c = (f_i^1, \ldots, f_i^c), \quad i = 1, \ldots, r
\]

with mass $f_i$. Similarly each column of $F$ is represented by its column-profile $f_j^r$ in $R^r$ space:

\[
f_j^r = (f_j^1, \ldots, f_j^r), \quad j = 1, \ldots, c
\]

with mass $f_j$.

The squared distance between two rows-profiles $f_i^c$ and $f_j^c$ is defined as

\[
d_i^2 = \frac{1}{c} \sum_{j=1}^{c} (f_i^j - f_j^j)^2
\]

where $D_c = \text{diag}(f_1, f_2, \ldots, f_c)$.

The square distance between two column-profiles $f_i^r$ and $f_j^r$ is

\[
d_j^2 = \frac{1}{r} \sum_{j=1}^{r} (f_j^i - f_i^i)^2
\]

where $D_r = \text{diag}(f_1, f_2, \ldots, f_r)$.

This choice of distance, called chi-square distance, is a fundamental ingredient of the method. The matrices $D_c^2$ and $D_r^2$ are the metrics in both spaces and distinguish this method from principal component analysis where the distance is euclidian and the associated metric is defined by the identity matrix. There are several reasons for the choice of a chi-square metric:

- symmetry: the analysis can be done in either space with similar results. This method is responsible for the recognition of the duality concept in statistical analysis.

- elimination of size effect: the analysis will not be dominated by a few rows (or columns) with large values of $n_i$ (or $n_j$). These will contribute equally with rows (or columns) with small values of $n_i$ (or $n_j$).
distributional equivalence: the distance between two columns remains invariant when we merge two rows with identical profiles. This has a stabilising influence: adding or eliminating a row with a nearly identical profile to an already existing row will not change the analysis.

The set of row-profiles is contained in a hyperplane defined by the equation
\[
\sum_{j=1}^{c} z_{j} a_{j} = 1
\]
with center of gravity \( z_{c} \).

After translating the origin to the center of gravity and taking into account the scaling factor introduced by the metric, the coordinates of the row-point can be written:
\[
f_c = (f_{j1}, \ldots, f_{jc})^T.
\]

Let \( \mathbf{u} = (u_1, \ldots, u_c)^T \) be a unit vector through \( z_c \) and \( a_1 \) the orthogonal projection of \( f_c \) onto \( \mathbf{u} \). Then
\[
a_1 = \frac{c}{\sqrt{\sum_{j=1}^{c} f_{j1}^2}} (f_{j1}, \ldots, f_{jc}) \mathbf{u}_j.
\]

We have to determine \( \mathbf{u} \), called the first principal axis (factor) of inertia so as to maximise the moment of inertia
\[
\max \sum_{i=1}^{c} f_{i1}^2 a_i^2.
\]

The solution is obtained by finding the eigenvector \( \phi \) associated with the largest eigenvalue \( \lambda_1 \) of the symmetric \( c \times c \) matrix \( T \),
\[
T = XX^T, \quad \text{where} \quad X = \left[ f_{i1}, \ldots, f_{ic} \right].
\]

This process can be continued. We obtain the factor space solution of row-profiles by finding the sequence of decreasing eigenvalues,
\[
\lambda_1 > \lambda_2 > \ldots > \lambda_{c-1} > \lambda_c,
\]
with their corresponding orthogonal vectors
\[
\phi_1, \phi_2, \ldots, \phi_{c-1}, \phi_c.
\]

It can be easily shown that \( \lambda_c = 0 \) is an eigenvalue with associated eigenvector \( \phi_c = z_c \).

The coordinate \( a_{ki} \) of a row-profile on axis \( k \) is
\[
a_{ki} = \frac{c}{\sqrt{\sum_{j=1}^{c} f_{j1}^2}} f_{ij},
\]
and the coordinates \( b_{kj} \) of column-profile is
\[
b_{kj} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{c} (f_{ij}) a_{ki}.
\]

A similar analysis could have been done in \( R^r \) space by working with the column-profiles and this would lead us to the eigenvalue problem for the \( r \times r \) matrix \( W \)
\[
W = XX^T.
\]

It is a well known result the matrix \( W \) has the same positive eigenvalues as \( T \). For obvious, practical reasons, the eigenvalue problem is solved in the space of smaller dimension.

