Fiche TD avec le logiciel R : course3

Correspondence Analysis (COA or CA)

A.B. Dufour

Contents

1 Introduction

Correspondence Analysis (CA) is a set of theoretical results, statistical uses and examples. Nishisato [2] calls the method 'dual scaling' but gives a lot of terms such as:
- the method of reciprocal averages
- additive scoring
- appropriate scoring
- canonical scoring
- Guttman weighting
- principal component analysis of qualitative data
- optimal scaling
- Hayashi's theory of quantification
- simultaneous linear regression
- correspondence factor analysis
- biplot.

The relationships between all these proposed methods are clear using the duality diagram as a mathematical framework. The correspondence analysis was described by Benzecri [1].

In ecology, CA is used for the study of abundance matrices such as sites × species tables. As for P.C.A., the objective is the reduction of the number of variables under the constraint of maximizing the kept inertia. One can provide a scatter plot of contingency tables.

2 Introducing the method by using an example

The data we will use here come from Snee (1974) [3].

> *"The data are the observed frequencies of hair color (black, brunette, red, blond) and eye color (brown, blue, hazel, green) of 592 subjects. These data were collected, as part of a class project, by students in an elementary statistics course taught by the author at the University of Delaware."*

```
snee74 <- read.table("snee74e.txt", h = T)
names(snee74)
```

```
[1] "hair_colour" "eye_colour"  "sex"
```

```
head(snee74)
```

```
  hair_colour eye_colour    sex
1       black      brown   male
2       blond       blue female
3       black       blue   male
4    brunette      brown female
5         red      brown   male
6    brunette       blue   male
```

The hair colour is a categorial variable (also called factor in the R software) with four modalities (levels): black, brunette, red and blond.

```
colhair <- snee74$hair
levels(colhair)
```

```
[1] "black"    "blond"    "brunette" "red"
```

```
summary(colhair)
```

```
   black    blond brunette      red
     108      127      286       71
```

The eye colour is a categorial variable with four modalities: brown, blue, hazel and green.

```
coleye <- snee74$eye
levels(coleye)
[1] "blue"  "brown" "green" "hazel"
summary(coleye)
 blue brown green hazel
  215   220    64    93
```

The link between these two qualitative variables is a two-way table or a contingency table.

```
(eyehair <- table(coleye, colhair))
       colhair
coleye  black blond brunette red
  blue     20    94       84  17
  brown    68     7      119  26
  green     5    16       29  14
  hazel    15    10       54  14
```

We now decide to associate a score to each column-level of the hair colour. The idea here is to separate dark colours (`black`, `brunette`) and light colours (`red`, `blond`).
We therefore choose to associate a score of 1 to dark colours, and a score of -1 to light colour.
For each row (eye colour), one can calculate observed relative frequencies linked to hair colours. If we choose the 'brown' colour (row 2), we obtain:

```
dfcolours <- data.frame(unclass(eyehair))
print(dfcolours[2, ]/sum(dfcolours[2, ]), digits = 4)
       black   blond brunette    red
brown 0.3091 0.03182   0.5409 0.1182
```

Based on this, one can therefore calculate an average score to the 'brown' eye colour:

```
browneye <- dfcolours[2, ]/sum(dfcolours[2, ])
scorehair <- c(1, -1, 1, -1)
sum(browneye * scorehair)
[1] 0.7
```

This average positive score shows that students with brown eyes tend to have dark hair.
This average score can be computed for all eye colours.

```
freqeye <- apply(dfcolours, 1, function(x) x/sum(x))
freqeye
               blue      brown    green     hazel
black    0.09302326 0.30909091 0.078125 0.1612903
blond    0.43720930 0.03181818 0.250000 0.1075269
brunette 0.39069767 0.54090909 0.453125 0.5806452
red      0.07906977 0.11818182 0.218750 0.1505376
t(freqeye)
           black      blond  brunette        red
blue  0.09302326 0.43720930 0.3906977 0.07906977
brown 0.30909091 0.03181818 0.5409091 0.11818182
green 0.07812500 0.25000000 0.4531250 0.21875000
hazel 0.16129032 0.10752688 0.5806452 0.15053763
scoreyes <- apply(t(freqeye), 1, function(x) sum(x * scorehair))
scoreyes
       blue       brown       green       hazel
-0.03255814  0.70000000  0.06250000  0.48387097
```

For blue eyes, we obtain an average score equal to -0.0326. The score is negative, meaning that dark hair are less common for this sub-population.

