Basic Methods of Data Analysis
Part 2

Sepp Hochreiter

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

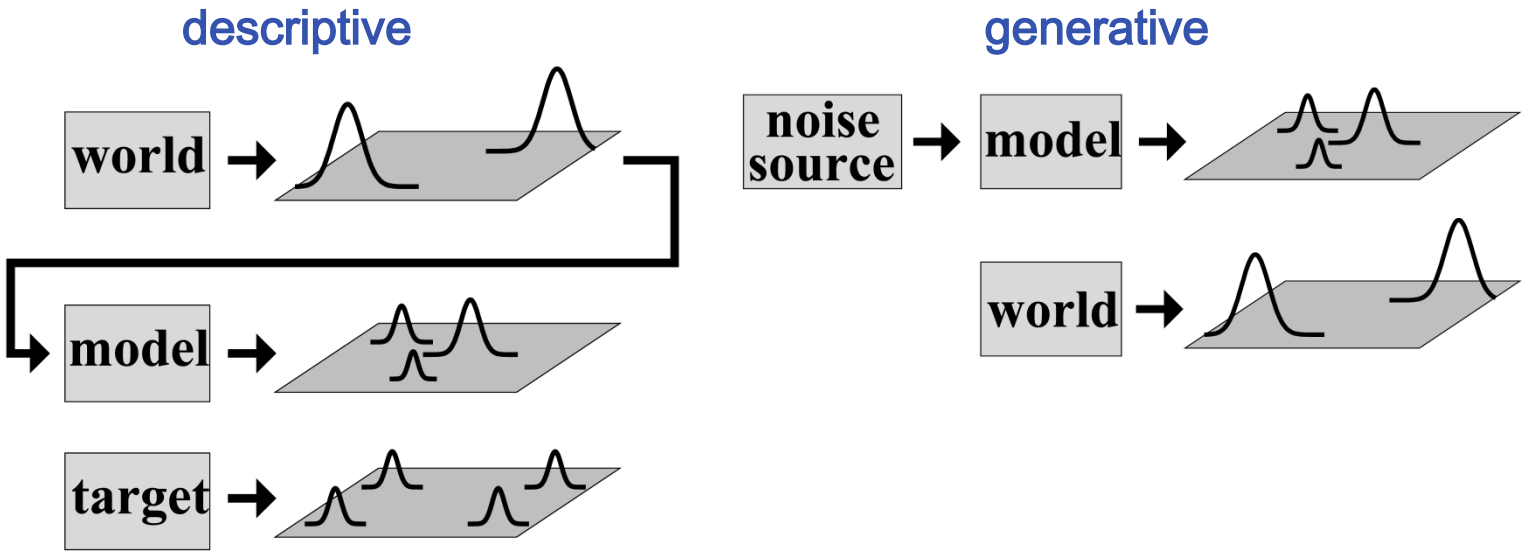
Chapter 4

Summarizing Multivariate Data

Summarizing Multivariate Data

multivariate data: data with more than two variables

Summarizing multivariate data:



- descriptive: model maps observations to another representation
- components with desired density, in a low dimensional space, with high variance, which are non-Gaussian, or statistically independent
- projection methods: principal component analysis, independent component analysis, projection pursuit

Summarizing Multivariate Data

most descriptive methods: map to lower dimensional space

descriptive:

- compact and non-redundant data storage or transmission
- data visualization
- feature selection
- preprocessing methods for subsequent data analysis

generative:

- model or to simulate the real world
- model samples same distribution as the real world observations
- describe the data generation process

Summarizing Multivariate Data

Advantages of generative models:

- determining model parameters like calcium concentration in a cell, rate of reaction, distribution of channels on a membrane, etc.
- generating new simulated observations,
- simulating in unknowns regimes, e.g. new parameters,
- assessing the noise and the signal in the data
- supplying distributions and error bars for latent variables
- detection of outliers as very unlikely observations,
- detection and correction of noise in the observations

- descriptive model with unique inverse → generative framework which selects inverse model, e.g. density estimation
- descriptive without inverse: principal curves, multidimensional scaling

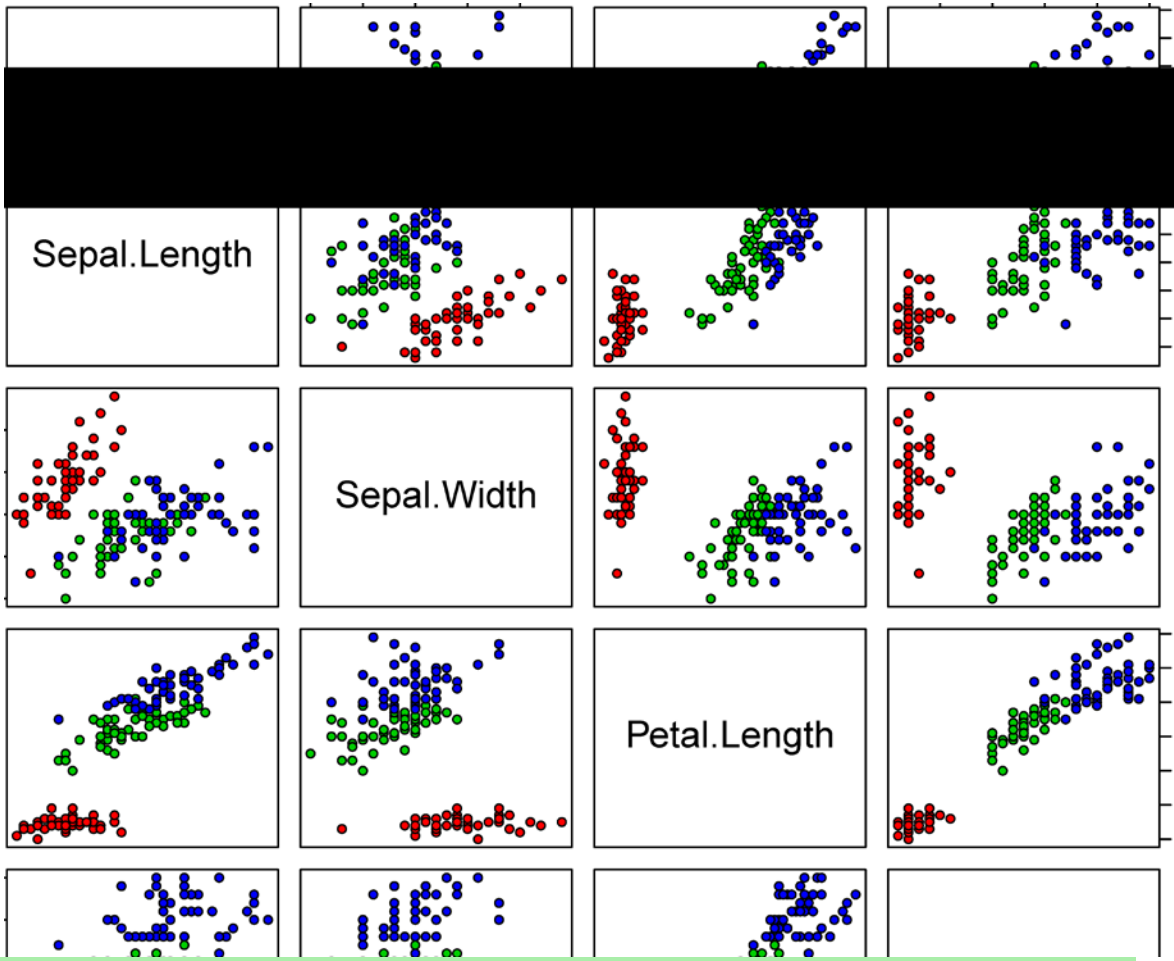
descriptive: principal component analysis, multidimensional scaling
generative: density estimation, factor analysis, independent component analysis, generative topographic mapping

Summarizing Multivariate Data

Idea: apply bivariate data summarization to all pairs of variables
→ increases quadratically with the number of features

Pairs of scatter plots for the iris data:

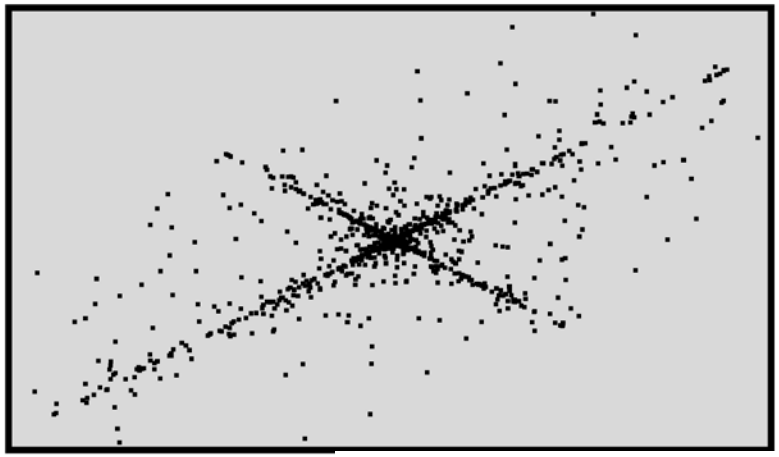
```
pairs(iris[1:4], pch =
21, bg = c("red",
"green3", "blue")
[unclass(iris$Species)])
```

