Multivariate Statistical Analysis

Dr. Prabhanjan Tattar
Lead Statistician
CustomerXPs Software Private Limited
Bangalore

Definitions and Notations

- $\mathbf{X} = (X_1, X_2, \ldots, X_q)$: a random vector
- $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = E(X), \boldsymbol{\Sigma} = var(X)$
- Typical problems
  - Estimation of the mean vector and the covariance matrix
  - Sample Correlation Coefficients
  - Testing for the mean vector and the covariance matrix
  - Inference problems in multiple samples

A Simple Example

- Hage: Age of husband
- Hheight: Height of husband
- Wage: Age of wife
- Wheight: Height of wife
- Hagefm: Age of husband at first marriage
- **Everitt (2005)**

Hage	Hheight	Wage	Wheight	Hagefm
49	1809	43	1590	25
25	1841	28	1560	19
40	1659	30	1620	38
52	1779	57	1540	26
58	1616	52	1420	30
32	1695	27	1660	23
43	1730	52	1610	33
47	1740	43	1580	26
31	1685	23	1610	26
26	1735	25	1590	23

```
huswif=read.csv("huswif.csv",header=T)
mean(huswif)
   Hage Hheight    Wage Wheight Hagefm
   40.3  1728.9    38.0  1578.0    26.9
sd(huswif)^2
      Hage    Hheight       Wage
Wheight     Hagefm
 130.23333 4706.98889  164.66667
4173.33333   29.87778
```

```
var(huswif)
```

	Hage	Hheight	Wage	Wheight	Hagefm
Hage	130.23	-192.19	128.56	-436	28.03
Hheight	-192.19	4706.99	25.89	876.44	-229.34
Wage	128.56	25.89	164.67	-456.67	21.67
Wheight	-436	876.44	-456.67	4173.33	-8
Hagefm	28.03	-229.34	21.67	-8	29.88

```
cor(huswif)
```

	Hage	Hheight	Wage	Wheight	Hagefm
Hage	1	-0.25	0.88	-0.59	0.45
Hheight	-0.25	1	0.03	0.2	-0.61
Wage	0.88	0.03	1	-0.55	0.31
Wheight	-0.59	0.2	-0.55	1	-0.02
Hagefm	0.45	-0.61	0.31	-0.02	1

Testing for μ with Σ known

- $X_1, X_2, \ldots, X_n, X_1 \sim N_p(\mu, \Sigma)$
- The test statistic is given by

$$Z^2 = n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0)$$

- Under the null hypothesis, Z^2 is distributed as chi-square variate with p degrees of freedom
- Assume that $\Sigma = [20\,100; 100\,1000]$ and the we are interested in testing

$$\mu = (70, 170)'$$

- Since the calculated chi-square value is greater than the tabulated, we reject the null hypothesis

The R codes which gives us Z^2:

```
library(MASS)
hw=read.csv("Height_Weight.csv",header
=T)
mu0=c(70,170)
sigma = matrix(c(20, 100, 100,
1000),nrow=2)
n=nrow(hw)
meanx=mean(hw[,2:3])
z2=n*t(meanx-mu0)%*%ginv(sigma)%*
%(meanx-mu0)
 z2
 8.4026
 qchisq(1-0.05,2)
[1] 5.991465
```

Testing for μ with Σ unknown

- The test statistic is given by

$$Z^2 = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0)$$

where S is the sampling covariance matrix

- Under the null hypothesis, Z^2 is distributed as Hotellings' T^2 distribution with p and $v = n - 1$ degrees of freedom
- Use the R package "ICSNP"

- Consider the "pulmonary" data set in the "ICSNP" R package
- The variables of interest are FVC, FEV, and CC
- The correlation matrix shows strong association among the variable

```
cor(pulmonary)
          FVC       FEV        CC
FVC  1.0000000 0.9269397 -0.5339726
FEV  0.9269397 1.0000000 -0.2390159
CC  -0.5339726 -0.2390159 1.0000000
```

- We are then interested in testing if the mean vector of the three variables is (0,0,0)
- The covariance matrix is unknown and hence Hotelling's T^2-test is appropriate

```
HotellingsT2(pulmonary)
        Hotelling's one sample T2-test
data:  pulmonary
T.2 = 3.8231, df1 = 3, df2 = 9, p-value = 0.05123
alternative hypothesis: true location is not equal to c(0,0,0)
```

Multivariate Two-sample Test

Consider

$$\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}, \boldsymbol{X}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$
$$\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}, \boldsymbol{X}_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

The hypothesis of interest is

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2$$

Assume that the samples are independent and

$$\Sigma_1 = \Sigma_2 = \Sigma, \Sigma \text{ unknown}$$

Define the matrix of sum squares and cross products

$$\boldsymbol{W}_1 = \sum_{i=1}^{n_1} (\boldsymbol{x}_{1i} - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1i} - \bar{\boldsymbol{x}}_1)' = n\boldsymbol{S}_1$$

$$\boldsymbol{W}_1 = \sum_{i=1}^{n_2} (\boldsymbol{x}_{2i} - \bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2i} - \bar{\boldsymbol{x}}_2)' = n_2\boldsymbol{S}_2$$

Define the pooled population covariance matrix

$$\boldsymbol{S}_{pl} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2]$$

- The test statistic is then given by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\boldsymbol{x}_1 - \boldsymbol{x}_2)' \boldsymbol{S}_{pl}^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

- The test statistic T^2, under the null hypothesis, is distributed as Hotellings $T^2_{p,n1+n2-2}$ distribution
- Some important properties:
  - $n_1 + n_2 - 2 > p$ is necessary condition for $\mathbf{S}_{\mathbf{pl}}^{-1}$ to be nonsingular
  - T^2 is skewed
  - For a two-sided alternative, the critical-region is one-tailed
  - An easy transformation of T^2 gives the F-distribution

