Illustrations, with comments and sources, for categorical data analysis

Begin w/ the caith example, here defined as a **matrix (i.e., as.matrix(caith))**

```
caithM:   F   R   M   D   B

blue     326  38 241 110   3

light    688 116 584 188   4

medium   343  84 909 412  26

dark      98  48 403 681  85   using my function: crossd.svd
```

```
crossd.svd(caithM,2)$obs      # For those who ask, I will provide this
   function. It is designed for correspondence analysis, including many
   auxiliary outputs, as will be seen below. Graphics are also easy
   following computation of certain outputs, esp. coefsRt below.

  canonical correlations are:   0.446 0.173 0.029 0.022

  Square roots of singular values for Cont. table analysis are:

       5.72 3.57 1.47 0.00

The chi squared statistic for Cont. table is: 1240.03 with d.f.= 12

    obs = observed frequencies matrix:

          F   R    M    D   B    blue light medium dark

F      1455   0    0    0   0     326   688    343   98

R         0 286    0    0   0      38   116     84   48

M         0   0 2137    0   0     241   584    909  403

D         0   0    0 1391   0     110   188    412  681

B         0   0    0    0 118       3     4     26   85

blue    326  38  241  110   3     718     0      0    0

light   688 116  584  188   4       0  1580      0    0

medium  343  84  909  412  26       0     0   1774    0

dark     98  48  403  681  85       0     0      0 1315
```

This matrix has been partitioned. The col. sums in diagonal above; the row sums in the diagonal below (right), i.e. marginal frequencies, and cross-tabs in the full matrix (symmetric). This is a raw sums of cross-products matrix for Eye and Hair Color data; it is readily converted into a variance covariance matrix, seen below as cv.

That is, if the preceding is called the 'observed' frequencies matrix, we need only subtract from it, `exp = (1/n)*outer(vsums,vsums)` where vsums refers to vector of sums given as the diagonal of the obs matrix above, and n is sum of all entries off-diagonal (i.e. no. of cases); the function outer( ) finds the outer product of this vector w/ itself.

The name 'exp' here refers to expected values for cells if rows and columns are independent. (I do not print this exp.)

So obs – exp, could be generated for the whole obs matrix; but we will only print the obs – exp (numerators of terms in Chi Square statistic):

```
blue    132.1   -0.119  -43.8  -75.4 -12.73   #so exp could be obtained

light   261.2   32.117  -42.8 -220.0 -30.61   # for each cell by adding

medium -136.1  -10.183  205.3  -46.1 -12.86   # observed values above.

dark   -257.2  -21.814 -118.7  341.4  56.20
```

The entries of the preceding matrix are the deviations from expected values under the assumption of row and column independence. The sum of all entries is zero, by row and by column.

Correspondence analysis, generally, aims to exhibit structure that may exist in the pattern of values in the preceding matrix. The larger the chi square statistic, the greater the structure; and that structure could be 1, 2, 3 or even higher dimensional (although 2 dimensional is often sufficient to capture the main structural information).

The chi square statistic, the sum of squares of ((obs – exp)/√exp), across rows & columns, here is: 1240.03 with d.f.= 12; i.e. very large, so it is reasonable to search for structure.

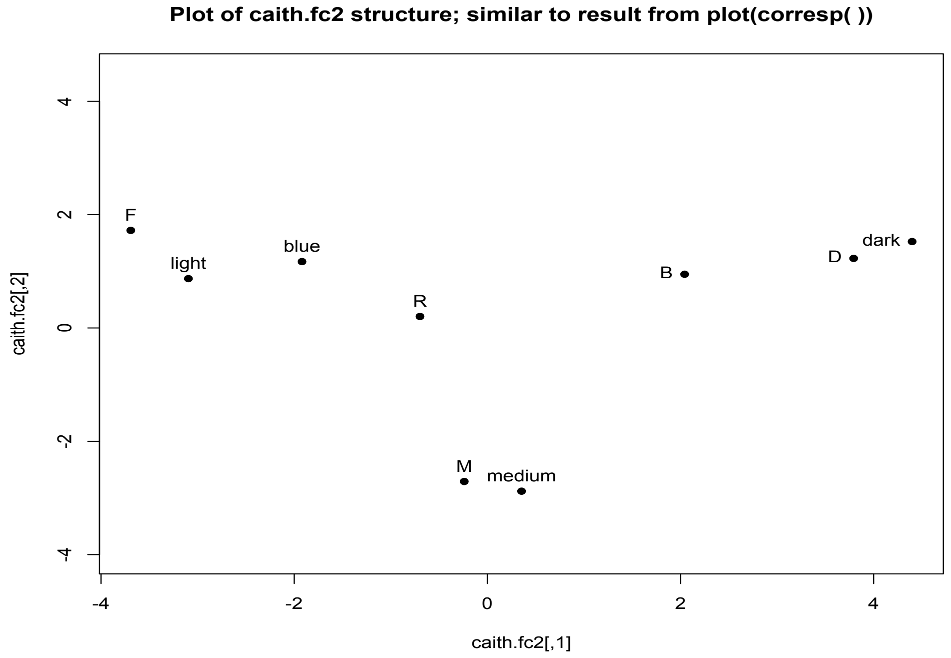
As seen above, the canonical correlations are: 0.446 0.173 0.029 0.022; I've underlined the first two, as they are both much larger than the last two, and so these data appear to have a 2 dimensional structure.

