Computers & Geosciences Vol. 19, No. 7, pp. 1007–1022, 1993
Printed in Great Britain.

0098-3004/93 $6.00 + 0.00

CORRESPONDENCE ANALYSIS WITH MATLAB

DAQUAN TIAN,[1] SOROOSH SOROOSHIAN,[1] and DONALD E. MYERS[2]
[1]Department of Hydrology and Water Resources and [2]Department of Mathematics,
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 11 November 1992; accepted 18 December 1992)

Abstract—A portable program MATCORS, for correspondence analysis is presented. To help understand the correspondence analysis method and computer algorithm, a detailed matrix algebraic description is used. Various diagnostic results are provided in the program to aid in interpreting the results, for example relative and absolute contributions, error profiles, and supplementary elementary projections. A series of graphs generated by this program is helpful for the analysis of the results. High-quality graphs are produced easily because the program operates within the user-friendly Matlab environment. Applications to soil sample data and precipitation data sets are given to verify and demonstrate this computer program.

Key Words: Correspondence analysis, Eigenvalue and eigenvector decomposition, Matlab, Cluster analysis.

INTRODUCTION

Correspondence analysis (Benzecri, 1973) is a multivariate statistical analysis method that simultaneously produces R (variable) and Q (sample) mode analyses. Similar to principal component analysis (PCA), correspondence analysis (CA) uses principal factors to extract the significant information from the variable space and from the sample space by projecting into lower dimensional space and by quantifying the correlation between the variation and between samples. The principal factors $G_1, G_2, \ldots, G_n$ in the variable space (R^p) have the same contribution to the total variation as the principal factors $F_1, F_2, \ldots, F_p$ in the sample space (R^n). Unlike principal component analysis, correspondence analysis treats variables and samples in a symmetrical fashion. These principal factors can be used to represent graphically the variables and samples in a manner that geometrically describes the correlation structure. In correspondence analysis the principal factors for variables are dual to the principal factors for samples, hence it is possible to combine the plots using the same axes. Several aspects of the correlation structure of a data set may be seen in such graphical representations:

(1) Variables that plot close together are correlated more highly than those far apart. Note that variables that plot close together with one pair of principal factors may not for another pair hence close proximity may not indicate always high correlation.
(2) Samples that plot close together are correlated more highly than those far apart. Note that samples that plot close together for one pair of principal factors may not for another pair, hence close proximity may not indicate always high correlation.
(3) The combined plots provide some indication of the degree of correlation between a variable and a group of samples or between a sample and a group of variables but this graphical indication of correlation is less reliable than that between samples or between variables. This is discussed in greater detail in LeBart, Morineau, and Warrick (1984).

There are several commercial programs (Dual3, MAPWISE, PC-MDS, SimCA) for correspondence analysis, but none of these are portable, and do not incorporate high quality graphics (Goldstein, 1991). The object of this paper is to present a computer program that can perform correspondence analysis easily and quickly with high-quality graphic plots. It is written with Matlab (version 3.5), a software package for matrix operations which shares many of the characteristics of a high-level programming language. It operates interactively on a UNIX system (workstation) in the user-friendly Matlab environment. It also can be used on a personal computer (IBM, Macintosh, etc.). Drivers for different hardcopy devices are available within Matlab. There is almost no limitation on matrix size (depending on the machine memory). This program checks the input data matrix to ensure there are no negative entries. It also allows users to specify supplementary rows and columns. Thus the CA solution is determined by the active points only. The supplementary variables and samples then will be projected on the factorial loading plots. The plotting feature, of Matlab, provides for an easy and immediate visual inspection of the results. This particular attraction is the result of the fact that

correspondence analysis is a geometric approach to data analysis (Goldstein, 1991). The point label also is included on the plots and the high-quality output plots are acceptable for use in a presentation or paper. The output graphics can be saved as a PostScript file to take advantage of the capability of a laser printer.

METHOD

Let the input matrix be represented by the nonnegative entries x_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, p)$. Every element in the input matrix is divided by the sum of the $(n \times p)$ data elements to form a new matrix which may be referred to as the relative frequency matrix $F = X/\Sigma_{i=1}^{n} \Sigma_{j=1}^{p} x_{ij}$. The variable weights are defined by the row vectors $f_{0j} = \Sigma_{i=1}^{n} f_{ij}$, and sample weights by column vectors $f_{i0} = \Sigma_{j=1}^{p} f_{ij}$, respectively. There are two normalizing matrices, $D_n = \mathrm{diag}[f_{i0}]$ and $D_p = \mathrm{diag}[f_{0j}]$. Thus the sample coordinate matrix is obtained by weighting the frequency matrix as $D_n^{-1}F = f_{ij}/f_{i0}$. The coordinates of every sample in this matrix are in proportion to the values of the variables in the samples. So the relationship between variables can be represented by the relative location of n sample points in the variable space (R^p). The usual Euclidean distance between two variables l and m is given by:

$$d(l, m) = \sqrt{\sum_{j=1}^{p} \left(\frac{f_{lj}}{f_{l0}} - \frac{f_{mj}}{f_{m0}} \right)^2}. \tag{1}$$

Because the distance (1) gives the same weight to each variable, the chi-squared distance is used in place of the usual Euclidean distance (Rhodes and Myers, 1991).

$$d(l, m) = \sqrt{\sum_{j=1}^{p} \left(\frac{f_{lj}}{f_{l0}\sqrt{f_{0j}}} - \frac{f_{mj}}{f_{m0}\sqrt{f_{0j}}} \right)^2}. \tag{2}$$

When the principal axes for these distances are computed and plotted on the usual Cartesian system, the relationship between the samples and variables can be demonstrated visually.

The probability average of jth variable in the sample space (R^n) is:

$$\sum_{i=1}^{n} \frac{f_{ij}}{f_{i0}\sqrt{f_{0j}}} f_{i0} = \frac{1}{\sqrt{f_{0j}}} \sum_{i=1}^{n} f_{ij}. \tag{3}$$

Let $\mathbf{W} = (w_{ij})_{n \times p}$ and $\mathbf{A} = \mathbf{W}^T\mathbf{W}_{(p \times p)}$;

$$W_{ij} = \frac{f_{ij} - f_{i0}f_{0j}}{\sqrt{f_{i0}f_{0j}}} = \frac{x_{ij} - x_{i0}x_{0j} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}}{\sqrt{x_{i0}x_{0j}}}. \tag{4}$$

The factors $\mathbf{U}$ for variables can be computed by the eigen decomposition of the symmetric matrix $\mathbf{A}([\mathbf{A}] = [\mathbf{U}][\mathbf{D}][\mathbf{U}^T])$. $\mathbf{U}$ is a matrix of eigenvectors and a diagonal matrix $\mathbf{D}$ contains the eigenvalues. The principal axes for variables (R-mode factor loading) then are obtained from the scaled factors $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{U}$.

Similarly, the principal axes for samples (Q-mode factor loading) are $\mathbf{Q} = \mathbf{D}^{1/2}\mathbf{V}$, where $\mathbf{V}$ is obtained from eigen decomposition of symmetric matrix $\mathbf{B} = \mathbf{W}\mathbf{W}^T_{(n \times n)}$.

The original element can be reconstructed from a bilinear form (Avila and Myers, 1991):

$$f_{ij} = f_{i0}f_{0j}\left(1 + \sum_{k=1}^{p-1} \sqrt{d_k} R_{ik} Q_{jk}\right). \tag{5}$$

This program calculates several quantities that aid in the interpretation of the output (Avila and Myers, 1991).

(1) *The cumulative percentage of variation*: it is a global measure of fit when K factors are retained. The term variation in here does not refer to "variance" in the normal statistical sense. It is expressed as a cumulative percentage of explained variation which is similar to the measure used in principal component analysis, and is given by:

$$\sum_{k}^{K} d_k \Big/ \sum_{k}^{p-1} d_k. \tag{6}$$

(2) *Relative contributions*: these are measures of sample or variable variation explained by a particular factor. If all factors are retained, the sum is equal to 100 for a particular sample or variable. It is computed by:

For every variable $j = 1, \ldots, p$

$$RC^k(j) = d_k u_{jk}^2 \Big/ \sum_{l=1}^{p-1} d_l u_{jl}^2 \quad k = 1, \ldots, p-1 \tag{7}$$

and, for every sample $i = 1, \ldots, n$

$$RC^k(j) = d_k v_{jk}^2 \Big/ \sum_{l=1}^{p-1} d_l v_{jl}^2 \quad k = 1, \ldots, p-1. \tag{8}$$

(3) *Absolute contributions*: these are the contributions of variables or samples to a factor. For a particular factor it is equal to 100.

