## 9.3 CORRESP— Correspondence analysis

This program performs a correspondence analysis, CA. Correspondence analysis is similar to principal components analysis, PCA, but it differs in several important aspects. First, PCA is appropriate when one has a matrix of measurements of a set of $p$ variables on a set of $n$ objects. In CA the data are in the form of a 2-way contingency table of observed frequencies. In PCA one can obtain a "biplot" in which the objects and the variables are superimposed on the same plot so that one can study their inter-relationships. In CA one obtains an analogous plot for the row and column variables of the contingency table. In PCA one judges proximities among the objects using Euclidean distances and among the variables using covariance (or correlation). In CA one uses Chi-square distances to judge proximities for both the row and for the column variables.

The Chi-square distance between two columns $i$ and $j$, for example, is:

$$d_{ij} = \sqrt{\sum_k \left(\frac{f_{ki}}{f_{\bullet i}} - \frac{f_{kj}}{f_{\bullet j}}\right)^2 \Big/ f_{k\bullet}}$$

This distance can be computed using the SIMINT program in NTSYS-pc. Note that a principal coordinates analysis of a Chi-square distance matrix does not yield the same results as CA— since in CA the objects are weighted according to their frequencies.

In correspondence analysis one fits the following model to the observed data:

$$f_{ij} = \sqrt{f_{i\bullet} f_{\bullet j}}\left(1 + \sum_k \sqrt{\lambda_k}\,\psi_{ik}\phi_{jk}\right)$$

where $f_{ij}$ are the observed relative frequencies, $x_{ij}/x_{\bullet\bullet}$, $f_{i\bullet}$ and $f_{\bullet j}$ are the relative frequencies of the rows and columns, $\lambda_k$ is the k*th* eigenvalue, and $\psi_{ik}$ and $\phi_{jk}$ are ele-

ments of the row and column factor matrices. Thus, the factors describe the patterns of deviations from independence. If the rows and columns were perfectly independent then the eigenvalues would all be equal to zero. Note: this program outputs matrices of factor *coordinates*, $\hat{\psi}_{ik} = \sqrt{\lambda_k}\psi_{ik}$ and $\hat{\phi}_{jk} = \sqrt{\lambda_k}\phi_{jk}$, rather than the factors themselves.

The program follows, in part, the computational outline given in Lebart *et al.* (1984). The program computes eigenvalues and eigenvectors for either the rows or columns (whichever is less) and then computes the other vectors by projection. For example, if the number of columns is less than the number of rows then the program computes the eigenvalues and eigenvectors of the matrix

$$\mathbf{S}_1 = \mathbf{D}_p^{-1/2}\hat{\mathbf{A}}\mathbf{D}_p^{-1/2}$$

where $\hat{\mathbf{A}} = \mathbf{F}^t\mathbf{D}_n^{1/2}\mathbf{F}$, $\mathbf{D}_n$ is a diagonal matrix of $f_{i\bullet}$, $\mathbf{D}_p$ is a diagonal matrix of $f_{\bullet j}$, and $\mathbf{F}$ is the $n$ by $p$ matrix of relative frequencies. If $n < p$ then a corresponding matrix is factored to give the row eigenvectors. Note: the first eigenvalue (always equal to 1) and its associated eigenvector are not displayed or saved.

Greenacre (1984) is a general text on correspondence analysis and furnishes many examples and extensive discussions. He also shows how it can be computed using the following singular-value decomposition:

$$\mathbf{D}_n^{\frac{-1}{2}}\mathbf{F}\mathbf{D}_p^{\frac{-1}{2}} = \mathbf{L}\Lambda^{\frac{1}{2}}\mathbf{R}^t$$

where

$$\hat{\Psi} = \mathbf{D}_n^{\frac{-1}{2}}\Psi\Lambda^{\frac{1}{2}}\sqrt{n} \text{ and } \hat{\Phi} = \mathbf{D}_p^{\frac{-1}{2}}\Phi\Lambda^{\frac{1}{2}}\sqrt{n}.$$