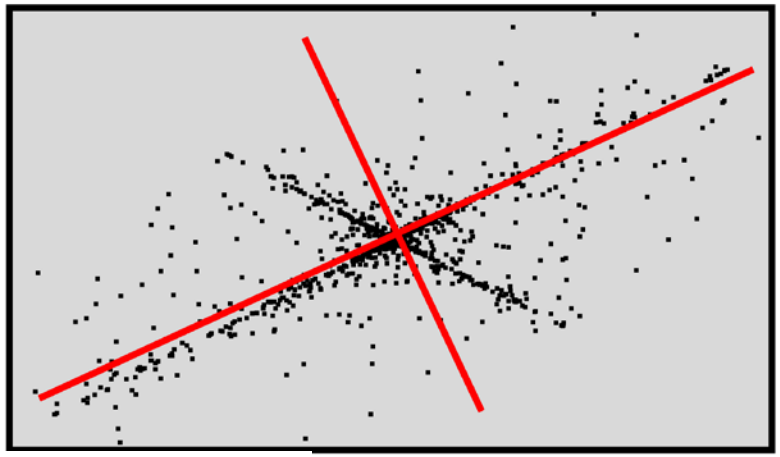


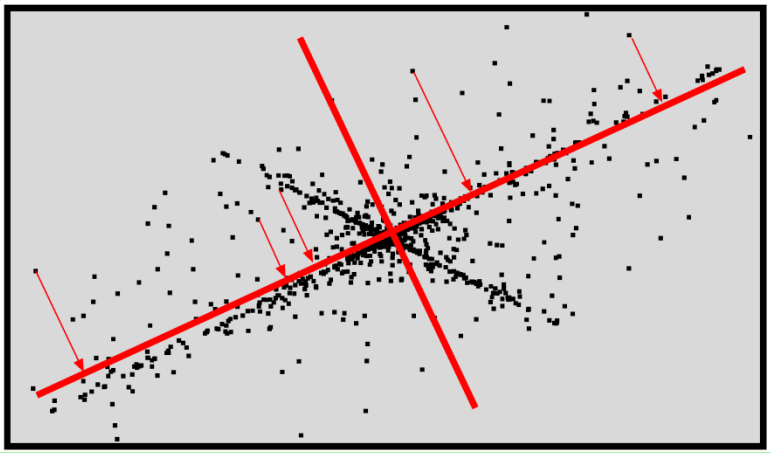
Summarizing Multivariate Data

Principal Component Analysis (PCA), Karhunen-Loéve transform (KTL), Hotelling transform makes a transformation of the coordinate system:

- data has largest variance along the first coordinate
- second largest data variance is along the second coordinate







Summarizing Multivariate Data

summarize multivariate data by PCA via projecting observations onto the first principal components: for visualization the first two

data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ summarized by $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$
data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$

rows of the data matrix contain the observations
columns contain the features

We assume that the features have zero sample mean
(otherwise, the feature mean must be subtracted)

Summarizing Multivariate Data

sample covariance matrix $\boldsymbol{C} \in \mathbb{R}^{m \times m}$ of features across observations is

$$C_{st} = \frac{1}{n} \sum_{i=1}^{n} x_{is}\, x_{it} \text{ , where } x_{is} = (\boldsymbol{x}_i)_s \text{ and } x_{it} = (\boldsymbol{x}_i)_t$$

$$\boldsymbol{C} = \frac{1}{n} \boldsymbol{X}^T \boldsymbol{X} = \frac{1}{n} \boldsymbol{U} \boldsymbol{D}_m \boldsymbol{U}^T$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is orthogonal and $\boldsymbol{D}_m \in \mathbb{R}^{m \times m}$ diagonal

This is the eigendecomposition or spectral decomposition of $\boldsymbol{C}$, which is a symmetric positive definite matrix

diagonal entries of $\boldsymbol{D}_m$: eigenvalues (positive, sorted decreasingly)
column vectors $\boldsymbol{u}_i = [\boldsymbol{U}]_i$: eigenvectors (principal components)

first principal component corresponds to the largest eigenvalue

assume that $n \geq m$ and at least m linear independent observations
→ $\boldsymbol{C}$ has full rang (often ensured by unsupervised feature selection)

Summarizing Multivariate Data

singular value decomposition (SVD) $\boldsymbol{X} = \boldsymbol{V}\,\boldsymbol{D}\,\boldsymbol{U}^T$

where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ are orthogonal, $\boldsymbol{D} \in \mathbb{R}^{n \times m}$ is diagonal with positive entries, the singular values, sorted decreasingly

Computing $\boldsymbol{X}^T\boldsymbol{X}$ we see that $\boldsymbol{D}_m = \boldsymbol{D}^T\,\boldsymbol{D}$ (the eigenvalues are the singular values squared) and $\boldsymbol{U}$ is the orthogonal matrix from PCA.

PCA projection: $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{U} = \boldsymbol{V}\,\boldsymbol{D}$

SVD automatically provides the PCA projections via $\boldsymbol{V}\,\boldsymbol{D}$
For single observations $\boldsymbol{x}$ the projection is $\boldsymbol{y} = \boldsymbol{U}^T\boldsymbol{x}$

PCA is a matrix decomposition problem: $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{U}^T$

where $\boldsymbol{U}$ is orthogonal, $\boldsymbol{Y}^T\,\boldsymbol{Y} = \boldsymbol{D}_m$ (the $\boldsymbol{y}$ are orthogonal, decorrelated), and the eigenvalues $\boldsymbol{D}_m$ are sorted decreasing; for single observations that is $\boldsymbol{x} = \boldsymbol{U}\,\boldsymbol{y}$

Summarizing Multivariate Data

outer product representation:

$$\boldsymbol{X} = \sum_{i=1}^{m} D_{ii}\ \boldsymbol{v}_i \boldsymbol{u}_i^T = \sum_{i=1}^{m} \boldsymbol{y}_i \boldsymbol{u}_i^T$$

$\boldsymbol{u}_i$ is the i-th orthogonal column vector of $\boldsymbol{U}$

$\boldsymbol{v}_i$ is the i-th orthogonal column vector of $\boldsymbol{V}$

$\boldsymbol{y}_i = D_{ii}\ \boldsymbol{v}_i$

Summarizing Multivariate Data

Iterative methods for PCA:

current projection is $t = \boldsymbol{u}^T \boldsymbol{x}$ then Oja's rule is

$$\boldsymbol{u}^{\text{new}} = \boldsymbol{u} + \eta \left(t\, \boldsymbol{x} - t^2\, \boldsymbol{u} \right)$$

where η is the learning rate

The eigenvectors of $\boldsymbol{C}$ are the fixed points of Oja's rule; only the eigenvector with largest eigenvalue is a stable fixed point

$$\begin{aligned} \mathrm{E}_{\boldsymbol{x}}(\boldsymbol{u}^{\text{new}}) &= \boldsymbol{u} + \eta\, \mathrm{E}_{\boldsymbol{x}} \left(\boldsymbol{x}(\boldsymbol{x}^T \boldsymbol{u}) - (\boldsymbol{u}^T \boldsymbol{x})(\boldsymbol{x}^T \boldsymbol{u})\, \boldsymbol{u} \right) = \\ & \boldsymbol{u} + \eta \left(\mathrm{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{u} - \left(\boldsymbol{u}^T E_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{u} \right) \boldsymbol{u} \right) = \\ & \boldsymbol{u} + \eta \left(\boldsymbol{C}\boldsymbol{u} - (\boldsymbol{u}^T \boldsymbol{C}\boldsymbol{u})\, \boldsymbol{u} \right) \end{aligned}$$

If $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{C}$ with eigenvalue λ then

$$\mathrm{E}_{\boldsymbol{x}}(\boldsymbol{u}^{\text{new}}) = \boldsymbol{u} + \eta \left(\lambda \boldsymbol{u} - \lambda\, \boldsymbol{u} \right) = \boldsymbol{u}$$

Summarizing Multivariate Data

The first principal component $\boldsymbol{u}_1$ is the direction of maximum variance:

$$\boldsymbol{u}_1 = \arg\max_{\|\boldsymbol{u}\|=1} \sum_{i=1}^{n} \left(\boldsymbol{u}^T \boldsymbol{x}_i\right)^2 \qquad \sum_{i=1}^{n} \left(\boldsymbol{u}^T \boldsymbol{x}_i\right)^2 = \sum_{i=1}^{n} \left(\boldsymbol{u}^T \boldsymbol{x}_i\right) \left(\boldsymbol{x}_i^T \boldsymbol{u}\right) =$$

$$\boldsymbol{u}^T \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u} = n\, \boldsymbol{u}^T \boldsymbol{C} \boldsymbol{u}$$

$$\boldsymbol{u}^T \boldsymbol{C} \boldsymbol{u} = \sum_{i=1}^{m} \lambda_i a_i^2$$

$$\boldsymbol{C} = \sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

$$\boldsymbol{u} = \sum_{i=1}^{m} a_i \boldsymbol{u}_i \qquad \sum_{i=1}^{m} a_i^2 = 1$$

This sum is maximal for $a_1 = 1, a_i = 0, i \neq 1$ because $\lambda_1 > \lambda_i > 0$

principal components are the direction of maximal variance orthogonal to all previous components:

$$\boldsymbol{x}_i^k = \boldsymbol{x}_i - \sum_{t=1}^{k-1} \left(\boldsymbol{u}_t^T \boldsymbol{x}_i\right) \boldsymbol{u}_t \qquad \boldsymbol{u}_k = \arg\max_{\|\boldsymbol{u}\|=1} \sum_{i=1}^{n} \left(\boldsymbol{u}^T \boldsymbol{x}_i^k\right)^2$$

inductively been proved analog to the first principal component

first l components span l-dimensional space of maximal variance

Summarizing Multivariate Data

Is there only one PCA solution? $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{U}^T$

$\boldsymbol{U}$ is orthogonal, $\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{D}_m$, $\boldsymbol{D}_m$ is diagonal with sorted values

PCA is unique up to signs, if the eigenvalues of the covariance matrix are different from each other (proof: see manuscript).