For every factor $k = 1, \ldots, p-1$

$$AC^k(j) = (f_{i0}(j))(\mathrm{diag}(f_{0j}(j))^{-1}u(j, k))^2 \quad \text{variable } j = 1, \ldots, p \tag{9}$$

and

$$AC^k(i) = (f_{0j}(i))(\mathrm{diag}(f_{i0}(i))^{-1}u(i, k))^2 \quad \text{sample } i = 1, \ldots, n. \tag{10}$$

(4) *Error profiles*: these provide a measure of the errors when the original data matrix is "reconstructed" by K principal factors. They are defined as:

Error profile for variable $j = 1, \ldots, p$

$$EP(j) = \sum_{i=1}^{n} f_{i0}\left(\sum_{k=K+1}^{p-1} \sqrt{d_k} v_{ik} u_{jk}\right)^2. \tag{11}$$

Error profile for sample $i = 1, \ldots, n$

$$EP(i) = \sum_{i=1}^{n} f_{0j}\left(\sum_{k=K+1}^{p-1} \sqrt{d_k} v_{ik} u_{jk}\right)^2. \tag{12}$$

(5) *Supplementary elements*: these are inactive elements in the data matrix, as they are not used in *CA* to determine the factors. Their projections onto the factors determined only by the active elements are computed by the formula:

For a supplementary row s

$$SP = \sum_{j=1}^{p} \frac{f_{sj}}{f_{s0}} u_{kj}. \tag{13}$$

For a supplementary column s

$$SV = \sum_{i=1}^{n} \frac{f_{is}}{f_{0s}} v_{ik}. \tag{14}$$

PROGRAM DESCRIPTION

A Matlab code, MATCORS, is composed of three functions as listed in Appendix 1. The main function MATCORS echoes with input options, checks negative entries in the input matrix, and calls the subfunction EIGV to form a symmetric matrix (A) and to perform eigen decomposition. Matlab built-in function (eig) can get the same eigenvectors, but opposite sign. If supplementary elements are selected, it will delete the supplementary elements to form an active data matrix, and calculate the supplementary projections. It also enables EIGV to operate again with a new active data matrix. Then MATCORS obtains the factor loading for both variables and samples. The program calculates the number of principal factors determined by the cumulative percent of variation specified by the user. Then the program compares this to the number provided by the user. The larger one is used for the factor loading. Several quantities used for interpreting the results are also calculated by this program. These quantities include the variable and sample weights, the relative and absolute contributions for every factor, as well as error profiles for each variable and sample, etc. Finally, a series of graphics is generated by this program, which includes a two-dimensional factor loading plot, a combined plane for both Q-mode and R-mode, and bar graphics for variable and sample weights, their absolute contributions and error profiles.

The subfunction EIGV constructs a real symmetric covariance matrix of variables and performs eigen decomposition based on the Jacobi's orthogonal rotation method. The subfunction INDEX, which is called by EIGV, locates the row and column number of maximum off-diagonal element, as well as its absolute value in a real symmetric matrix.

The data matrix is read in free-format. Several parameters are input from the keyboard:

pre—cumulative percent variation to determine the number of principal factors retained.
nb—desired number of principal factors to be retained.
nc—the number of supplementary variables.
cc—list of supplementary variables.
nr—the number of supplementary samples.
cr—list of supplementary samples.
f1 & f2—two principal factors to be plotted.

The program output consists of the following matrices, which can be saved as eight MATLAB files or ASCII files for later use.

r—R-mode factor loading $p \times p$ matrix.
q—Q-mode factor loading $n \times p$ matrix.
ev—Eigenvalues, relative and cumulative variation explained by factors.

Table 1. Example 1 soil sample data (Luo and Xing, 1987)

Sample No.	Sand Content	Silt Content	Clay Content	Organic Matter	pH
1	77.3	13.0	9.7	1.5	6.4
2	82.5	10.0	7.5	1.5	6.5
3	66.9	20.6	12.5	2.3	7.0
4	47.2	33.3	19.0	2.8	5.8
5	65.3	20.5	14.2	1.9	6.9
6	83.3	10.0	6.7	2.2	7.0
7	81.6	12.7	5.7	2.9	6.7
8	47.8	36.5	15.7	2.3	7.2
9	48.6	37.1	14.3	2.1	7.2
10	61.6	25.5	12.9	1.9	7.3
11	58.6	26.5	14.9	2.4	6.7
12	69.3	22.3	8.4	4.0	7.0
13	61.8	30.8	7.4	2.7	6.4
14	67.7	25.3	7.0	4.8	7.3
15	57.2	31.2	11.6	2.4	6.3
16	67.2	22.7	10.1	3.3	6.2
17	59.2	31.2	9.6	2.4	6.0
18	80.2	13.2	6.6	2.0	5.8
19	82.2	11.1	6.7	2.2	7.2
20	69.7	20.7	9.6	3.1	5.9

c_ra—p by $2p-2$ variables contribution matrix; the first $p-1$ columns are the relative contributions, and the rest are the absolute contributions.

s_ra—n by $2p-2$ samples contribution matrix; the first $p-1$ columns are the relative contributions, and the rest are the absolute contributions.

evv—weights and error profiles of variables.

ess—weights and error profiles of samples.

sv—supplementary variable projections.

sp—supplementary sample projections.

APPLICATION

Two examples are presented in this section. Table 1 lists an example data set from Luo and Xing (1987) for which a detailed analysis is given to verify the program. Figure 1 shows the plots using the original data set, whereas Figure 2 is a graphic plot in which variable 4 and samples 4 and 8 are made supplementary. Appendix 2 lists the output for this example. Both the factors loading matrix and factor plane plot are identical to Luo's results, except for the supplementary projection plots (his program does not have this type of function).