### 9.3.1 Program parameters

| Batch code | Description |
|---|---|
| O | Name of the input data matrix to be operated upon. Normally this will be a contingency table. |
| N | The number of factors to be extracted (default is 3). |
| RF | Name to be given to the matrix of coordinates for the row factors, $\hat{\Psi}$. |
| CF | Name to be given to the matrix of coordinates for the column factors, $\hat{\Phi}$. |
| VAL | Name to be given to a diagonal matrix of the eigenvalues (optional). |
| RA | Name to be given to the matrix of "absolute contributions" for the row factors (optional). These give the proportion of the variance of a factor explained by each variable. |
| CA | Name to be given to the matrix of absolute contributions for the column factors (optional). |
| RC | Name to be given to the resultant matrix of "squared correlations" for the row factors (optional). These give the proportion of the variance of a variable explained by each factor (optional). |
| CC | Name to be given to the resultant matrix of squared correlations for the column factors (optional). |
| LST | Name of the listing device or file. |

### 9.3.2 Output

The input parameters are written to the listing device or file. The trace of the matrix $\mathbf{S}_1$, the $X^2$ value for a "Chi-square" test of independence of the rows and columns (the trace times the total sample size), and the eigenvalues of the $\mathbf{S}_1$ matrix are output. In addition, the relative frequencies and chi-squared distances to the centroid are listed for both the row and column variables.

The coordinates of the row and column variables on the factors are usually superimposed on the same scatter diagram. This allows one to not only appreciate the relationships among each set of variables (closer points are more similar), but also the relationships among the row and column variables. Note that Lebart *et al.* (1984, p. 46) warn that one must normally interpret the relative positions of points in one set (rows or columns) only in terms of the set of *all* points in the other set. It is dangerous to interpret the proximity of a particular row and column point. The interpretation must be in terms of projections (see below).

The matrices of absolute contributions can be examined to see which variables are the most important contributors to the variance explained by each factor. The matrices of squared correlations can be examined to determine on which factors a variable has the most influence.

### 9.3.3 Example

The data from page 198 of Lebart *et al.* (1984) serves as a convenient example. The rows correspond to 23 occupations and the columns to 15 "advantages" of each.

| | VARI | FREE | HUMA | SCHE | SALA | SECU | COMP | INTE |
|---|---|---|---|---|---|---|---|---|
| FARM1 | 4 | 189 | 0 | 3 | 2 | 2 | 9 | 3 |
| FARM2 | 1 | 13 | 3 | 10 | 17 | 12 | 4 | 1 |
| ENER | 1 | 9 | 1 | 0 | 4 | 13 | 0 | 2 |
| STEE | 5 | 5 | 2 | 9 | 18 | 5 | 3 | 2 |
| CHEM | 2 | 7 | 1 | 4 | 15 | 5 | 2 | 1 |
| WOOD | 2 | 5 | 0 | 4 | 1 | 0 | 3 | 0 |
| AUT | 2 | 3 | 1 | 8 | 16 | 17 | 1 | 8 |
| TEXT | 3 | 18 | 0 | 6 | 16 | 5 | 4 | 4 |
| PHAR | 3 | 7 | 3 | 6 | 6 | 0 | 0 | 2 |
| MANU | 0 | 18 | 1 | 12 | 31 | 7 | 0 | 8 |
| CONS | 7 | 63 | 2 | 9 | 31 | 9 | 4 | 6 |
| FOOD | 2 | 43 | 16 | 7 | 6 | 4 | 7 | 1 |
| SBUS | 8 | 95 | 23 | 15 | 15 | 2 | 13 | 7 |
| MBUS | 5 | 32 | 9 | 9 | 17 | 4 | 5 | 4 |
| TELE | 1 | 7 | 2 | 11 | 3 | 14 | 2 | 6 |
| SO.S | 4 | 10 | 10 | 8 | 2 | 1 | 6 | 4 |
| HE.S | 3 | 31 | 16 | 15 | 11 | 19 | 5 | 19 |
| TEAC | 2 | 33 | 27 | 31 | 9 | 18 | 27 | 24 |
| TRAN | 2 | 19 | 2 | 12 | 12 | 21 | 0 | 1 |
| BANK | 8 | 12 | 4 | 8 | 13 | 21 | 2 | 10 |
| DOME | 0 | 8 | 0 | 4 | 5 | 2 | 7 | 1 |
| O.SE | 8 | 35 | 14 | 13 | 16 | 10 | 6 | 25 |