At most one eigenvalue can be zero, which can be removed.

Summarizing Multivariate Data

- first l principal components span l-dim. space of maximal variance

$$\sum_{i=1}^{l} \boldsymbol{u}_i^T \, \boldsymbol{C} \, \boldsymbol{u}_i \;\; \text{s.t.} \;\; \boldsymbol{u}_i^T \, \boldsymbol{u}_j = \delta_{ij}$$

- projections onto PCs have zero means:

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_k^T \, \boldsymbol{x}_i \;=\; \boldsymbol{u}_k^T \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \right) \;=\; \boldsymbol{u}_k^T \boldsymbol{0} \;=\; 0$$

- projections onto PCs are mutually uncorrelated (orthogonal):

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{u}_t^T \, \boldsymbol{x}_i) \, (\boldsymbol{u}_s^T \, \boldsymbol{x}_i) \;&=\; \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{u}_t^T \, \boldsymbol{x}_i) \, (\boldsymbol{x}_i^T \, \boldsymbol{u}_s) \\
&=\; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_t^T \, (\boldsymbol{x}_i \, \boldsymbol{x}_i^T) \, \boldsymbol{u}_s \\
&=\; \boldsymbol{u}_t^T \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right) \boldsymbol{u}_s \\
&=\; \boldsymbol{u}_t^T \, \boldsymbol{C} \, \boldsymbol{u}_s \;=\; \lambda_s \, \boldsymbol{u}_t^T \, \boldsymbol{u}_s \;=\; 0
\end{aligned}$$

Summarizing Multivariate Data

- the sample variance of the k-th projection is equal to the k-th eigenvalue of the sample covariance matrix:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{u}_k^T\boldsymbol{x}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_k^T\left(\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{u}_k = \boldsymbol{u}_k^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\,\boldsymbol{x}_i^T\right)\boldsymbol{u}_k = \boldsymbol{u}_k^T\boldsymbol{C}\boldsymbol{u}_k = \lambda_k\,\boldsymbol{u}_k^T\boldsymbol{u}_k = \lambda_k$$

- PCs are ranked decreasingly according to their eigenvalues

- The first l PCs minimize the mean-squared error: $\hat{\boldsymbol{x}} = \sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\boldsymbol{x}$
  mean-squared error is

$$\mathrm{E}\left(\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2\right) = \mathrm{E}\left(\boldsymbol{x}^T\boldsymbol{x} - 2\,\boldsymbol{x}^T\,\hat{\boldsymbol{x}} + \hat{\boldsymbol{x}}^T\hat{\boldsymbol{x}}\right)$$

$$= \mathrm{E}\left(\mathrm{Tr}\left(\boldsymbol{x}\boldsymbol{x}^T\right) - 2\,\mathrm{Tr}\left(\sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\boldsymbol{x}\boldsymbol{x}^T\right) + \mathrm{Tr}\left(\sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\boldsymbol{x}\boldsymbol{x}^T\right)\right)$$

$$= \mathrm{Tr}\left(\mathrm{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right) - 2\sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\mathrm{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right) + \sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\mathrm{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right) = \mathrm{Tr}\left(\boldsymbol{C} - \sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\boldsymbol{C}\right)$$

$$= \mathrm{Tr}\left(\boldsymbol{C} - \sum_{k=1}^{l}\boldsymbol{u}_k\,\boldsymbol{u}_k^T\sum_{k=1}^{m}\lambda_k\boldsymbol{u}_k\,\boldsymbol{u}_k^T\right) = \mathrm{Tr}\left(\sum_{k=1}^{m}\lambda_k\,\boldsymbol{u}_k\,\boldsymbol{u}_k^T - \sum_{k=1}^{l}\lambda_k\boldsymbol{u}_k\,\boldsymbol{u}_k^T\right)$$

$$= \mathrm{Tr}\left(\sum_{k=l+1}^{m}\lambda_k\,\boldsymbol{u}_k\,\boldsymbol{u}_k^T\right) = \sum_{k=l+1}^{m}\lambda_k\,\mathrm{Tr}\left(\boldsymbol{u}_k\,\boldsymbol{u}_k^T\right) = \sum_{k=l+1}^{m}\lambda_k\,\mathrm{Tr}\left(\boldsymbol{u}_k^T\,\boldsymbol{u}_k\right) = \sum_{k=l+1}^{m}\lambda_k$$

Summarizing Multivariate Data

Iris Data Set

```
xp <- princomp(iris[,1:4],scores=TRUE)
summary(xp)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
```

the first principal component explains 92% of the variance in the data
→ features are correlated which is captured by PC1

projections:

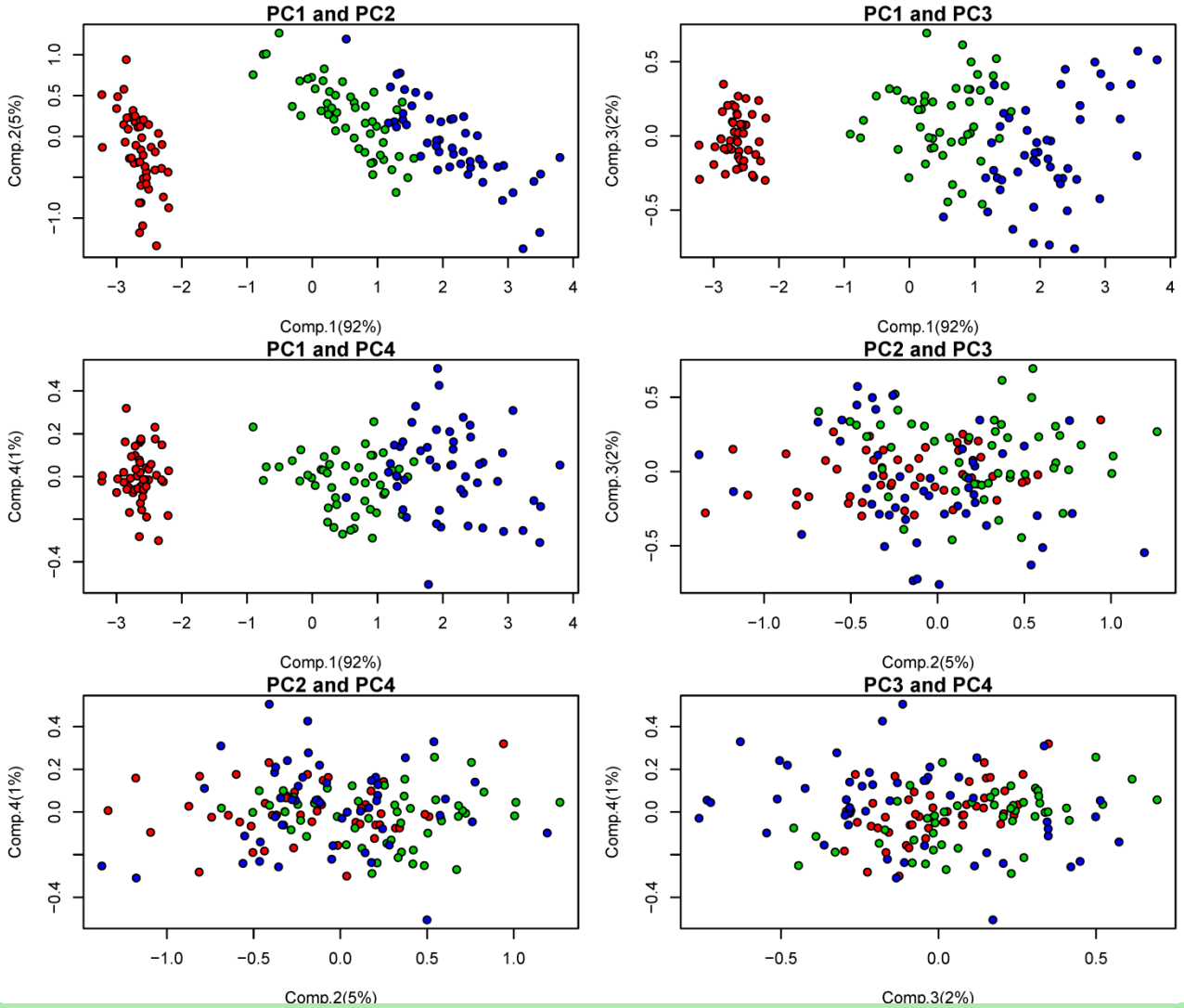
```
irisPC <- xp$scores
```

OR

```
irisPC <- sweep(as.matrix(iris[,1:4]),2,xp$center)%*%xp$loadings
```

Summarizing Multivariate Data

Only PC1 helps to separate the iris species:



