Multivariate Analysis, MT2004

Correspondence Analysis

©1994–2004 B. D. Ripley

In discrete multivariate analysis we suppose we have several categorical factors measured for each case. Unlike logistic regression and (most instances) of log-linear models we will be interested in problems where most of the factors are *responses* and not *history* variables.

1 Correspondence Analysis

Correspondence analysis is applied to two-way tables of counts, especially large tables. A small example is Fisher's (1940) data on colours of eyes and hair of people in Caithness, Scotland:

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

which Fisher used for an early example of correspondence analysis. However, the names comes from the Francophone school of *analyses des données* who called it *l'analyse factorielle de correspondences*; the principal developer is Benzecri. Because like PCA the method can be seen in several different ways, there has been a lot of confusion in the literature until recently.

Suppose we have an $r \times c$ table N of counts. Fisher sought 'scores' f and g for the rows and columns which are maximally correlated. Each individual has a row label and a column label, and we can associate a score with each label, so each individual has two numeric variables, and we look at the sample correlation of these. (Just to be confusing, in this field variances are normally computed without subtracting the mean.) Clearly the maximum 'correlation' is one, attained by constant scores, so we seek the largest non-trivial solution.

Let R and C be matrices of the group indicators[1] of the rows and columns, so $R^T C = N$. Then the 'correlation' matrix is

$$X_{ij} = \frac{n_{ij}/n - (n_{i\cdot}/n)(n_{\cdot j}/n)}{\sqrt{(n_{i\cdot}/n)(n_{\cdot j}/n)}} = \frac{n_{ij} - n\, r_i\, c_j}{n\sqrt{r_i\, c_j}}$$

where $r_i = n_{i\cdot}/n$ and $c_j = n_{\cdot j}/n$ are the proportions in each row and column. This should be familiar, as $\sum_{i,j} X_{ij}^2$ is the chi-squared test of no association in the table, that is of independence of the row and column attributes. In the terminology of correspondence analysis, the *row profile* is (r_i) and the *column profile* is (c_i), and the chi-squared test is of differences in the row profiles (or equivalently, in the column profiles). Let D_r and D_c be the diagonal matrices of (r_i) and (c_j) respectively.

Fisher's analysis corresponds to forming the singular value decomposition $X = U\Lambda V^T$ and selecting the first singular value and left and right singular vectors of X_{ij} and rescaling by $(r_i)^{-1/2}$ and $(c_j)^{-1/2}$, respectively. So this is a sort of weighted decomposition of N, adjusting for the difference prevalences of different rows and columns.

This is done by function `corresp` in library `MASS`:

[1] so R is a $n \times r$ matrix of zeroes and ones, with a single one in each row, and analogously for C.

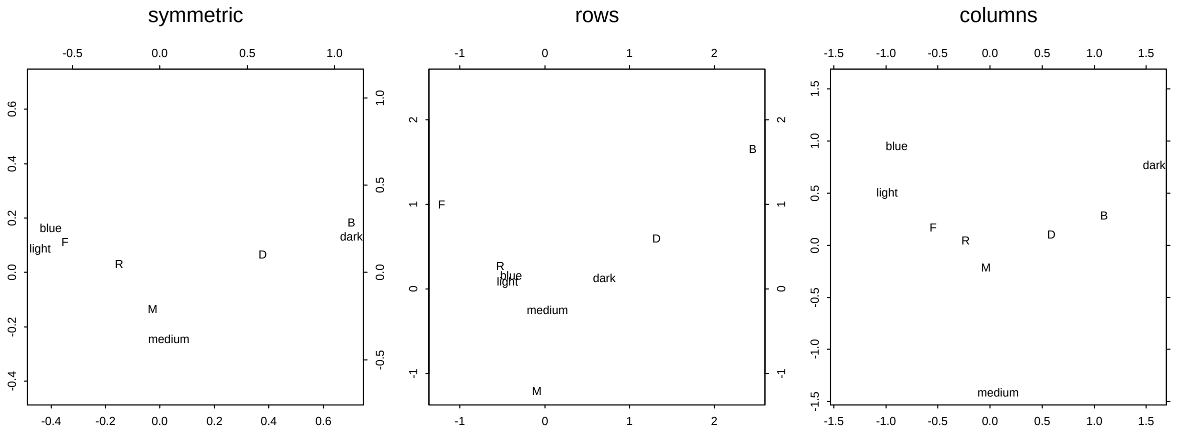

Figure 1: Three variants of correspondence analysis plots from Fisher’s data on people in Caithness: (left) ‘symmetric”, (middle) ‘row asymmetric’ and (right) ‘column asymmetric’.

```
> corresp(caith)
First canonical correlation(s): 0.44637

 eyes scores:
     blue    light   medium   dark
 -0.89679 -0.98732 0.075306 1.5743

 hair scores:
    fair      red    medium   dark  black
 -1.2187 -0.52258 -0.094147 1.3189 2.4518
```