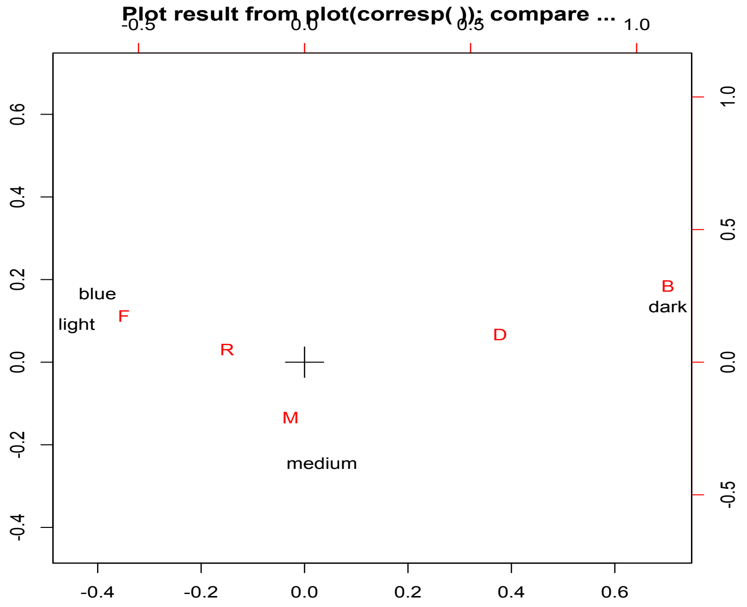
The key structural information is seen in the following coefficients matrix, where the two columns correspond to two 'derived factors' that account for more than 90% of the variance in the obs – exp matrix.

```
F      -3.692  1.723

R      -0.697  0.203

M      -0.239 -2.710

D       3.793  1.228

B       2.043  0.949

blue   -1.919  1.172

light  -3.095  0.871

medium  0.355 -2.881

dark    4.399  1.525
```

Define this 9 x 2 matrix as coefs2; then coefs2 %*% t(coefs2), the product of the matrix times its transpose, is what shows how much of the structure of the matrix ((obs – exp)/√exp) is recovered in this case. But it is the plot of these two columns that provides the key structural information, as seen here: (I use slightly different names, but the matrix caith.fc2 is just coefs2. plot(caith.fc2,ylim=c(-4,4.5))

```
#ylim not essential (will explain) then
identify(caith.fc2,labels=rownames(caith.fc2)) #to identify points.
```





Different algorithms yield (slightly) different results; but the structures are clearly similar. The interpretation follows from that of a conventional x,y scatterplot; values close proximity to one another for show how particular hair colors and eye colors ‘go together’ for this sample of data. The next example entails a larger matrix; I chose it because there is a detailed examination in the Intro2C.A.pdf.

The data here pertain to frequencies of doctorates granted in 8 selected years (columns) for 12 different disciplines in the U.S. I follow the same procedures outlined above, w/ little commentary, until the end. The matrix of frequencies is called doctx:

	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	1340	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