We can separate the four eye colours using the scoring proposed for the hair colour. But,
- Can we find a better hair score to discriminate eye colours ?
- How can we characterize a score for understanding the data structure when we have no idea about the subject ?

The Correspondence Analysis (COA or CA) is the appropriate method to find row-scores (or column-scores) such that the average row-scores (or the average column-scores) are separated as much as possible.

The correspondence analysis provides the following optimal scores for columns (i.e. hair):

```
 library(ade4)
 coa0 <- dudi.coa(dfcolours, scannf = F, nf = 3)
 rownames(coa0$c1)
[1] "black"    "blond"    "brunette" "red"
 print(coa0$c1[, 1], digits = 4)
[1]  1.1043 -1.8282  0.3245  0.2835
```

Exercise. Let's find the average scores of rows (i.e. eyes) using the optimal scores (i.e hair) given by the correspondence analysis.

```
      blue      brown      green      hazel 
-0.5474139  0.4921577 -0.1617534  0.2125969 
```

These results can be find directly with the correspondence analysis:

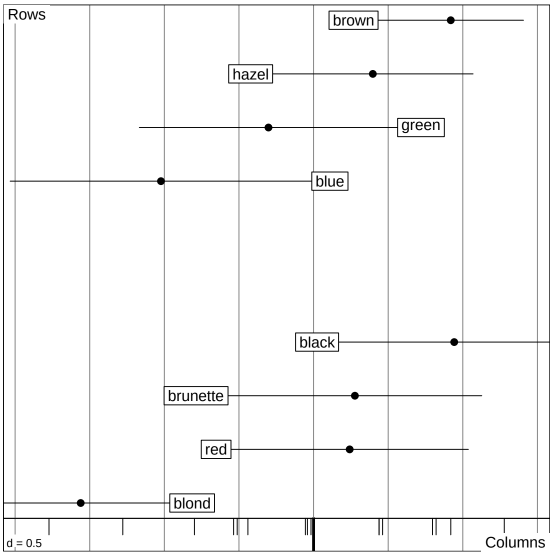
```
 rownames(coa0$li)
[1] "blue"  "brown" "green" "hazel"
 print(coa0$li[, 1])
[1] -0.5474139  0.4921577 -0.1617534  0.2125969
```

Entering the hair colour or the eye colour first has no importance. So one can apply the same process on the hair colour looking for optimal scores based on the eyes criteria. The analysis provides:

```
 rownames(coa0$co)
[1] "black"    "blond"    "brunette" "red"
 print(coa0$co[, 1], digits = 4)
[1]  0.5046 -0.8353  0.1483  0.1295
```

These two average scores (column-hair colour and row-eye colour) can be displayed on the same graph.

```
 score(coa0)
```



3 The contingency table

3.1 Dataset

The two-way table of the two factors is called an observed contingency table and,

i) an individual gets one and only one level by variable,

ii) each level must be observed once (deleted otherwise).

The information required are the total number of individuals (n), the number of levels for the first variable (I) and the number of levels for the second variable (J).

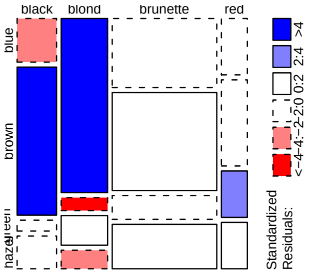
```
n <- sum(eyehair)
I <- 4
J <- 4
```

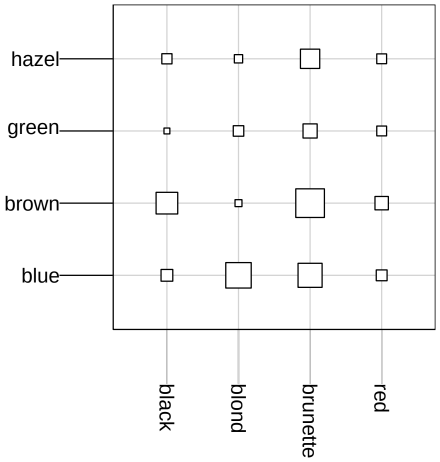
One can build the table of relative frequencies $f_{ij} = \frac{n_{ij}}{n}$.

```
freqcolours <- eyehair/n
round(freqcolours, digits = 4)
      colhair
coleye   black  blond brunette    red
  blue  0.0338 0.1588   0.1419 0.0287
  brown 0.1149 0.0118   0.2010 0.0439
  green 0.0084 0.0270   0.0490 0.0236
  hazel 0.0253 0.0169   0.0912 0.0236
```

One can represent the contingency table.

```
par(mfrow = c(1, 2))
mosaicplot(t(dfcolours), shade = T, main = "")
table.cont(dfcolours, csize = 2)
```





3.2 Row-profile and column-profile tables

One can compute the conditional frequencies. Let's call V_1 and V_2 the two categorical variables.

