STAT 441: Lecture 26
Multivariate analysis with qualitative variables
Correspondence analysis

Venables and Ripley, 11.4

χ^2-statistic

Let us, say, test independence in a $r \times c$ contingency table. The probabilities of observations in cells are p_{ij}; under the hypothesis of independence, $p_{ij} = p_{i\cdot}p_{\cdot j}$, where $p_{i\cdot}$ and $p_{\cdot j}$ are column and row sums - marginal probabilities.

We observe cells frequencies n_{ij}; the estimates for $p_{i\cdot}$, and $p_{\cdot j}$ are $\widehat{p}_{i\cdot} = n_{i\cdot}/n$ and $\widehat{p}_{\cdot j} = n_{\cdot j}/n$, respectively; n is the total number of observations. Under the hypothesis of independence, the estimate for the cell probability is $\widehat{p}_{ij} = \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = (n_{i\cdot}/n)(n_{\cdot j}/n)$ and therefore the predicted number of observations is $P = n\ \dfrac{n_{i\cdot}n_{\cdot j}}{n^2} = \dfrac{n_{i\cdot}n_{\cdot j}}{n}$ and the observed number is $O = n_{ij}$

The test statistic can be written as $\displaystyle\sum_{\text{all cells}} \frac{(O_k - P_k)^2}{P_k}$

We use χ^2 distribution with $(r-1)(c-1)$ degrees of freedom to assess how large is this statistic is large enough, via its right tail value, which gives the p-value for the hypothesis of independence. If this p-value is low, we may reject the hypothesis - but what then?

Note: the statistic sums squares of "Pearson residuals" $\dfrac{O_k - P_k}{\sqrt{P_k}}$

Correspondence analysis

Suppose that E is the matrix formed from n_{ij}/n (the estimates of cell probabilities not assuming the independence hypothesis) and R, C are diagonal matrices formed from vectors r and c, with elements $r_i = n_{i\cdot}/n$ and $c_j = n_{\cdot j}/n$, respectively.

Consider the matrix $R^{-1/2}EC^{-1/2}$ (it is easy to form the square roots of diagonal matrices with positive elements). This matrix has elements

$$\frac{e_{ij}}{\sqrt{r_i c_j}}$$

SVD of this matrix can be viewed as returning "scores giving maximal correlations for rows and columns". The largest singular value is always one, corresponding to constant scores; hence we dismiss it, and look only for nontrivial solutions corresponding to singular values beginning with the second largest. That is, we form the SVD of

$R^{-1/2}(E - rc^\top)C^{-1/2}$

instead, and then we may take first one or two singular values.

Aggregated hair/eye color data: $r = c = 4$

```
> haireye=apply(HairEyeColor,1:2,sum)
> haireye
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16

> r = apply(haireye,1,sum)/sum(haireye)
> c = apply(haireye,2,sum)/sum(haireye)
> E = haireye/sum(haireye)
```

```
> svd(diag(1/sqrt(r)) %*% E %*% diag(1/sqrt(c)))
$d
[1] 1.00000000 0.45691646 0.14908593 0.05097489
$u
           [,1]        [,2]       [,3]       [,4]
[1,] -0.4271211  0.47166009  0.6154461  0.4651134
[2,] -0.6950598  0.22552151 -0.1522951 -0.6654608
[3,] -0.3463126  0.09817011 -0.7424993  0.5649115
[4,] -0.4631706 -0.84678181  0.2161646  0.1473309
$v
           [,1]       [,2]       [,3]        [,4]
[1,] -0.6096078  0.6566258  0.3611439  0.25844921
[2,] -0.6026406 -0.7220003  0.3353208 -0.05567605
[3,] -0.3963516  0.1844169 -0.4450167 -0.78157274
[4,] -0.3287980 -0.1163980 -0.7477267  0.56502056
> svd(diag(1/sqrt(r)) %*% (E - r %*% t(c)) %*% diag(1/sqrt(c)))
$d
[1] 4.569165e-01 1.490859e-01 5.097489e-02 2.929785e-19
$u
            [,1]       [,2]       [,3]      [,4]
[1,] -0.47166009  0.6154461 -0.4651134 0.4271211
[2,] -0.22552151 -0.1522951  0.6654608 0.6950598
[3,] -0.09817011 -0.7424993 -0.5649115 0.3463126
[4,]  0.84678181  0.2161646 -0.1473309 0.4631706
$v
           [,1]       [,2]        [,3]       [,4]
[1,] -0.6566258  0.3611439 -0.25844921 -0.6096078
[2,]  0.7220003  0.3353208  0.05567605 -0.6026406
[3,] -0.1844169 -0.4450167  0.78157274 -0.3963516
[4,]  0.1163980 -0.7477267 -0.56502056 -0.3287980
```