Can we make use of the second and further singular values? In what Gower & Hand (1996) call ‘classical CA’ we consider $A = D_r^{-1/2} U\Lambda$ and $B = D_c^{-1/2} V\Lambda$. Then the first columns of A and B are what we have termed the row and column scores *scaled by* ρ, the first canonical correlation. More generally, we can see distances between the rows of A as approximating the distances between the row profiles, and analogously for the rows of B and the column profiles.

Classical CA plots the first two columns of A and B on the same figure. This is a form of a biplot, sometimes known as a ‘symmetric’ plot. Note that unlike PCA biplots, the Λ enters in both A and B. Other authors (for example, Greenacre, 1992) advocate ‘asymmetric’ plots. The asymmetric plot for the rows is a plot of the first two columns of A with the column labels plotted at the first two columns of $\Gamma = D_c^{-1/2} V$; the corresponding plot for the columns has columns plotted at B and row labels at $\Phi = D_r^{-1/2} U$. These are much closer to PCA biplots for the rows and columns respectively.

The two-dimensional forms of the plot are shown in Figure 1. These were produced by

```
caith2 <- caith  # make shorter labels
dimnames(caith2)[[2]] <- c("F", "R", "M", "D", "B")
par(mfcol = c(1, 3))
plot(corresp(caith2, nf = 2)); title("symmetric")
plot(corresp(caith2, nf = 2), type = "rows"); title("rows")
plot(corresp(caith2, nf = 2), type = "col"); title("columns")
```

Note that the symmetric plot (left) has the row points from the asymmetric row plot (middle) and the column points from the asymmetric column plot (right) superimposed on the same plot (but with different scales).

2 Multiple correspondence analysis

Multiple correspondence analysis (MCA) is (confusingly!) a method for visualizing the joint properties of $p \geqslant 2$ categorical variables that does *not* reduce to correspondence analysis (CA) for $p = 2$, although the methods are closely related (see, for example, Gower & Hand, 1996, §10.2).

Suppose we have n observations on the p factors with ℓ total levels. Consider G, the $n \times \ell$ indicator matrix whose rows give the levels of each factor for each observation. Then all the row sums are p. MCA is often (Greenacre, 1992) defined as CA applied to the table G, that is the singular-value decomposition of $D_r^{-1/2}(G/\sum_{ij} g_{ij})D_c^{-1/2} = U\Lambda V^T$. Note that $D_r = pI$ since all the row sums are p, and $\sum_{ij} g_{ij} = np$, so this amounts to the SVD of $p^{-1/2}GD_c^{-1/2}/pn$.

An alternative point of view is that MCA is a principal components analysis of the data matrix $X = G(pD_c)^{-1/2}$; with PCA it is usual to centre the data, but it transpires that the largest singular value is one and the corresponding singular vectors account for the means of the variables. A simple plot for MCA is to plot the first two principal components of X (which correspond to the second and third singular vectors of X). This is a form of biplot, but it will not be appropriate to add axes for the columns of X as the possible values are only $\{0, 1\}$, but it is usual to add the positions of 1 on each of these axes, and label these by the factor level. (The 'axis' points are plotted at the appropriate row of $(pD_c)^{-1/2}V$.) The point plotted for each observation is the vector sum of the 'axis' points for the levels taken of each of the factors. Gower and Hand seem to prefer (e.g., their Figure 4.2) to rescale the plotted points by p, so they are plotted at the centroid of their levels. This is exactly the asymmetric row plot of the CA of G, apart from an overall scale factor of $p\sqrt{n}$.

We can apply this to the example of (Gower & Hand, 1996, p. 75) by

```
farms.mca <- mca(farms, abbrev = T)  # Use levels as names
plot(farms.mca, cex = rep(0.7, 2), axes = F)
```

References

Fisher, R. A. (1940) The precision of discriminant functions. *Annals of Eugenics (London)* **10**, 422–429.

Gower, J. C. and Hand, D. J. (1996) *Biplots*. London: Chapman & Hall.

Greenacre, M. (1992) Correspondence analysis in medical research. *Statistical Methods in Medical Research* **1**, 97–117.