Summarizing Multivariate Data

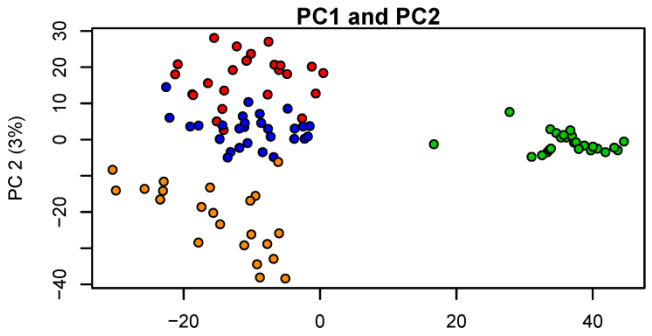
Multiple Tissue Data Set

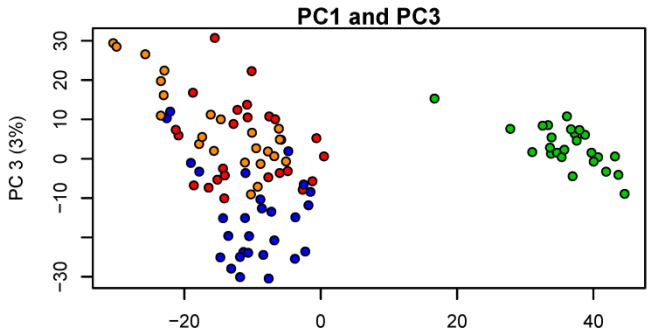
- gene expression values microarray
- human and mouse
- 102 samples
- 5,565 genes
- different tissue types
  - breast (Br)
  - prostate (Pr)
  - lung (Lu)
  - colon (Co)

PCA via a singular value decomposition (SVD):

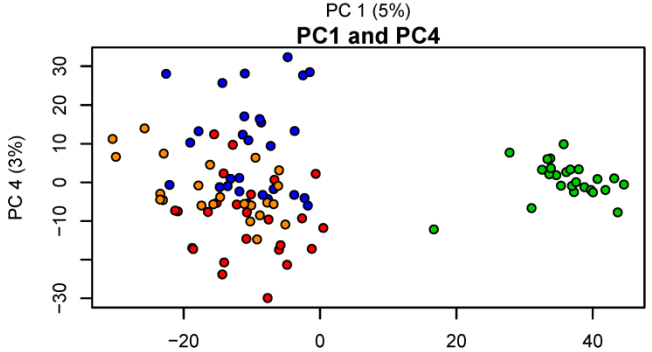
```
sv <- svd(t(XMulti))
PCS <- sv$u%*%diag(sv$d)
```

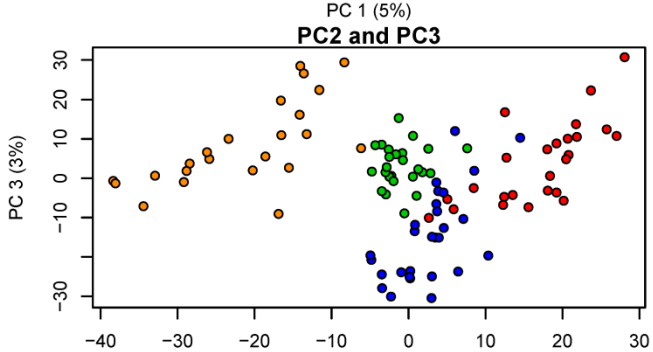
Summarizing Multivariate Data



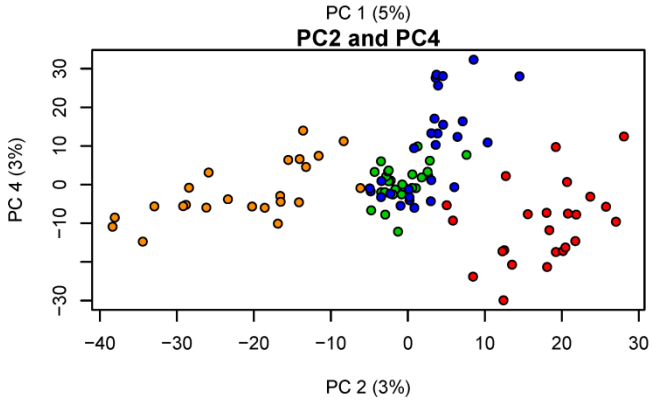


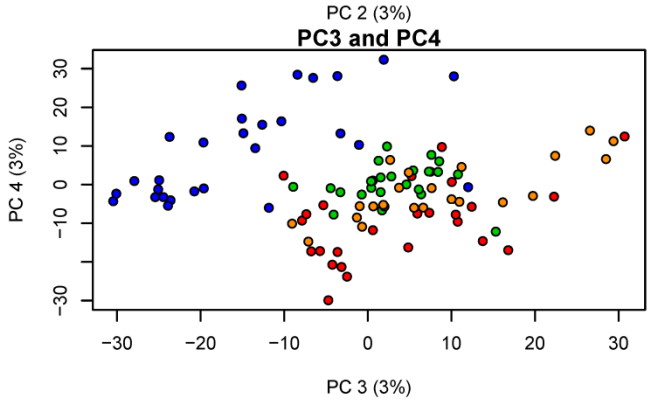
PC1 separates the prostate samples (green) from the rest.





PC2 separates the colon samples (orange) but also breast samples (red).





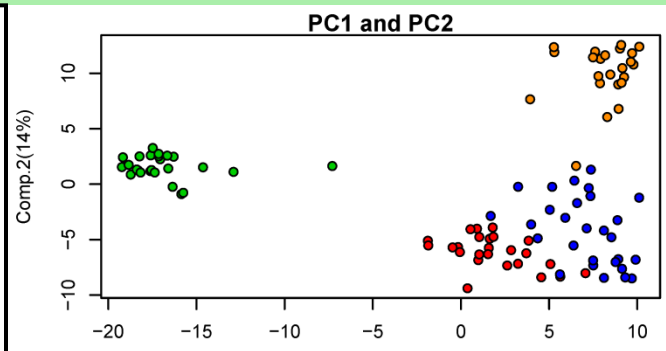
PC3 separates some lung samples (blue).

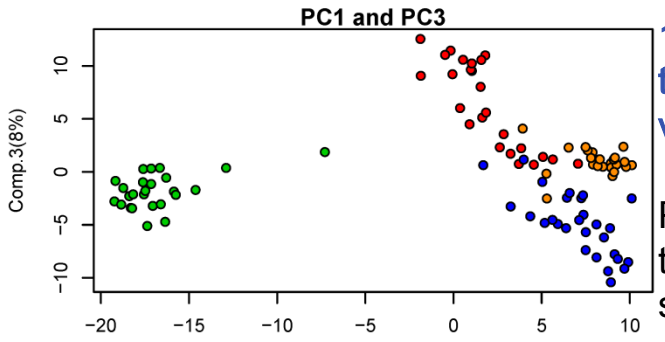
Summarizing Multivariate Data

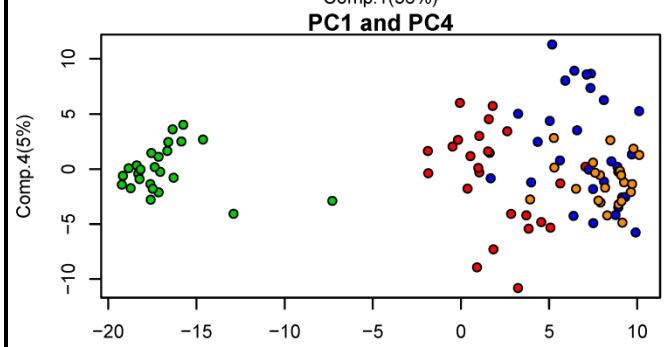
variance filtering before PCA: justified for microarray data

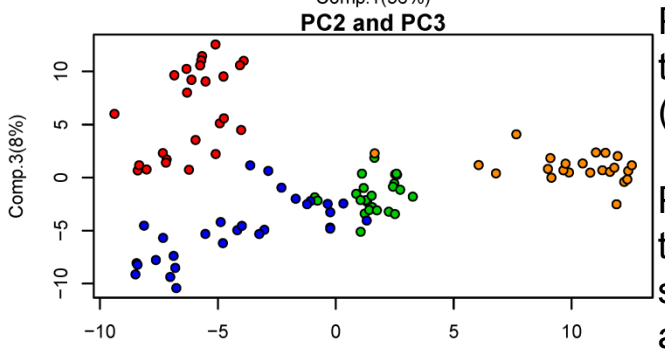
```
vv <- diag(var(t(XMulti)))
length(which(vv>2))
 101
length(which(vv>4))
 13
length(which(vv>5))
 5
XMultiF1 <- t(XMulti[which(vv>2),])   # 101
XMultiF2 <- t(XMulti[which(vv>4),])   # 13
XMultiF3 <- t(XMulti[which(vv>5),])   # 5
```