**DATA TYPES TO BE SUBMITTED TO THE ANALYSIS**

Besides a contingency table, several other data matrix structure can be submitted to a correspondence analysis since the only requirement is a matrix with positive numbers. Depending on the type of data, a derived table is submitted to the analysis. Two important operations in this respect are the complete disjunctive form and the doubling of the data. Here are some typical data structure:

**a) Table of answers to a questionnaire**

Let
\[
x_{ij} = \text{answer } i \text{ on question } j,
\]
where each question has a fixed set of possible choices. The initial table of answers \( X \) is transformed into the complete disjunctive form. The idea is to generate as many dichotomous 0-1 variables as there are choices on all the questions. Suppose a questionnaire has 3 questions such that:

question 1 has 2 choices: Y or N
question 2 has 3 choices: A or B or C
question 3 has 4 choices: 1 or 2 or 3 or 4

then the original 3 columns matrix \( X \) is expanded into a 9 columns \( N \). For example,
\[
X = \begin{bmatrix} Y & B & 4 \\ N & A & 1 \\ Y & C & 2 \end{bmatrix}, \quad N = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.
\]

The matrix \( N \) is then submitted to the analysis.

**b) Table of ratings**

Let
\[
x_{ij} = \text{rating } x_j \text{ on subject } i
\]
where \( 0 \leq x_{ij} \leq c \) is a bounded scale variable.

For this type of data, it is recommended that the initial \( X \) matrix is "doubled" by adding the complementary value \( x'_{ij} \) of \( x_{ij} \):

\[ x'_{ij} = c - x_{ij} \]

The analysis is then performed on the \( r \times 2c \) matrix \( N \):

\[ N = [x_{ij} \ x'_{ij}] \]

c) Table of homogenous positive measurements

The table as such can be submitted to the analysis if all variables use the same measurement units.

d) Table of heterogenous variables

With a mixture of qualitative and quantitative measurements, the initial table has to undergo preliminary transformations before analysis:

- qualitative variables are put in disjunctive form,
- positive continuous variables on a bounded scale are "doubled",
- variables that can take negative values are "doubled" by replacing them with their positive part and their negative part; i.e., \( x_{ij} \) is replaced by \( (x^+_{ij}, x^-_{ij}) \) where

\[
x^+_{ij} = \begin{cases} x_{ij} \text{ if } x_{ij} > 0 \\ 0 \text{ if } x_{ij} < 0 \end{cases}
\]

\[
x^-_{ij} = \begin{cases} 0 \text{ if } x_{ij} > 0 \\ -x_{ij} \text{ if } x_{ij} < 0 \end{cases}
\]

Successful application of correspondence analysis will be due, in part, to the preliminary transformation performed on the data matrix before the analysis.

AIDS FOR INTERPRETATION

The following quantities should be included in any program of correspondence analysis:

- table and histogram of the eigenvalues,
- quality of the representation expressed as a percentage of the total inertia, \( \frac{\lambda_1}{\lambda_1} \),
- relative inertia of a row-point:

\[ \text{INER}(i) = \frac{d^2(i, g_0)}{\sum_{i=1}^{c} \sum_{j=1}^{r} d^2(i, j)} \]

where

\[ d^2(i, g_0) = \frac{\sum_{j=1}^{r} x_{ij} (\frac{1}{\lambda_1} x^+_{ij} + \frac{1}{\lambda_2} x^-_{ij})}{\sum_{j=1}^{r} x_{ij}} \]

- coordinates of a row-point:

\[ (a_{i1}, a_{i2}, \ldots, a_{ic-1}, a_{ic}) \]

- relative contribution (CTREL) expressed as a percentage of the total inertia:

\[ \text{CTREL}(i) = \frac{\sum_{k=1}^{m} \cos^2(i, k) \text{CTREL}(i, k)}{\sum_{k=1}^{m} \text{CTREL}(i, k)} \]

- square cosines (\( \cos^2 \)) of the angles between elements and the axis (this shows which elements are well represented by the axis):

\[ \cos^2(i, k) = \frac{\sum_{j=1}^{r} x_{ij} (\frac{1}{\lambda_1} x^+_{ij} + \frac{1}{\lambda_2} x^-_{ij})}{\sum_{j=1}^{r} x_{ij}} \]

Note: similar quantities are computed for column-point.

- axis interpretation:

To interpret an axis, find the points with positive coordinates and largest C'TREL values, identify those points with negative coordinates and large C'TREL, and then express, in a concise manner, the opposition between these two groups. Points with C'TREL in excess of, say, 0.50 should be made supplementary so as not to contribute to the solution of the eigenvalue problem. In other words, we want the factor space to be determined by the bulk of the data and not by a few points. Usually between 2 and 4 factors explain between 70 and 90% of the data.

THE SAS PROGRAM

The french statistical literature contains several well known computer programs written in FORTRAN for correspondence analysis. For example, AC103 in Lebart, Morineau, Tabard (1977) or TABET in Benzécri (1980). In the 1983 SUGI Proceedings, J.M. Gautier proposed a library of french methods in a package called SYSPAD. SYSPAD is available only as a licensed product through COREF in France and is not readily available in U.S.A. or Canada.

