NRES_798_19_201501

Multivariate statistics

Univariate vs multivariate

- Single response variable

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Multiple response variable

$$y = c(y_1, y_2, y_3, y_4)$$

Aims of multivariate statistics

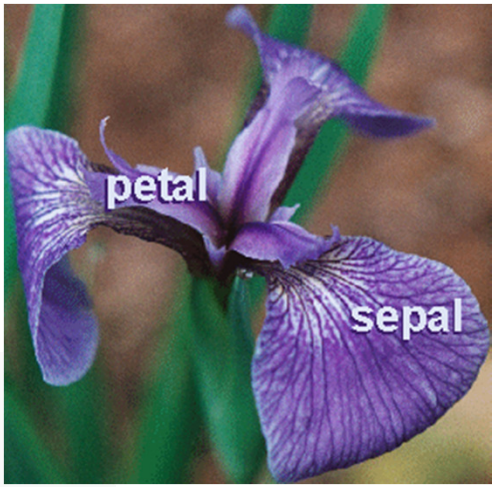
- Simplification, data reduction (PCA)
- Inference (non-observed variable: Factor analysis)
- Classification (Cluster analysis, discriminant analysis)
- Hypothesis testing (MANOVA)

PCA

- Original data : multiple measurements per object
- Calculate a reduced set of variables (PC1, PC2) that captures the essence of the original data (in a simpler data structure)
  - Algorithms focus on variance (i.e. PC1 accounts for the most variance, PC2 the second most)
  - Principle components are linearly uncorrelated (orthogonal)

PCA: Principle Components Analysis

Can these variables be reduced to a simpler set?
(covariance matrix)




```
head(iris)
```

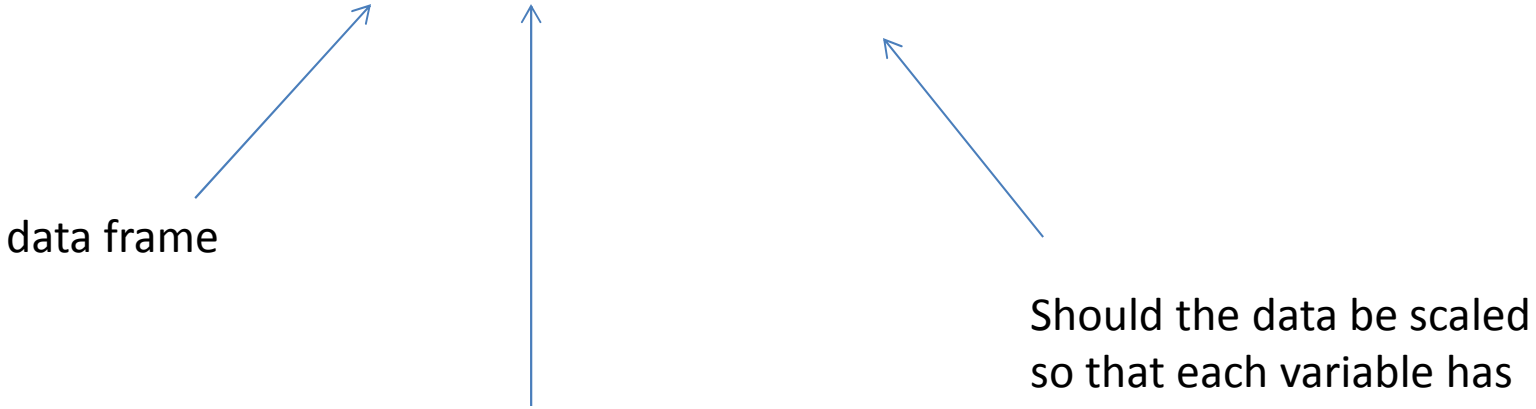
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1
1 1 1 1 1 1 1 ...
```

PCA

pca <- **prcomp**(ir, center = TRUE, scale = TRUE)



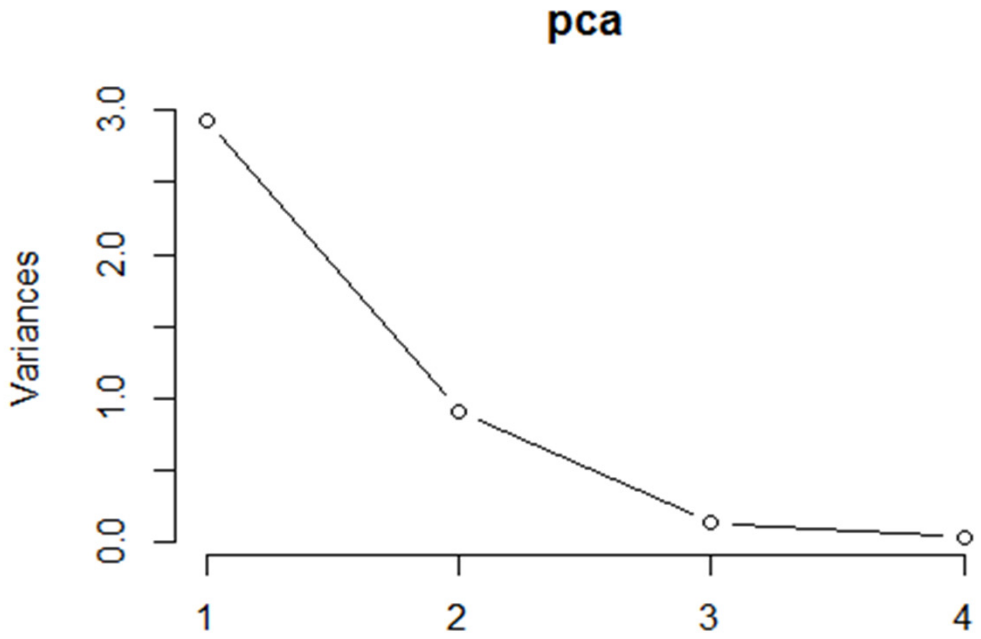
pca <- **princomp**(ir, cor = TRUE)

PCA

- Variance (in the original data) accounted for by the new components

```
summary(pca,loadings=TRUE)

## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7125 0.9524 0.36470 0.16568
## Proportion of Variance 0.7331 0.2268 0.03325 0.00686
## Cumulative Proportion  0.7331 0.9599 0.99314 1.00000
```

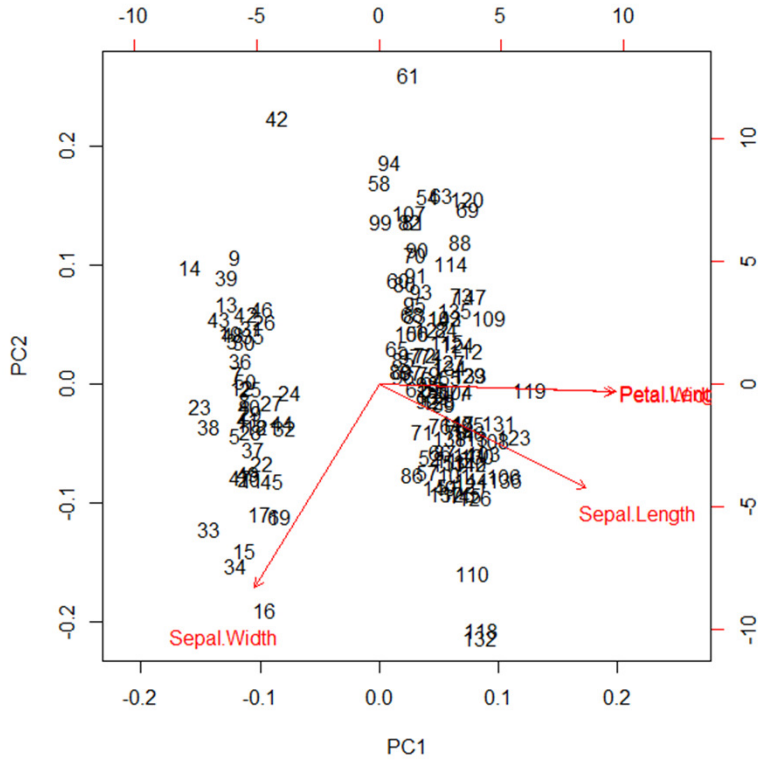
```
plot(pca,type="l")
```



PCA: eigenvector

```
print(pca)

## Standard deviations:
## [1] 1.7124583 0.9523797 0.3647029 0.1656840
##
## Rotation:
##                      PC1         PC2        PC3         PC4
## Sepal.Length  0.5038236 -0.45499872  0.7088547  0.19147575
## Sepal.Width  -0.3023682 -0.88914419 -0.3311628 -0.09125405
## Petal.Length  0.5767881 -0.03378802 -0.2192793 -0.78618732
## Petal.Width   0.5674952 -0.03545628 -0.5829003  0.58044745
```

biplot(pca)



What do these components represent

- PC1?
- PC2?

PCA: example 2

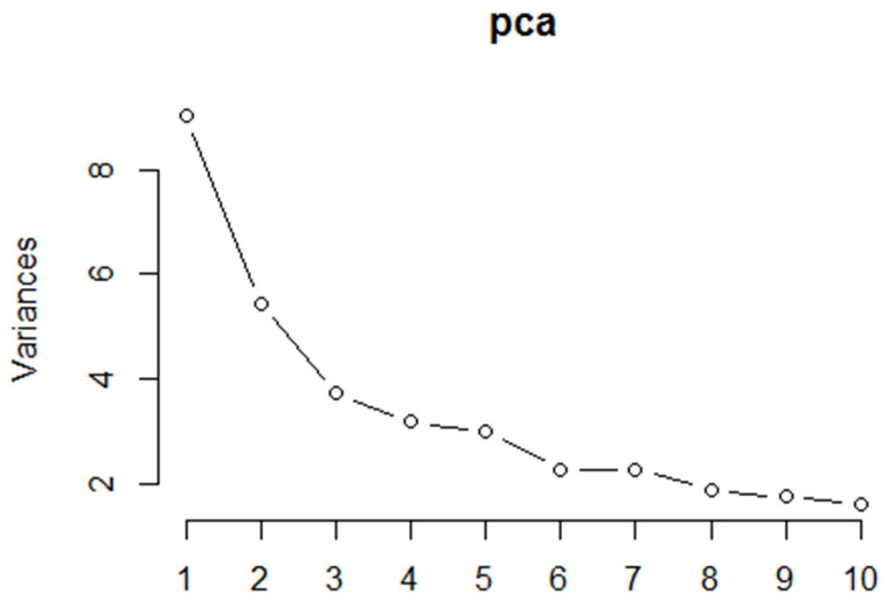
- 54 species (AC to VK), dry weight (g)
- What are the principle components?
- What environmental factors are associated with them?

```
names(pgdata)
##  [1] "AC"     "AE"     "AM"     "AO"     "AP"     "AR"     "AS"
##  [8] "AU"     "BH"     "BM"     "CC"     "CF"     "CM"     "CN"
## [15] "CX"     "CY"     "DC"     "DG"     "ER"     "FM"     "FP"
## [22] "FR"     "GV"     "HI"     "HL"     "HP"     "HS"     "HR"
## [29] "KA"     "LA"     "LC"     "LH"     "LM"     "LO"     "LP"
## [36] "OR"     "PL"     "PP"     "PS"     "PT"     "QR"     "RA"
## [43] "RB"     "RC"     "SG"     "SM"     "SO"     "TF"     "TG"
## [50] "TO"     "TP"     "TR"     "VC"     "VK"     "plot"   "lime"
## [57] "species" "hay"    "pH"

pgd <- pgdata[,1:54]
head(pgd)
##    AC   AE   AM   AO   AP   AR AS   AU BH   BM   CC   CF   CM   CN CX
## 1 2.51 1.18 0.45 0.91 0.47 0.00  0 0.00  0 0.06 0.01 0.32 0.15 2.12  0
## 2 6.85 0.10 0.58 1.02 0.35 0.00  0 0.00  0 0.36 0.04 0.00 0.35 3.90  0
```

PCA: example 2

```
pca <- prcomp(pgd,scale = TRUE)

summary(pca,loadings=TRUE)
```

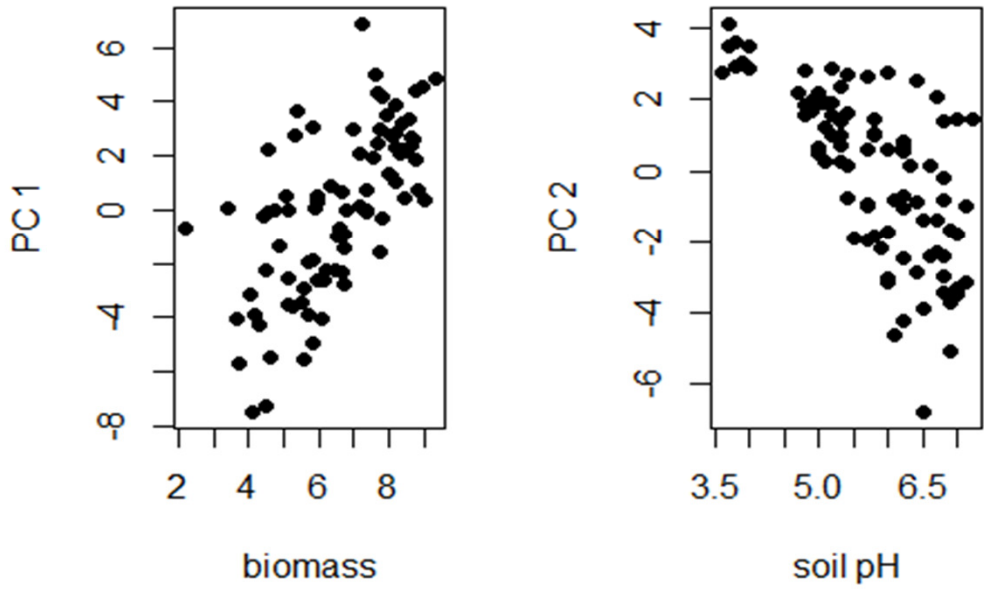
```
## Importance of components:
##                             PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     3.0048 2.3358 1.9317 1.78562 1.73303 1.51187
## Proportion of Variance 0.1672 0.1010 0.0691 0.05904 0.05562 0.04233
## Cumulative Proportion  0.1672 0.2682 0.3373 0.39639 0.45201 0.49434
```



PCA: example 2

Examining how raw variables relate to PCs
(what environmental factors relate to the PCs)

```
yv1 <- predict(pca)[,1]
yv2 <- predict(pca)[,2]
```



Factor analysis

- Using measured variables to estimate non-observed factors?
  - Based on correlations between variables
- Observed variables
  - E.g. 54 species
  - E.g. income, education, occupation
- Unobserved, underlying “factors”
  - Latent variables
  - E.g. Community assembly rules
  - E.g. Historically: intelligence, social status

Factor analysis

```
factanal(ir,factors=1)

##
## Call:
## factanal(x = ir, factors = 1)
##
## Uniquenesses:
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        0.292        0.780        0.005        0.064
##
## Loadings:
##              Factor1
## Sepal.Length  0.842
## Sepal.Width  -0.469
## Petal.Length  0.998
## Petal.Width   0.968
##
##                Factor1
## SS loadings      2.860
## Proportion Var   0.715
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 69.05 on 2 degrees of freedom.
## The p-value is 1.01e-15
```

Number of factors specified

How much a factor explains a variable

Null hypothesis is that “perfect” model and test model are the same

Cluster analysis

- Multiple variables

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

- Do the variables naturally produce cluster (groups)?
- Number of clusters unknown
  - (vs. discriminant analysis)

Cluster analysis

```
# Cluster analysis
ir <- iris[,1:4]
ca <- kmeans(ir,3)
ca
```

Try a range of cluster #'s

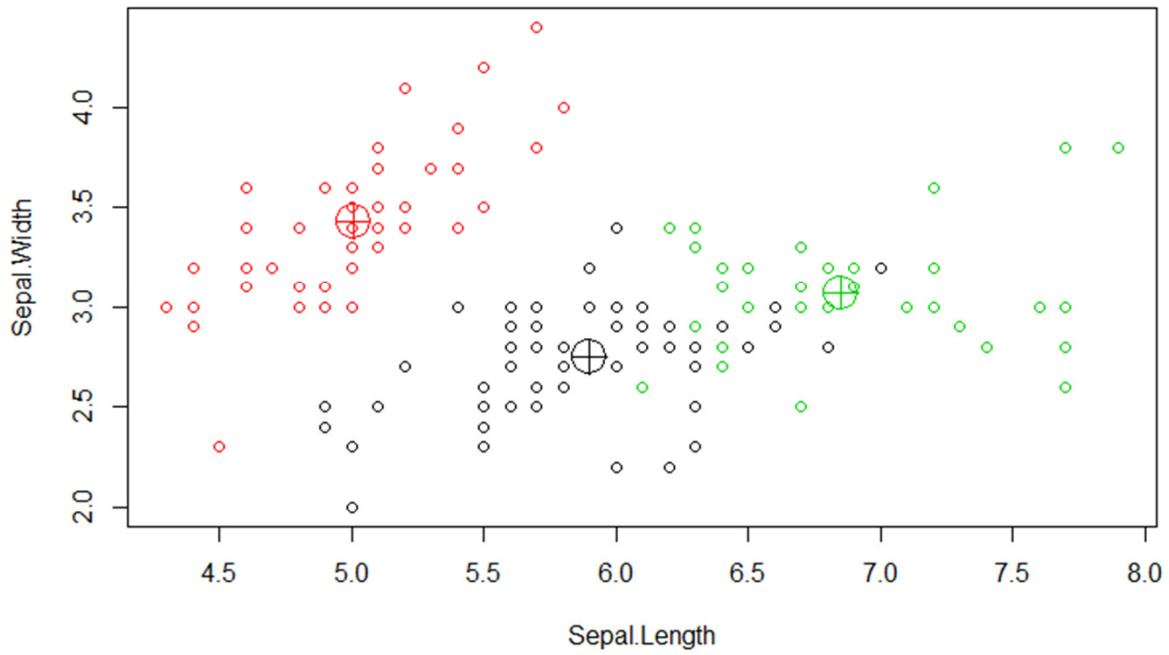
```
## K-means clustering with 3 clusters of sizes 38, 62, 50
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     6.850000    3.073684     5.742105    2.071053
## 2     5.901613    2.748387     4.393548    1.433871
## 3     5.006000    3.428000     1.462000    0.246000
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [36] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [106] 1 2 1 1 1 1 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 1 1 2 1 1 2
## [141] 1 1 2 1 1 1 2 1 1 2
##
## Within cluster sum of squares by cluster:
## [1] 23.87947 39.82097 15.15100
##  (between_SS / total_SS =  88.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Separation of clusters by variables

Remaining variation within groups

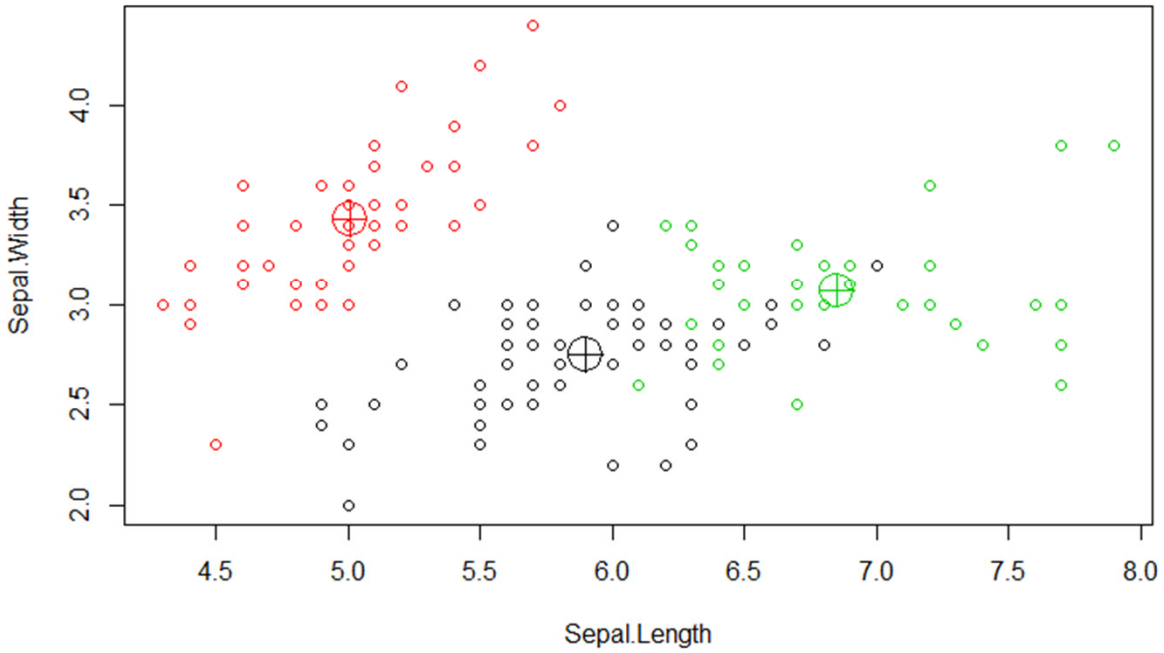
Cluster Analysis

```
plot(iris[c("Sepal.Length", "Sepal.Width")], col=ca$cluster)
points(ca$centers[,c("Sepal.Length", "Sepal.Width")], col=1:3, pch=10,
cex=3)
```



Is this correct?

Cluster Analysis



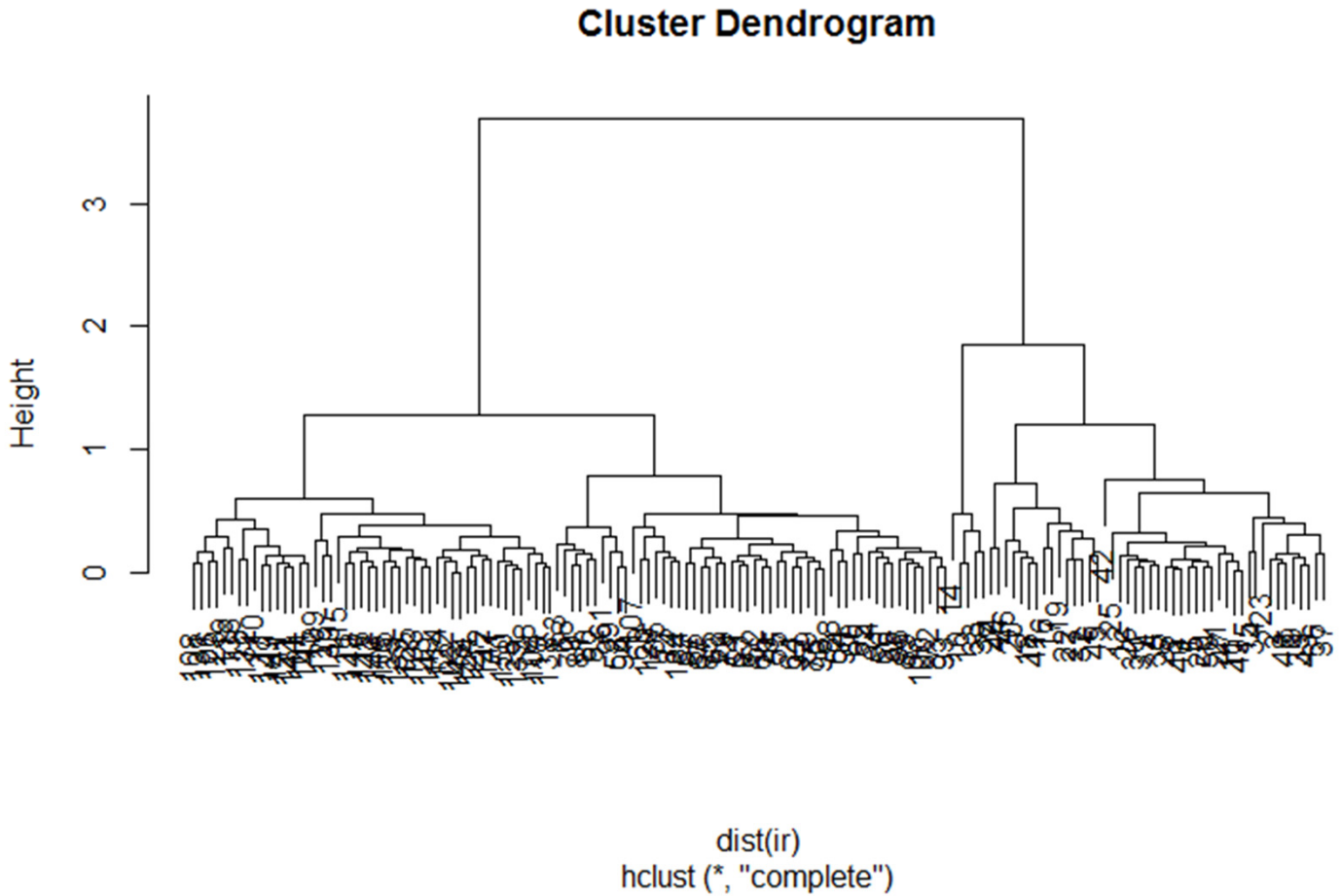

```
table(iris$Species,ca$cluster)
```

```
##
##                1  2  3
##    setosa      0  0 50
##    versicolor  2 48  0
##    virginica  36 14  0
```

Hierarchical Cluster Analysis

```
# Hierarchical cluster analysis
hca <- hclust(dist(ir))
plot(hca,labels=iris[,5])
```

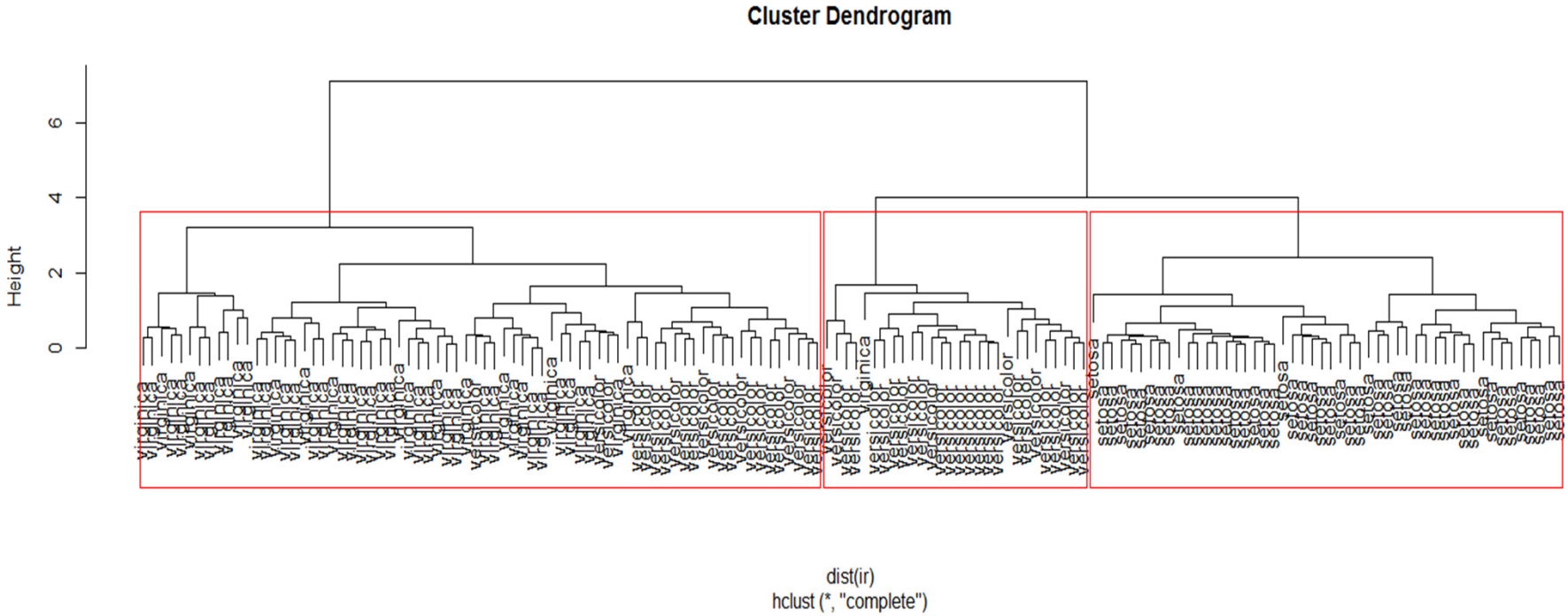
Distance matrix of data frame



Hierarchical Cluster Analysis

```
hca2 <- hclust(dist(ir))
plot(hca2,labels=iris[,5])
rect.hclust(hca2, 3)
```

Illustrate where clustering of set order occurs



Cluster Analysis example 2

7 variables

```
taxa <- read.table(paste(datpath,"taxon.txt",sep=""),header=TRUE)
head(taxa)
```

```
##     Petals Internode    Sepal    Bract  Petiole     Leaf    Fruit
## 1 5.621498  29.48060 2.462107 18.20341 11.27910 1.128033 7.876151
## 2 4.994617  28.36025 2.429321 17.65205 11.04084 1.197617 7.025416
## 3 4.767505  27.25432 2.570497 19.40838 10.49072 1.003808 7.817479
## 4 6.299446  25.92424 2.066051 18.37915 11.80182 1.614052 7.672492
## 5 6.489375  25.21131 2.901583 17.31305 10.12159 1.813333 7.758443
## 6 5.785868  25.52433 2.655643 17.07216 10.55816 1.955524 7.880880
```

Cluster Analysis 2

Try 4 groups

Number of individuals allocated to each group

```
ca1 <- kmeans(taxa,4)
ca1
```

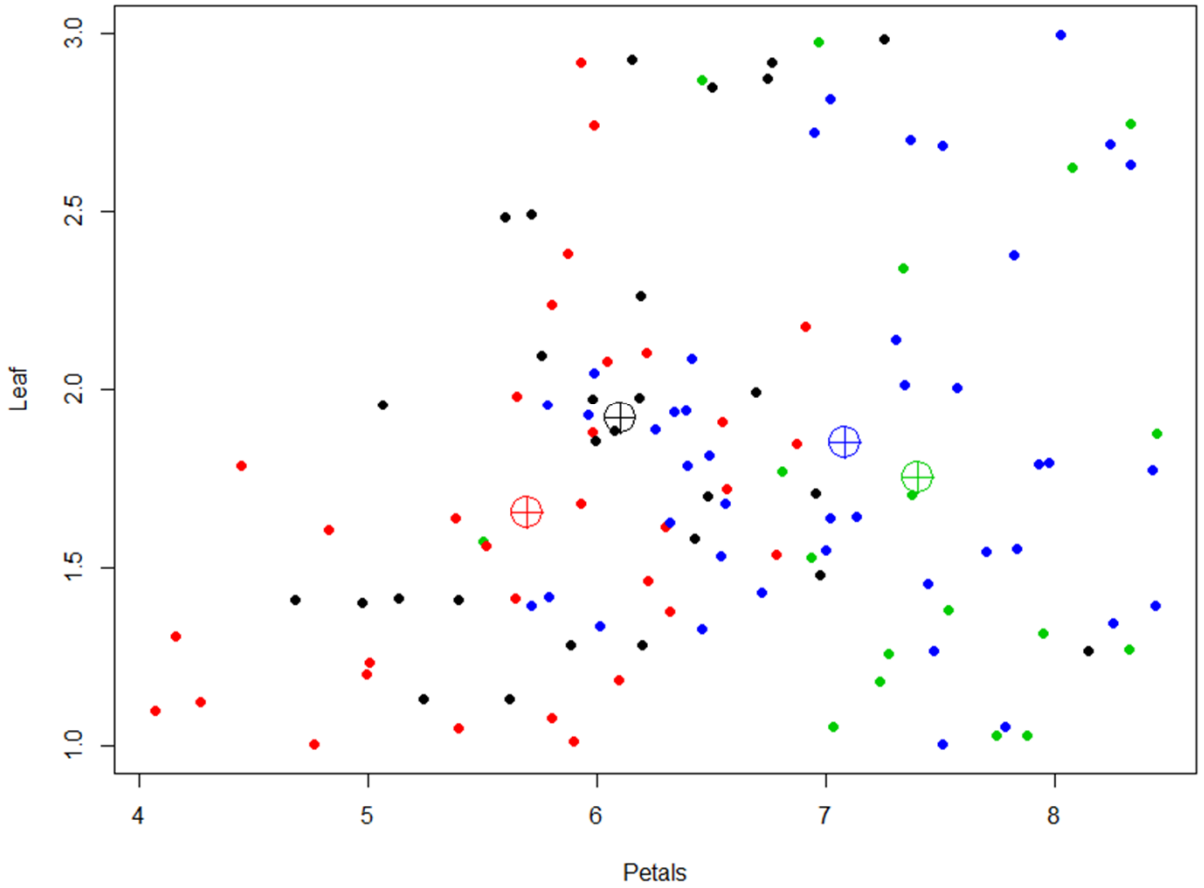
```
## K-means clustering with 4 clusters of sizes 29, 34, 30, 27
##
## Cluster means:
##     Petals Internode    Sepal    Bract   Petiole     Leaf    Fruit
## 1 5.433583  27.74826 2.560491 18.73832 10.898197 1.557074 7.543210
## 2 6.746014  29.99127 3.079527 18.30375  9.717186 2.036219 7.514235
## 3 6.846712  26.85746 2.431017 18.53379  8.725473 1.755100 7.414451
## 4 7.094087  26.36046 4.011740 18.19925 10.142092 1.805335 7.468062
```

Better

Poor

Cluster Analysis 2

```
plot(taxa$Petals,taxa$Leaf,col=ca1$cluster,pch=19,xlab ="Petals",ylab="Leaf")
points(ca1$centers[,"Petals"],ca1$centers[,"Leaf"],pch=10,cex=3,col=c(1:4))
```



Clusters poorly separate by the variables

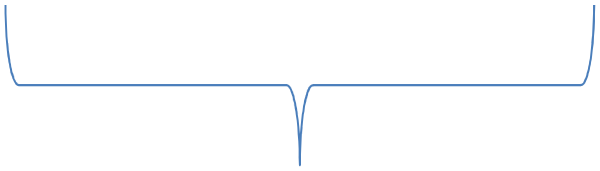
Cluster Analysis 2b

Use PCA to simplify data prior to cluster analysis

```
pca1 <- prcomp(taxa,center=TRUE,scale=TRUE)
summary(pca1)

## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6     PC7
## Standard deviation     1.2333 1.1085 1.0398 0.9924 0.9740 0.9289 0.61052
## Proportion of Variance 0.2173 0.1755 0.1544 0.1407 0.1355 0.1233 0.05325
## Cumulative Proportion  0.2173 0.3928 0.5473 0.6880 0.8235 0.9467 1.00000
```

Cluster Analysis 2b

```
pca1
## Standard deviations:
## [1] 1.2333082 1.1085079 1.0397586 0.9923526 0.9739864 0.9289335 0.6105184
##
## Rotation:
##              PC1         PC2          PC3       PC4        PC5
## Petals     0.62197308  -0.08929246  -0.2169767304  0.4160531 -0.25107393
## Internode -0.14616202  -0.46151428  -0.0861616066  0.3914963  0.54835635
## Sepal     -0.01901835   0.40986759  -0.7939107182  0.1848241 -0.05548272
## Bract     -0.17607514   0.32192813   0.4339579784  0.3831308 -0.50186154
## Petiole   -0.63323621  -0.11436469  -0.3287959034 -0.2138901 -0.28904510
## Leaf       0.17920062  -0.64904503  -0.1370283587 -0.2494848 -0.52151584
## Fruit     -0.35682074  -0.27030267  -0.0000576287  0.6202335 -0.16061764
```



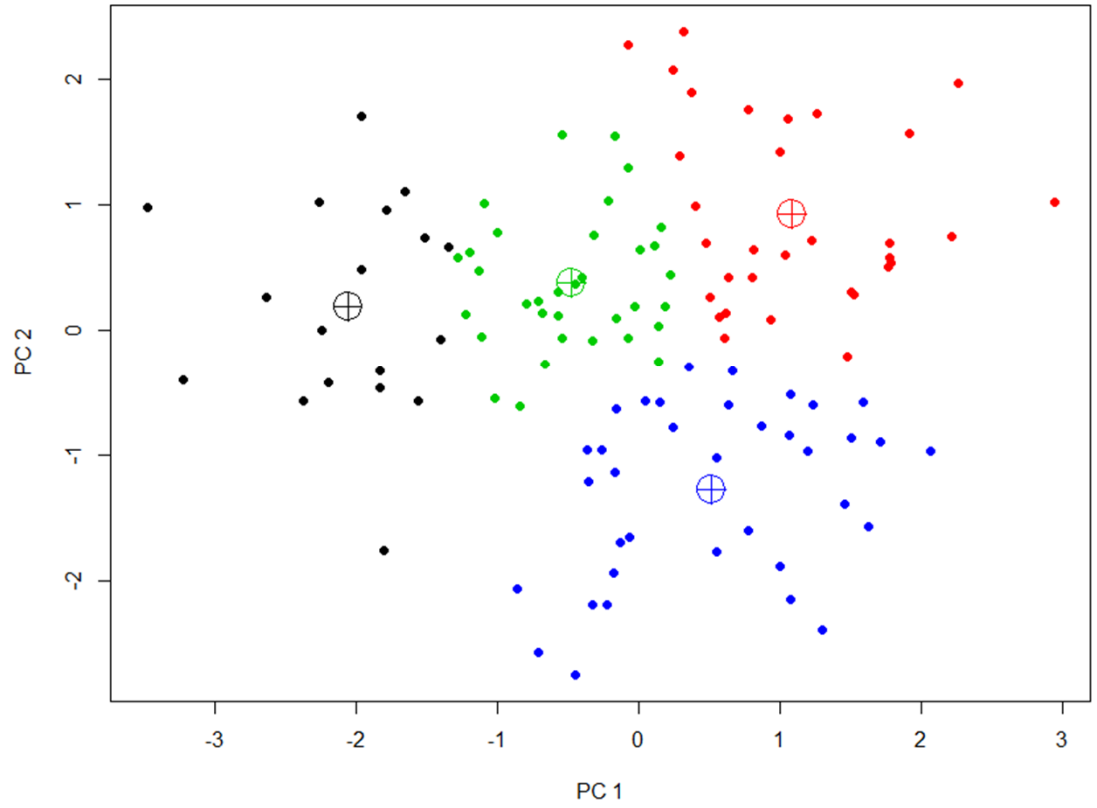
PC interpretation

Cluster Analysis 2b

```
pc.comp1 <- predict(pca1)[,1]
pc.comp2 <- predict(pca1)[,2]
pc.bind <- cbind(pc.comp1,pc.comp2)


ca2 <- kmeans(pc.bind,4)
```

Get PC values for each observation

```
plot(pc.comp1,pc.comp2,col=ca2$cluster,pch=19,xlab="PC 1",ylab="PC 2")
points(ca2$centers,pch=10,cex=3,col=c(1:4))
```



Discriminant analysis

- Determining how variables discriminate between known categorical groups
- Pattern recognition and interpretation

- Predict which group an observation belongs to based on measured variables
- Determine optimal separation of groups

Discriminant analysis

```
# Discriminant Analysis
#install.packages("MASS")
library(MASS)

Taxon <- rep(c("I","II","III","IV"),each=30)
nTaxon <- rep(c(1,2,3,4),each=30)

da1 <- lda(Taxon~.,taxa)
da1
## Call:
## lda(Taxon ~ ., data = taxa)
##
## Prior probabilities of groups:
##    I       II      III    IV
## 0.25 0.25 0.25 0.25
##
## Group means:
##       Petals          Internode    Sepal          Bract   Petiole     Leaf    Fruit
## I   5.476128        27.91886    2.537955        18.60268 10.864184 1.508029 7.574642
## II  7.035078        27.69834    2.490336        18.47557  8.541085 1.450260 7.418702
## III 6.849666        27.99308    2.446003        18.26330  9.866983 2.588555 7.482349
## IV  6.768464        27.78503    4.532560        18.42953 10.128838 1.645945 7.467917
```

Create group identifier

Initial assumptions of group membership
Null expectation represented

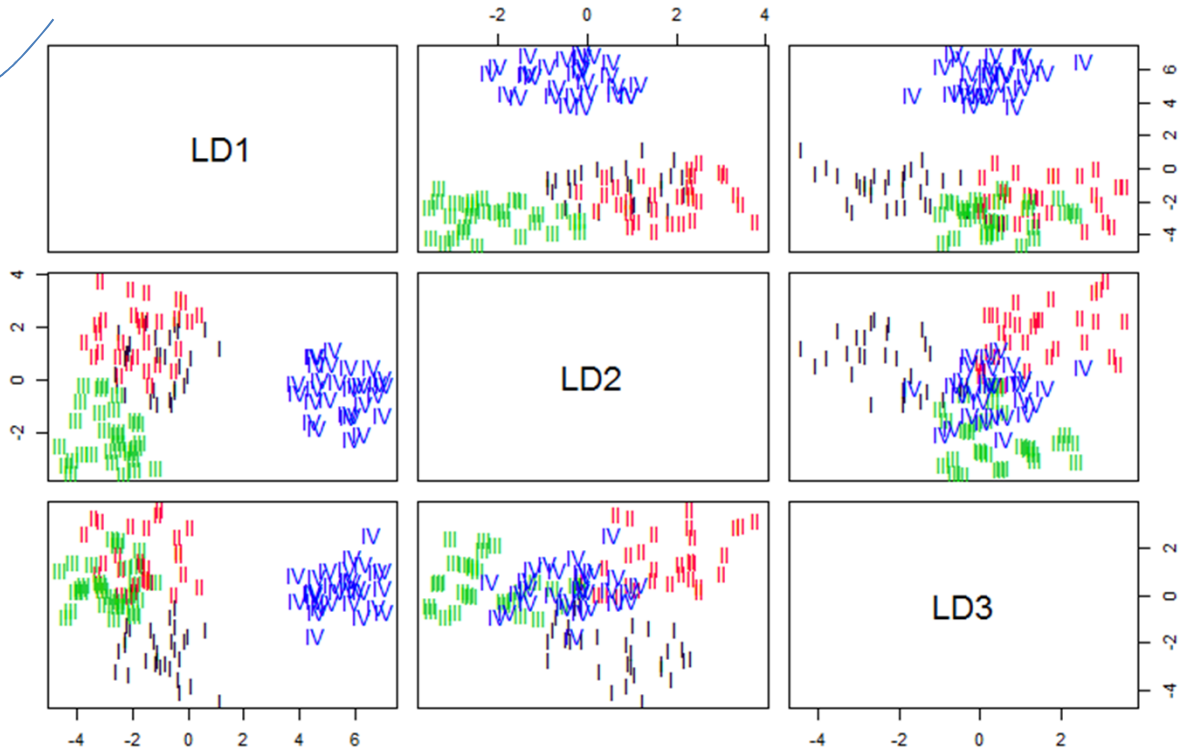
Discriminant analysis

```
## Coefficients of linear discriminants:
##                    LD1          LD2          LD3
## Petals     -0.01891137  0.034749952  0.559080267
## Internode   0.03374178  0.009670875  0.008808043
## Sepal       3.45605170 -0.500418135  0.401274694
## Bract       0.07557480  0.068774714 -0.024930728
## Petiole     0.25041949 -0.343892260 -1.249519047
## Leaf       -1.13036429 -3.008335468  0.647932763
## Fruit       0.18285691 -0.208370808 -0.269924935
```

```
## Proportion of trace:
##    LD1    LD2    LD3
## 0.7268 0.1419 0.1313
```

```
plot(da1,col=nTaxon,cex=1.2)
```

Similar to PCA



Discriminant analysis

Using training and sample data

```
train <- sample(1:120,60)
da2 <- lda(Taxon~.,taxa,prior=c(1,1,1,1)/4,subset=train)
da2
```

```
## Call:
## lda(Taxon ~ ., data = taxa, prior = c(1, 1, 1, 1)/4, subset =
train)
##
## Prior probabilities of groups:
##    I   II  III   IV
## 0.25 0.25 0.25 0.25
##
## Group means:
##       Petals Internode    Sepal    Bract   Petiole     Leaf
Fruit
## I   5.579682  27.75082 2.464828 18.53039 10.861420 1.514005
7.549624
## II  7.112912  27.62329 2.556901 18.51235  8.533198 1.531630
7.376543
## III 6.732470  28.03853 2.497793 18.09914  9.837361 2.611742
7.526170
## IV  6.690065  27.67323 4.418367 18.25902 10.144911 1.571427
7.394307
```

Discriminant analysis

Full

```
## Coefficients of linear discriminants:
##                    LD1          LD2          LD3
## Petals    -0.01891137  0.034749952  0.559080267
## Internode  0.03374178  0.009670875  0.008808043
## Sepal      3.45605170 -0.500418135  0.401274694
## Bract      0.07557480  0.068774714 -0.024930728
## Petiole    0.25041949 -0.343892260 -1.249519047
## Leaf      -1.13036429 -3.008335468  0.647932763
## Fruit      0.18285691 -0.208370808 -0.269924935
```

Train

```
## Coefficients of linear discriminants:
##                    LD1         LD2         LD3
## Petals    -0.05601362 -0.05517384  0.60511409
## Internode  0.02966780 -0.07217842 -0.01563392
## Sepal      3.29181737 -0.80726007  0.14060861
## Bract      0.23042540  0.12601267  0.03360850
## Petiole    0.14190294  0.01036077 -1.18814879
## Leaf      -1.04426196 -2.94334180 -0.09678861
## Fruit     -0.32832024 -0.63902109 -0.69961093
```

Linear discriminants differ depending on the size and nature of the data set

Discriminant analysis

Using the discriminants from the test data to predict group occurrence in new data

```
pda3 <- predict(da2,newdata=taxa[-train,])
pda3
```

```
## $class
##  [1] I   I   I   I   I   I   I   I   I   I   I   I   I   II  II  II
II
## [18] II  II  II  II  II  II  II  II  II  I   III III III III III
III III
## [35] III III III III III III III III III III IV  IV  IV  IV  IV  IV
IV
## [52] IV  IV  IV  IV  IV  IV  IV  IV  IV
## Levels: I II III IV
##
## $posterior
##                  I            II           III            IV
## 6   8.259463e-01 1.596169e-03 1.724576e-01 3.562881e-10
## 8   9.963463e-01 3.472212e-03 1.814437e-04 1.698120e-11
## 11  9.999963e-01 3.609032e-07 3.306926e-06 5.137116e-10
## 13  9.954814e-01 3.839969e-03 6.729999e-04 5.642025e-06
## 16  9.999331e-01 4.789522e-05 1.898459e-05 2.410358e-10
## 17  9.994300e-01 3.756456e-04 1.943737e-04 6.133810e-11
```

Probability of association

Discriminant analysis

```
pda3 <- predict(da2,newdata=taxa[-train,])
pda3
```

```
## $x
##             LD1         LD2         LD3
## 6   -1.5932167 -0.38947647 -1.65951713
## 8   -1.5673857  1.79547249 -1.43791620
## 11  -0.4381238  1.25926595 -3.82701080
## 13   0.1788501  0.74066716 -1.40601050
## 16  -0.8576699  1.59071082 -2.54886842
## 17  -1.3035107  1.35796650 -2.02743075
## 18  -1.0070411  0.45722081 -1.59007690
## 19   1.0494098  2.12577218 -3.99013353
## 20  -1.3582022  2.69232221 -2.09465507
## 22  -0.2162561  2.65907962 -1.52424203
## 25   0.7907364  2.05679403 -0.92694946
## 26  -0.8719132 -0.83903073 -2.29196970
```

Coefficients used in assignment

MANOVA

```
d1 <- data.frame(height,weight,volume,cal)

m5 <- manova(d1 ~ N + P + N*K)
```

Multivariate statistics

- PCA
  - Representing complex multivariate data as smaller number of key variables (principle components)
- Factor analysis
  - Estimating unmeasurable “factors” based on multiple observed variables
- Cluster analysis
  - Grouping items based on multiple variables
  - Unknown (latent) real grouping structure
- Discriminant analysis
  - Determining how explanatory variables discriminate between (known) groups
- MANOVA
  - Multivariate model testing