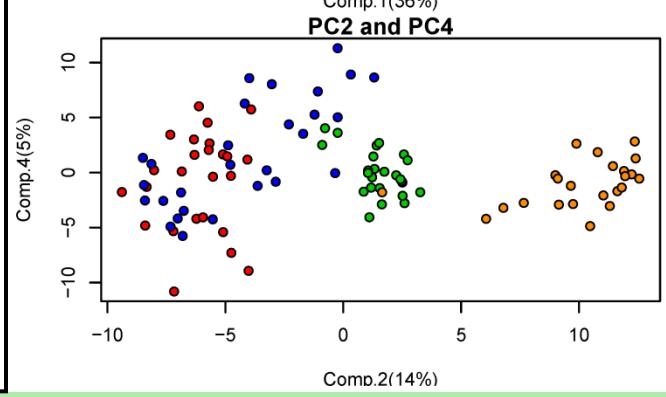
Summarizing Multivariate Data

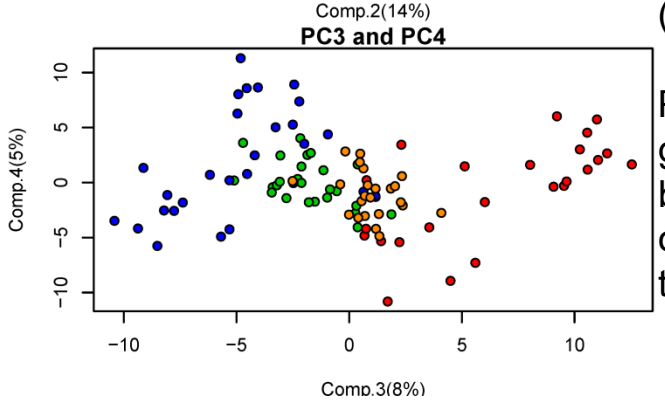












101 genes with the highest variance

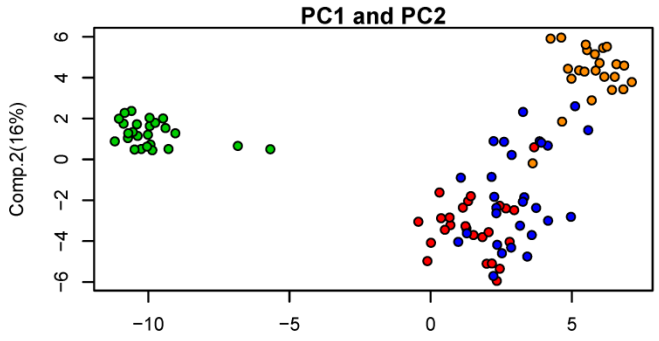
PC1 separates the prostate samples (green)

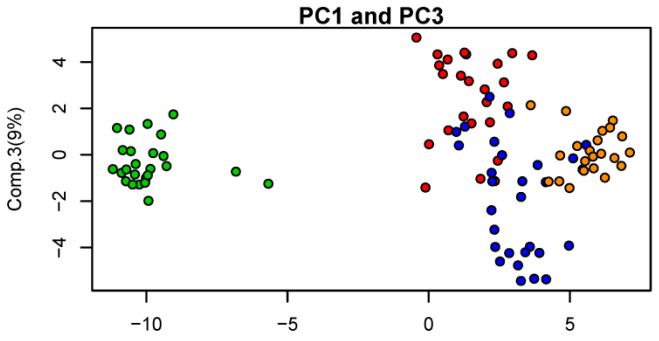
PC2 separates the colon samp. (orange)

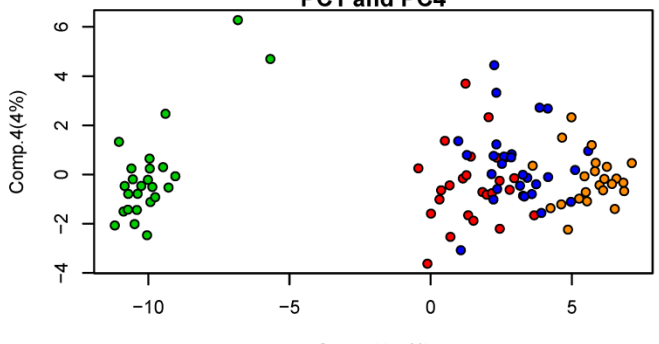
PC3 separates the breast samples (red) and lung samp. (blue)

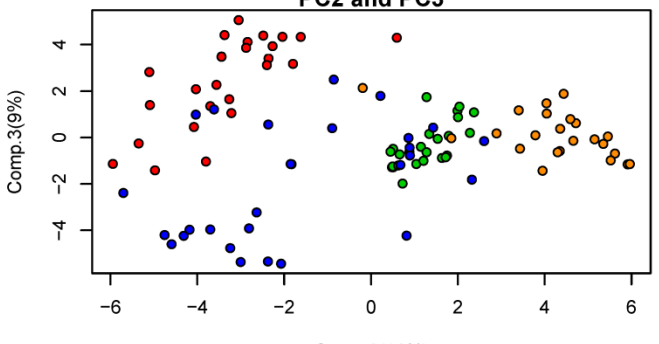
PCA on filtered genes performs better than PCA on all genes for tissue separat.

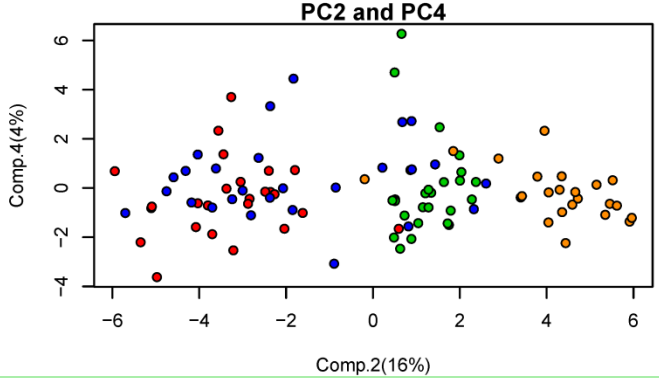
Summarizing Multivariate Data

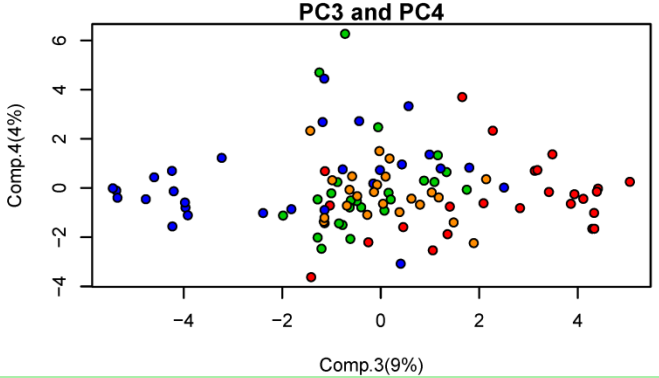












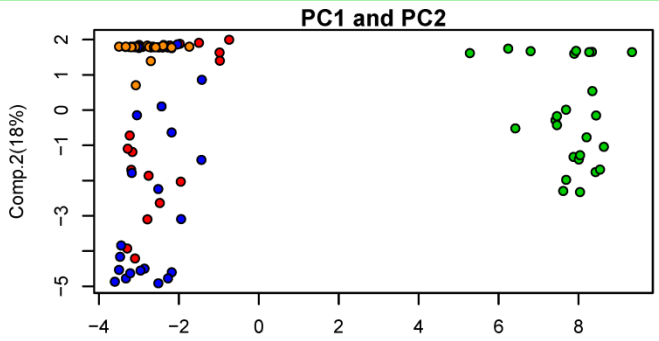
13 genes with largest variance

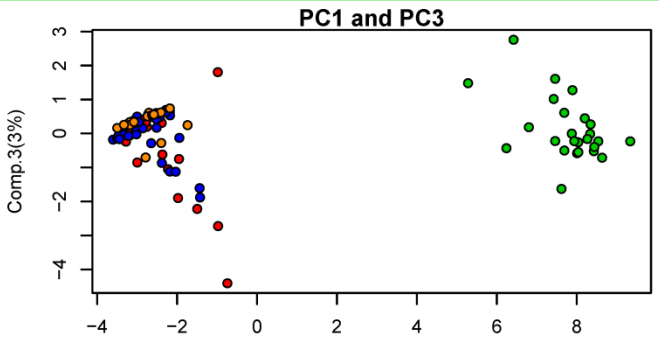
PC1 separates the prostate samples (green)