Row-profiles

Let's call $f_{i|j}$ the conditional frequencies linked to row-profiles:

$$f_{j|i} = P(V_2 = j|V_1 = i) = \frac{P(V_2 = j \cap V_1 = i)}{P(V_1 = i)}$$

$$f_{j|i} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i.}}{n}} = \frac{n_{ij}}{n_{i.}}$$

```
 RowProf <- prop.table(eyehair, 1)
 RowProf
       colhair
coleye       black      blond   brunette        red
  blue  0.09302326 0.43720930 0.39069767 0.07906977
  brown 0.30909091 0.03181818 0.54090909 0.11818182
  green 0.07812500 0.25000000 0.45312500 0.21875000
  hazel 0.16129032 0.10752688 0.58064516 0.15053763
```

Within the sub-population of individuals observed that had brown eyes, 30.91% had black hair. One can verify that all the row sums equal to 1.

```
 apply(RowProf, 1, sum)
 blue brown green hazel
    1     1     1     1
```

Column-profiles

Let's $f_{i|j}$ the conditional frequencies linked to column-profiles:

$$f_{i|j} = P(V_1 = i|V_2 = j) = \frac{P(V_1 = i \cap V_2 = j)}{P(V_2 = j)}$$

$$f_{i|j} = \frac{\frac{n_{ij}}{n}}{\frac{n_{.j}}{n}} = \frac{n_{ij}}{n_{.j}}$$

```
ColProf <- prop.table(eyehair, 2)
ColProf
      colhair
coleye      black      blond   brunette        red
  blue  0.18518519 0.74015748 0.29370629 0.23943662
  brown 0.62962963 0.05511811 0.41608392 0.36619718
  green 0.04629630 0.12598425 0.10139860 0.19718310
  hazel 0.13888889 0.07874016 0.18881119 0.19718310
```

Within the sub-population of individuals observed that had black hair, 62.96% had brown eyes. One can verify that all the column sums equal to 1.

```
apply(ColProf, 2, sum)
   black    blond brunette      red
       1        1        1        1
```

3.3 Linking this to the χ^2 test

The null and alternative hypotheses of the Chi-square test between two categorial variables are:

- ⋆ H_0 : The two categorial variables are independent.
- ⋆ H_1 : The two ctegorial variables are dependent.

Under the null hypothesis H_0, $P(V_2 = j \cap V_1 = i) = P(V_1 = i) \times P(V_2 = j)$.
Therefore, under H_0, the expected frequencies are $\frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$.
One can deduce the expected or theoretical contingency table keeping the observed marginal counts.

$$\frac{n_{i.} \times n_{.j}}{n}$$

If we go back to our example, and want to test whether hair and eye colour are related, we get:

- ⋆ H_0 : Hair colour is independent of eye colour
- ⋆ H_1 : Hair colour is dependent of eye colour

```
reschi <- chisq.test(eyehair)
reschi$expected
      colhair
coleye    black    blond  brunette       red
  blue  39.22297 46.12331 103.86824 25.785473
  brown 40.13514 47.19595 106.28378 26.385135
  green 11.67568 13.72973  30.91892  7.675676
  hazel 16.96622 19.95101  44.92905 11.153716
```

The chi-squared test statistic is:

$$\chi^2 \; = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

and approximatively approaches a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom. The results of the chi-square test is:

```
reschi
```

```
        Pearson's Chi-squared test
data:  eyehair
X-squared = 138.2898, df = 9, p-value < 2.2e-16
```

The p-value is small enough to reject the null-hypothesis. Hair colour is thus significantly linked to eye colour. The Chi-square test thus tells us that the two categorical variables are linked - but how are they link?

Definition
The following quantity is called the **link** between the V_1 i-level and the V_2 j-level:

$$\frac{1}{n}\frac{\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$

Pairs of modalities (i, j), corresponding to the most important links, are responsible for the V_1 and V_2 dependency.

Conclusion. One can compute a Correspondence Analysis to link the modalities for two variable, even if the Chi-square test is not statistically significant. This might especially helpful in the case of numerous modalities per variable.

4 Understanding the C.A. results

```
coa0 <- dudi.coa(dfcolours, scannf = F, nf = 3)
names(coa0)
```

```
 [1] "tab"  "cw"   "lw"   "eig"  "rank" "nf"   "c1"   "li"   "co"   "l1"   "call"
[12] "N"
```

4.1 The analysed data `tab`

```
coa0$tab
```

```
           black      blond    brunette         red
blue  -0.4900947  1.0380150 -0.19128314 -0.34071405
brown  0.6942761 -0.8516822  0.11964399 -0.01459667
green -0.5717593  0.1653543 -0.06206294  0.82394366
hazel -0.1158901 -0.4987723  0.20189488  0.25518704
```

`coa0$tab` is actually the table providing the relationship between the observed and expected counts!

```
(dfcolours - reschi$expected)/reschi$expected
```

```
           black      blond    brunette         red
blue  -0.4900947  1.0380150 -0.19128314 -0.34071405
brown  0.6942761 -0.8516822  0.11964399 -0.01459667
green -0.5717593  0.1653543 -0.06206294  0.82394366
hazel -0.1158901 -0.4987723  0.20189488  0.25518704
```

4.2 The weightings `lw` and `cw`

The row-weights and column-weights are the marginal frequencies of the observed contingency table.

```
coa0$cw
```

```
    black     blond  brunette       red 
0.1824324 0.2145270 0.4831081 0.1199324 
```

```
apply(dfcolours, 2, function(x) sum(x)/n)
```

```
    black     blond  brunette       red 
0.1824324 0.2145270 0.4831081 0.1199324 
```

18.24% of the sampled population had black hair.

```
coa0$lw
```

```
     blue     brown     green     hazel 
0.3631757 0.3716216 0.1081081 0.1570946 
```

```
apply(dfcolours, 1, function(x) sum(x)/n)
```

```
     blue     brown     green     hazel 
0.3631757 0.3716216 0.1081081 0.1570946 
```

36.32% of the sampled population had blue eyes.

4.3 The matrix to be diagonalized

Let's call **H** the needed matrix to compute the eigenanalysis (i.e. looking for eigenvalues and eigenvectors). **H** comes from the three following matrices:

- ⋆ **Z** the contingency table,
- ⋆ $\mathbf{D_I}$ the matrix of row-weights, diagonal,
- ⋆ $\mathbf{D_J}$ the matrix of column-weights, diagonal.

This triplet of matrices defines the duality diagram of the Correspondence Analysis.

$$\mathbf{H} \;=\; \mathbf{D}_J^{1/2}\mathbf{Z}^T\mathbf{D}_I\mathbf{Z}\mathbf{D}_J^{1/2}$$

In a classical use of the correspondence analysis, one can just give the contigency table, the row and columns weigthings are automatically calculated.

```
matZ <- as.matrix(coa0$tab)
DI <- diag(coa0$lw)
DrJ <- diag(sqrt(coa0$cw))
matH <- DrJ %*% t(matZ) %*% DI %*% matZ %*% DrJ
```

The rank (`rank`) of the matrix **H** is $min(I-1, J-1)$:

```
min(I - 1, J - 1)
```

```
[1] 3
```

```
coa0$rank
```

```
[1] 3
```

The eigenvalues provided by `dudi.coa` are the same as the ones calculated by the eigen function:

```
eigen(matH)
```

```
$values
[1]  2.087727e-01  2.222661e-02  2.598439e-03 -1.347873e-17
$vectors
            [,1]       [,2]       [,3]       [,4]
[1,]  0.47166009  0.6154461  0.4651134 -0.4271211
[2,] -0.84678181  0.2161646  0.1473309 -0.4631706
[3,]  0.22552151 -0.1522951 -0.6654608 -0.6950598
[4,]  0.09817011 -0.7424993  0.5649115 -0.3463126
```

```
coa0$eig
```

```
[1] 0.208772652 0.022226615 0.002598439
```

Let's call $\mathbf{\Lambda}$ the matrix containing the eigenvalues and **U** the matrix containing the eigenvectors.

4.4 The row coordinates li

The row coordinates also called **principal axes** are $\mathbf{ZD}_J^{1/2}\mathbf{U}$. These coordinates are centered, their variances equal to λ and covariances equal to zero.

```
 matZ %*% DrJ %*% reseigen$vectors[, 1:3]

            [,1]        [,2]         [,3]
blue  -0.5474139  0.08295428 -0.004709408
brown  0.4921577  0.08832151  0.021611305
green -0.1617534 -0.33903957  0.087597437
hazel  0.2125969 -0.16739109 -0.100518284
```

```
 coa0$li

           Axis1       Axis2        Axis3
blue  -0.5474139  0.08295428 -0.004709408
brown  0.4921577  0.08832151  0.021611305
green -0.1617534 -0.33903957  0.087597437
hazel  0.2125969 -0.16739109 -0.100518284
```

coa0$li provides the scores of the eye colour modalities on the retained axes. From this, we can see that the first axis opposes blue and green eyes to brown and hazel eyes.

Let's call Axis1 and Axis2 the two first row coordinates. The means of Axis1 and Axis2 are equal to 0 ($e-17$ being interpreted as 0).

```
 Axis1 <- coa0$li$Axis1
 Axis2 <- coa0$li$Axis2
 sum(Axis1 * coa0$lw)

[1] -3.122502e-17

 sum(Axis2 * coa0$lw)

[1] -2.775558e-17
```

The variances of Axis1 and Axis2 are equal to λ_1 and λ_2, respectively.

```
 sum(Axis1 * Axis1 * coa0$lw)

[1] 0.2087727

 coa0$eig[1]

[1] 0.2087727

 sum(Axis2 * Axis2 * coa0$lw)

[1] 0.02222661

 coa0$eig[2]

[1] 0.02222661
```

The covariance between Axis1 and Axis2 equals zero.

```
 sum(Axis1 * Axis2 * coa0$lw)

[1] -9.454243e-17
```

li contains the row coordinates normed to the eigenvalues and l1 the row coordinates normed to 1.

4.5 The column coordinates `co`

The column coordinates also called **principal components** are $\mathbf{D}_J^{-1/2}\mathbf{U}\mathbf{\Lambda^{1/2}}$. These coordinates are centered, their variances equal to λ and covariances equal to zero.

```
diag(1/sqrt(coa0$cw)) %*% reseigen$vectors[, 1:3] %*% diag(sqrt(coa0$eig))
```

```
           [,1]        [,2]        [,3]
[1,]  0.5045624  0.21482046  0.05550909
[2,] -0.8353478  0.06957934  0.01621471
[3,]  0.1482527 -0.03266635 -0.04880414
[4,]  0.1295233 -0.31964240  0.08315117
```

```
coa0$co
```

```
              Comp1       Comp2       Comp3
black     0.5045624  0.21482046  0.05550909
blond    -0.8353478  0.06957934  0.01621471
brunette  0.1482527 -0.03266635 -0.04880414
red       0.1295233 -0.31964240  0.08315117
```

`coa0$co` provides the scores of the hair colour modalities on the retained axes. From this, we can see that the first axis opposes blond hair to the other categories.

Let's call `Comp1` and `Comp2` the two first columns coordinates. The means of `Comp1` and `Comp2` are equal to 0.

```
Comp1 <- coa0$co$Comp1
Comp2 <- coa0$co$Comp2
sum(Comp1 * coa0$cw)
```

```
[1] 1.231654e-16
```

```
sum(Comp2 * coa0$cw)
```

```
[1] -1.023487e-16
```

The variances of `Comp1` and `Comp2` are equal to λ_1 and λ_2 respectively.

```
sum(Comp1 * Comp1 * coa0$cw)
```

```
[1] 0.2087727
```

```
coa0$eig[1]
```

```
[1] 0.2087727
```

```
sum(Comp2 * Comp2 * coa0$cw)
```

```
[1] 0.02222661
```

```
coa0$eig[2]
```

```
[1] 0.02222661
```

The covariance between `Comp1` and `Comp2` equals zero.

```
sum(Comp1 * Comp2 * coa0$cw)
```

```
[1] 3.469447e-18
```

`co` contains the column coordinates normed to the eigenvalues and `c1` the column coordinates normed to 1.

4.6 Link between the correspondence analysis and the chi-square statistic

If I_T is the total inertia and χ^2 the value of the chi-square statistic calculated on the observed contingency table, we have:

$$I_T = \frac{\chi^2}{n}$$

```
 reschi$statistic
X-squared
 138.2898
 reschi$statistic/n
X-squared
0.2335977
```

n provides the number of students sampled ($n = 592$ in this example)

```
 sum(coa0$eig)
[1] 0.2335977
```

`sum(coa0$eig)` provides the sum of the eigenvalues (i.e., the total inertia): we can see here that this sum equals `reschi$statistic/n`.