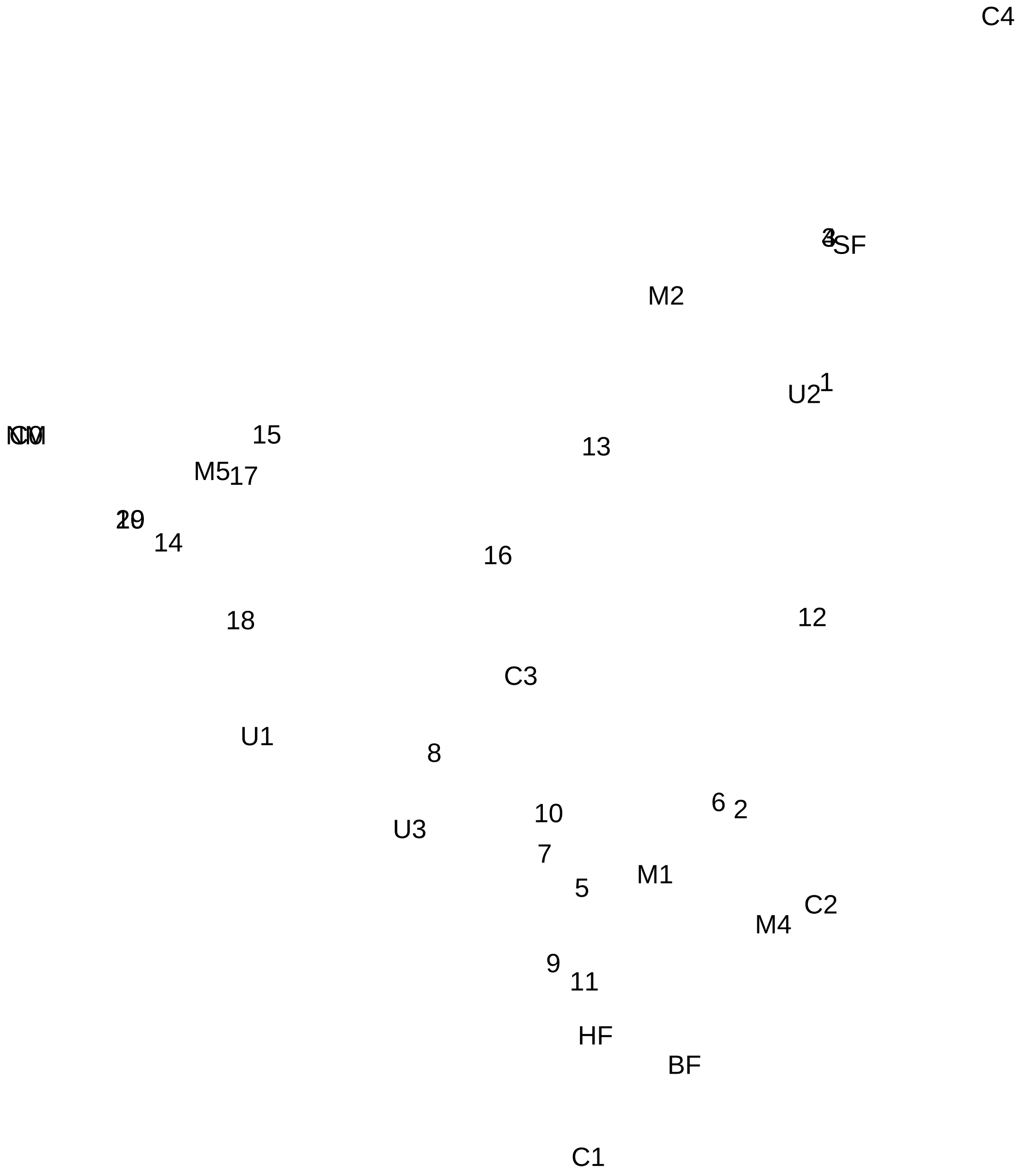

Figure 2: Multiple correspondence analysis plot of dataset `farms` on 20 farms on the Dutch island of Terschelling. Numbers represent the farms and labels levels of moisture (`M1`, `M2`, `M4` and `M5`), grassland usage (`U1`, `U2` and `U3`), manure usage (`C0` to `C4`) and type of grassland management (`SF`: standard, `BF`: biological, `HF`: hobby farming, `NM`: nature conservation). Levels `C0` and `NM` are coincident (on the extreme left), as are the pairs of farms 3 & 4 and 19 & 20.