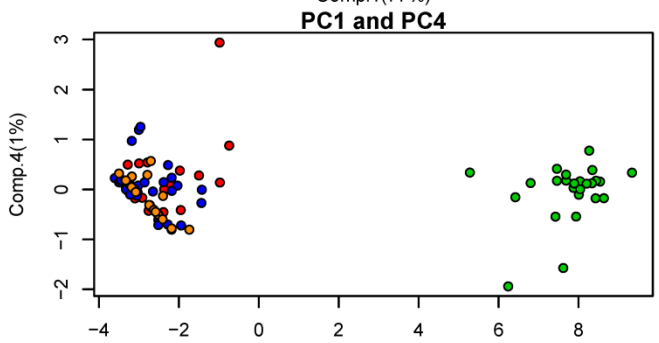
PC2 separates the colon samples (orange)

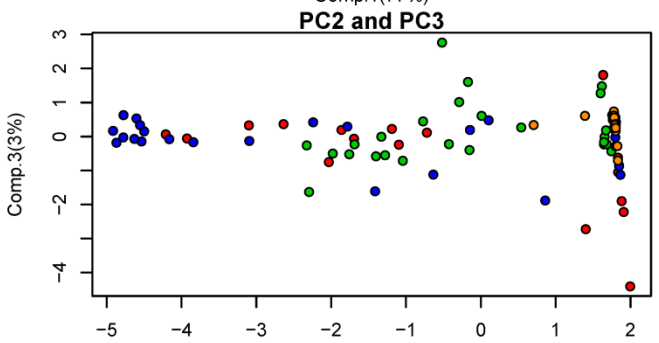
PC3 separates the breast (red) and lung (blue) samples

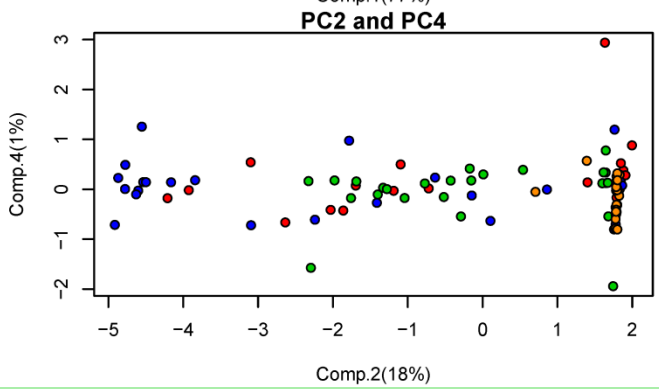
Summarizing Multivariate Data

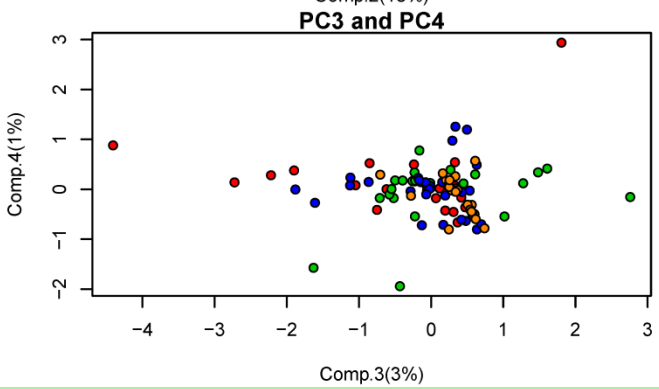












5 genes with largest variance

Still PC1 separates the prostate samples (green)

However other tissues are difficult to separate

Summarizing Multivariate Data

4 out of 5 genes are highly correlated:

```
cor(XMultiF3)
```

	ACPP	KLK2	KRT5	MSMB	TRGC2
ACPP	1.000000000	0.97567890	-0.004106762	0.90707887	0.947433227
KLK2	0.975678903	1.00000000	-0.029900946	0.89265825	0.951841913
KRT5	-0.004106762	-0.02990095	1.000000000	-0.05565599	0.008877815
MSMB	0.907078869	0.89265825	-0.055655985	1.00000000	0.870922667
TRGC2	0.947433227	0.95184191	0.008877815	0.87092267	1.000000000

GeneCards database:

- ACPP “is synthesized under androgen regulation and is secreted by the epithelial cells of the prostate gland”
- KLK2 “is primarily expressed in prostatic tissue and is responsible for cleaving pro-prostate-specific antigen into its enzymatically active form” (KLK3 is the PSA gene)
- MSMB “is synthesized by the epithelial cells of the prostate gland and secreted into the seminal plasma”

Summarizing Multivariate Data

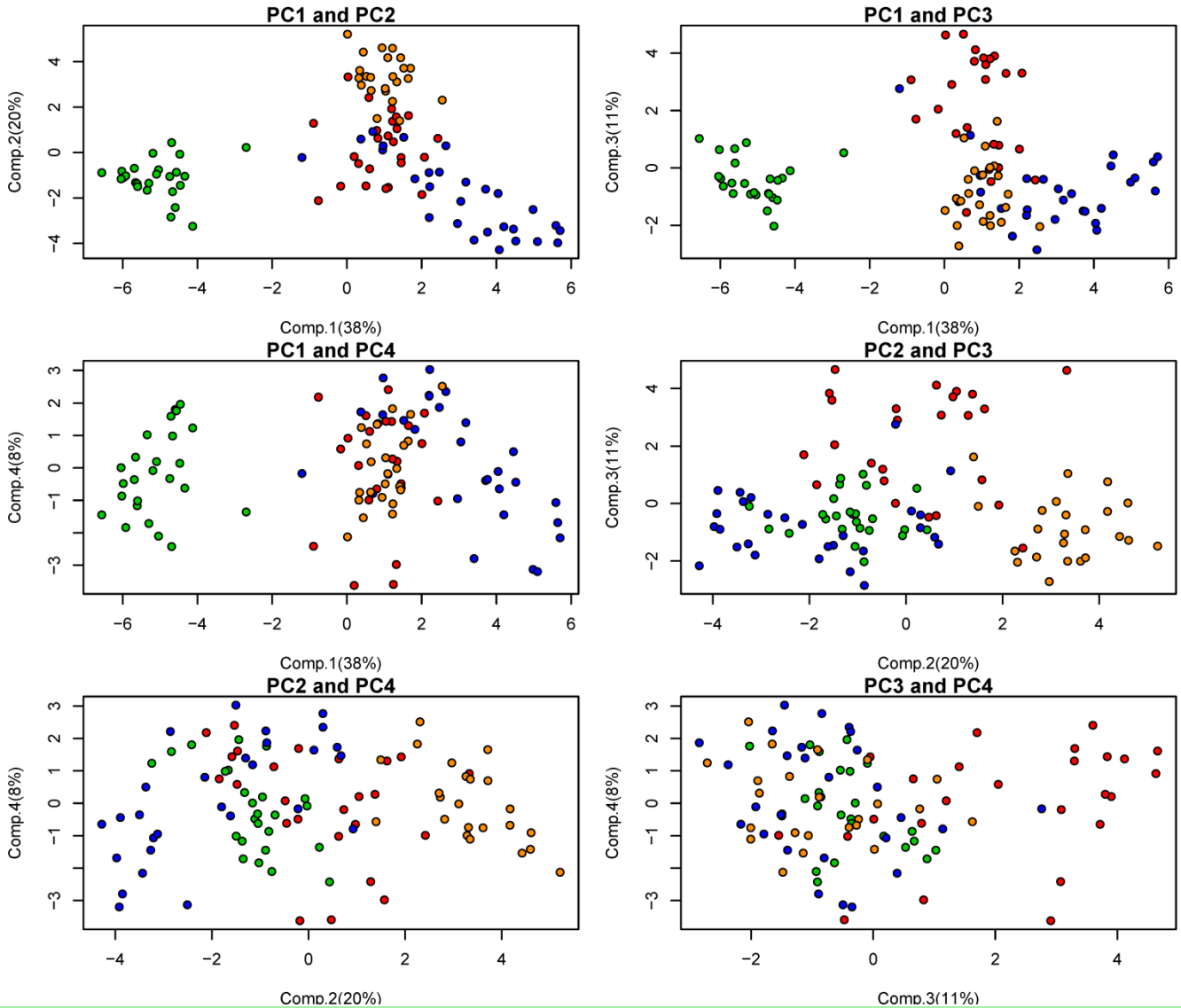
genes which are not correlated to each other → clustering & prototype

```
hc1 <-hclust(dist(t(XMultiF1)))
ct <- cutree(hc1,h=25)
table(ct)
ct
 1  2  3  4  5  6  7  8  9 10
21 14 12 21  6  4  2  9  3  9

l1 <- length(table(ct))
sel <- c()
for(i in 1:l1) {
+ sel <- c(sel,which(ct==i)[1])
+ }
XMultiF4 <- XMultiF1[,sel]


cor(XMultiF4)
              ABP1       ACPP      AKR1C1     ALDH1A3      ANXA8          APOD
ABP1     1.00000000 -0.1947766 -0.04224634 -0.21577195 -0.2618053 -0.3791812658
ACPP    -0.19477662  1.0000000 -0.22929893  0.88190657 -0.2978638  0.4964638048
AKR1C1  -0.04224634 -0.2292989  1.00000000 -0.07536066  0.4697886 -0.1793466620
ALDH1A3 -0.21577195  0.8819066 -0.07536066  1.00000000 -0.1727669  0.4113925823
ANXA8   -0.26180526 -0.2978638  0.46978864 -0.17276688  1.0000000 -0.1863923785
APOD    -0.37918127  0.4964638 -0.17934666  0.41139258 -0.1863924  1.0000000000
```