4.7 Other elements of the `dudi.coa` object

`nf` is an integer giving the number of axes kept.

```
 coa0$nf
[1] 3
```

`call` keeps a record of the realised correspondence analysis.

```
 coa0$call
dudi.coa(df = dfcolours, scannf = F, nf = 3)
```

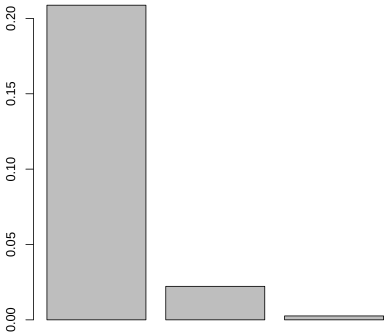
`N` is an integer giving the total number of individuals.

```
 coa0$N
[1] 592
```

4.8 Summing up the example

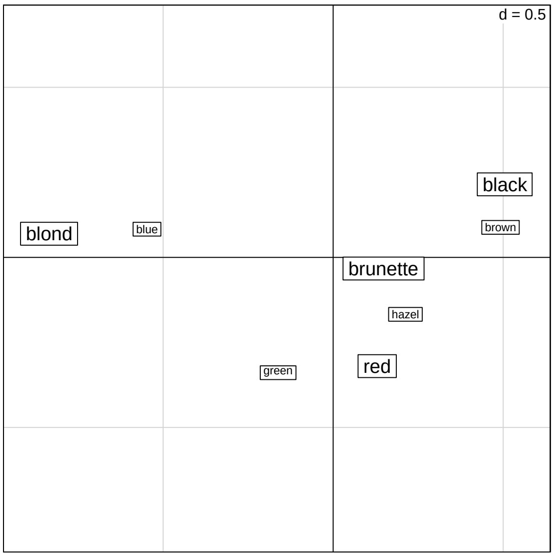
The main topic of the correspondence analysis is to reveal structures between levels of the two categorical variables: in our case, hair colours and eye colours.

```
 barplot(coa0$eig)
 coa0$eig/sum(coa0$eig)
[1] 0.89372732 0.09514911 0.01112356
```



The two first axes of the correspondence analysis keep 98.89% of the total inertia.

```
 scatter(coa0, posieig = "none")
NULL
```



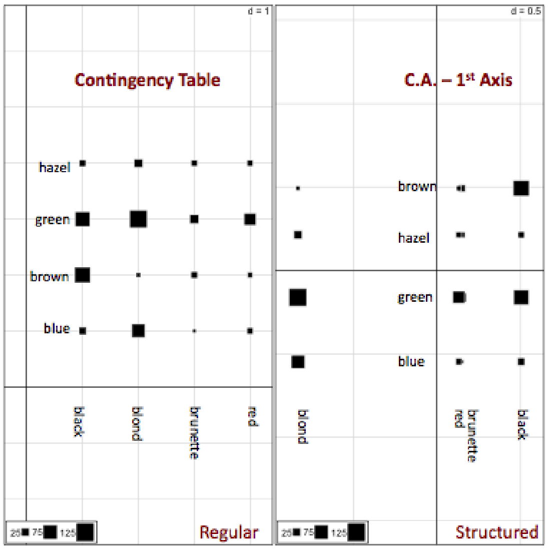
The first axis reveals an opposition between ('blue eyes','blond hair') and('brown eyes','black hair'). The second axis, which explains a lot less variability, speficies the uniqueness of ('green eyes','red hair') individuals.

5 Your turn!

Two exercises are proposed here:
- an example about wine tasting (`bordeaux`)
- an ecological application (`doubs`).
For both datasets, information is available using the function `help()` or `?`.

6 Conclusion

In a correspondence analysis, one transforms an observed contingency table (equal repartition of levels) into a structured contingency table, showing correspondences between the levels of two categorial variables.



References

[1] J.P. Benzecri. *L'analyse des données. T.2 : L'analyse des correspondances.* Dunod, Paris, 1973.

[2] S. Nishisato. *Analysis of categorical data : dual scaling and its applications.* University of Toronto Press, Toronto, 1980.

[3] R.D. Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28(1):9–12, 1974.