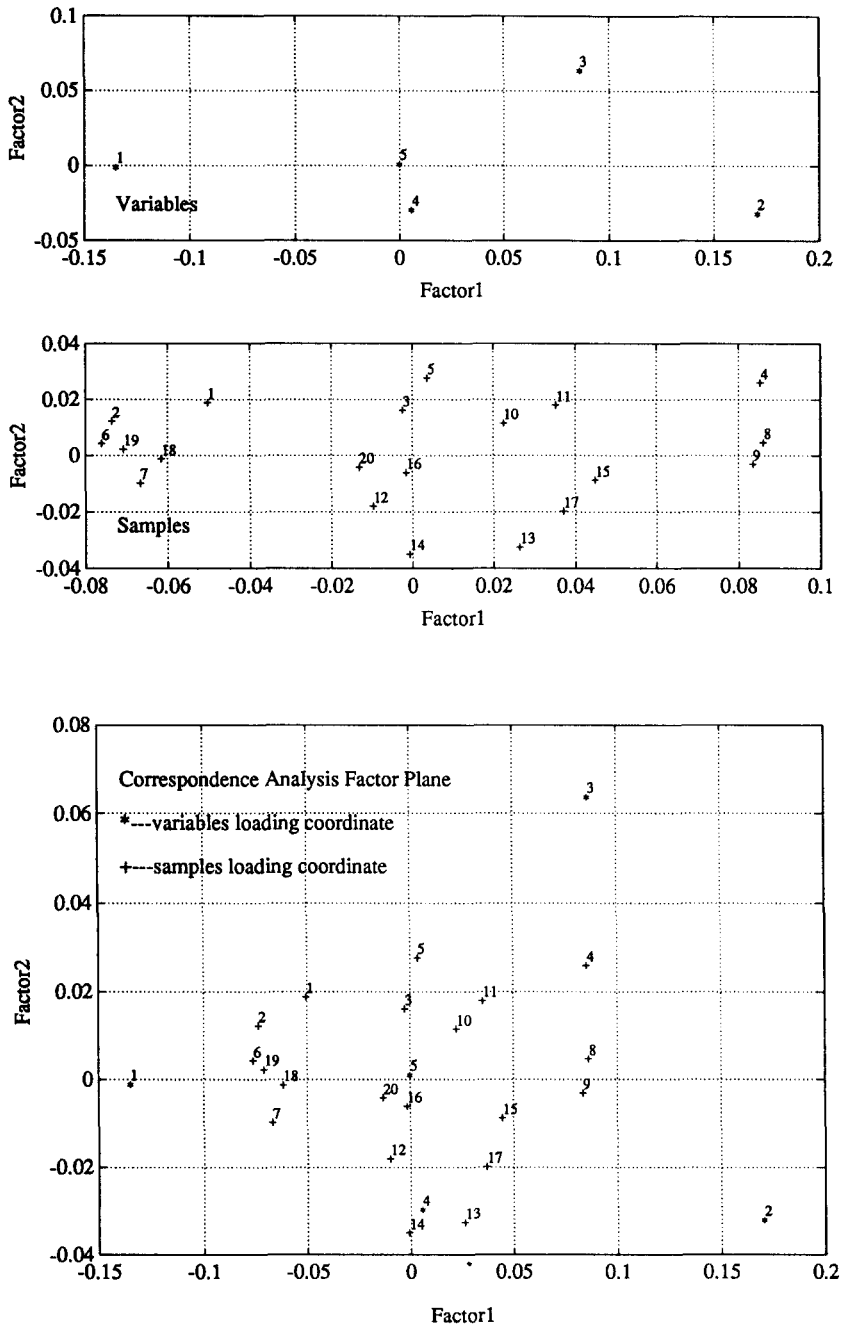

Figure 1. Example 1 factor loading and factor plane.

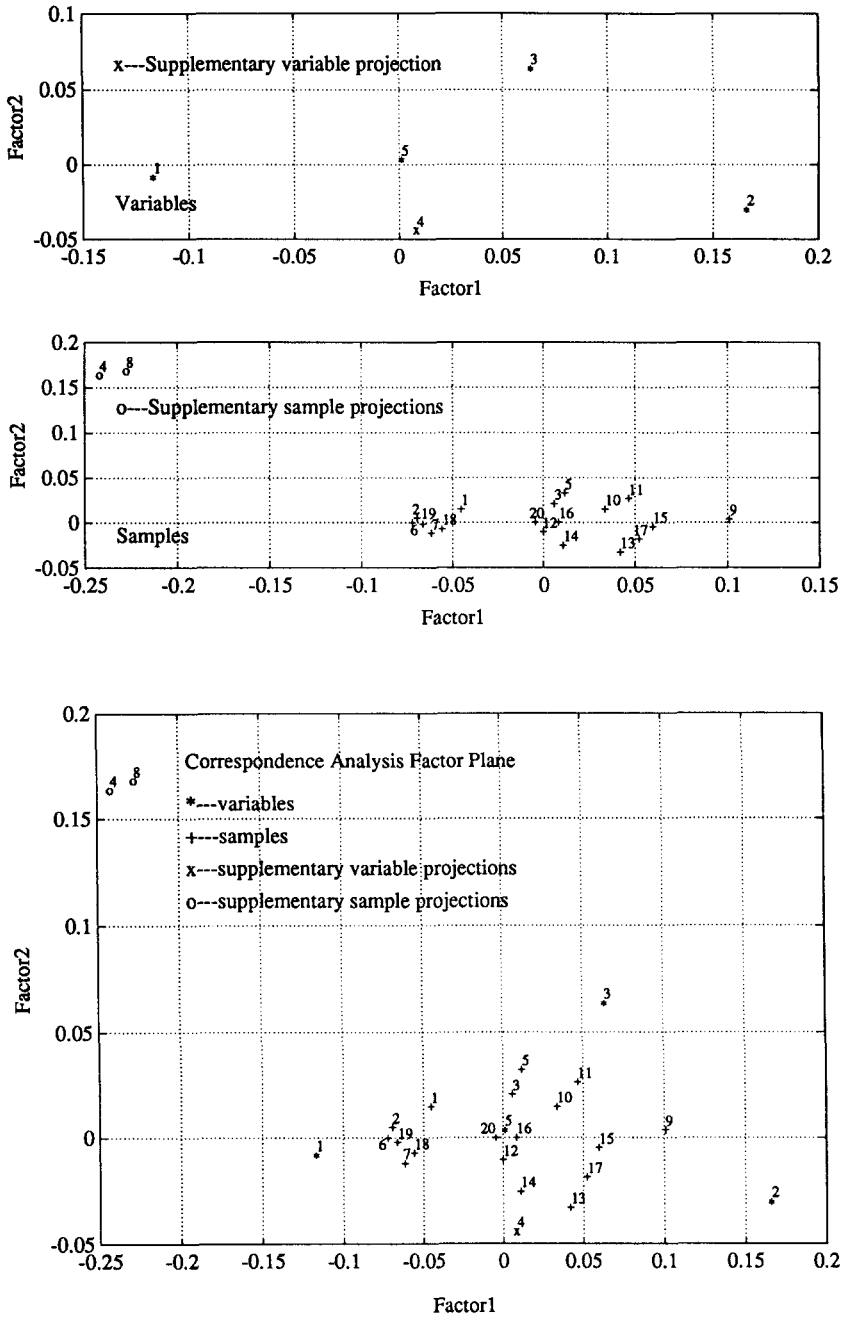

Figure 2. Example 1 factor loading and factor plane with supplementary elements.

Table 2 lists the monthly average precipitation in inches for 64 weather stations in Arizona (EarthInfo, 1992). The columns represent the months January–December, and the rows are the station numbers as samples. All of the data in this table have at least a 10 yr record. Table 3 lists the weather station descriptions.

Computer output for example 2 is omitted. Sellers and Hill (1974) indicated that there are two wet seasons per year in Arizona. The monsoon season starts in July and ends in August; the winter season extends from December through the middle of March. We can determine spatial patterns of precipitation from correspondence analysis. From the factor plane plots (Fig. 3), we can see the precipitation is approximately divided into 4 types statewide. Group 1 (including stations 6, 7, 8, 11, 13, 14, 26, 35, 36, 37, 40, 50, 53, 54, 58, and 61) is the monsoon precipitation. Though the summer precipitation is the prime precipitation statewide, in this region more than 60% of annual precipitation falls during the monsoon season. The summer moisture normally flows from