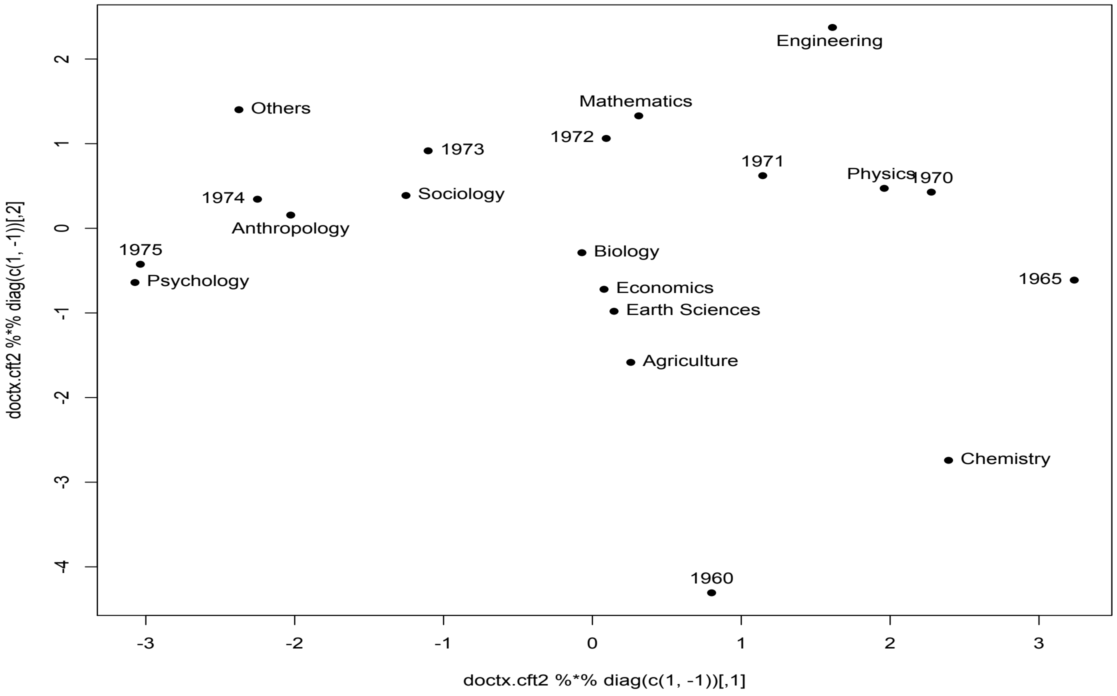
```
crossd.svd(doctx)

canonical correlations are:0.096 0.057 0.017 0.014 . .
                   2 dimensions appear to be indicated.

Square roots of singular values for Cont. table analysis are:"

5.87 4.52 2.47 2.22 1.71 1.53 1.37 0.01
```

The chi squared statistic is: 1684.37 with d.f.= 77 #again, LARGE, so we may very well find structure. We shall go immediatedly to the coefficients matrix, here (20 x 2), but arrayed on left/right here:

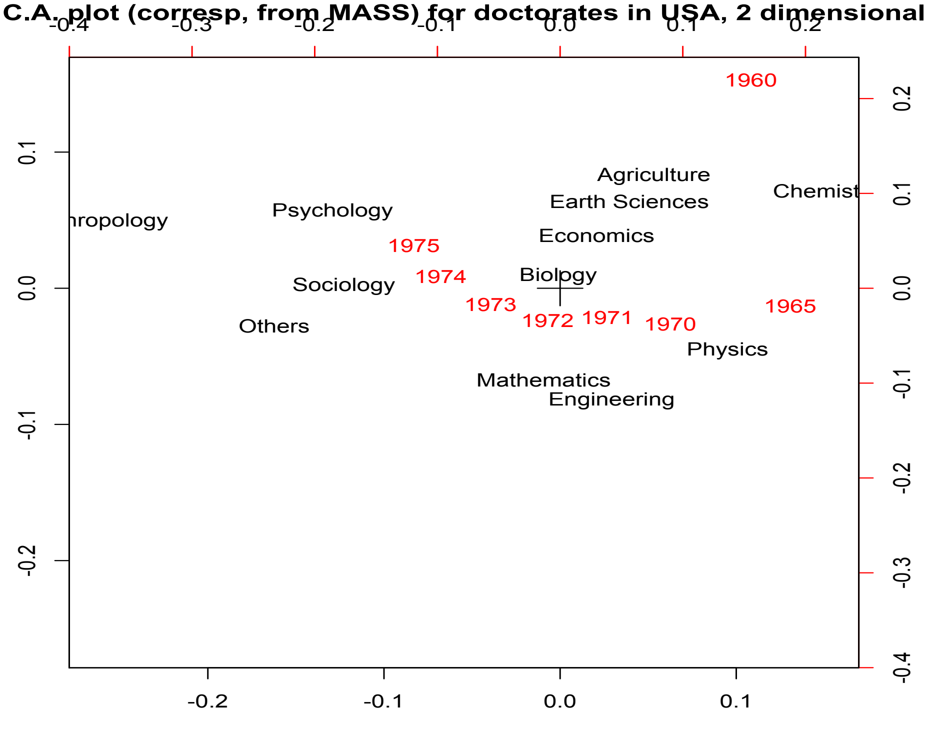
```
          [,1]   [,2]        Engineering      1.6126 -2.373
1960      .8007  4.307       Mathematics      0.3112 -1.329
1965     3.2370  0.612       Physics          1.9604 -0.472
1970     2.2767 -0.427       Chemistry        2.3925  2.742
1971     1.1441 -0.622       Earth Sciences   0.1452  0.980
1972     0.0929 -1.062       Biology         -0.0707  0.288
1973    -1.1035 -0.916       Agriculture      0.2577  1.584
1974    -2.2500 -0.344       Psychology      -3.0731  0.640
1975    -3.0368  0.425       Sociology       -1.2525 -0.387
                             Economics        0.0781  0.721
                             Anthropology    -2.0267 -0.156
                             Others          -2.3748 -1.401
```