| PRIV | 3 | 26 | 9 | 3 | 12 | 5 | 8 | 8 |
|---|---|---|---|---|---|---|---|---|
| | NEAR | ATMO | SOCI | AUTO | LIKE | NONE | OUTD | |
| FARM1 | 12 | 2 | 1 | 4 | 11 | 12 | 8 | |
| FARM2 | 8 | 3 | 5 | 1 | 9 | 11 | 0 | |
| ENER | 2 | 0 | 2 | 1 | 4 | 6 | 1 | |
| STEE | 6 | 5 | 5 | 0 | 2 | 22 | 0 | |
| CHEM | 6 | 1 | 2 | 2 | 3 | 5 | 0 | |
| WOOD | 2 | 1 | 1 | 1 | 1 | 3 | 0 | |
| AUT | 7 | 2 | 4 | 3 | 6 | 24 | 0 | |
| TEXT | 13 | 4 | 2 | 3 | 6 | 26 | 0 | |
| PHAR | 6 | 3 | 3 | 0 | 2 | 8 | 0 | |
| MANU | 19 | 11 | 3 | 2 | 10 | 26 | 0 | |
| CONS | 9 | 10 | 3 | 4 | 14 | 35 | 2 | |
| FOOD | 8 | 2 | 0 | 1 | 6 | 7 | 0 | |
| SBUS | 9 | 5 | 2 | 3 | 13 | 18 | 1 | |
| MBUS | 7 | 4 | 3 | 0 | 8 | 18 | 0 | |
| TELE | 3 | 1 | 1 | 2 | 1 | 5 | 0 | |
| SO.S | 2 | 3 | 1 | 0 | 3 | 1 | 0 | |
| HE.S | 10 | 2 | 3 | 7 | 24 | 5 | 0 | |
| TEAC | 3 | 4 | 43 | 8 | 18 | 11 | 1 | |
| TRAN | 4 | 5 | 5 | 1 | 3 | 13 | 0 | |
| BANK | 4 | 2 | 5 | 6 | 3 | 10 | 0 | |
| DOME | 5 | 7 | 2 | 1 | 2 | 11 | 1 | |
| O.SE | 6 | 4 | 10 | 9 | 11 | 14 | 0 | |
| PRIV | 4 | 4 | 2 | 3 | 10 | 8 | 0 | |

The listing output from CORRESP using these data is furnished below.

```
========== CORRESP ===== 4/17/88 22:50 ==========
Input matrix:LMWS.DTA
" SUBSET of Test data from Lebart, Morineau, & Warwich 1984,
" page 198, 26 occupations, 22 "advantages" of each.
type=1, size=23 by 15

Trace of matrix =    0.52594
X2              =  1445.809 df = 308
```

| i | Eigenvalue | Percent | Cumulative |
|---|---|---|---|
| 1 | 0.21108 | 40.13 | 40.13 |
| 2 | 0.12641 | 24.03 | 64.17 |
| 3 | 0.06108 | 11.61 | 75.78 |
| 4 | 0.03940 | 7.49 | 83.27 |
| 5 | 0.02359 | 4.48 | 87.76 |
| 6 | 0.01787 | 3.40 | 91.16 |
| 7 | 0.01309 | 2.49 | 93.64 |
| 8 | 0.00998 | 1.90 | 95.54 |
| 9 | 0.00763 | 1.45 | 96.99 |
| 10 | 0.00636 | 1.21 | 98.20 |
| 11 | 0.00394 | 0.75 | 98.95 |
| 12 | 0.00342 | 0.65 | 99.60 |
| 13 | 0.00167 | 0.32 | 99.92 |
| 14 | 0.00043 | 0.08 | 100.00 |

Next the relative frequency and Chi-squared distance (a squared distance) to the centroid is printed for each row and column variable.

Column variables:

| Variable | Rel. freq | Dist^2 |
|---|---|---|
| VARI | 0.02765 | 0.48314 |
| FREE | 0.25027 | 0.53052 |
| HUMA | 0.05311 | 0.79649 |
| SCHE | 0.07530 | 0.25945 |
| SALA | 0.10113 | 0.38033 |
| SECU | 0.07130 | 0.86253 |
| COMP | 0.04292 | 0.58735 |
| INTE | 0.05347 | 0.56097 |
| NEAR | 0.05638 | 0.30045 |
| ATMO | 0.03092 | 0.54342 |
| SOCI | 0.03929 | 1.28161 |
| AUTO | 0.02255 | 0.44892 |
| LIKE | 0.06184 | 0.19149 |
| NONE | 0.10877 | 0.36402 |
| OUTD | 0.00509 | 3.36656 |