We undertook to write a correspondence analysis program that any user of SAS could immediately implement and run. The program permits a wide range of options, especially for the graphical output. The computer code contains 915 lines and is too long to be included in this paper (however it is available from the senior author at the address given at the end of this paper).
The user must create 6 SAS datasets containing the user controls the choice of axes, the width of the graph (1 page or 2 page wide), the length of the graph, and whether or not far-away points are put on the frame. The points are put on the graph with 3 characters and superimposed points are identified.

**EXAMPLE:** Top US industrial companies in 1978.

The data represent the financial results of the top 20 US industrial companies in 1978 are taken from Lowi (1982), p. 12 where

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAL</td>
<td>Sales</td>
<td>1229</td>
</tr>
<tr>
<td>NPR</td>
<td>Negative Profit</td>
<td>22101</td>
</tr>
<tr>
<td>CAP</td>
<td>Capital</td>
<td>11049</td>
</tr>
<tr>
<td>GMO</td>
<td>General Motors</td>
<td>11861</td>
</tr>
<tr>
<td>FOR</td>
<td>Ford</td>
<td>7455</td>
</tr>
<tr>
<td>TEX</td>
<td>Texaco</td>
<td>8730</td>
</tr>
<tr>
<td>IBM</td>
<td>IBM</td>
<td>2296</td>
</tr>
<tr>
<td>GUL</td>
<td>Gulf</td>
<td>1018</td>
</tr>
<tr>
<td>ITT</td>
<td>ITT</td>
<td>2831</td>
</tr>
<tr>
<td>ATL</td>
<td>Atlantic Richfield</td>
<td>11110</td>
</tr>
<tr>
<td>USS</td>
<td>US Steel</td>
<td>6586</td>
</tr>
<tr>
<td>NES</td>
<td>Western Electric</td>
<td>423</td>
</tr>
<tr>
<td>TEN</td>
<td>Tenneco</td>
<td>8099</td>
</tr>
</tbody>
</table>

**CORRESPONDENCE ANALYSIS**

**TABLE OF ACTIVE EIGENVALUES**

<table>
<thead>
<tr>
<th>No</th>
<th>Lambda</th>
<th>Percent</th>
<th>Cumul</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.032</td>
<td>53.999</td>
<td>53.999</td>
</tr>
<tr>
<td>2</td>
<td>0.012</td>
<td>25.755</td>
<td>79.755</td>
</tr>
<tr>
<td>3</td>
<td>0.005</td>
<td>11.683</td>
<td>91.437</td>
</tr>
<tr>
<td>4</td>
<td>0.003</td>
<td>7.447</td>
<td>99.883</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>1.217</td>
<td>101.000</td>
</tr>
</tbody>
</table>

**EDITING OF COORDINATES SORTED BY CTRAEL**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mass</th>
<th>INER</th>
<th>COORD</th>
<th>CTRAEL</th>
<th>CSG2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASS</td>
<td>2.34</td>
<td>56.8</td>
<td>24.6</td>
<td>28.2</td>
<td>66.4</td>
</tr>
<tr>
<td>NPR</td>
<td>1.14</td>
<td>21.7</td>
<td>17.4</td>
<td>14.9</td>
<td>71.7</td>
</tr>
<tr>
<td>CAP</td>
<td>0.60</td>
<td>12.6</td>
<td>8.6</td>
<td>7.7</td>
<td>27.7</td>
</tr>
<tr>
<td>GMO</td>
<td>0.50</td>
<td>10.5</td>
<td>6.6</td>
<td>6.0</td>
<td>23.7</td>
</tr>
<tr>
<td>FOR</td>
<td>0.42</td>
<td>9.5</td>
<td>6.0</td>
<td>5.3</td>
<td>21.9</td>
</tr>
<tr>
<td>TEX</td>
<td>0.41</td>
<td>8.7</td>
<td>6.2</td>
<td>5.9</td>
<td>21.8</td>
</tr>
<tr>
<td>IBM</td>
<td>0.38</td>
<td>7.9</td>
<td>6.7</td>
<td>6.2</td>
<td>20.8</td>
</tr>
<tr>
<td>GUL</td>
<td>0.33</td>
<td>7.9</td>
<td>6.7</td>
<td>6.2</td>
<td>20.8</td>
</tr>
<tr>
<td>ITT</td>
<td>0.31</td>
<td>7.8</td>
<td>6.7</td>
<td>6.2</td>
<td>20.8</td>
</tr>
<tr>
<td>ATL</td>
<td>0.29</td>
<td>7.8</td>
<td>6.7</td>
<td>6.2</td>
<td>20.8</td>
</tr>
<tr>
<td>USS</td>
<td>0.28</td>
<td>7.8</td>
<td>6.7</td>
<td>6.2</td>
<td>20.8</td>
</tr>
<tr>
<td>TEN</td>
<td>0.27</td>
<td>7.8</td>
<td>6.7</td>
<td>6.2</td>
<td>20.8</td>
</tr>
</tbody>
</table>

**EDITING OF COORDINATES SORTED BY CTRAEL**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mass</th>
<th>INER</th>
<th>COORD</th>
<th>CTRAEL</th>
<th>CSG2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASS</td>
<td>9.7</td>
<td>1.7</td>
<td>0.3</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>NPR</td>
<td>8.3</td>
<td>1.4</td>
<td>0.8</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>CAP</td>
<td>7.7</td>
<td>1.2</td>
<td>0.7</td>
<td>0.4</td>
<td>1.2</td>
</tr>
<tr>
<td>GMO</td>
<td>7.5</td>
<td>1.1</td>
<td>0.7</td>
<td>0.4</td>
<td>1.2</td>
</tr>
<tr>
<td>FOR</td>
<td>7.4</td>
<td>1.0</td>
<td>0.7</td>
<td>0.4</td>
<td>1.2</td>
</tr>
<tr>
<td>TEX</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>IBM</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>GUL</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>ITT</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>ATL</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>USS</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
<tr>
<td>TEN</td>
<td>7.3</td>
<td>0.9</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
</tr>
</tbody>
</table>

---

**VARTYP=PRINC**

**VARTYP=IPRINC**

---
**EDITING OF COORDINATES SORTED BY CTVEL**

<table>
<thead>
<tr>
<th>VAR</th>
<th>MASS</th>
<th>INCR</th>
<th>COORD</th>
<th>CTREL</th>
<th>CUS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP</td>
<td>0.042</td>
<td>4.7</td>
<td>-61.6</td>
<td>63.4</td>
<td>81.4</td>
</tr>
<tr>
<td>NPR</td>
<td>0.000</td>
<td>13.7</td>
<td>-304.2</td>
<td>7.3</td>
<td>20.1</td>
</tr>
<tr>
<td>SAL</td>
<td>0.455</td>
<td>14.4</td>
<td>5.2</td>
<td>6.5</td>
<td>24.2</td>
</tr>
<tr>
<td>ASS</td>
<td>0.321</td>
<td>12.4</td>
<td>12.4</td>
<td>12.4</td>
<td>57.4</td>
</tr>
<tr>
<td>CAP</td>
<td>0.150</td>
<td>10.2</td>
<td>3.6</td>
<td>3.3</td>
<td>49.1</td>
</tr>
<tr>
<td>PPR</td>
<td>0.323</td>
<td>7.9</td>
<td>18.0</td>
<td>9.9</td>
<td>5.9</td>
</tr>
</tbody>
</table>

**GRAPH #1**

**AXIS:1 AXEY:2 TYPE: STEM & JOFT COR=+0.0 CORVOL=0.0 WIDTH= 1 P. WIDTH=50 SAMP=1**

**PROJECT PROJECTION PLAN VS XS ALONG THE REQUIRED AXIS**

**GRAPH #2**

**AXIS:1 AXEY:2 TYPE: STEM & JOFT COR=+0.0 CORVOL=0.0 WIDTH= 1 P. WIDTH=50 SAMP=1**

**POINTS DISTANT BY 3 STANDARD DEVIATIONS FROM AXIS ORIGINS (PUT ON THE FRAME)**

<table>
<thead>
<tr>
<th>POINT</th>
<th>COORDX</th>
<th>COORDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP</td>
<td>-63.6</td>
<td>29.7</td>
</tr>
<tr>
<td>NPR</td>
<td>-304.2</td>
<td>-144.4</td>
</tr>
<tr>
<td>CHR</td>
<td>-40.9</td>
<td>-16.4</td>
</tr>
</tbody>
</table>

**HORIZONTAL HISTOGRAMS OF MULTIPLE POINTS**

<table>
<thead>
<tr>
<th>LINE</th>
<th>APPROX. POSITION</th>
<th>HORIZ. HISTOGR.</th>
</tr>
</thead>
<tbody>
<tr>
<td>###50</td>
<td>(-18.0,-16.4)</td>
<td>NPR CHR</td>
</tr>
</tbody>
</table>
CONTACT AUTHORS

For further information about this paper contact:

Bernard Clément, Ph.D.
Département de mathématiques appliquées
Ecole Polytechnique de Montréal
C.P. 6079, Succursale "A"
Montréal, Québec H3C 3A7
Tel: (514) 344-4944

REFERENCES


Lewi, P.J. (1982). Multivariate Data Analysis in Industrial Practice, Research Studies Press, A Division of John Wiley & Sons Ltd.