Examination of patterns here will show which cols go w/ one another, as well as which rows are similar; also which rows and columns. But the graphic shows this far more vividly. I present it in two forms below.

C.A. plot (crossd.svd) for doctorates in USA, 2 dimensional



C.A. plot (corresp, from MASS) for doctorates in USA, 2 dimensional

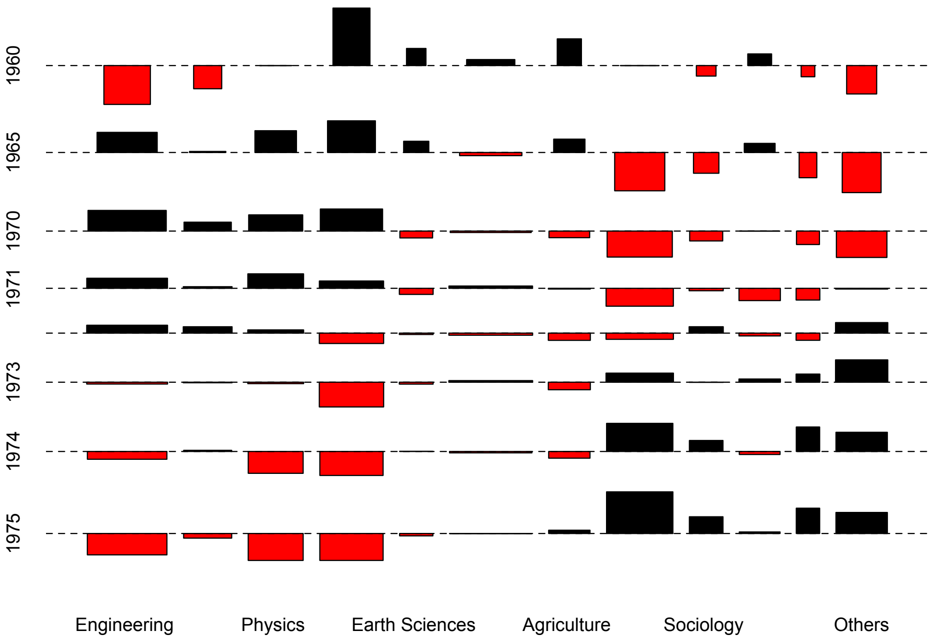


Note that the vertical dimension is reversed when comparing the two graphics. This is to be expected, as signs of columns of coefficients are always arbitrary. There are other differences too, but there is a general similarity. Do you prefer one to the other? Which?

General remarks: for 3 dimensions and more, plots can be done for columns as pairs, or dynamic graphics could be employed (possibly scatter3d, but I have not tried it). Clearly there are many ways to go, but in these two cases, 2 dimensional solutions see, to have been revealing of structure that 'makes sense.' See the Intro2CA.pdf where this example is described in more detail, and yet another graphic is shown, and it differs from each of these, at least to some extent.

An assocplot for these data is also revealing: (dput version of data below; dget can be used to recover this object)