Mechanized way

```
> library(MASS)
> hc1 = corresp(haireye)
First canonical correlation(s): 0.4569165
 Hair scores:
     Black      Brown        Red      Blond
-1.1042772 -0.3244635 -0.2834725  1.8282287
 Eye scores:
     Brown       Blue      Hazel      Green
-1.0771283  1.1980612 -0.4652862  0.3540108

> hc2 = corresp(haireye,nf=2)
First canonical correlation(s): 0.4569165 0.1490859
 Hair scores:
            [,1]       [,2]
Black -1.1042772  1.4409170
Brown -0.3244635 -0.2191109
Red   -0.2834725 -2.1440145
Blond  1.8282287  0.4667063
 Eye scores:
            [,1]       [,2]
Brown -1.0771283  0.5924202
Blue   1.1980612  0.5564193
Hazel -0.4652862 -1.1227826
Green  0.3540108 -2.2741218
```

Interpretation I

Can be thought of as analysis of the χ^2 statistic for independence, because the elements of the matrix $R^{-1/2}(E - rc^\top)C^{-1/2}$ are Pearson residuals, up to a factor $\sqrt{n}$.

```
> class(haireye)='table'
> matrix(residuals(glm(Freq~Hair+Eye,family=poisson,
data=as.data.frame(haireye)),type='pearson'),4,4)/sqrt(sum(haireye))
             [,1]        [,2]        [,3]        [,4]
[1,]  0.180773066 -0.12615064 -0.01961905 -0.08029590
[2,]  0.050694815 -0.08012300  0.05561963 -0.01418351
[3,] -0.003081574 -0.07110772  0.03502737  0.09381990
[4,] -0.240474512  0.28973637 -0.09156384  0.02518174
> diag(1/sqrt(r)) %*% (E - r %*% t(c)) %*% diag(1/sqrt(c))
             [,1]        [,2]        [,3]        [,4]
[1,]  0.180773066 -0.12615064 -0.01961905 -0.08029590
[2,]  0.050694815 -0.08012300  0.05561963 -0.01418351
[3,] -0.003081574 -0.07110772  0.03502737  0.09381990
[4,] -0.240474512  0.28973637 -0.09156384  0.02518174
```

Another interpretation - and plotting

Can be viewed also as a search for the linear combination giving maximal contingency ("correlation") - not accounting for the trivial constant solution

can be a comparison of distances between "profiles", rowwise or columnwise conditional distributions.

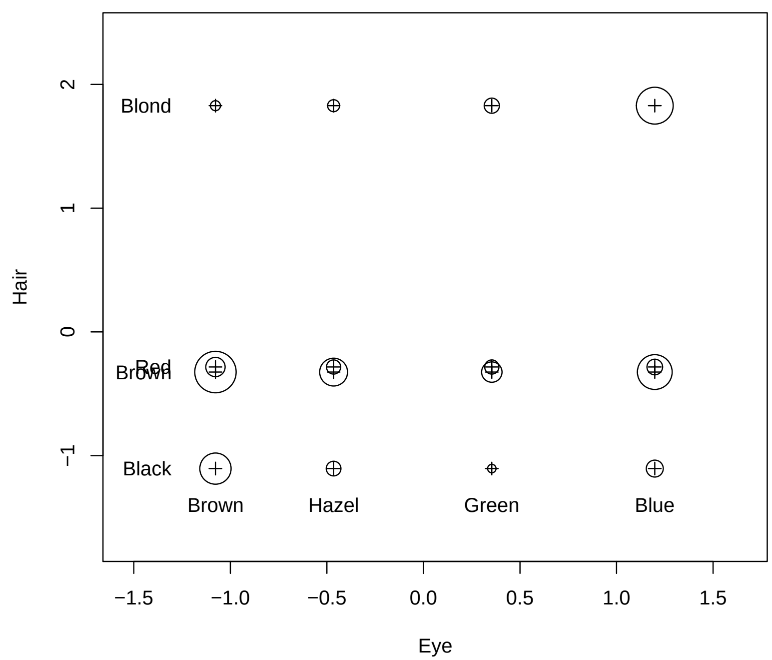
If the resulting SVD is $U\Lambda V^\top$, then of interest are first columns of $A = R^{-1/2}U\Lambda$ and $B = C^{-1/2}V\Lambda$

"Classical correspondence analysis": first two columns of A and B are plotted on the same figure. "Asymmetric approach" plots either first two columns of A with first two columns of $C^{-1/2}V$ (rows) or B with first two columns of $R^{-1/2}U$ (columns). Row plot can be viewed as that A is a convex combination of row profiles, given by $C^{-1/2}V$

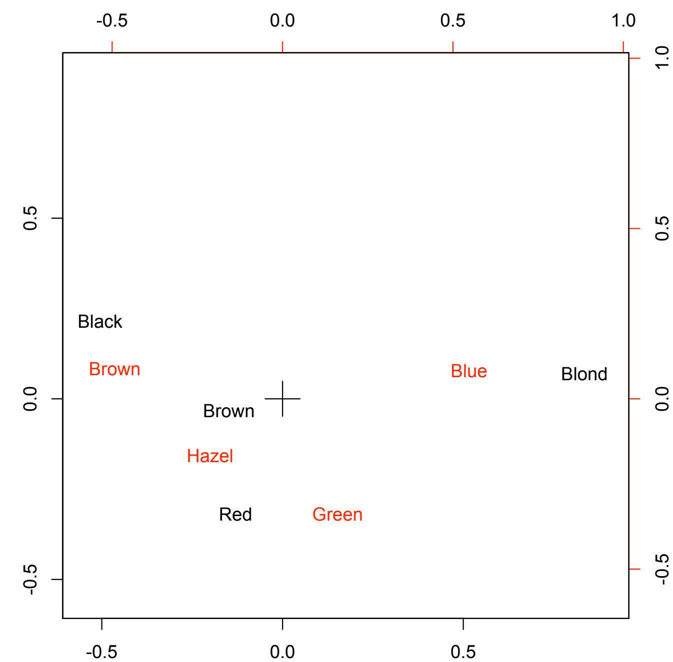
Inertia: the sum of squares of omitted singular values.

Symmetric view

First column

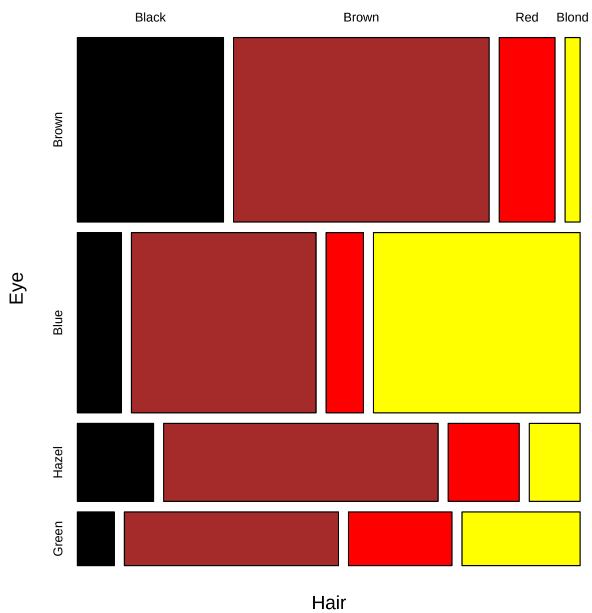


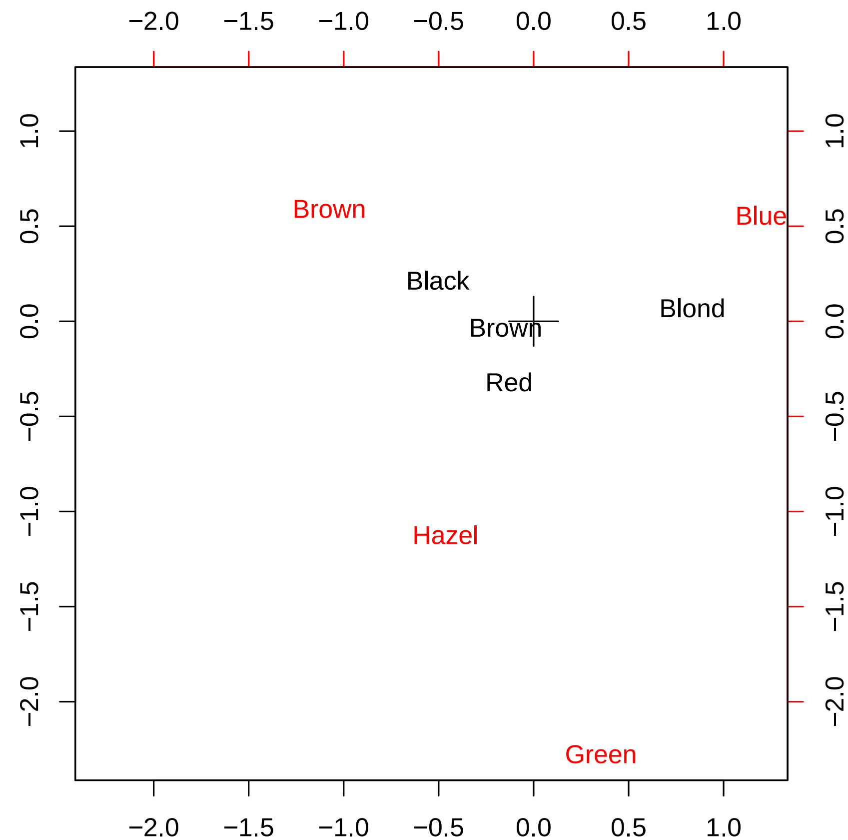
First two columns




```
> plot(hc2)
> biplot(hc2)
> biplot(hc2,xlim=c(-0.55,0.9),ylim=c(-0.55,0.9))
```

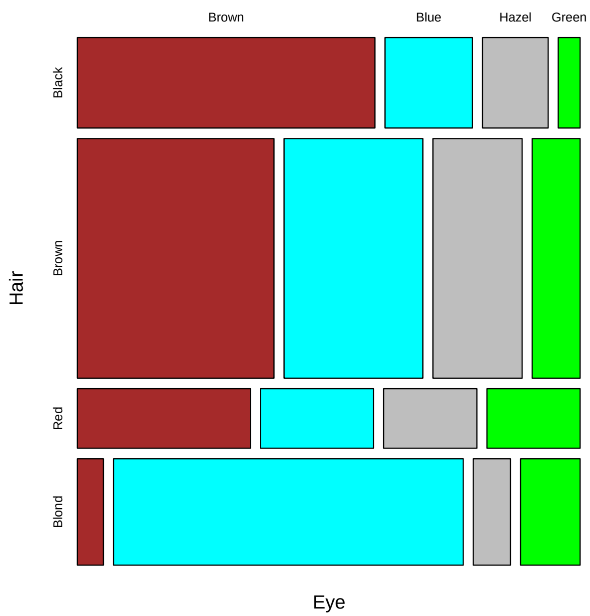
Asymmetric view: rows

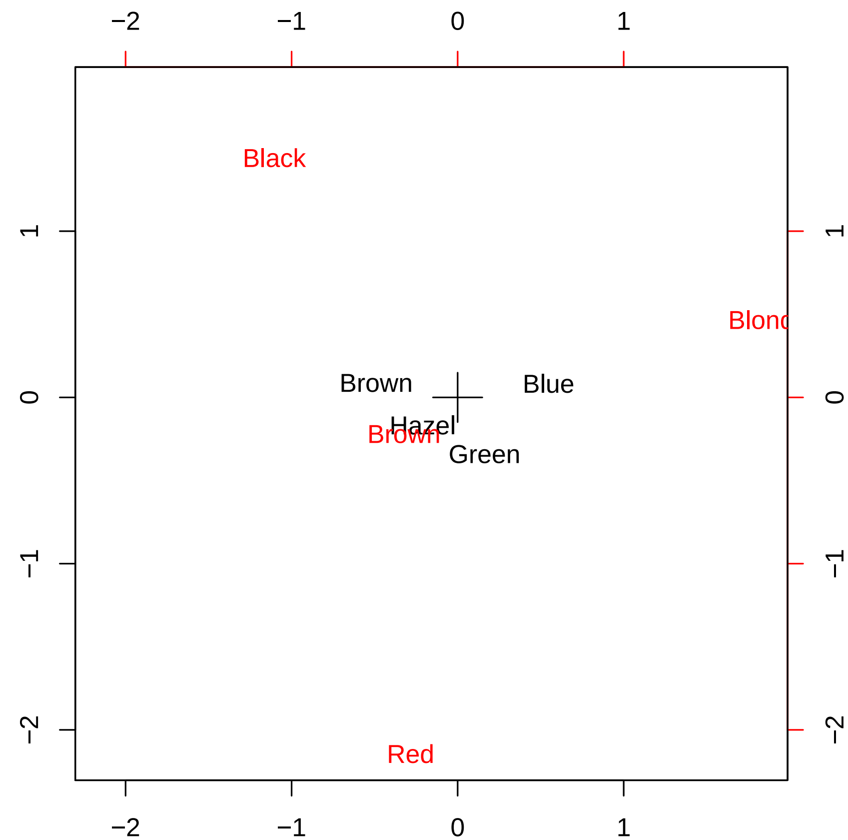





```
> plot(hc2,type="rows")
> biplot(hc2,type="rows")
> mosaicplot(haireye,sort=c(2,1),dir=c('v','h'))
```

Asymmetric view: columns






```
> plot(hc2,type="columns")
> biplot(hc2,type="columns")
> mosaicplot(haireye,sort=c(1,2),dir=c('h','v'))
```