Summarizing Multivariate Data



10 uncorrelated genes

tissues are not as well separated as with maximal variance

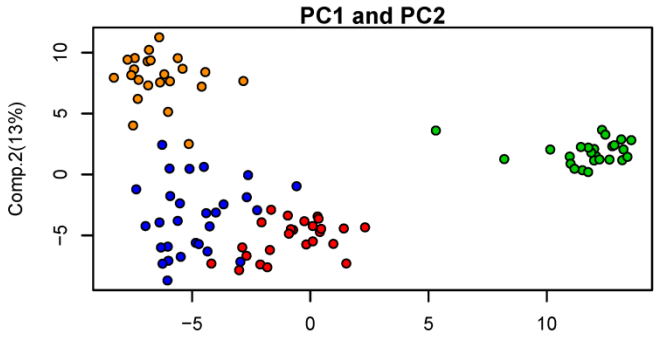
→ highly variable genes missed

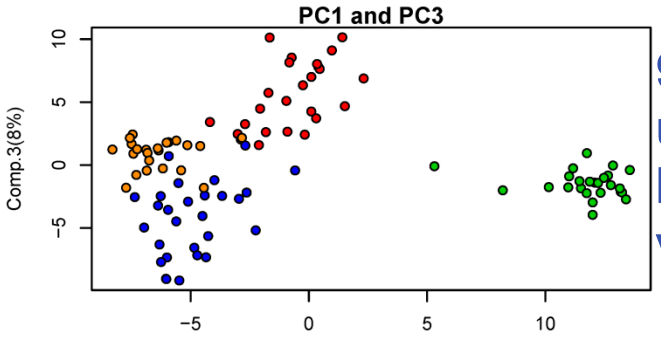
Summarizing Multivariate Data

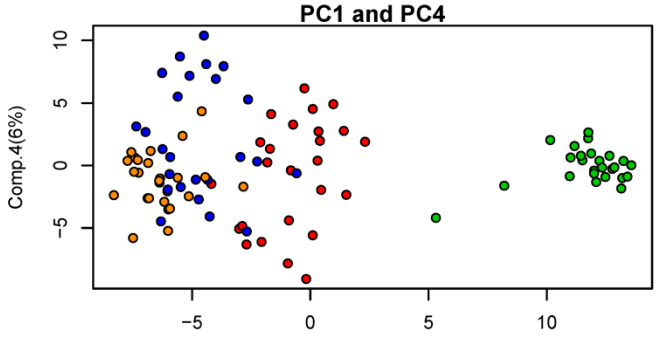
hierarchical clustering and variance maximization within one cluster:

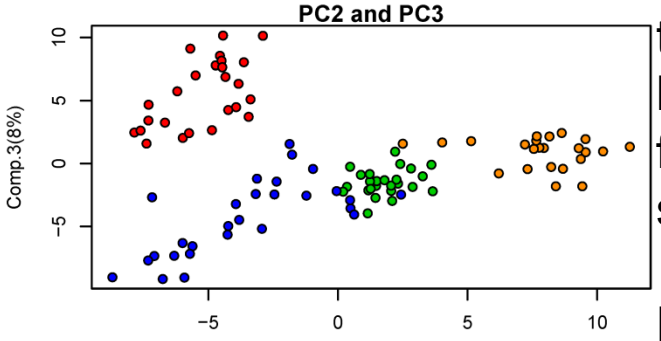
```
hc1 <-hclust(dist(XMulti))
ct <- cutree(hc1,h=16)
table(ct)
ct
  1   2    3   4   5   6   7   8  9 10 11 12 13 14 15 16 17 18
682 126 1631 742 347 797 196 104 44 35  5  8 12  5 12 14  5 71
 19  20   21  22  23  24  25  26 27 28 29 30 31 32 33 34 35 36
 22   8   16  32  48  72   2  93 22 22 56  9 54  7  4  2 16 26
 37  38   39  40  41  42  43  44 45 46 47 48 49 50 51 52 53 54
  3   8   42   1   9   1   7  14  1  2  8  3  2 20  3  2  9  7
 55  56   57  58  59  60  61  62 63 64 65 66 67 68 69 70 71 72
  3   2    1   5   2   2   1   1  1  3  9  3  3  3  3  1  2  3
 73  74   75  76  77  78  79  80 81 82 83 84 85 86 87 88 89 90
  1   1    1   1   2   2   1   3  1  2  1  1  2  1  2  2  1  1
 91  92
  1   1
l1 <- length(table(ct))
sel <- c()
for(i in 1:l1) {
clas <- which(ct==i)
M <- which.max(diag(var(t(XMulti[clas,]))))
sel <- c(sel,clas[M]) }
XMultiF5 <- t(XMulti[sel,])
```

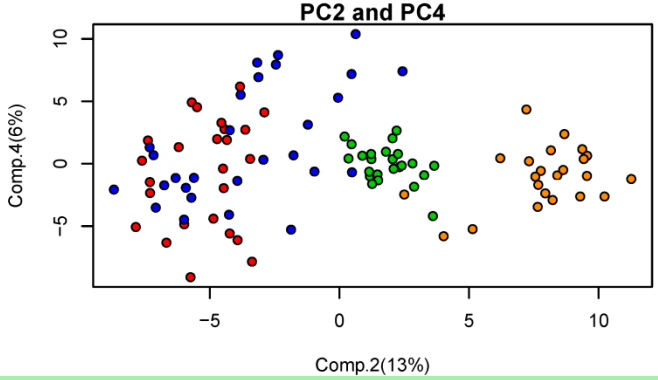
Summarizing Multivariate Data

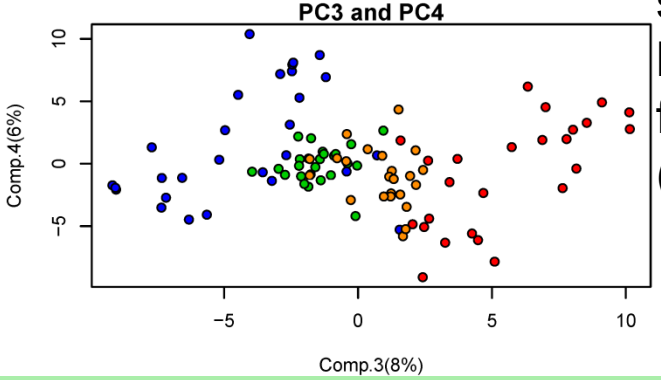












92 genes uncorrelated but maximal variance

very similar to variance based feature selection

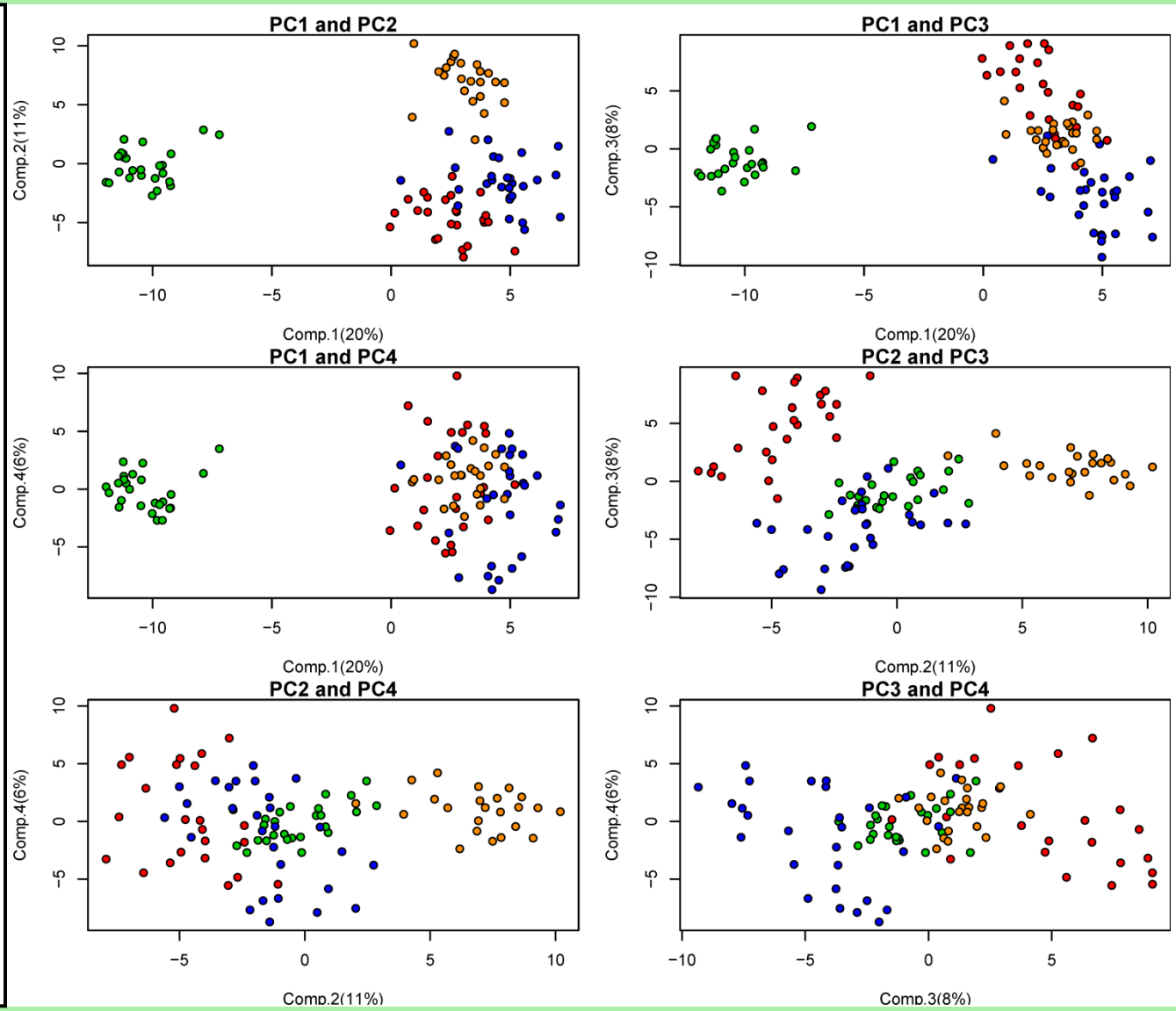
PC3 separates breast (red) from lung (blue) samp.