Table 2. Example 2 monthly average rainfall data

Station	Jan.	Feb.	Mar.	Apr.	May.	June	July	Aug.	sept.	Oct.	Nov.	Dec.
1	0.63	0.48	0.66	0.21	0.09	0.07	1.16	1.84	0.68	0.56	0.49	0.81
2	0.25	0.22	0.82	0.69	0.05	0.18	0.72	0.64	0.23	0.47	0.93	0.35
3	0.63	0.51	0.59	0.35	0.13	0.03	0.59	1.21	0.58	0.57	0.66	0.81
4	0.90	1.03	1.12	0.55	0.45	0.37	1.56	2.44	1.15	0.94	1.03	1.02
5	0.93	1.03	0.41	1.30	0.27	0.48	1.41	0.78	0.56	0.36	1.44	0.72
6	1.17	0.83	0.77	0.33	0.18	0.53	4.05	3.74	1.57	0.70	0.55	1.65
7	1.64	1.37	1.63	0.71	0.44	0.56	3.22	3.01	1.76	1.30	1.14	1.63
8	0.76	0.59	0.49	0.17	0.09	0.57	2.03	2.54	0.87	0.76	0.33	0.84
9	0.75	0.63	0.86	0.30	0.11	0.12	0.86	1.01	0.67	0.78	0.62	1.02
10	1.47	1.38	1.91	0.81	0.38	0.40	1.83	3.27	2.88	1.57	2.23	1.57
11	0.76	0.48	0.47	0.24	0.23	0.38	2.53	2.33	1.22	0.71	0.41	0.85
12	2.46	2.00	3.02	1.34	0.46	0.52	2.82	3.94	2.63	1.31	3.26	3.04
13	0.72	0.49	0.53	0.24	0.27	0.60	3.75	2.80	1.39	0.99	0.52	0.86
14	1.08	0.99	0.64	0.34	0.34	0.46	2.93	1.60	1.32	1.02	0.79	1.13
15	2.04	1.91	2.21	1.36	0.68	0.56	2.63	2.85	1.78	1.57	1.87	2.13
16	1.38	1.17	1.52	0.28	0.22	0.17	0.56	1.48	0.88	1.36	0.75	0.78
17	1.04	1.25	1.24	1.03	0.57	0.42	1.16	1.88	1.15	1.08	1.05	1.36
18	1.77	1.48	1.87	0.90	0.57	0.48	1.66	1.12	1.00	0.96	1.22	1.04
19	1.23	0.49	0.88	0.68	0.25	0.25	1.61	1.71	0.49	0.59	0.46	0.63
20	0.65	1.08	0.71	0.57	0.16	0.29	0.95	1.70	0.80	0.51	0.83	1.02
21	0.72	0.69	0.82	0.61	0.36	0.34	1.06	1.40	0.75	0.91	0.66	0.66
22	1.06	0.90	0.95	0.64	0.23	0.13	0.85	1.34	0.64	0.62	0.67	0.81
23	0.69	0.93	1.02	0.50	0.23	0.18	0.86	1.36	0.68	0.52	0.64	0.73
24	1.54	1.89	2.07	0.51	0.41	0.36	2.47	2.54	1.77	0.96	1.44	1.25
25	1.17	1.37	1.46	0.80	0.27	0.32	1.48	2.46	1.43	1.31	1.02	1.18
26	0.91	0.74	0.77	0.24	0.15	0.37	4.49	3.48	1.57	1.09	0.63	1.18
27	1.96	1.43	1.82	0.87	0.30	0.32	3.18	3.13	1.74	1.85	1.52	2.25
28	0.34	0.43	0.66	0.35	0.39	0.20	0.44	0.60	0.53	0.72	0.52	0.34
29	0.56	0.54	0.52	0.18	0.08	0.06	0.55	1.17	0.72	0.58	0.47	0.76
30	1.91	1.69	2.12	0.97	0.51	0.40	2.57	3.04	1.84	1.43	1.57	1.72
31	1.04	0.76	1.36	0.82	0.42	0.54	1.46	2.06	0.52	0.67	1.72	0.27
32	0.52	0.55	0.66	0.33	0.38	0.34	1.33	1.56	0.99	1.05	0.51	0.56
33	0.76	0.63	0.80	0.23	0.11	0.14	0.85	0.99	0.74	0.68	0.53	0.91
34	0.73	0.53	0.63	0.37	0.05	0.14	0.83	1.26	0.57	0.53	0.44	0.62
35	0.77	0.52	0.66	0.32	0.12	0.37	2.28	2.10	0.89	0.62	0.46	0.82
36	1.40	0.93	1.03	0.60	0.23	0.55	2.33	2.97	1.27	0.92	0.95	1.85
37	0.97	0.71	0.77	0.61	0.33	0.44	2.74	2.48	1.01	0.62	0.65	1.01
38	2.10	1.71	1.52	0.81	0.35	0.64	3.23	3.98	1.90	2.35	1.51	2.21
39	0.91	0.74	0.39	0.35	0.15	0.13	0.87	0.87	0.62	0.14	0.95	0.36
40	1.51	1.09	1.20	0.62	0.18	0.65	4.54	3.74	1.87	1.45	0.97	1.78
41	2.18	2.27	2.23	1.22	0.69	0.33	1.59	1.46	1.69	1.37	1.97	1.02
42	1.96	1.14	1.90	0.76	0.28	0.33	2.01	2.73	0.87	1.46	0.94	1.28
43	3.38	2.79	2.88	1.12	0.95	0.24	2.22	3.25	2.53	1.69	2.98	2.92
44	1.00	0.54	0.65	0.26	0.15	0.01	0.43	1.13	0.68	0.21	0.70	0.52
45	1.31	1.00	1.32	0.47	0.20	0.37	1.99	2.41	1.31	1.46	1.08	1.78
46	2.56	1.77	2.13	1.09	0.33	0.33	2.09	2.59	2.17	1.91	3.35	2.43
47	1.55	1.38	1.65	0.52	0.27	0.22	1.47	2.54	1.33	1.54	1.14	2.41
48	3.15	2.40	2.95	0.82	0.48	0.14	2.05	2.86	1.67	1.47	2.39	2.66
49	1.30	0.86	0.91	0.46	0.13	0.30	1.24	1.14	0.39	0.75	0.56	1.41
50	0.58	0.48	0.39	0.16	0.05	0.31	1.71	1.85	0.81	0.47	0.21	0.74
51	1.27	0.56	0.38	0.33	0.56	0.13	0.79	0.72	0.33	0.26	2.67	0.52
52	1.01	0.97	1.40	0.61	0.37	0.32	1.36	1.50	1.10	0.77	0.95	0.78
53	0.87	0.59	0.58	0.48	0.05	0.55	2.52	2.05	0.88	0.68	0.49	0.68
54	0.88	0.66	0.70	0.33	0.16	0.24	2.44	2.23	1.40	0.93	0.62	0.94
55	2.18	2.21	1.89	1.00	0.29	0.58	3.70	4.12	1.81	2.29	1.74	2.83
56	1.14	0.82	1.14	0.60	0.46	0.40	1.30	1.65	0.83	0.76	0.94	1.14
57	2.70	1.83	2.45	1.11	0.33	0.47	2.52	3.73	2.00	1.67	1.87	2.65
58	1.61	1.16	1.09	0.46	0.39	0.26	2.45	3.23	1.64	1.42	0.99	1.11
59	1.51	1.64	1.72	0.68	0.56	0.26	1.89	2.81	1.53	0.94	1.20	1.44
60	1.58	1.23	1.72	0.85	0.44	0.51	2.65	3.19	1.59	1.59	1.26	1.54
61	0.51	0.44	0.32	0.24	0.05	0.16	1.55	1.24	0.93	0.48	0.22	0.59
62	0.46	0.48	0.52	0.31	0.31	0.36	1.24	1.42	0.85	0.86	0.49	0.61
63	4.07	2.61	3.52	1.50	0.57	0.56	3.29	4.13	2.68	2.27	3.35	3.66
64	0.42	0.24	0.20	0.12	0.04	0.01	0.26	0.49	0.22	0.32	0.17	0.36

Table 3. Weather station description

No.	Station	Elevation(ft)	Latitude	Longitude
1	AJO	1800.00	32:22:00	112:52:00
2	ALAMO	1060.00	34:16:00	113:34:00
3	ALAMO DAM	1290.00	34:14:00	113:35:00
4	ASH FORK 2	5080.00	35:13:00	112:29:00
5	BAGDAD	3710.00	34:34:00	113:10:00
6	BISBEE	5310.00	31:26:00	109:55:00
7	BLACK RIVER PUMPS	6040.00	33:29:00	109:46:00
8	BOWIE JCT R15 ON W5	4720.00	32:26:00	109:42:00
9	CASA GRANDE RUINS N M	1420.00	33:00:00	111:32:00
10	CIBECUE	5050.00	34:02:00	110:29:00
11	COCHISE 4 SSE	4180.00	32:04:00	109:54:00
12	CROWN KING	5920.00	34:12:00	112:20:00
13	DOUGLAS	4040.00	31:21:00	109:32:00
14	DUNCAN	3660.00	32:45:00	109:07:00
15	FLAGSTAFF WSO AP	7010.00	35:08:00	111:40:00
16	FLORENCE JUNCTION	1880.00	33:17:00	111:22:00
17	GRAND CANYON N P	6950.00	36:03:00	112:08:00
18	GRAND CANYON NATL PK 2	6790.00	36:03:00	112:09:00
19	HACKBERRY	3580.00	35:22:00	113:44:00
20	HACKBERRY 2 SE	3700.00	35:21:00	113:41:00
21	KEAMS CANYON	6210.00	35:49:00	110:12:00
22	KINGMAN	3360.00	35:11:00	114:03:00
23	KINGMAN 2	3540.00	35:12:00	114:01:00
24	MAYER 3 NNW	4640.00	34:26:00	112:15:00
25	MONTEZUMA CASTLE N W	3180.00	34:37:00	111:50:00
26	NOGALES	3810.00	31:21:00	110:55:00
27	ORACLE 2 SE	4510.00	32:36:00	110:44:00
28	PAGE	4270.00	36:56:00	111:27:00
29	PAINTED ROCK DAM	550.00	33:05:00	113:02:00
30	PAYSON	4910.00	34:14:00	111:20:00
31	PERNER RANCH	5600.00	35:22:00	113:17:00
32	PETRIFIED FOREST NAT PK	5450.00	34:49:00	109:53:00
33	PHOENIX WSFO AP	1110.00	33:26:00	112:01:00
34	PHOENIX CITY	1080.00	33:27:00	112:04:00
35	PIMA R4 ON W2	3770.00	32:50:00	110:01:00
36	POLAND JUNCTION	4900.00	34:27:00	112:16:00
37	PRESCOTT FAA AP	5020.00	34:39:00	112:26:00
38	ROCK CREEK R S	3630.00	33:49:00	109:48:00
39	ROUND VALLEY	3740.00	35:06:00	113:38:00
40	SANTA RITA EXP RANGE	4300.00	31:46:00	110:51:00
41	SEDONA R S	4220.00	34:52:00	111:46:00
42	SENECA 3 NW	4920.00	33:47:00	110:32:00
43	SIERRA ANCHA	5100.00	33:48:00	110:58:00
44	SIGNAL 13 SW	2510.00	34:22:00	113:48:00
45	SUMMIT	3650.00	33:33:00	110:57:00
46	SUNFLOWER 3 NNW	3720.00	33:54:00	111:29:00
47	SUPERIOR	3000.00	33:18:00	111:06:00
48	SUPERIOR 2 ENE	4160.00	33:18:00	111:04:00
49	SUPERSTITION MTN	1960.00	33:22:00	111:26:00
50	TANQUE R9 ON W4	3560.00	32:37:00	109:37:00
51	TROUT CREEK STORE	2850.00	34:53:00	113:39:00
52	TRUXTON CANYON	3820.00	35:23:00	113:40:00
53	TUCSON NURSERY 4 NW	2250.00	32:18:00	111:03:00
54	TUCSON WSO AP	2580.00	32:08:00	110:57:00
55	TURKEY CREEK 1	6750.00	33:45:00	109:48:00
56	TUWEEP	4780.00	36:17:00	113:04:00
57	UPPER PARKER CREEK	5500.00	33:48:00	110:57:00
58	VAIL	3230.00	32:03:00	110:43:00
59	WALNUT CREEK	5090.00	34:56:00	112:49:00
60	WHITERIVER 1 SW	5120.00	33:50:00	109:58:00
61	WHITLOCK VLY R2 ON W1	3290.00	32:49:00	109:31:00
62	WINSLOW WSO AP	4890.00	35:01:00	110:44:00
63	WORKMAN CREEK 1	6970.00	33:49:00	110:55:00
64	YUMA WSO AP	210.00	32:40:00	114:36:00

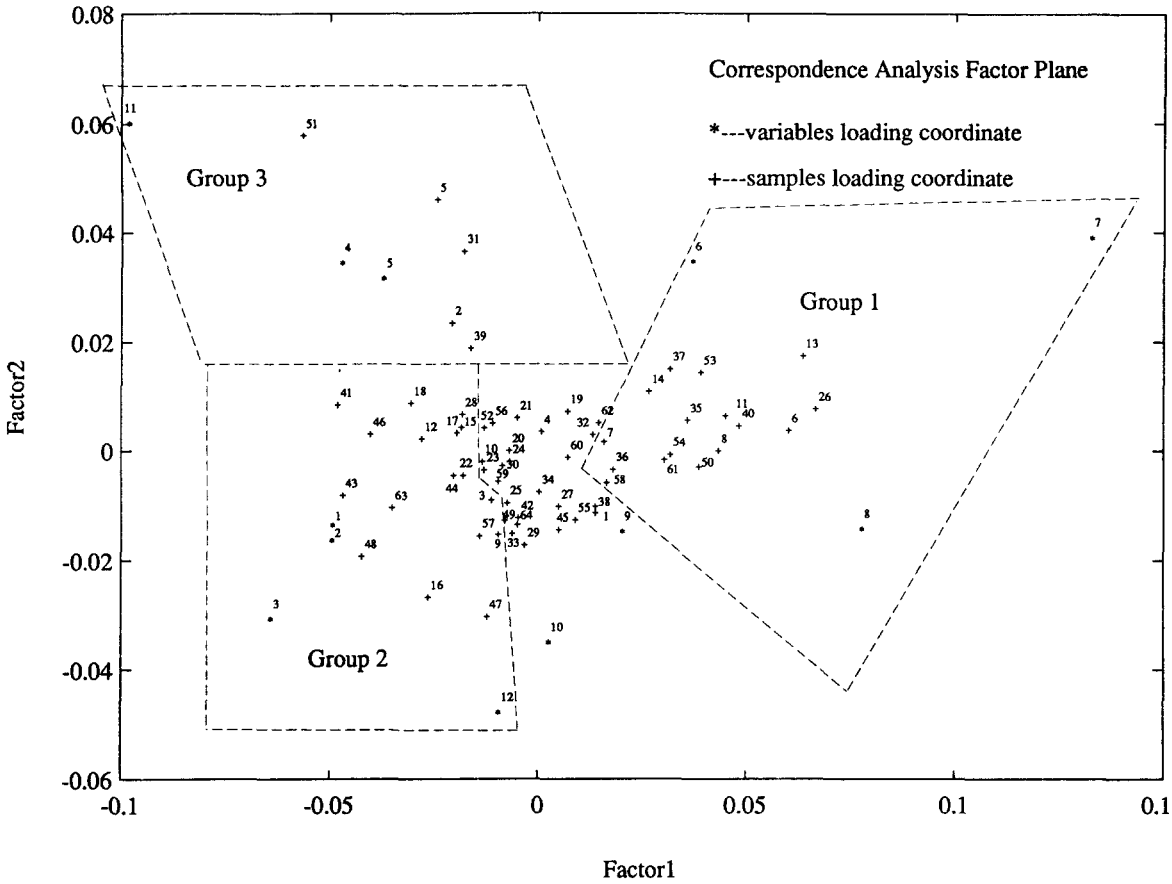

Figure 3. Example 2 factor plane.

the Gulf of Mexico over strongly heated mountainous terrain, which generates convective thundershowers. All group 1 stations (except for stations 36 and 37, which are located in central Arizona between the forested plateaus to the northeast and the arid desert region to the southwest) are located in southeastern Arizona. This result tallies with Sellers and Hill's (1974) conclusion. Most of Arizona has a second rainy season in midwinter. The correspondence analysis results tell us that group 2 (stations 3, 9, 12, 15, 16, 17, 18, 22, 23, 28, 41, 43, 44, 46, 47, 48, 49, 57, and 63) weather stations are associated mostly with the winter rainy season. Figure 3 shows that winter season (December–February) precipitation is more important at these stations. This precipitation is generated by the eastward movement of storm systems from the northern Pacific Ocean, and some of it falls in the form of snow. Most of the stations in this group are located between the northwest and northeast. The third group (stations 2, 5, 31, 39, and 51) is an intermediary precipitation type. These stations are associated with November, April, and May precipitation (Fig. 4) and are located in the northwestern part of Arizona. Station 52 is anomalous, at which November's average precipitation is three times greater than all other monthly averages. The rest of the stations do not present a clear pattern from the view of monthly average rainfall depth. Figure 4 are graphics for absolute contributions of the first four factors respect with variables and samples, whereas Figure 5 plots variable and sample weights and error profiles, respectively.

CONCLUSION

"Graphics are arguably the most important part of the program, because CA is a geometric approach to data analysis" (Goldstein, 1991). MATCORS provides varied and high-quality graphic plot functions. This program works in the user-friendly Matlab environment, which makes the program portable and easy to use. It can operate with any number of variables and with any number of samples. This program allows users to specify any number of variables or samples as supplementary elements, and then will project them on the plots. If these elements are not desired, they can be deleted within Matlab to form a new data matrix.

Acknowledgments—The financial support for this research was provided by grant BCS8920851 from NSF and NA90AA-H-HY505 cooperative agreement of Hydrologic Research Laboratory of NOAA. The comments of two anonymous reviewers which helped improve the quality of this paper are also gratefully acknowledged.

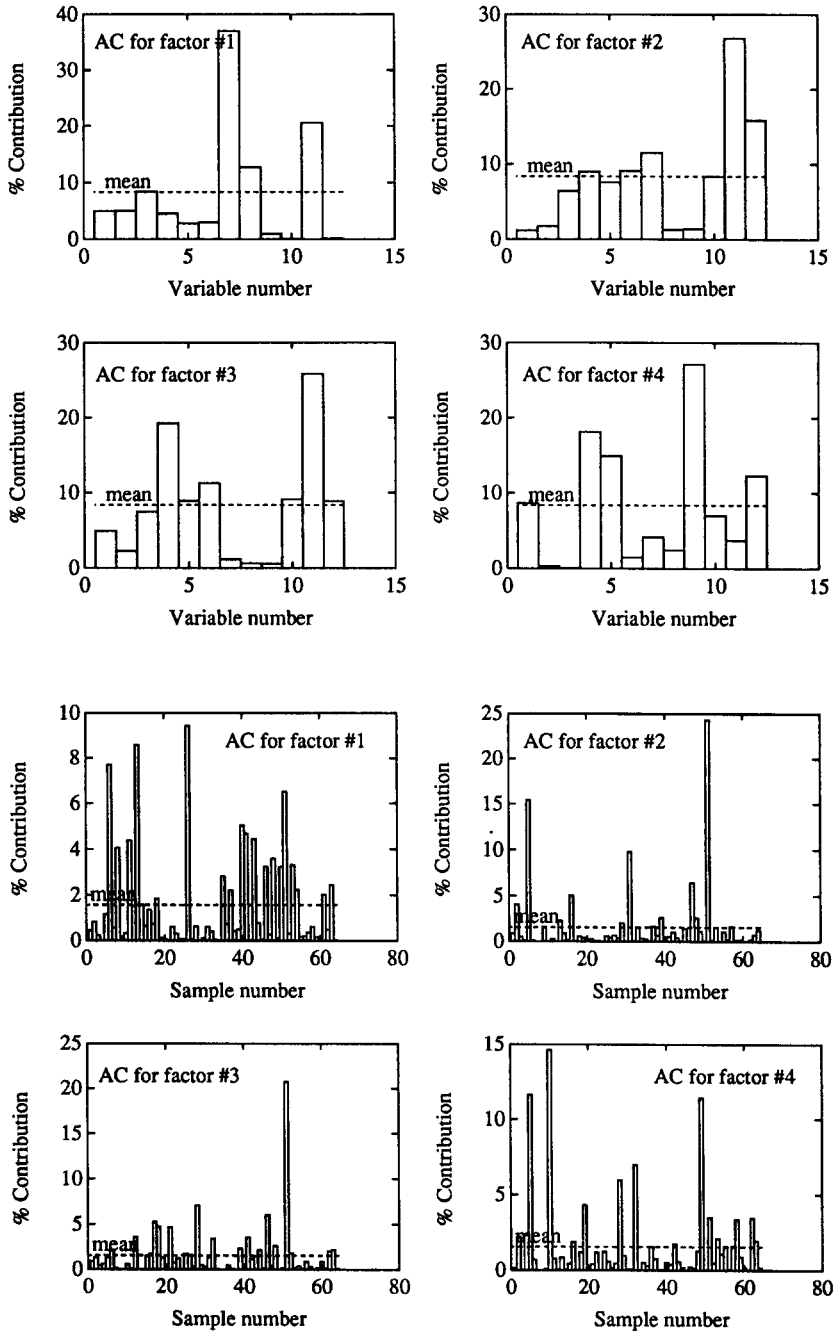

Figure 4. Example 2 absolute contributions.

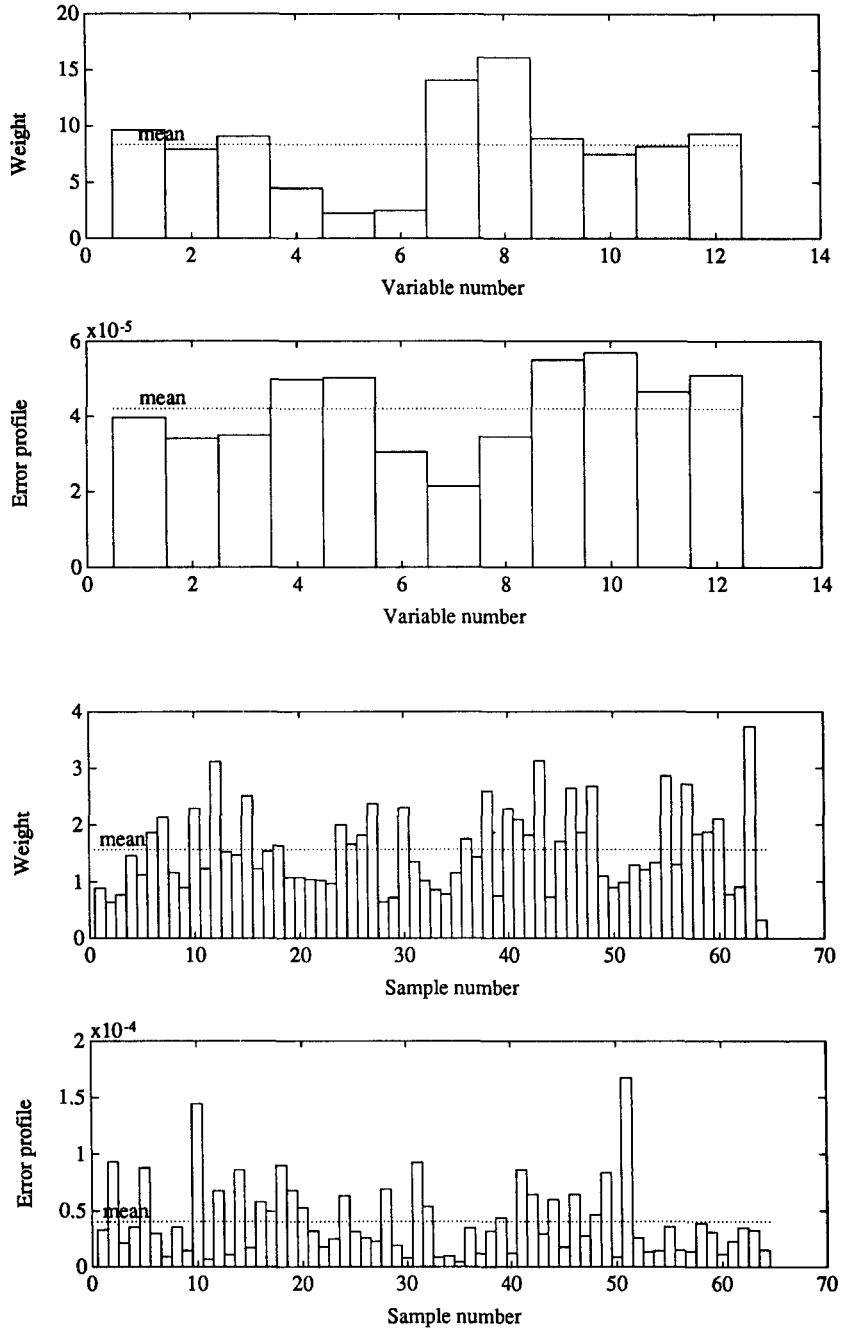

Figure 5. Example 2 weights and error profiles.

REFERENCES

Avila, F., and Myers, D. E., 1991, Correspondence analysis applied to environmental data sets: a study of Chautauqua lake sediments: Chemometrics and Intelligent Laboratory Systems, v. 11, p. 229–249.

Benzecri, J. P., 1973, l'Analyse des donnees, Tomm II, l'Andyse des correspondances: Dunod, Paris, 619 p.

EarthInfo, eds., 1992, NOAA-NCDC, hourly precipitation—Western Region, CD-ROM, Vol. 2.0, Published as CLIMATEDATA by EarthInfo Inc., Boulder, Colorado.

Goldstein, R., 1991, Statistical computing software reviews: The American Statistician, v. 45, no. 4, p. 305–311.

LeBart, L., Morineau, A., and Warrick, K., 1984, Multivariate descriptive statistical analysis: John Wiley & Sons, New York, 231 p.

Luo, J., and Xing, Y., 1987, Economic statistics analysis method and forecast (in Chinese): Qing Hua Univ. Press, Peking, China, p. 270–299.

Rhodes, H. R., and Myers, D. E., 1991, Correspondence analysis used in the evaluation of Lakewater chemistry in the Adirondacks: Jour. Chemometrics, v. 5, p. 273–290.

Sellers, W. D., and Hill, R. H., 1974, Arizona climate: The Univ. of Arizona Press, Tucson, Arizona, p. 6–19.

APPENDIX 1

```
function [r,q,ev,c_ra,s_ra,evv,ess,sv,sp]=matcors(x)
%  MATCORS performs correspondence analysis for a contingency matrix x with n
%  samples and p variables.

%  Syntax [r,q,ev,c_ra,s_ra,evv,ess,sv,sp]=matcors(x)


%  Missing values are coded as 'NaN' (not-a-number).

%  Input Description
%  x  : contingency matrix with n samples and p variables
%  Echo input
%  pec: Cumulative percent variation to determine the amount of principal factors.
%  nb : Selected number of principal factors.
%  nc : The number of supplementary variables.
%  cc : Supplementary variables.
%  nr : The number of supplementary samples.
%  cr : Supplementary samples.
%  f1 & f2 : Two principal factors to be plotted.

%  Output description
%  r   : R-type main factors loading matrix.
%  q   : Q-type main factors loading matrix.
%  ev  : Eigenvalues, relative and cumulative variation explained by factors
%  c_ra : Relative and absolute contributions of variables.
%  s_ra : Relative and absolute contributions of samples.
%  evv  : Weights and error profiles of variables.
%  ess  : Weights and error profiles of samples.
%  sv   : Supplementary variables projection values.
%  sp   : Supplementary samples projection values.

% Echo of data and input options

disp('Input data matrix')
disp(x);

tot=sum(sum(x));x=x/tot;

% Check negative entries in the data matrix
nz=find(x<0);
if nz>0
error('Data matrix has negative entries !!!')
end

disp('Strike any key when ready'), pause

% Call function EIGV to form a real symmetric covariance matrix of variables
% and compute all eigenvalues and eigenvectors by Jacobi rotation method.
[n,p,evl,evc,evr,tco,tro]=eigv(x);
nn=n;pp=p;

kk=menu('Do you want to select Supplementary Elements?','Yes','No');
if kk==1
   nc=input('How many variables (column)?');
   for i=1:nc
       cc(i)=input(['Please give supplementary column no.',num2str(i),'number:  ' ]
          % Check input choice
```

```
        if (nc>p) | (cc(i)>p)
        error('Wrong number--Larger than maximal column number')
        end
  end
  nr=input('How many samples (row)?');
  for i=1:nr
      cr(i)=input(['Please give supplementary row no.',num2str(i),'number:  ' ])
        % Check input choice
        if (nr>n) | (cr(i)>n)
        error('Wrong number--Larger than maximal row number')
        end
  end

%  Supplementary element projection values
    sv=((x(:,cc))'*evr)./(diag(tco(cc))*ones(nc,pp));
    sp=(x(cr,:)*evc')./(diag(tro(cr))*ones(nr,pp));

%  Eliminate the supplementary elements from original data matrix [x]
%  to form new matrix [xc]
xc=x;xc(:,cc)=[];xc(cr,:)=[];

%  Performs EIGV function again for new matrix [xc].
[n,p,evl,evc,evr,tco,tro]=eigv(xc);

end


%  Calculate and display eigenvalues and their relative and cumulative
%  percentage of variation.
ev(:,1)=evl;ev(:,2)=100*evl/sum(evl);ev(:,3)=100*cumsum(evl)/sum(evl);
disp('Eigvenvalues  Rel.--%    Cul.--%');
disp(ev);

  pre=input('How many cumulative percent variations do you want? ');
     if pre>100
     error('Wrong number -- Cumulative percentage larger than 100')
     end

  nb=input('How many principal factors do you want? ');
     if nb>p-1
     error(['Wrong number -- larger than the total number ',num2str(p)-1])
     end

% The number of principal factors are determined by cumulative percent variation
    for i=1:p-1
    if ev(i,3) >= pre
    m=i;
    break,end
    end

% The amount of principal factors are determined by the selected number
    if nb > m
    m=nb;
    end

% R-mode factor loadings;
r=evc*diag(sqrt(evl));r=r(:,1:p-1);

% Q-mode factor loadings;
q=evr*diag(sqrt(evl));q=q(:,1:p-1);

% Relative and Absolute contributions of variables and samples for main factors.
for k=1:p-1
  % Relative contributions
  c_ra(:,k)=100*(evl(k)*evc(:,k).^2)./(evc.^2*evl);
  s_ra(:,k)=100*(evl(k)*evr(:,k).^2)./(evr.^2*evl);
  % Absolute contributions
  c_ra(:,p+k-1)=100*(tco'.*(inv(diag(tco))*evc(:,k))).^2;
  s_ra(:,p+k-1)=100*(tro.*(inv(diag(tro))*evr(:,k))).^2;
end

% Error profiles of all variables and samples for the remaining factors.
for j=1:p

epv(j)=(((diag(sqrt(evl(m+1:p))*evc(j,m+1:p)))'*(evr(:,m+1:p))').^2)*tro;
end
```

```
for i=1:n
epps(i)=(((diag(sqrt(evl(m+1:p))*evr(i,m+1:p)))'*(evc(:,m+1:p))').^2)*tco';
end

% Variables and samples' weights and their error profiles.
      vw=(tco/sum(tco))';sw=tro/sum(tro);
      evv(:,1)=100*vw;evv(:,2)=epv';
      ess(:,1)=100*sw;ess(:,2)=epps';

% Echo of principal factors to be plotted
f1=input('Which two main factors do you want to plot? First--please!');
f2=input('Second--please!');
   if (f1 >m) | (f2 >m)
   error('Wrong number -- larger than the number of principal factors')
   end

%  R-mode factor loading plot
if nc>0
tt=1:pp;tt(cc)=[];
rm(cc,1:m)=sv(:,1:m);rm(tt,1:m)=r(:,1:m);
subplot(211),plot(r(:,f1),r(:,f2),'*',sv(:,f1),sv(:,f2),'x');
gtext('x -- Supplementary variable projections');
else
rm=r;
subplot(211),plot(r(:,f1),r(:,f2),'*')
end
xlabel(['Factor',num2str(f1)]);
ylabel(['Factor',num2str(f2)]);
gtext('Variables');

%  Mark down the variables number.
    for i=1:pp
    text(rm(i,f1),rm(i,f2),num2str(i));
    end
grid;

%  Q-mode factor loading plot
if nr>0
tt=1:nn;tt(cr)=[];
qm(cr,1:m)=sp(:,1:m);qm(tt,1:m)=q(:,1:m);
subplot(212),plot(q(:,f1),q(:,f2),'+',sp(:,f1),sp(:,f2),'o');
gtext('o -- Supplementary sample projections');
else
qm=q;
subplot(212),plot(q(:,f1),q(:,f2),'+')
end
xlabel(['Factor',num2str(f1)]);
ylabel(['Factor',num2str(f2)]);
gtext('Samples');
    for i=1:nn
    text(qm(i,f1),qm(i,f2),num2str(i));
    end
grid;pause;clg;

%  factor plane
if kk==2  % No supplementary elements
plot(r(:,f1),r(:,f2),'*',q(:,f1),q(:,f2),'+');
else
if (nc>0) & (nr>0) % Both variables and samples have supplementary elements.
plot(r(:,f1),r(:,f2),'*',q(:,f1),q(:,f2),'+',sv(:,f1),sv(:,f2),'x',sp(:,f1),sp(:,f2
elseif (nc>0) & (nr==0) % Only variables have supplementary elements.
plot(r(:,f1),r(:,f2),'*',q(:,f1),q(:,f2),'+',sv(:,f1),sv(:,f2),'x');
elseif (nr>0) & (nc==0) % Only samples have supplementary elements.
plot(r(:,f1),r(:,f2),'*',q(:,f1),q(:,f2),'+',sp(:,f1),sp(:,f2),'o');
end;end;end;end

%  Mark down both variable and sample number on the plot;
     for i=1:pp
     text(rm(i,f1),rm(i,f2),num2str(i));
     end
     for j=1:nn
     text(qm(j,f1),qm(j,f2),num2str(j));
     end;grid;
xlabel(['Factor',num2str(f1)]);
ylabel(['Factor',num2str(f2)]);
gtext('Correspondence Analysis Factor Plane');
```

```
gtext('*---variables loading coordinate');
gtext('+---samples loading coordinate');
if kk==1
gtext('x---supplementary variables projection');
gtext('o---supplementary samples projection');
end
pause;clg;

% Plot bar graph of varibles' and samples' absolute contributions for
% first 4 factors
if p>=5
np=4;
else
np=p-1;
end
for nk=1:np;
tw1=['Variable number';'Sample number  ';'% Contribution '];
subplot(220+nk),bar(1:p,c_ra(:,p-1+nk))
xlabel(tw1(1,:));
ylabel(tw1(3,:));
gtext(['AC for factor #',num2str(nk)]);
hold on
me=mean(c_ra(:,p-1+nk));
plot([.5 p+.5],[me me],'--b');
text(1,me,'mean');
hold off
end
pause;clg

for nk=1:np;
subplot(220+nk),bar(1:n,s_ra(:,p-1+nk));
xlabel(tw1(2,:));
ylabel(tw1(3,:));
gtext(['AC for factor #',num2str(nk)]);
hold on
me=mean(s_ra(:,p-1+nk));
plot([.5 n+.5],[me me],'--b');
text(1,me,'mean');
hold off
end
pause;clg

% Plot bar graph of variables' and samples' weights and error profiles
tw2=['Weight       ';'Error profile'];
for nk=1:2;
subplot(210+nk),bar(1:p,evv(:,nk));
ylabel(tw2(nk,:));
xlabel(tw1(1,:));
me=mean(evv(:,nk));
hold on
plot([.5 p+.5],[me me],':w');
text(1,me,'mean');
hold off
end
pause;clg

for nk=1:2;
subplot(210+nk),bar(1:n,ess(:,nk));
ylabel(tw2(nk,:));
xlabel(tw1(2,:));
me=mean(ess(:,nk));
hold on
plot([.5 n+.5],[me me],':w');
text(1,me,'mean');
hold off
end
pause;clg



function [n,p,evl,evc,evr,tco,tro]=eigv(x)

% Forms a real symmetric  covariance matrix of variables, and computes
% all eigenvalues and eigenvectors by Jacobi rotation mathod.
```

```
pre=1.0e-6;%Jacobi iteration precision;
total=pre+1; %Initial first rotation;

[n,p]=size(x);
tot=sum(sum(x)); % the sum of every element in matrix [x];
tco=sum(x);   % a row vector with the sum over each column of matrix [x];
      tro=sum(x')';  % a column vector with the sum over each row of matrix [x];

      z=(x-tro*tco/tot)./sqrt(tro*tco);
      a=z'*z;       % covariance matrix of variables

p=length(a);
v=eye(p);u=v;

while total > pre


          [cul,row,total]=index(a);

          st=(a(cul,cul)-a(row,row))./(2.*a(row,cul));

          if st==0
          c=1/sqrt(2);
          s=c;

          else

             t=sign(st)./(abs(st)+sqrt(st.^2+1));
             c=1/sqrt(t.^2+1);
             s=t.*c;

          end

       v(row,row)=c; v(row,cul)=s;
       v(cul,row)=-s;v(cul,cul)=c;

a=v'*a*v;
u=u*v;v=eye(p);
end

eigva=a;
eigvc=u;

% Sort the eigenvalues and eigenvectors in descending order;
      [h,k]=sort(diag(-eigva));
      evl=(-h); % eigenvalues;
      evc=eigvc(:,k); % eigenvectors;

%Compute eigenvectors of covariance matrix of samples
    evr=(z*evc)./(ones(n,p)*diag(sqrt(evl)));


return



function [cul,row,tot]=index(x)
% Find the row and column number of maximum·off-diagonal element of a
% real symmetric matrix [x]

z=triu(x,1);
p=length(x);


[y,i]=sort(abs(z));
[a,b]=sort(abs(y(p,:)));
m=max(a);
cul=b(p);
row=i(p,cul);

tot=abs(x(row,cul));

return
```

APPENDIX 2

Example 1 output with variable 4 and sample 4 and 8 as supplementary elements

a. (r) R-mode factor loadings (variables)

factor1	factor2	factor3	factor4
-0.117	-0.008	-0.003	0.000
0.166	-0.030	-0.002	0.000
0.064	0.064	-0.002	0.000
0.001	0.003	0.016	0.000

b. (q) Q-mode factor loadings (samples)

-0.045	0.014	-0.003	0.000
-0.069	0.005	-0.001	0.000
0.006	0.020	0.002	0.000
0.012	0.032	0.000	0.000
-0.072	-0.001	0.003	0.000
-0.061	-0.012	0.001	0.000
0.101	0.003	0.004	0.000
0.034	0.015	0.005	0.000
0.047	0.026	-0.001	0.000
0.000	-0.011	0.003	0.000
0.042	-0.033	-0.001	0.000
0.011	-0.026	0.007	0.000
0.060	-0.005	-0.003	0.000
0.009	-0.000	-0.004	0.000
0.052	-0.019	-0.005	0.000
-0.056	-0.007	-0.007	0.000
-0.066	-0.002	0.005	0.000
-0.004	-0.000	-0.006	0.000

c. (ev)

Eigenvalue	Relative variation	Cumulative variation
0.045	89.440	89.440
0.005	9.997	99.437
0.000	0.563	100.000
0.000	0.000	100.000

d. (c_ra)

Variable rela contri.			Variable abs. contri.		
99.425	0.493	0.082	30.120	1.335	3.944
96.778	3.212	0.010	60.919	18.088	0.970
49.913	50.032	0.055	8.958	80.336	1.570
0.498	4.354	95.149	0.003	0.241	93.516

e. (s_ra)

Sample rela contri.			Sample abs. contri.		
90.328	9.354	0.318	4.485	4.156	2.506
99.502	0.470	0.029	10.603	0.448	0.488
8.251	90.977	0.772	0.083	8.226	1.239
11.979	88.008	0.013	0.312	20.512	0.054
99.774	0.005	0.221	11.467	0.005	4.033
96.070	3.883	0.047	8.323	3.009	0.649
99.734	0.117	0.149	22.573	0.237	5.346
83.138	15.338	1.524	2.529	4.174	7.361
75.888	24.060	0.052	4.809	13.642	0.522
0.010	90.438	9.552	0.000	2.287	4.288
61.711	38.269	0.020	3.904	21.659	0.200
14.805	79.638	5.557	0.268	12.890	15.966
99.042	0.684	0.275	7.889	0.487	3.474
82.241	0.026	17.733	0.160	0.000	5.497
87.801	11.406	0.794	6.068	7.052	8.713
96.863	1.745	1.392	6.841	1.103	15.616
99.265	0.130	0.605	9.648	0.113	9.341
28.643	0.040	71.317	0.037	0.000	14.706

f. (evv)

Variable weight	Error profile
64.60047900	0.00000055
20.02291900	0.00000013
9.13636840	0.00000022
6.24023340	0.00001301

g. (ess)

Sample weight	Error profile
5.54224400	0.00000055
5.54745290	0.00000011
5.57349720	0.00000027
5.56828840	0.00000001
5.57349720	0.00000088
5.55787060	0.00000014
5.58391500	0.00000117
5.58912390	0.00000161
5.55787060	0.00000011
5.57349720	0.00000094
5.54224400	0.00000004
5.58912390	0.00000349
5.53703510	0.00000076
5.53182620	0.00000120
5.52140850	0.00000190
5.51099070	0.00000341
5.58391500	0.00000204
5.51619960	0.00000321

h. Supplementary variable projection (sv)

0.008	-0.044	0.051	-0.000	0.032

i. Supplementary sample projections (sp)

-0.242	0.164	0.479	0.059	0.044
-0.228	0.168	0.495	0.022	0.042