Example: Psychological Tests for Males and Females

x_1: pictorial inconsistencies

x_2: paper form board

x_3: tool recognition

x_4: vocabulary

Variables observed for 32 males and females

We are interested in testing if the mean vectors for males and females are equal.

```
mfp=read.csv("MF_Psycho_Test_Scores.csv",header=T)
males=mfp[,1:4]; females=mfp[,5:8]
nm=nrow(males);nf=nrow(females)
meanm=mean(males); meanf=mean(females)
sigmam=var(males); sigmaf=var(females)
sigmapl=(1/(nm+nf-2))*((nm-1)*sigmam+(nf-1)*sigmaf)
```

```
t2=((nm*nf)/(nm+nf))*(t(meanm-meanf)
%*%ginv(sigmapl)%*%(meanm-meanf))
nm;nf;meanm;meanf;sigmapl;t2
```

```
[1] 32
[1] 32
   M_y1     M_y2     M_y3     M_y4
15.96875 15.90625 27.18750 22.75000
   F_y1     F_y2     F_y3     F_y4
12.34375 13.90625 16.65625 21.93750
```

```
          M_y1      M_y2      M_y3      M_y4
M_y1 7.164315  6.047379  5.693044  4.700605
M_y2 6.047379 15.894153  8.492440  5.855847
M_y3 5.693044  8.492440 29.356351 13.980847
M_y4 4.700605  5.855847 13.980847 22.320565
       [,1]
[1,] 97.6015
```

Comparing the sample T^2 value with

$$T^2_{.01,4,62} = 15.373$$

we reject the null hypothesis

Multivariate Paired Two-sample Test

Tests on Covariance Matrices

Testing for $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$

- First calculate the sample variance matrix ***S***
- The test statistic, a modification of the likelihood ratio test,

$$u = v\left[\ln|\Sigma_0| - \ln|S| + tr\left(S\Sigma^{-1}\right) - p\right]$$

where v is the d.f. of ***S***

- For large v, the above test statistic u is approximately distributed as chi-square distribution with $p(p+1)/2$ d.f
- For moderate v, a modification of u is given by

$$u' = \left[1 - \frac{1}{6v-1}\left(2p+1 - \frac{2}{p+1}\right)\right]u$$

Height-Weight Example Continued

- In the height-weight example earlier, we assumed $\Sigma = \begin{bmatrix} 20 & 100 \\ 100 & 1000 \end{bmatrix}$
- Lets test if thats the real too

```
hw=read.csv("Height_Weight.csv",header=T)
sigma0 = matrix(c(20, 100, 100, 1000),nrow=2)
sigma = var(hw[,2:3])
v = nrow(hw)-1
p = ncol(hw)-1
u = v*(log(det(sigma0))-log(det(sigma)) + sum(diag(sigma%*%ginv(sigma0)))-p)
u1 = (1- (1/(6*v-1))*(2*p+1 - 2/(p+1)))*u
u,u1,qchisq(1-0.05,p*(p+1)/2)
[1] 11.09374
[1] 10.66832
[1] 7.814728
```

Since the calculated test statistic value exceeds 7.81, we reject the null hypothesis.

Multivariate Analysis of Variance

	Sample 1 from $N_p(\mu_1, \Sigma)$	Sample 2 from $N_p(\mu_2, \Sigma)$	...	Sample k from $N_p(\mu_k, \Sigma)$
	$\mathbf{y}_{11}$	$\mathbf{y}_{21}$	...	$\mathbf{y}_{k1}$
	$\mathbf{y}_{12}$	$\mathbf{y}_{22}$	...	$\mathbf{y}_{k2}$
	$\vdots$	$\vdots$		$\vdots$
	$\mathbf{y}_{1n}$	$\mathbf{y}_{2n}$	...	$\mathbf{y}_{kn}$
Total	$\mathbf{y}_{1.}$	$\mathbf{y}_{2.}$	...	$\mathbf{y}_{k.}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	...	$\bar{y}_{k.}$

The model for each observation is

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$
$$= \mu_i + \epsilon_{ij}$$

The hypothesis of interest is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

The "between" and "within" sum of squares matrices, denoted **H** and **E** respectively, are defined by

$$\boldsymbol{H} = n \sum_{i=1}^{k} (\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})' \qquad \boldsymbol{E} = \sum_{i=1}^{k} \sum_{j=1}^{n} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'$$

$$= \sum_{i=1}^{k} \frac{1}{n} \boldsymbol{y}_{i.} \boldsymbol{y}_{i.}' - \frac{1}{kn} \boldsymbol{y}_{..} \boldsymbol{y}_{..}' \qquad = \sum_{ij} \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}' - \sum_{i} \frac{1}{n} \boldsymbol{y}_{i.} \boldsymbol{y}_{i.}'$$

Wilks Test Statistic

- v_H, v_E : the rank of **H** and **E**
- Wilks Test Statistic: for H_0 is given by

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}$$

- The test procedure is to reject H_0 if $\Lambda \le \Lambda_{\alpha, p, v_E, v_H}$
- The test statistic can be equivalently obtained in terms of the eigen values of the matrix $\mathbf{E}^{-1}\mathbf{H}$: $\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$
- The range of Wilks Λ is between 0 to 1
- Transforms to *F-test* when either v_H = *1 or 2*, or when p = *1 or 2*

Pillai's Test Procedure

- $\lambda_1, \lambda_2, \ldots, \lambda_s$ the eigen values of $\mathbf{E^{-1}H}$
- The Pillai test statistic is then given by

$$V^{(s)} = tr\left[(\boldsymbol{E}+\boldsymbol{H})^{-1}\boldsymbol{H}\right] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

- Reject H_0 if $V^{(s)} \geq V_{\alpha}^{(s)}$

Example: Apple of Different Rootstock

- y_1 = trunk girth at 4 years (mm ×100)
- y_2 = extension growth at 4 years (m)
- y_3 = trunk girth at 15 years (mm ×100)
- y_4 = weight of tree above ground at 15 years (lb ×1000)

The goal is to test if the mean vector of the four variables is same across 6 stratas of the experiment.

```
# rootstock.dta is available at http://www.stata-press.com/data/r10/rootstock.dta
library(foreign)
rootstock=read.dta("/home/prabhanjan/Desktop/rootstock.dta")
rootstock1=rootstock[rootstock[,1]==1,2:5]
rootstock2=rootstock[rootstock[,1]==2,2:5]
rootstock3=rootstock[rootstock[,1]==3,2:5]
rootstock4=rootstock[rootstock[,1]==4,2:5]
rootstock5=rootstock[rootstock[,1]==5,2:5]
rootstock6=rootstock[rootstock[,1]==6,2:5]
n=8; p=4; vh=5; ve=6*(8-1); k=6
```

```
ymm=colSums(rootstock[,2:5])
y1m=colSums(rootstock1)
y2m=colSums(rootstock2)
y3m=colSums(rootstock3)
y4m=colSums(rootstock4)
y5m=colSums(rootstock5)
y6m=colSums(rootstock6)
H = ((y1m%*%t(y1m))/n) + ((y2m%*%t(y2m))/n)+((y3m%*%t(y3m))/n)+((y4m%*
%t(y4m))/n)+((y5m%*%t(y5m))/n)+((y6m%*%t(y6m))/n) - (ymm%*%t(ymm))/(k*n)
E = matrix(0,nrow=4, ncol=4);
for(i in 1:nrow(rootstock))    {
a = as.numeric(rootstock[i,2:5])
E = E + a%*%t(a)
}
E = E - (((y1m%*%t(y1m))/n) + ((y2m%*%t(y2m))/n)+((y3m%*%t(y3m))/n)+((y4m
%*%t(y4m))/n)+((y5m%*%t(y5m))/n)+((y6m%*%t(y6m))/n))
E_H=E+H
wlambda=det(E)/(det(E+H))
options(digits=3)
E;H;E_H;wlambda
```

```
       y1    y2    y3    y4
[1,] 0.320  1.70 0.554 0.217
[2,] 1.697 12.14 4.364 2.110
[3,] 0.554  4.36 4.291 2.482
[4,] 0.217  2.11 2.482 1.723
```

```
       y1    y2     y3    y4
[1,] 0.394  2.23  0.886 0.426
[2,] 2.234 16.34  6.719 3.747
[3,] 0.886  6.72 10.405 6.263
[4,] 0.426  3.75  6.263 4.216
```

```
       y1     y2    y3    y4
[1,] 0.0736 0.537 0.332 0.208
[2,] 0.5374 4.200 2.355 1.637
[3,] 0.3323 2.355 6.114 3.781
[4,] 0.2085 1.637 3.781 2.493
```

```
[1] 0.154
```

```
# Towards Pillai's test statistic
EH = solve(E) %*% H
eveh=eigen(EH)$values
pillaivs=sum(eveh/(1+eveh))
pillaivs
```
[1] 1.31

The calculated values of Wilks lambda 0.154 is lesser than the theoretical value of 0.455 (corresponding to $p = 4$, $v_H = 5$, $v_E = 42$). Thus, we reject the null hypothesis. Similarly the Pillai's test statistic 1.31 is greater than the theoretical value of 0.645 and leads to the same conclusion.

Using “manova” Function in R

```
attach(rootstock)
rs=rootstock[,1];
rs=factor(rs,ordered=is.ordered(rs)) # Too important a step
root.manova=manova(cbind(y1,y2,y3,y4)~rs)
summary(root.manova, test = "Pillai")
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
rs	5	1.31	4.07	20	168	1.983e-07	***
Residuals	42						

```
summary(root.manova, test = "Wilks")
```

	Df	Wilks	approx F	num Df	den Df	Pr(>F)	
rs	5	0.15	4.94	20	130.3	7.714e-09	***
Residuals	42						

```
summary(root.manova, test = "Hotelling")
```

	Df	Hotelling-Lawley	approx F	num Df	den Df	Pr(>F)	
rs	5	2.92	5.48	20	150	2.568e-10	***
Residuals	42						

```
summary(root.manova, test = "Roy")
```

	Df	Roy	approx F	num Df	den Df	Pr(>F)	
rs	5	1.88	15.76	5	42	1.002e-08	***
Residuals	42						

Testing for Sphericity $H_0: \Sigma = \sigma^2 I$

- The test for the independence of the components of a random vector is same as testing $H_0: \Sigma = \sigma^2 I$
- Under H_0, the ellipsoid $(x-\mu)'\Sigma(x-\mu)=c^2$ becomes $(x-\mu)'(x-\mu)=\sigma^2 c^2$ which is the equation of a sphere, and thus the name *sphericity*

The log likelihood ratio test is given by

$$LR=\left[\frac{|S|}{(tr(S)/p)^p}\right]^{n/2}$$

and the resulting LR test statistic is

$$-2\ln(LR)=-n\ln\left[\frac{|S|}{(tr\,S/p)}\right]=-n\ln(u),$$

$$\text{where } u=(LR)^{2/n}$$

- If λ_i is the i-th eigen value of the correlation matrix $\boldsymbol{S}$, we can re-write u in terms of the eigen-values:

$$u = \frac{p^p \prod_{i=1}^{p} \lambda_i}{\left(\sum_{i=1}^{p} \lambda_i\right)^p}$$

- An improvement over u is given by

$$u' = -\left(v - \frac{2p^2 + p + 1}{6p}\right) \ln(u)$$

- The statistic u', under the null hypothesis, has a chi-square distribution with $p(p+1)/2 - 1$ degrees of freedom.

Example: Tests of sphericity

- Response time to 5 probe words in a sentence
- Probe words are used to test recall of words in various linguistic contexts
- The interest is in testing if the response times to different probe words are independent
- If we fail to reject the sphericity hypothesis, we can compare the mean response times using ANOVA

```
pw=read.csv("Probe_Word.csv",header=T
)
sigma = var(pw[2:6])
p=ncol(pw)-1; v = nrow(pw)-1
u = p^p*(det(sigma))/
(sum(diag(sigma))^p)
u1 = -(v-(2*p^2+p+2)/(6*p))*log(u)
u;u1
[1] 0.03948874
[1] 26.17709
> qchisq(1-.05,df)
[1] 23.68479
```

Since the calculated chi-square value is greater than 23.68479, we reject the sphericity hypothesis.

Multivariate Tests of Equality of Covariance Matrices $H_0: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$

- $n_1, n_2, \ldots, n_k$: sizes of the k-samples
- $\boldsymbol{S}_i$: the sample covariance matrix of the i-the sample
- Define $v_i = n_i - 1,\ i=1, 2, \ldots, k$
- We require that $v_i > p,\ i=1, 2, \ldots, k$
- Define the pooled sample covariance matrix

$$\boldsymbol{S}_{pl} = \frac{\sum_{i=1}^{k} v_i \boldsymbol{S}_i}{\sum_{i=1}^{k} v_i}$$

- The test statistic is then given by

$$M = \frac{|\boldsymbol{S}_1|^{v_1/2} |\boldsymbol{S}_2|^{v_2/2} \ldots |\boldsymbol{S}_k|^{v_k/2}}{|\boldsymbol{S}_{pl}|^{\sum_{i=1}^{k} v_i/2}}$$

Box's Chi-square Approximation

- The Box's M-test is the way out
- Define c_1 as follows

$$c_1 = \left[\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \right]$$

- $u = -2\,(1 - c_1)\,ln(M)$ is distributed as chi-square with $(k-1)p(p+1)/2$ d.f.
- In the above step

$$\ln(M) = \frac{1}{2} \sum_{i=1}^{k} v_i \ln\left(|S_i|\right) - \frac{1}{2} \left(\sum_{i=1}^{k} v_i \right) \ln |S_{pl}|$$

Box's *F*- Approximation

- Define the following quantities:

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k}\frac{1}{v_i^2} - \frac{1}{\left(\sum_{i=1}^{k} v_i\right)^2}\right]$$

$$a_1 = \frac{1}{2}(k-1)\,p(p+1), a_2 = \frac{a_1+2}{\left|c_2 - c_1^2\right|}$$

$$b_1 = \frac{1-c_1-a_1/a_2}{a_1}, b_2 = \frac{1-c_1+2/a_2}{a_2}$$

- If $c_2 > c_1^2$, $F = -2b_1\, ln(M)$ is approximately $F_{a1,a2}$
- If $c_2 < c_1^2$, $F = \frac{-2a_2 b_2 \ln(M)}{a_1(1+2b_2\ln(M))}$ is approximately $F_{a1,a2}$

Example: Return to Psychological Tests for Males and Females

We need to test if the covariance matrices for males and females are identical or not

```
# Testing for Equality of Covariance Matrices
mfp=read.csv("MF_Psycho_Test_Scores.csv",header=T)
males=mfp[,1:4]; females=mfp[,5:8]
nm=nrow(males);nf=nrow(females)
p=4; k=2
vm=nm-1; vf=nf-1
meanm=mean(males); meanf=mean(females)
sigmam=var(males); sigmaf=var(females)
sigmapl=(1/(nm+nf-2))*((nm-1)*sigmam+(nf-1)*sigmaf)
ln_M = .5*(vm*log(det(sigmam))+vf*log(det(sigmaf))) -.5*(vm+vf)*log(det(sigmapl))
exact_test = -2*ln_M # the Exact Test
[1] 14.5606
```

The calculated Exact Test value is less than the critical value 19.74, and thus we fail to reject the null hypothesis

Example: Return to Psychological Tests for Males and Females

```
# The Box's chi-square approximation
c1 = (sum(c(1/vm,1/vf))- (1/sum(c(vm,vf))))*((2*p^2+3*p-1)/(6*(p+1)*(k-1)))
u = -2*(1-c1)*ln_M
qchisq(1-0.05,(k-1)*p*(p+1)/2)
u; qchisq(1-0.05,(k-1)*p*(p+1)/2)
[1] 13.55075
[1] 18.30704
c2 = ((p-1)*(p+2)/(6*(k-1)))*(sum(c(1/vm,1/vf)^2)- (1/(sum(c(vm,vf))^2)))
a1 = (k-1)*p*(p+1)/2; a2 = (a1+2)/(abs(c2-c1^2))
b1 = (1-c1-a1/a2)/a1; b2 = (1-c1+2/a2)/a2
if(c2>c1^2) {Ftest = -2*b1*ln_M} else {Ftest = (2*a2*b2*ln_M)/
(a1*(1+2*b2*ln_M))}
Ftest; qf(1-.05,10,Inf)
[1] 1.354283
[1] 1.830704
```

Both the Chi-square and F- approximation yield the same conclusion as the Exact test.

Testing for Independence of Sub-Vectors

- Rencher (261-263)
- **Seishu Wine Data**

y_1 : Taste

y_2 : Odor

x_1 : pH

x_2 : Acidity_1

x_3 : Acidity_2

x_4 : Sake_meter

x_5 : Direct_reducing_sugar

x_6 : Total_sugar

x_7 : Alcohol

x_8 : Formyl_nitrogen

Consider the problem of testing for the independence of the sub-vectors:

$$(y_1, y_2), (x_1, x_2, x_3), (x_4, x_5, x_6), (x_7, x_8)$$

Towards this, we need

$$S = \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{pmatrix}$$

Theoretically, we need the following

$$u = \frac{|(S)|}{|(S_{11})||(S_{22})|\ldots|(S_{kk})|} = \frac{|(R)|}{|(R_{11})||(R_{22})|\ldots|(R_{kk})|}$$

$$u' = -vc \ln u$$

$$c = 1 - \frac{1}{12fv}(2a_3 + 3a_2),$$

$$f = \frac{1}{2}a_2, \qquad a_2 = p^2 - \sum_{i=1}^{k} p_i^2, \qquad a_3 = p^3 - \sum_{i=1}^{k} p_i^3$$

and in R we need

```
sheishu=read.csv("Seishu_wine.csv",head
er=T)
noc=c(2,3,3,2)
nov=10
v=nrow(sheishu)-1
varsheishu=var(sheishu)
s11 = varsheishu[1:2,1:2]
s22 = varsheishu[3:5,3:5]
s33 = varsheishu[6:8,6:8]
s44 = varsheishu[9:10,9:10]
u = det(varsheishu)/
(det(s11)*det(s22)*det(s33)*det(s44))
a2 = nov^2 - sum(noc^2)
a3  = nov^3 - sum(noc^3)
f = a2/2
cc = 1 - (2*a3 + 3*a2)/(12*f*v)
u1 = -v*cc*log(u)
```

- The R program returns the following values

```
u; a2; a3; f; cc; u1
[1] 0.01627025
[1] 74
[1] 930
[1] 37
[1] 0.8383038
[1] 100.1221
qchisq(1-0.001,37)
[1] 69.34645
```

which is in agreement with the values reported on page 264, Rencher (2002)

Since the u1 value exceeds $\chi^2_{.001,37} = 69.35$ we reject the null hypothesis of independence of sub-vectors

Principal Component Analysis (PCA)

- An effective method for data reduction
- Suppose, we have large number of correlated variables, say q, $x_1, x_2, \ldots, x_q$
- PCA returns a new set of variables $y_1, y_2, \ldots, y_q$, with each the new variables as a linear combination of the x's
- The y's are in decreasing order of importance in the sense that y_i has more information about x's than y_j, whenever $i > j$
- Further, the $y's$ are designed to be uncorrelated
- The central theme being that a few first y_i's capture a lot of information about the x's

- PCA may be useful in the following two cases:
  - Too many explanatory variables relative to the number of observations
  - The explanatory variables are highly correlated
- The first principal component y_1 is a combination of the x's

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1q}x_q$$

A useful restriction on the vector $\mathbf{a_1}=(a_{11}, a_{12}, \ldots, a_{1q})$ is $\mathbf{a_1}^T\mathbf{a_1}=1$

- The second principal component y_2 is a combination of the x's

$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2q}x_q \text{, with}$$

$$\mathbf{a_2}^T\mathbf{a_2} = \mathbf{1} \text{ and } \mathbf{a_2}^T\mathbf{a_1} = 0$$

- And so on

- We need to find $\mathbf{a}_1$ which will maximize the variance of y_1 subject to the constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$.
- The Lagrangian multiplier helps us out, and leads us to the solution that $\mathbf{a}_1$ is the eigen vector of the sample covariance matrix $\mathbf{S}$ corresponding to the maximum eigen value. Similarly, $\mathbf{a}_j$ corresponds to the eigen vector of $q\text{-}j\text{+}1$ – ordered eigen value
- Its further easy to see that the variance of the i-th principal component is just the i-th ordered eigen value
- The j-*th* principal component accounts for a proportion P_j of the total variation of the $x's$

$$P_j = \frac{\lambda_j}{tr(\boldsymbol{S})}$$

- Finally, the variation accounted by the first *m* principal components is

$$P^{(m)} = \frac{\sum_{j=1}^{m} \lambda_j}{tr(\boldsymbol{S})}$$

Note: *If the variables are on different scales, use the correlation matrix instead of the Covariance Matrix*

- The covariance between variable *i* and component *j* is given by

$$Cov(x_i, y_j) = \lambda_j a_{ji}$$

- and the correlation is

$$r_{x_i, y_j} = \frac{\lambda_j a_{ji}}{\sqrt{Var(x_i)\,Var(y_j)}} = \frac{\lambda_j a_{ji}}{s_i \sqrt{(\lambda_j)}} = \frac{a_{ji}\sqrt{(\lambda_j)}}{s_i}$$

- If the components are extracted from the correlation matrix

$$r_{x_i, y_j} = a_{ji}\sqrt{(\lambda_j)}$$

Loadings

- Loadings are defined as the correlation between the *i*-th variables and the *j*-th principal component, that is

$$L_{ij} = Corr\left(X_i, PC_j\right)$$

- The loadings are easily obtained in almost all the statistical software
- We can see its utility as it helps understand the relationships between the variables and the principal components.

Rescaling Principal Components

$\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_q$: the vectors of the principal components

Define $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_q]$

Define $\Lambda = diag\{\lambda_1, \lambda_2, ..., \lambda_q\}$

Then

$$S = A \Lambda A'$$

A useful rescaling that can be seen from the above expression is

$$\boldsymbol{a}_i^* = \lambda_i^{1/2} \boldsymbol{a}_i$$

leading to

$$\boldsymbol{S} = \boldsymbol{A}^* (\boldsymbol{A}^*)'$$

where $A^* = [a_1^*, a_2^*, ..., a_q^*]$

This rescaling plays a very important role in *factor analysis.*

Principal Component Scores

$\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m$ **:** the vectors of the first *m* principal components

For an individual *i* with variable $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iq})$, the *principal component scores* are defined as below:

$$y_{i1} = \boldsymbol{a}_1' \boldsymbol{x}_i$$

$$y_{i2} = \boldsymbol{a}_2' \boldsymbol{x}_i$$

$$\vdots$$

$$y_{iq} = \boldsymbol{a}_m' \boldsymbol{x}_i$$

Example: PCA for Air Pollution Data

- S02: Sulphur dioxide content of air in micrograms per cubic meter
- Temp: Average annual temperature in OF
- Manu/: Number of manufacturing enterprises employing 20 or more workers
- Pop: Population size (1970 census) in thousands
- Wind: Average annual wind speed in miles per hour
- Precip: Average annual precipitation in inches
- Days: Average number of days with precipitation per year

We believe that, when ever possible, writing programs from the scratch gives more insight when entering a new paradigm. Of course, once mastered the art the in-built functions can then be used quite freely.

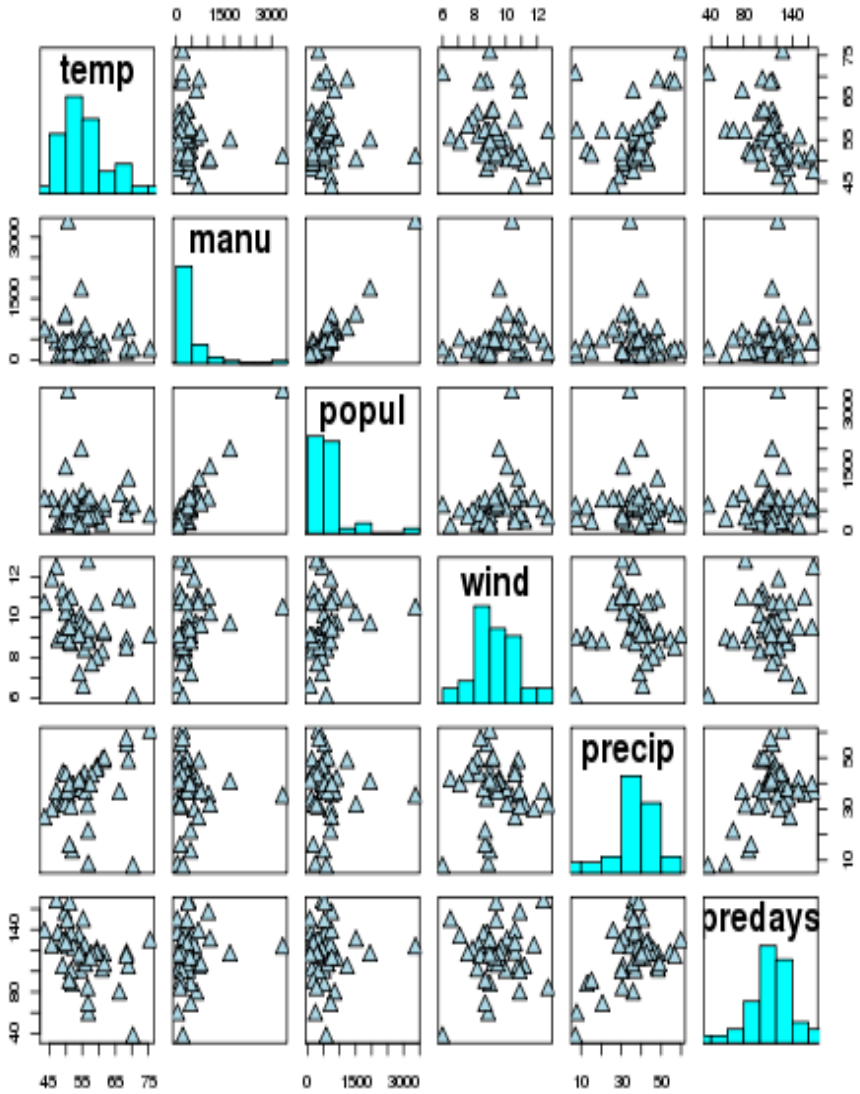
- The “pairs” command is a very useful too
- Its extensions are just even better

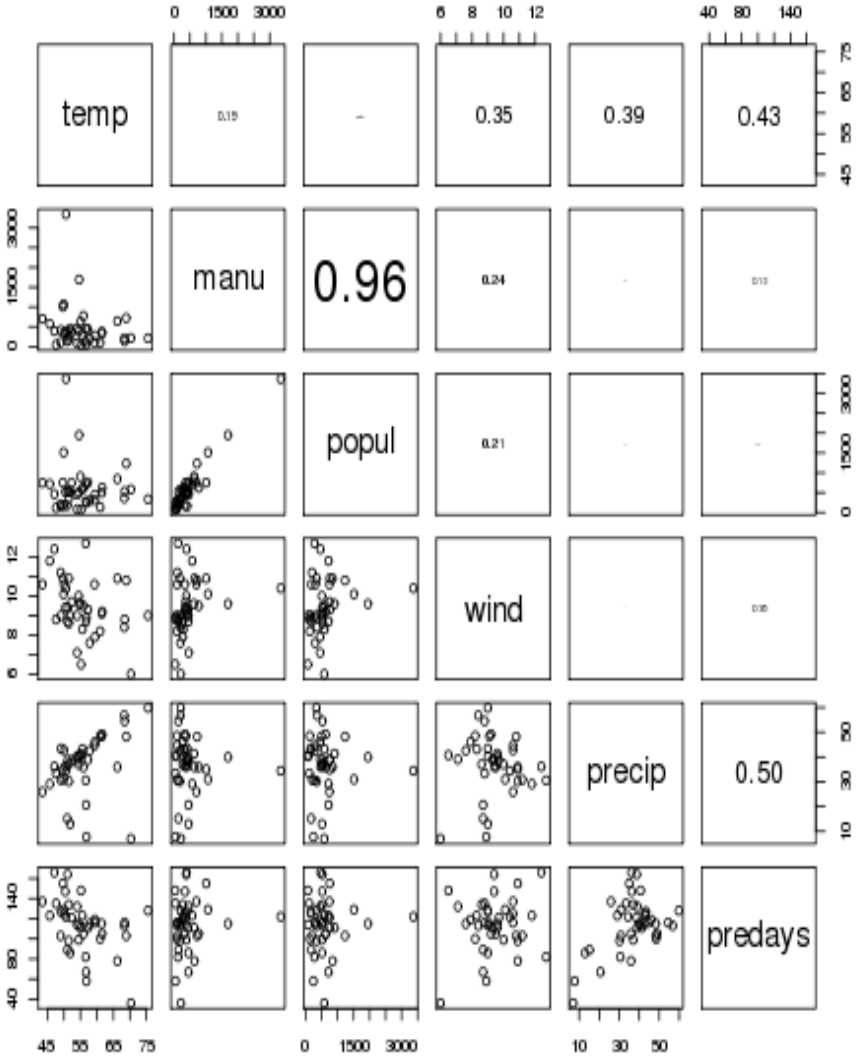
```
library(HSAUR2)
data(USairpollution)
panel.hist <- function(x, ...){
usr <- par("usr"); on.exit(par(usr))
par(usr = c(usr[1:2], 0, 1.5) )
h <- hist(x, plot = FALSE)
breaks <- h$breaks; nB <- length(breaks)
y <- h$counts; y <- y/max(y)
rect(breaks[-nB], 0, breaks[-1], y,
    col="cyan", ...)}
pairs(USairpollution[,-1],cex = 1.5, pch =
    24, bg="light blue", diag.panel =
    panel.hist, cex.labels = 2,
    font.labels=2)
```

Gives histogram in the diagonal

```
panel.cor <- function(x, y, digits=2,
prefix="", cex.cor, ...){
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789),
digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex.cor <-
0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r) }
pairs(USairpollution[,-1],
upper.panel=panel.cor)
```

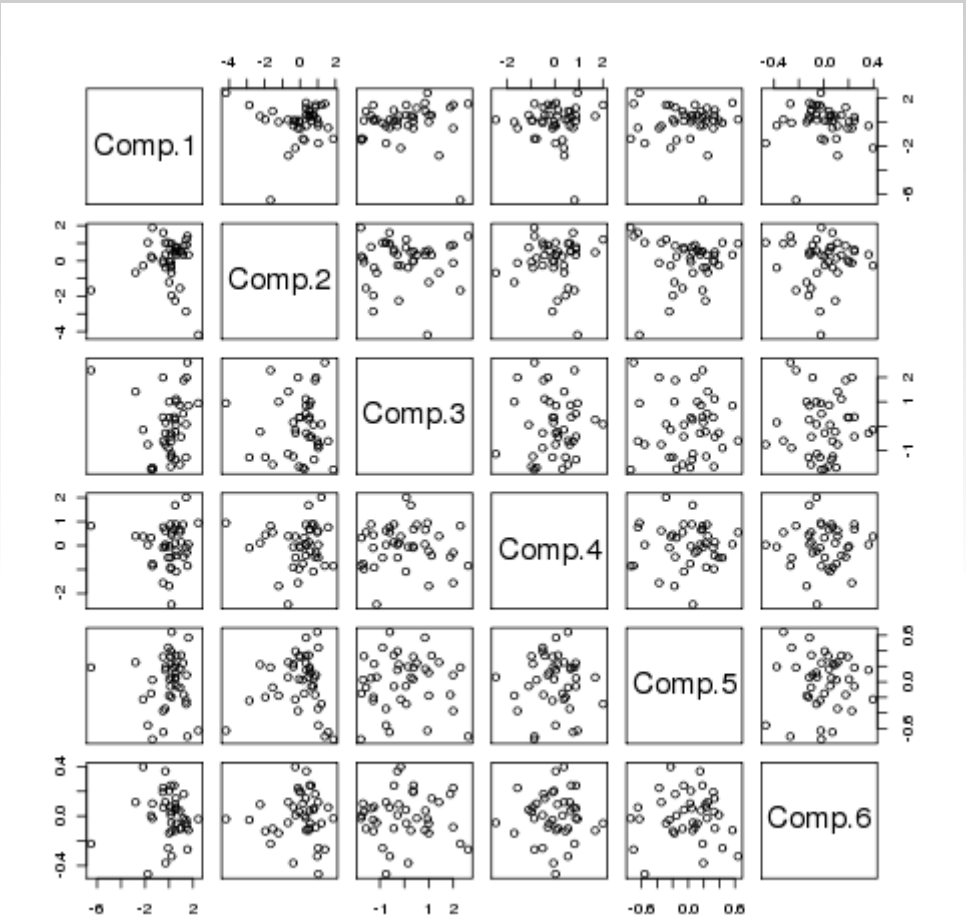
Gives Correlation Coefficient in the Upper Pannel


temp
manu
popul
wind
precip
predays


temp
0.15
0.35
0.39
0.43
manu
0.96
0.24
popul
0.21
wind
precip
0.50
predays

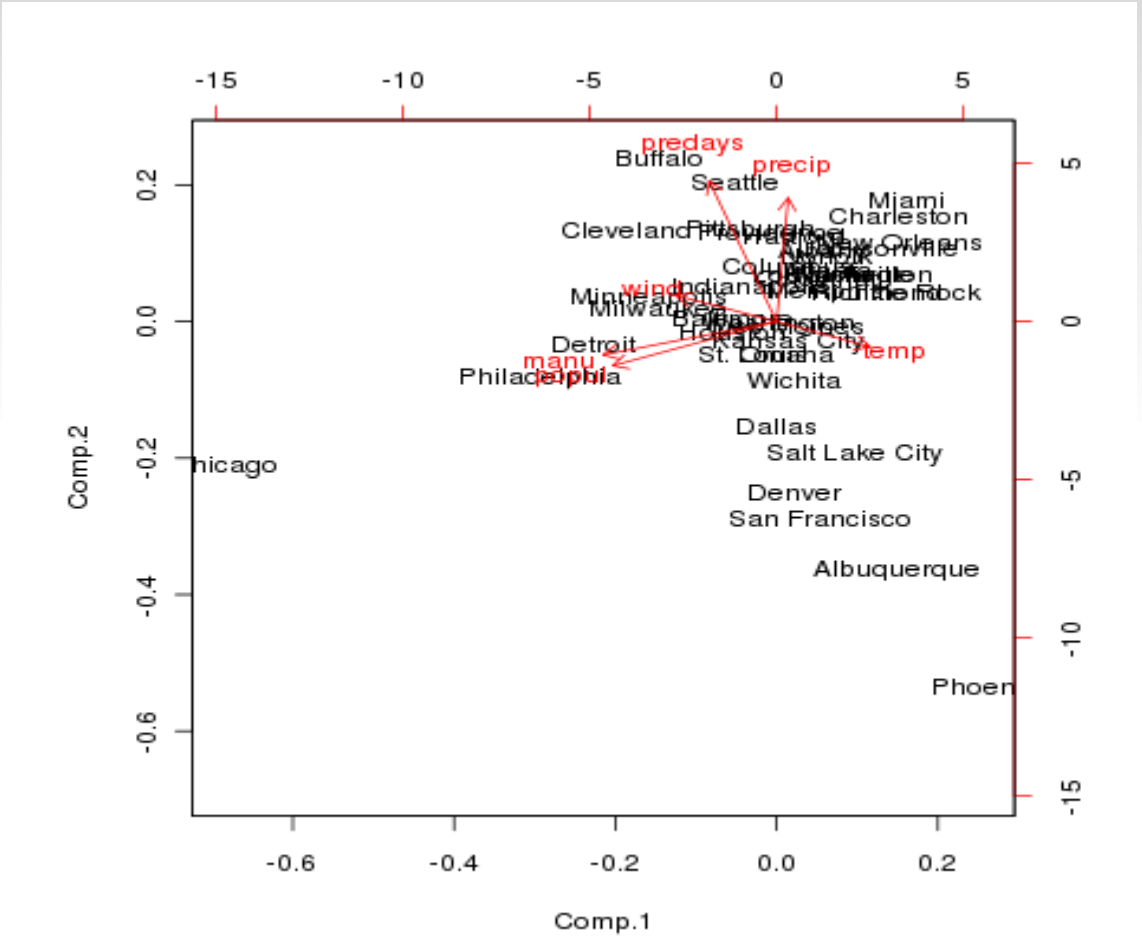
- From the above pictures its clear that the variables are on different scales
- Outliers are clearly present
- We need to use Correlation Matrix and not the Covariance Matrix for PCA

```
usair.pc=princomp(USairpollution[,-1],cor=T)
summary(usair.pc)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287
Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511
Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
                           Comp.6
Standard deviation     0.185599752
Proportion of Variance 0.005741211
Cumulative Proportion  1.000000000
```



Plot to Check of the Principal Components are Orthogonal or Not

Biplot


predays
precip
Buffalo
Seattle
Miami
Charleston
Cleveland
Detroit
Philadelphia
manu
temp
Wichita
Dallas
Salt Lake City
Denver
San Francisco
Albuquerque
Phoen
hicago
Comp.1
Comp.2
-15
-10
-5
0
5
0.2
0.0
-0.2
-0.4
-0.6

Discriminant Analysis

Discriminant functions are linear combinations of variables that best separate groups. (Rencher, 2002)

$$X_{11}, X_{12}, \ldots, X_{1n_1}, X_1 \sim N_p(\mu_1, \Sigma)$$

$$X_{21}, X_{22}, \ldots, X_{2n_2}, X_2 \sim N_p(\mu_2, \Sigma)$$

Covariance matrix is assumed to be same

Consider linear combinations

$$z_{1i} = \boldsymbol{a}' \boldsymbol{x}_{1i} = a_1 x_{1i1} + a_2 x_{1i2} + \ldots + a_p x_{1ip}, i = 1, 2, \ldots, n_1$$

$$z_{2i} = \boldsymbol{a}' \boldsymbol{x}_{1i} = a_1 x_{2i1} + a_2 x_{2i2} + \ldots + a_p x_{2ip}, i = 1, 2, \ldots, n_2$$

Define

$$\bar{z}_1 = \sum_{i=1}^{n_1} z_{1i}/n_1 = \boldsymbol{a}' \bar{\boldsymbol{x}}_1 ; \bar{z}_2 = \sum_{i=1}^{n_2} z_{2i}/n_2 = \boldsymbol{a}' \bar{\boldsymbol{x}}_2$$

The problem of DA is to find **a** which maximizes the standardized difference

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{\left[\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\right]^2}{\boldsymbol{a}' \boldsymbol{S}_{pl} \boldsymbol{a}}$$

The solution is given by

$$a = S_{pl}^{-1}(\bar{x}_1 - \bar{x}_2)$$

```
# Program for Discriminant Analysis
kj=read.csv("kj_69.csv",header=T)
attach(kj)
kjm=colMeans(kj,na.rm=T)
n1=5; n2=7;
m1=colMeans(cbind(t1y1,t1y2),na.rm=T)
m2=colMeans(cbind(t2y1,t2y2),na.rm=T)
sigma1=var(cbind(t1y1,t1y2),na.rm=T)
sigma2=var(cbind(t2y1,t2y2),na.rm=T)
sigmapl=(1/(n1+n2-2))*((n1-1)*sigma1+(n2-1)*sigma2)
discriminant=solve(sigmapl)%*%(m1-m2)
          [,1]
t1y1 -1.633377
t1y2  1.819779
```

Discriminant Analysis with *k*- Groups

- We won't go in the theory of discriminant analysis for *k*-groups
- Simply illustrate with “IRIS” data set

```
library(MASS)
irlda=lda(iris[,5]~iris[,1]+iris[,2]+iris[,3]+iris[,4])
ir_pred=predict(irlda,iris[,1:4])$class
table(iris[,5],ir_pred)
            ir_pred
             setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          1        49
sum(diag(table(iris[,5],ir_pred)))/150
[1] 0.98
```