Row variables:

| Variable | Rel. freq | Dist^2 |
|---|---|---|
| FARM1 | 0.09531 | 1.40142 |
| FARM2 | 0.03565 | 0.25020 |
| ENER | 0.01673 | 0.88334 |
| STEE | 0.03238 | 0.58971 |
| CHEM | 0.02037 | 0.46192 |
| WOOD | 0.00873 | 0.64689 |
| AUT | 0.03710 | 0.58948 |
| TEXT | 0.04001 | 0.36076 |
| PHAR | 0.01782 | 0.42082 |
| MANU | 0.05384 | 0.51420 |
| CONS | 0.07566 | 0.18560 |
| FOOD | 0.04001 | 0.41203 |
| SBUS | 0.08330 | 0.27977 |
| MBUS | 0.04547 | 0.10015 |
| TELE | 0.02146 | 0.76943 |
| SO.S | 0.02001 | 0.79993 |
| HE.S | 0.06184 | 0.36385 |
| TEAC | 0.09422 | 0.80694 |
| TRAN | 0.03638 | 0.46750 |
| BANK | 0.03929 | 0.50650 |
| DOME | 0.02037 | 0.74730 |
| O.SE | 0.06584 | 0.24100 |
| PRIV | 0.03820 | 0.14660 |

The matrix of row factors is:

| | 1 | 2 | 3 |
|---|---|---|---|
| FARM1 | 1.144 | -0.097 | -0.205 |
| FARM2 | -0.320 | -0.168 | -0.033 |
| ENER | -0.219 | -0.131 | -0.789 |
| STEE | -0.477 | -0.424 | 0.292 |
| CHEM | -0.331 | -0.330 | 0.024 |
| WOOD | -0.001 | 0.018 | 0.362 |
| AUT | -0.612 | -0.316 | -0.242 |

| | | | |
|---|---|---|---|
| TEXT | -0.191 | -0.473 | 0.157 |
| PHAR | -0.225 | -0.210 | 0.382 |
| MANU | -0.327 | -0.518 | 0.187 |
| CONS | 0.123 | -0.361 | 0.038 |
| FOOD | 0.402 | 0.169 | 0.168 |
| SBUS | 0.446 | 0.112 | 0.160 |
| MBUS | 0.035 | -0.127 | 0.206 |
| TELE | -0.450 | 0.140 | -0.535 |
| SO.S | -0.002 | 0.550 | 0.422 |
| HE.S | -0.138 | 0.313 | -0.211 |
| TEAC | -0.288 | 0.755 | 0.115 |
| TRAN | -0.314 | -0.202 | -0.372 |
| BANK | -0.435 | 0.047 | -0.430 |
| DOME | -0.123 | -0.285 | 0.398 |
| O.SE | -0.145 | 0.283 | -0.035 |
| PRIV | 0.052 | 0.136 | 0.111 |

The matrix of column factors is:

| | 1 | 2 | 3 |
|---|---|---|---|
| VARI | -0.075 | -0.023 | 0.067 |
| FREE | 0.722 | -0.025 | -0.076 |
| HUMA | -0.024 | 0.699 | 0.278 |
| SCHE | -0.328 | 0.184 | 0.078 |
| SALA | -0.357 | -0.426 | 0.107 |
| SECU | -0.551 | -0.003 | -0.706 |
| COMP | 0.060 | 0.509 | 0.376 |
| INTE | -0.343 | 0.433 | -0.147 |
| NEAR | -0.109 | -0.388 | 0.145 |
| ATMO | -0.225 | -0.364 | 0.388 |
| SOCI | -0.517 | 0.684 | 0.108 |
| AUTO | -0.204 | 0.254 | -0.292 |
| LIKE | -0.051 | 0.124 | -0.011 |
| NONE | -0.255 | -0.476 | 0.121 |
| OUTD | 1.432 | -0.210 | -0.486 |

In the matrices of "absolute contributions" one sees for each factor what proportion of its variation is explained by the variables. The matrix of row absolute contributions is given first. In this example one sees that the variable that contributes most to factor 1 is FARM1 with none of the other variables even close. TEAC is the largest contributor to factor 2 and no single row variable has a large contribution to factor 3.

| | 1 | 2 | 3 |
|---|---|---|---|
| FARM1 | 0.590 | 0.007 | 0.066 |
| FARM2 | 0.017 | 0.008 | 0.001 |
| ENER | 0.004 | 0.002 | 0.171 |
| STEE | 0.035 | 0.046 | 0.045 |
| CHEM | 0.011 | 0.018 | 0.000 |
| WOOD | 0.000 | 0.000 | 0.019 |
| AUT | 0.066 | 0.029 | 0.035 |
| TEXT | 0.007 | 0.071 | 0.016 |

| | | | |
|---|---|---|---|
| PHAR | 0.004 | 0.006 | 0.043 |
| MANU | 0.027 | 0.114 | 0.031 |
| CONS | 0.005 | 0.078 | 0.002 |
| FOOD | 0.031 | 0.009 | 0.019 |
| SBUS | 0.078 | 0.008 | 0.035 |
| MBUS | 0.000 | 0.006 | 0.032 |
| TELE | 0.021 | 0.003 | 0.101 |
| SO.S | 0.000 | 0.048 | 0.058 |
| HE.S | 0.006 | 0.048 | 0.045 |
| TEAC | 0.037 | 0.425 | 0.020 |
| TRAN | 0.017 | 0.012 | 0.083 |
| BANK | 0.035 | 0.001 | 0.119 |
| DOME | 0.001 | 0.013 | 0.053 |
| O.SE | 0.007 | 0.042 | 0.001 |
| PRIV | 0.000 | 0.006 | 0.008 |

The matrix of column absolute contributions is:

| | 1 | 2 | 3 |
|---|---|---|---|
| VARI | 0.001 | 0.000 | 0.002 |
| FREE | 0.618 | 0.001 | 0.024 |
| HUMA | 0.000 | 0.205 | 0.067 |
| SCHE | 0.038 | 0.020 | 0.008 |
| SALA | 0.061 | 0.145 | 0.019 |
| SECU | 0.102 | 0.000 | 0.581 |
| COMP | 0.001 | 0.088 | 0.099 |
| INTE | 0.030 | 0.079 | 0.019 |
| NEAR | 0.003 | 0.067 | 0.020 |
| ATMO | 0.007 | 0.032 | 0.076 |
| SOCI | 0.050 | 0.145 | 0.007 |
| AUTO | 0.004 | 0.011 | 0.032 |
| LIKE | 0.001 | 0.007 | 0.000 |
| NONE | 0.033 | 0.195 | 0.026 |
| OUTD | 0.049 | 0.002 | 0.020 |

The matrix of correlations squared is used to show for each variable the importance of each factor. The matrix of row correlations squared is given first. In this example one can see that most of the variation in FARM1 is explained by factor 1.

| | 1 | 2 | 3 |
|---|---|---|---|
| FARM1 | 0.933 | 0.007 | 0.030 |
| FARM2 | 0.409 | 0.113 | 0.004 |
| ENER | 0.054 | 0.020 | 0.705 |
| STEE | 0.386 | 0.305 | 0.145 |
| CHEM | 0.237 | 0.235 | 0.001 |
| WOOD | 0.000 | 0.001 | 0.202 |
| AUT | 0.636 | 0.170 | 0.099 |
| TEXT | 0.102 | 0.620 | 0.068 |
| PHAR | 0.120 | 0.105 | 0.346 |
| MANU | 0.208 | 0.522 | 0.068 |
| CONS | 0.082 | 0.703 | 0.008 |
| FOOD | 0.393 | 0.069 | 0.069 |
| SBUS | 0.710 | 0.045 | 0.091 |

| | | | |
|---|---|---|---|
| MBUS | 0.012 | 0.162 | 0.424 |
| TELE | 0.263 | 0.025 | 0.372 |
| SO.S | 0.000 | 0.378 | 0.222 |
| HE.S | 0.052 | 0.270 | 0.122 |
| TEAC | 0.103 | 0.707 | 0.016 |
| TRAN | 0.211 | 0.087 | 0.296 |
| BANK | 0.373 | 0.004 | 0.365 |
| DOME | 0.020 | 0.109 | 0.211 |
| O.SE | 0.087 | 0.331 | 0.005 |
| PRIV | 0.019 | 0.126 | 0.083 |

The matrix of column correlations squared is:

| | 1 | 2 | 3 |
|---|---|---|---|
| VARI | 0.012 | 0.001 | 0.009 |
| FREE | 0.983 | 0.001 | 0.011 |
| HUMA | 0.001 | 0.613 | 0.097 |
| SCHE | 0.415 | 0.130 | 0.024 |
| SALA | 0.335 | 0.477 | 0.030 |
| SECU | 0.352 | 0.000 | 0.578 |
| COMP | 0.006 | 0.441 | 0.240 |
| INTE | 0.210 | 0.335 | 0.039 |
| NEAR | 0.039 | 0.502 | 0.070 |
| ATMO | 0.094 | 0.244 | 0.278 |
| SOCI | 0.208 | 0.365 | 0.009 |
| AUTO | 0.093 | 0.143 | 0.190 |
| LIKE | 0.013 | 0.080 | 0.001 |
| NONE | 0.178 | 0.623 | 0.040 |
| OUTD | 0.609 | 0.013 | 0.070 |