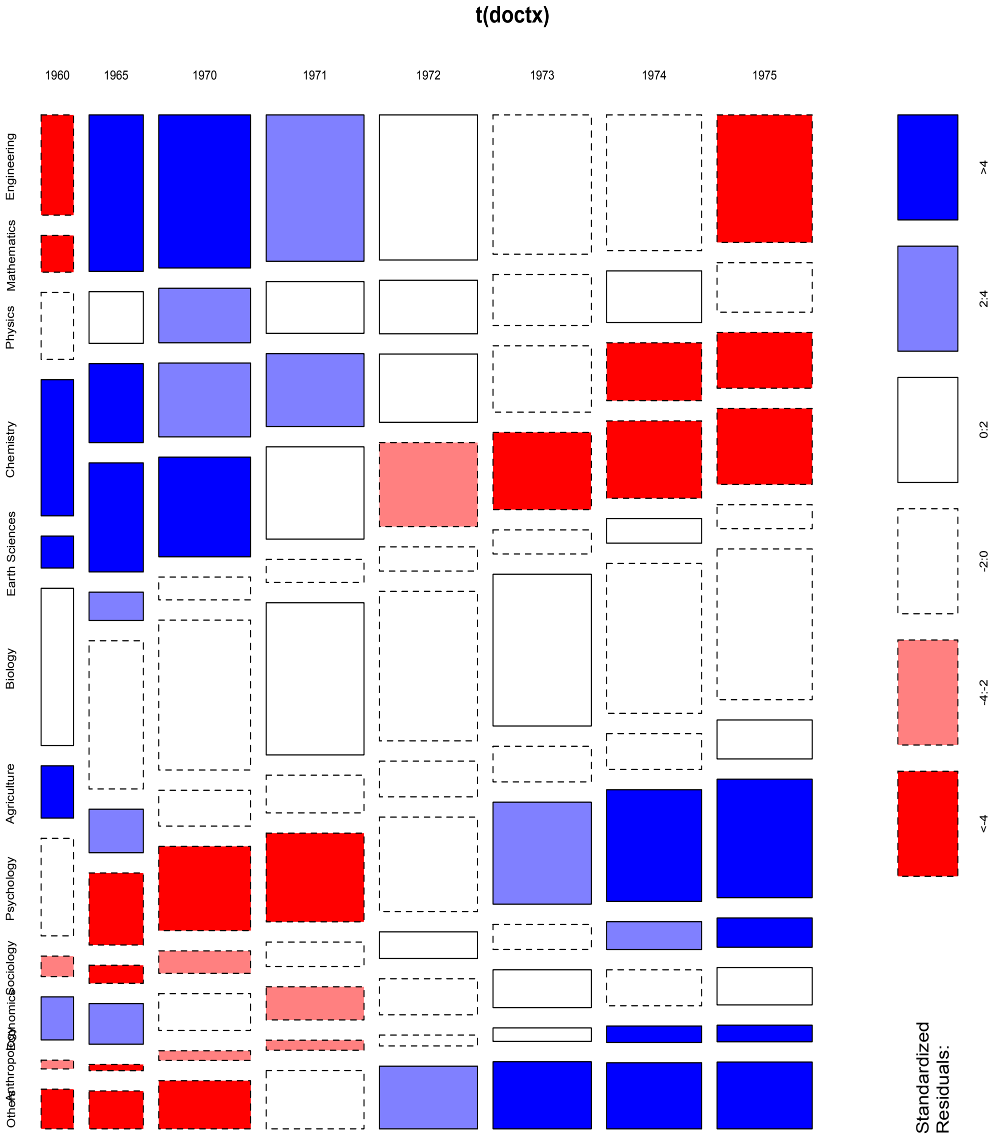
An assocplot for the doctx data; note the patterns in deviations from independence



doctx:

```
structure(c(794, 291, 530, 1078, 253, 1245, 414, 772, 162, 341,
69, 314, 2073, 685, 1046, 1444, 375, 1963, 576, 954, 239, 538,
82, 502, 3432, 1222, 1655, 2234, 511, 3360, 803, 1888, 504, 826,
217, 1079, 3495, 1236, 1740, 2204, 550, 3633, 900, 2116, 583,
791, 240, 1392, 3475, 1281, 1635, 2011, 580, 3580, 855, 2262,
638, 863, 260, 1500, 3338, 1222, 1590, 1849, 577, 3636, 853,
2444, 599, 907, 324, 1609, 3144, 1196, 1340, 1792, 570, 3473,
830, 2587, 645, 833, 381, 1531, 2959, 1149, 1293, 1762, 556,
3498, 904, 2749, 680, 867, 385, 1550), .Dim = c(12L, 8L), .Dimnames =
list(
c("Engineering ", "Mathematics ", "Physics ", "Chemistry ",
"Earth Sciences ", "Biology ", "Agriculture ", "Psychology ",
"Sociology ", "Economics ", "Anthropology", "Others"), c("1960",
"1965", "1970", "1971", "1972", "1973", "1974", "1975")))
```

Finally, let us use the mosaicplot function (after transposing the matrix doctx). Interpret in relation to what you have seen above.


t(doctx)
Engineering
Mathematics
Physics
Chemistry
Earth Sciences
Biology
Agriculture
Psychology
Sociology
Economics
Anthropology
Others
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals: