Correspondence Analysis Examples

Smoke data example

STAT 32950-24620

Spring 2025 (wk4)

Contingency table

Contingency tables:

Display the cell counts (n_{ij} or x_{ij}) of the row and column variables.

Example

Employee smoke status

```
library(ca); data(smoke); smoke
```

```
##    none light medium heavy
## SM    4     2      3     2
## JM    4     3      7     4
## SE   25    10     12     4
## JE   18    24     33    13
## SC   10     6      7     2
```


Cell percentages

$\sum_{i,j} p_{ij} = 1$, where

$$p_{ij} = \frac{x_{ij}}{n}, \qquad i = 1, \cdots, I; \; j = 1, \cdots, J$$

```
X = as.matrix(smoke); #sum(X) # 193
P = X/sum(X); round(P,2) # cell percent table
```

```
##    none light medium heavy
## SM 0.02  0.01   0.02  0.01
## JM 0.02  0.02   0.04  0.02
## SE 0.13  0.05   0.06  0.02
## JE 0.09  0.12   0.17  0.07
## SC 0.05  0.03   0.04  0.01
```


Marginal percentages

Row variable marginal percents (sum = 1)

$$r_i = \sum_{j=1}^{J} p_{ij} = \frac{n_{i\bullet}}{n}, \qquad i = 1, \cdots, I$$

```
round(t(P%*%c(1,1,1,1)),2)
```

```
##        SM   JM   SE   JE   SC
## [1,] 0.06 0.09 0.26 0.46 0.13
```

Column variable marginal percents (sum = 1)

$$c_j = \sum_{i=1}^{I} p_{ij}, \qquad j = 1, \cdots, J$$

```
round(c(1,1,1,1,1)%*%P,2)
```

```
##      none light medium heavy
## [1,] 0.32  0.23   0.32  0.13
```

Row profile matrix

$$P_r = \left[\frac{p_{ij}}{r_i}\right] = \left[\frac{n_{ij}}{n_{i\bullet}}\right] = D_r^{-1}P$$

Row profile matrix (row sum = 1) for comparing rows

```
smokerow = X%*%c(1,1,1,1) # row sum 11 18 51 88 25
round(diag(c(1/smokerow))%*%X,2)
```

```
##      none light medium heavy
## [1,] 0.36  0.18   0.27  0.18
## [2,] 0.22  0.17   0.39  0.22
## [3,] 0.49  0.20   0.24  0.08
## [4,] 0.20  0.27   0.38  0.15
## [5,] 0.40  0.24   0.28  0.08
```


Column profile matrix

$$P_c = \left[\frac{p_{ij}}{c_j}\right] = \left[\frac{n_{ij}}{n_{\bullet j}}\right] = PD_c^{-1}$$

Column profile matrix (column sum = 1) for comparing columns

```
smokecol = c(1,1,1,1,1)%*%X    # col sum 61 45 62 25
round(X%*%diag(c(1/smokecol)),2)
```

```
##    [,1] [,2] [,3] [,4]
## SM 0.07 0.04 0.05 0.08
## JM 0.07 0.07 0.11 0.16
## SE 0.41 0.22 0.19 0.16
## JE 0.30 0.53 0.53 0.52
## SC 0.16 0.13 0.11 0.08
```


Expected values under independence

$$E_{ij} = nr_ic_j = \frac{x_{i\bullet}x_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}, \qquad i = 1, \cdots, I;\ j = 1, \cdots, J.$$

where

$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij} = \sum_{j=1}^{J} n_{ij}, \qquad x_{\bullet j} = \sum_{i=1}^{I} x_{ij}$$

are the row sum and column sum.

```
E = smokerow%*%smokecol/193 # expected counts under indep.
round(E,1)
```

```
##    none light medium heavy
## SM  3.5   2.6    3.5   1.4
## JM  5.7   4.2    5.8   2.3
## SE 16.1  11.9   16.4   6.6
## JE 27.8  20.5   28.3  11.4
## SC  7.9   5.8    8.0   3.2
```


Test of independence

Assuming the $n = I \times J$ observations are independent.

H_o: The row variable and the column variable are independent.

Under H_o, the test statistic (sometime written as X^2)

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

is of χ^2 distribution with degrees of freedom $df = (I-1)(J-1)$

Chi-square test

```
chisq.test(smoke)  #16.442, df = 12, p-value = 0.1718
```

```
##
##  Pearson's Chi-squared test
##
## data:  smoke
## X-squared = 16, df = 12, p-value = 0.2
```

Total "inertia" $\frac{X^2}{n} = \frac{16.442}{193} = 0.085$

Cell level contributions

Individual cell contributions to the chi-square test statistic:

$$\frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

```
round((smoke-E)^2/E,2)
```

```
##    none light medium heavy
## SM 0.08  0.12   0.08  0.23
## JM 0.50  0.34   0.26  1.19
## SE 4.89  0.30   1.17  1.03
## JE 3.46  0.59   0.79  0.22
## SC 0.56  0.01   0.13  0.47
```

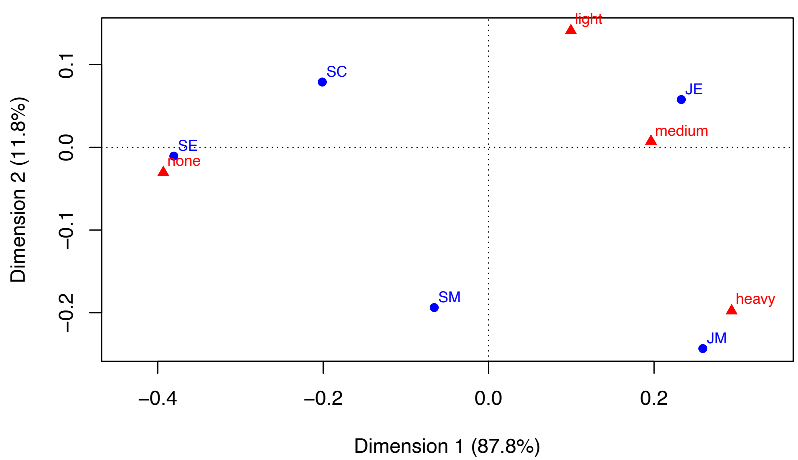
Overall

```
sum((smoke-E)^2/E)  # 16.44164
```

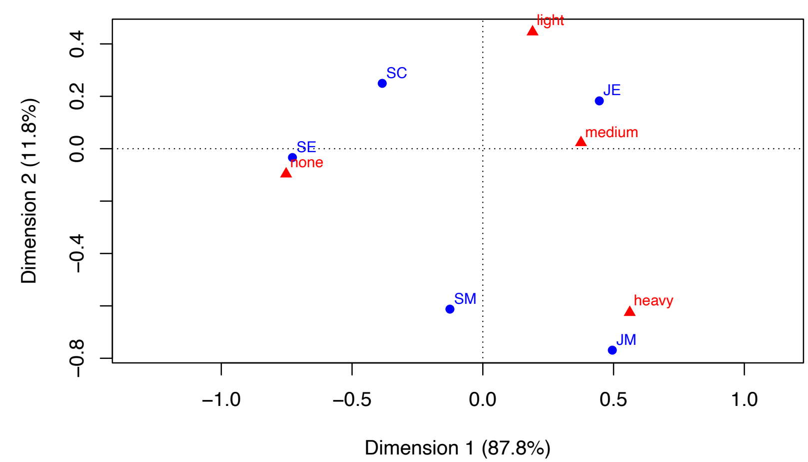
```
## [1] 16.44
```


Graphical representation by CA

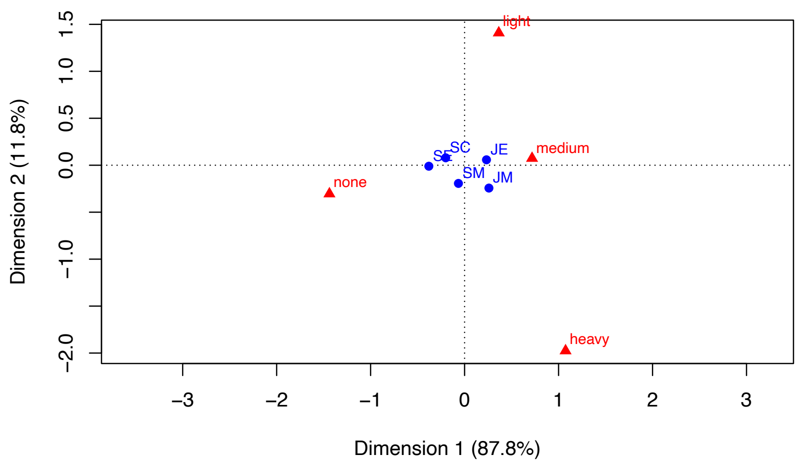
```
plot(ca(smoke),map="symmetric") # default of (ca(smoke))
```



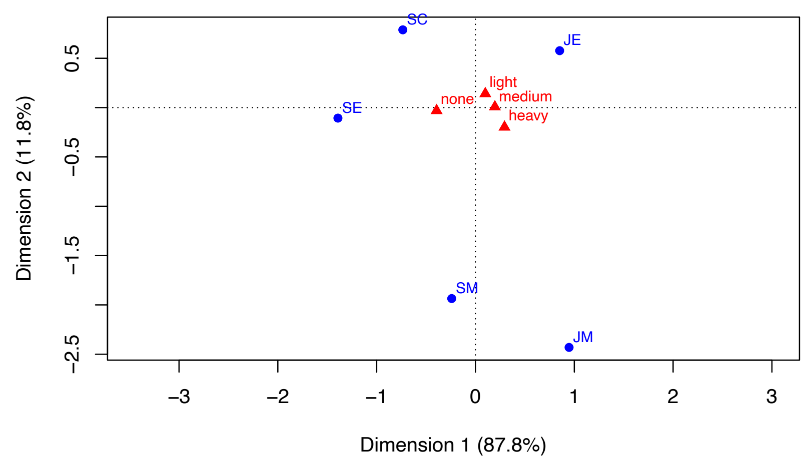

```
plot(ca(smoke), map="symbiplot") #rowgab, colgab
```



```r
plot(ca(smoke), map="rowprincipal") #rowgreen
```




```r
plot(ca(smoke), map="colprincipal") #colgreen
```



CA derivations

The chi-square statistic (divided by overall counts = total inertia)

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(x_{ij}-E_{ij})^2}{E_{ij}} = trace(SS^T) = \sum_k \lambda_k^2$$

where λ_k's are singular values of the $I \times J$ matrix

$$S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

$$D_r = diag\{r_1, \cdots, r_I\}, \qquad D_c = diag\{c_1, \cdots, c_J\}$$

are $I \times I$ and $J \times J$ diagonal matrices,

$$r = [r_1 \ \cdots \ r_I]', \qquad c = [c_1 \ \cdots \ c_J]'$$

are vectors of length I and J respectively. Recall

$$P = [p_{ij}],\ p_{ij} = \frac{x_{ij}}{n},\ r_i = \sum_{j=1}^{J} p_{ij},\ c_j = \sum_{i=1}^{I} p_{ij},\ i = 1, \cdots, I;\ j = 1, \cdots, J$$

Verify CA derivations

```r
# Verify by hand
Drow=diag(c(sqrt(smokerow/sum(X))))
Dcol=diag(c(sqrt(smokecol/sum(X))))
S = solve(Drow)%*%(as.matrix(smoke-E)/193)%*%solve(Dcol)
S
```

```
##          [,1]      [,2]     [,3]     [,4]
## [1,]  0.02020 -0.025384 -0.02044  0.03468
## [2,] -0.05098 -0.042054  0.03645  0.07865
## [3,]  0.15922 -0.039477 -0.07795 -0.07299
## [4,] -0.13394  0.055330  0.06404  0.03413
## [5,]  0.05374  0.005098 -0.02619 -0.04953
```

```r
193*sum(diag(S%*%t(S)))  # 16.4 = chisq = total mass
```

```
## [1] 16.44
```

Singular value decomp. $S = U\Sigma V^T$

```
svd(S)$d;svd(S)$u;svd(S)$v   # svd(S)
```

```
## [1] 2.734e-01 1.001e-01 2.034e-02 1.548e-17
```

```
##            [,1]      [,2]    [,3]    [,4]
## [1,] -0.05743 -0.46212  0.8333 -0.2604
## [2,]  0.28924 -0.74240 -0.5061 -0.3194
## [3,] -0.71555 -0.05475 -0.1303 -0.4492
## [4,]  0.57530  0.38958  0.1098 -0.6396
## [5,] -0.26470  0.28376 -0.1430 -0.4683
```

```
##          [,1]     [,2]     [,3]   [,4]
## [1,] -0.8087 -0.17128 -0.02462 0.5622
## [2,]  0.1756  0.68057  0.52232 0.4829
## [3,]  0.4070  0.04167 -0.71512 0.5668
## [4,]  0.3867 -0.71116  0.46387 0.3599
```


Ortho-normal U

```
round(t(svd(S)$u)%*%(svd(S)$u))  # U'U
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```

```
round((svd(S)$u)%*%t(svd(S)$u)) #UU'
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    0
```


Ortho-normal V

```
round(t(svd(S)$v)%*%(svd(S)$v))  #V'V
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```

```
round((svd(S)$v)%*%t(svd(S)$v))  #VV'
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```


Principal coordinates of rows

Principal coordinates of rows: $F = D_r^{-1/2} U\Sigma$

```
Fmat = solve(Drow)%*%(svd(S)$u)%*%diag(c(svd(S)$d))
round(Fmat,4)
```

```
##         [,1]    [,2]    [,3] [,4]
## [1,] -0.0658 -0.1937  0.0710    0
## [2,]  0.2590 -0.2433 -0.0337    0
## [3,] -0.3806 -0.0107 -0.0052    0
## [4,]  0.2330  0.0577  0.0033    0
## [5,] -0.2011  0.0789 -0.0081    0
```

Principal coordinates of columns

Principal coordinates of columns: $G = D_c^{-1/2} V \Sigma$

```r
Gmat = solve(Dcol)%*%(svd(S)$v)%*%diag(c(svd(S)$d))
round(Gmat,4)
```

```
##         [,1]    [,2]    [,3] [,4]
## [1,] -0.3933 -0.0305 -0.0009    0
## [2,]  0.0995  0.1411  0.0220    0
## [3,]  0.1963  0.0074 -0.0257    0
## [4,]  0.2938 -0.1978  0.0262    0
```


Coordinates of rows $D_r^{-1/2} U$ (standard)

Coordinates of rows using $D_r^{-1/2} U$

```r
round((solve(Drow)%*%(svd(S)$u)),4)
```

```
##         [,1]    [,2]    [,3]    [,4]
## [1,] -0.2405 -1.9357  3.4903 -1.0908
## [2,]  0.9471 -2.4310 -1.6574 -1.0459
## [3,] -1.3920 -0.1065 -0.2535 -0.8739
## [4,]  0.8520  0.5769  0.1625 -0.9472
## [5,] -0.7355  0.7884 -0.3974 -1.3011
```

```r
ca(smoke)$rowcoord
```

```
##       Dim1    Dim2    Dim3
## SM -0.2405 -1.9357  3.4903
## JM  0.9471 -2.4310 -1.6574
## SE -1.3920 -0.1065 -0.2535
## JE  0.8520  0.5769  0.1625
## SC -0.7355  0.7884 -0.3974
```


Coordinates of columns $D_c^{-1/2} V$ (standard)

Coordinates of columns using $D_c^{-1/2} V$

```r
round(solve(Dcol)%*%(svd(S)$v),4)
```

```
##         [,1]    [,2]    [,3] [,4]
## [1,] -1.4385 -0.3047 -0.0438    1
## [2,]  0.3637  1.4094  1.0817    1
## [3,]  0.7180  0.0735 -1.2617    1
## [4,]  1.0744 -1.9760  1.2889    1
```

```r
ca(smoke)$colcoord
```

```
##           Dim1     Dim2     Dim3
## none   -1.4385 -0.30466 -0.04379
## light   0.3637  1.40943  1.08170
## medium  0.7180  0.07353 -1.26172
## heavy   1.0744 -1.97596  1.28886
```



```r
ca(smoke)
```

```
##
##  Principal inertias (eigenvalues):
##            1        2        3
## Value      0.074759 0.010017 0.000414
## Percentage 87.76%   11.76%   0.49%
##
##
##  Rows:
##                 SM       JM       SE      JE        SC
## Mass      0.056995  0.09326  0.26425 0.45596  0.129534
## ChiDist   0.216559  0.35692  0.38078 0.24002  0.216169
## Inertia   0.002673  0.01188  0.03831 0.02627  0.006053
## Dim. 1   -0.240539  0.94710 -1.39197 0.85199 -0.735456
## Dim. 2   -1.935708 -2.43096 -0.10651 0.57694  0.788435
##
##
##  Columns:
##              none    light  medium     heavy
```

```r
ca(smoke)$rowcoord
```

```
##        Dim1    Dim2    Dim3
## SM -0.2405 -1.9357  3.4903
## JM  0.9471 -2.4310 -1.6574
## SE -1.3920 -0.1065 -0.2535
## JE  0.8520  0.5769  0.1625
## SC -0.7355  0.7884 -0.3974
```

```r
ca(smoke)$sv
```

```
## [1] 0.27342 0.10009 0.02034
```

```r
round(ca(smoke)$rowcoord[,1]*ca(smoke)$sv[1],4)
```

```
##      SM      JM      SE      JE      SC
## -0.0658  0.2590 -0.3806  0.2330 -0.2011
```

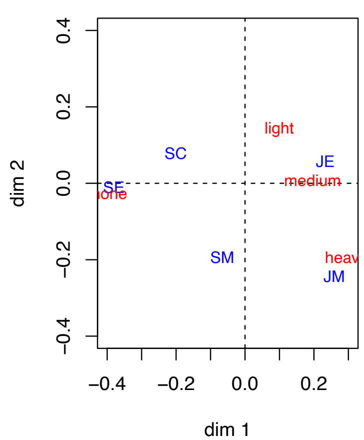
```r
round(Fmat[,1],4)
```

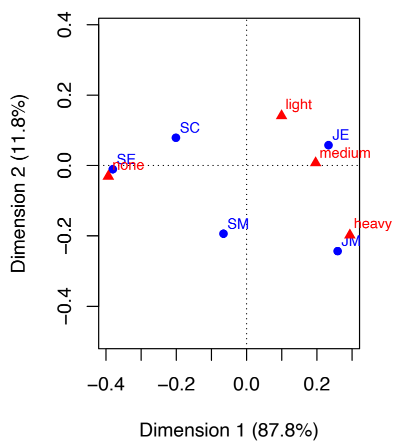
```
## [1] -0.0658  0.2590 -0.3806  0.2330 -0.2011
```


Verify CA by hand

```r
par(mfrow=c(1,2))
plot(Fmat[,1:2],type="n",xlab="dim 1",ylab="dim 2",
     xlim=c(-.4,.3), ylim=c(-.4,.4))
text(Fmat[,1:2],labels=rownames(smoke),
     cex=.8,col="blue",lwd=3)
points(Gmat[,1:2],type="n");
abline(v=0,lty=2); abline(h=0,lty=2)
text(Gmat[,1:2],labels=colnames(smoke),
     cex=.8,col="red",lwd=2)
plot(ca(smoke),map="symmetric")
```

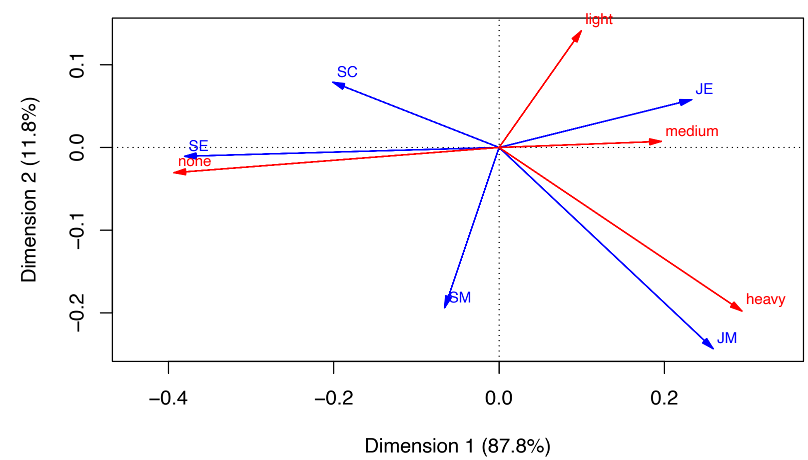

Plots verifying CA





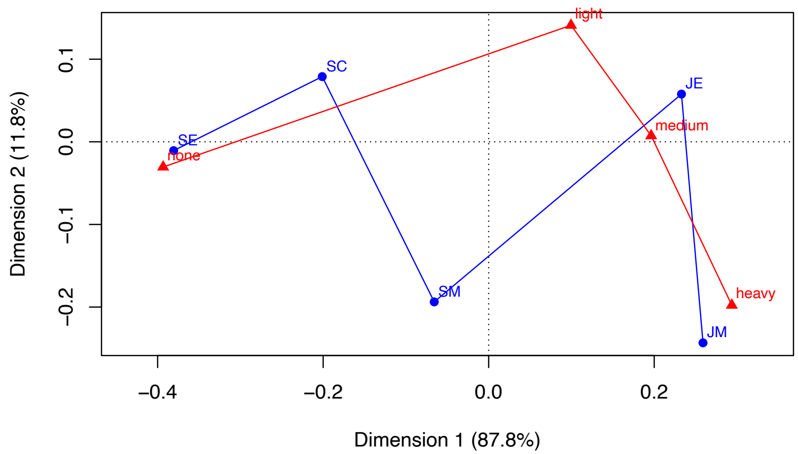
CA Plot with vector notations

```r
plot(ca(smoke),arrows=c(TRUE,TRUE))
```



CA Plot connecting row/col variables

```r
plot(ca(smoke),lines=TRUE)
```



Comparison to independent case

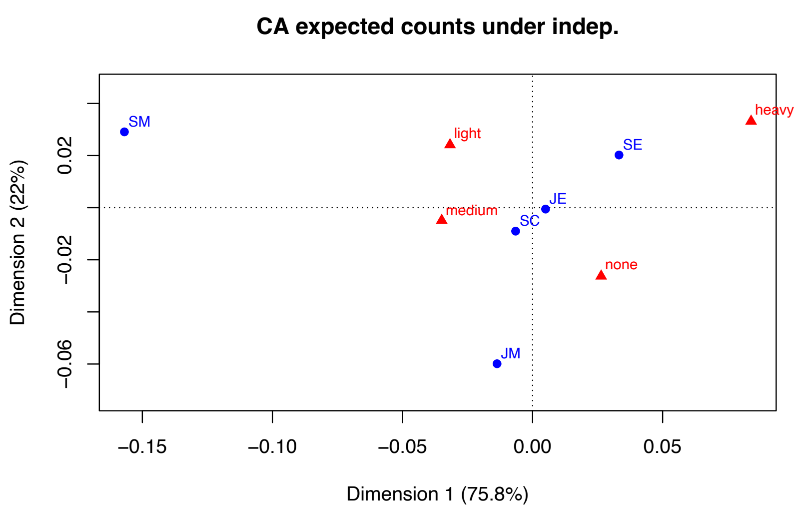
If the row variables and column variables almost independent:

- The data agrees with the expected counts
- The total mass is small
- The chisquare test is not significant

However, the ca picture is still available.

CA plot under independence

```r
Edata = round(E)  # rounded expected counts
plot(ca(Edata),main="CA expected counts under indep.")
```



ca(data) vs ca(expected counts) in comparable scales