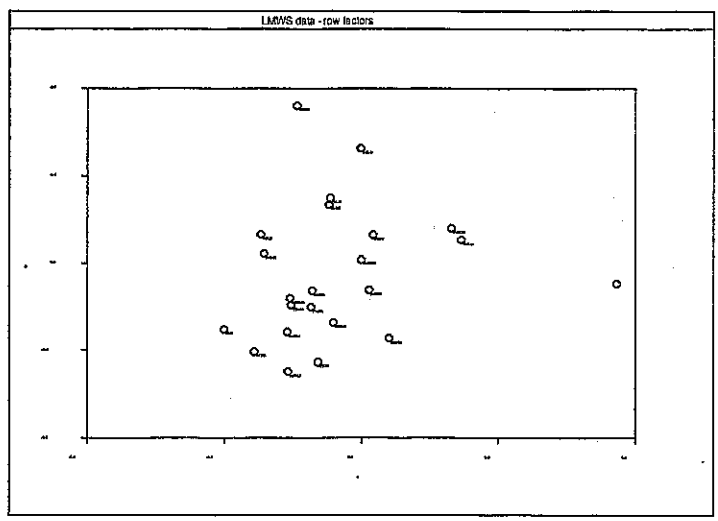


Figure 9.1.1. Plot of row factors 1 against 2 for the LMWS dta.

Next, row factors 1 and 2 were plotted against one another using the program MXPLOT (see Figure 9.1.1). The statistics and the plot are given below.

```
n= 23 r= 0.16944
 X: min=   -0.61233 max=   1.14356 mean=  -0.1041 var=   0.1419
 Y: min=   -0.51822 max=   0.75518 mean=  -0.0487 var=   0.1058
```

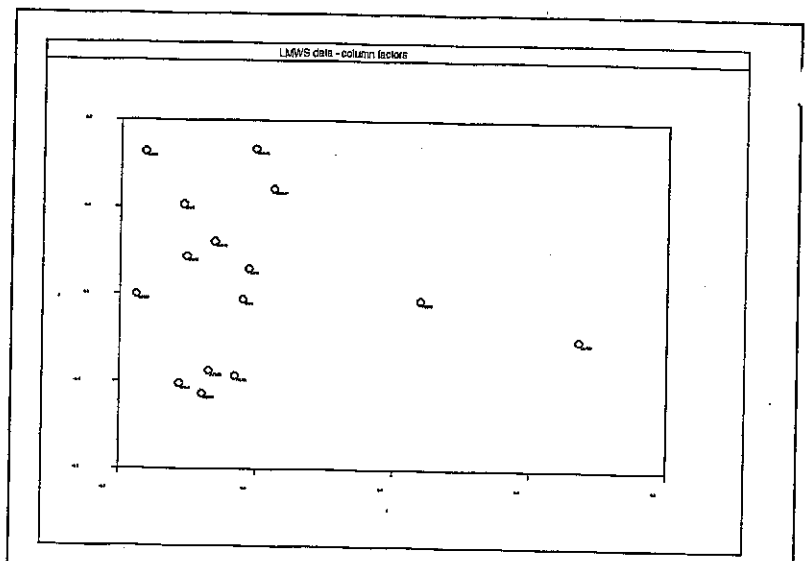


Figure 9.1.2. Plot of column factors 1 against 2 for the LMWS data.

The corresponding results for the column factors are given next (see Figure 9.1.2). The statistics were:

```
n= 15 r=-0.16737
 X: min=   -0.55077 max=    1.43197 mean=  -0.0550 var=    0.2594
 Y: min=   -0.47619 max=    0.69900 mean=   0.0647 var=    0.1564
```

In addition to examining each of the above two plots by themselves, it is of interest to study their interrelationships. The eigenvectors are the same in both plots so variation in the same directions in the two plots correspond. For example, the two row points, *e.g.*, AUT and FOOD, differ in a direction of about 45° in the first plot. Hence those two row variables should differ in column variables such as SALA since a vector connecting it and the origin is more or less parallel. In a symmetrical manner, one should expect column variables SALA and FREE to differ with respect to a row variable such as FOOD.