Summarizing Multivariate Data

Correlation as distance measure for clustering

```
# Genes 2964 and 4663 are constant !!
# First remove these genes
XMultiC <-  XMulti[-c(2964,4663),]
D <- 1 - abs(cor(t(XMultiC)))
D <- as.dist(D)
hc1 <-hclust(D)
ct <- cutree(hc1,h=0.999)
l1 <- length(table(ct))
sel <- c()
for(i in 1:l1) {
clas <- which(ct==i)
M <- which.max(diag(var(t(XMultiC[clas,]))))
sel <- c(sel,clas[M])
}
XMultiF7 <- t(XMultiC[sel,])
```

Summarizing Multivariate Data



clustering with correlation coefficient.
95 clusters: gene with maximal variance

very similar to variance based feature selection

PC3 separates breast (red) from lung (blue) samp.

Summarizing Multivariate Data

Clustering is one of the most popular unsupervised techniques

Clusters in the data are regions where observations group together
→ regions of high data density

clusters may correspond to a prototype from which observations are obtained via noise perturbations

Clustering extracts structures and can identify new data classes

important application of clustering: data visualization

observations are represented by prototypes: vector quantization

Summarizing Multivariate Data

k-means clustering is probably the best known clustering algorithm

Only parameters are the cluster centers.

A sample x_i belongs to the cluster with the closest center:

$$c_{\boldsymbol{x}_i} = \arg\min_k \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|$$

Center updates is mean of its members:

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{1}{n_j} \sum_{i=1,\ j=c_{\boldsymbol{x}_i}}^{n} \boldsymbol{x}_i$$

$$n_j = \sum_{i=1,\ j=c_{\boldsymbol{x}_i}}^{n} 1$$

Summarizing Multivariate Data

k-means algorithm

Given: data $\{\boldsymbol{x}\} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, number of clusters l

BEGIN initialization
initialize the cluster centers $\boldsymbol{\mu}_j$, $1 \leq j \leq l$
END initialization

BEGIN Iteration

Stop=false
while Stop=false **do**
for $(i = 1\ ;\ i\ \geq\ n\ ;\ i++)$ **do**
assign $\boldsymbol{x}_i$ to the nearest $\boldsymbol{\mu}_j$
end for
for $(j = 1\ ;\ j\ \geq\ l\ ;\ j++)$ **do**

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{1}{n_j} \sum_{i=1,\ j=c_{\boldsymbol{x}_i}}^{n} \boldsymbol{x}_i$$

end for
if stop criterion fulfilled **then**
Stop=true
end if
end while
END Iteration

Summarizing Multivariate Data

k-means clustering

- fast
- robust (outliers)
- simple (advantage or disadvantage)
- prone to initialization

center near some outliers → center will stay on the outliers even if some cluster are not modeled

Summarizing Multivariate Data

membership continuous:
(softmax)

$$w_j(\boldsymbol{x}_i) = \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^{-2/(b-1)}}{\sum_{k=1}^{l} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^{-2/(b-1)}}$$

update rule

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{\sum_{i=1}^{n} w_j(\boldsymbol{x}_i)\ \boldsymbol{x}_i}{\sum_{i=1}^{n} w_j(\boldsymbol{x}_i)}$$

This algorithm is called fuzzy k-means clustering

Summarizing Multivariate Data

Given: data $\{\boldsymbol{x}\} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, number of clusters l, parameter b

BEGIN initialization
initialize the cluster centers $\boldsymbol{\mu}_j$, $1 \leq j \leq l$, and
$w_j(\boldsymbol{x}_i)$ so that $\sum_{j=1}^{l} w_j(\boldsymbol{x}_i) = 1$, $w_j(\boldsymbol{x}_i) \geq 0$.
END initialization

BEGIN Iteration

Stop=false
while Stop=false **do**

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{\sum_{i=1}^{n} w_j(\boldsymbol{x}_i)\, \boldsymbol{x}_i}{\sum_{i=1}^{n} w_j(\boldsymbol{x}_i)}$$

$$w_j(\boldsymbol{x}_i) = \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^{-2/(b-1)}}{\sum_{k=1}^{l} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^{-2/(b-1)}}$$

if stop criterion fulfilled **then**
Stop=true
end if
end while
END Iteration

fuzzy k-means algorithm

Summarizing Multivariate Data

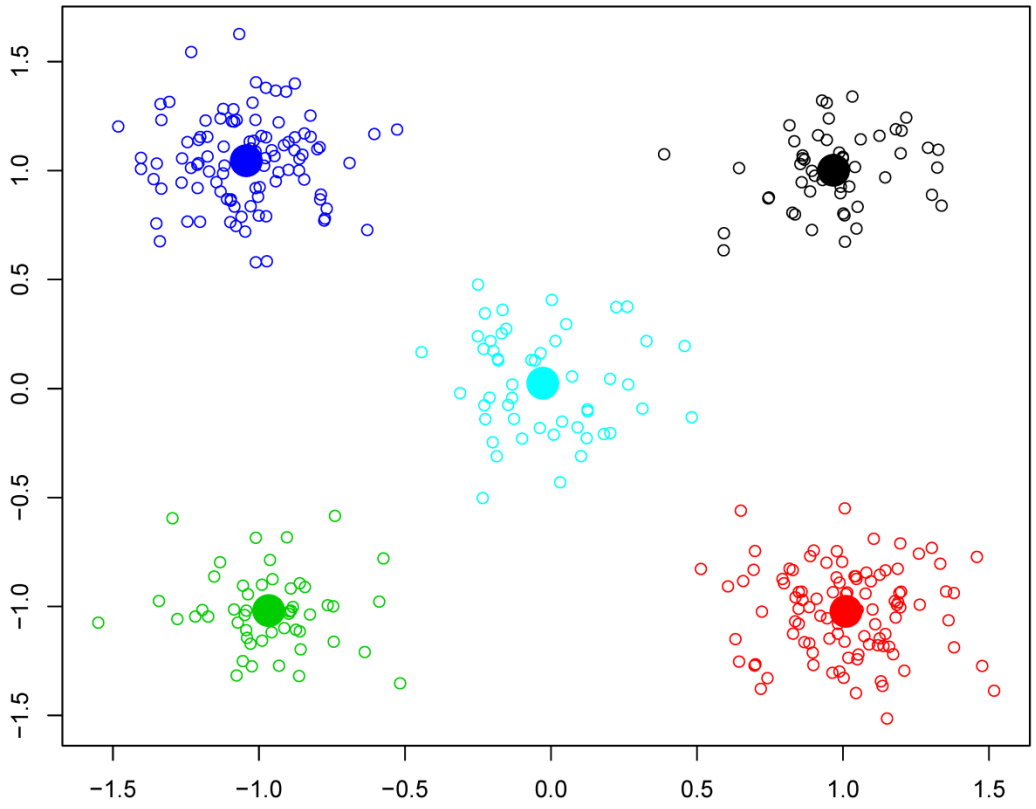
artificial data set in two dimensions with five clusters:

```
x <- rbind( matrix(rnorm(100, sd=0.2), ncol=2),
matrix(rnorm(100, mean=1, sd=0.2), ncol=2),
matrix(rnorm(100, mean=-1, sd=0.2), ncol=2),
matrix(c(rnorm(100, mean=1, sd=0.2),rnorm(100, mean=-1, sd=0.2)), ncol = 2),
matrix(c(rnorm(100, mean=-1, sd=0.2),rnorm(100, mean=1, sd=0.2)), ncol = 2))
colnames(x) <- c("x", "y")
kmeans(x, 5)
```

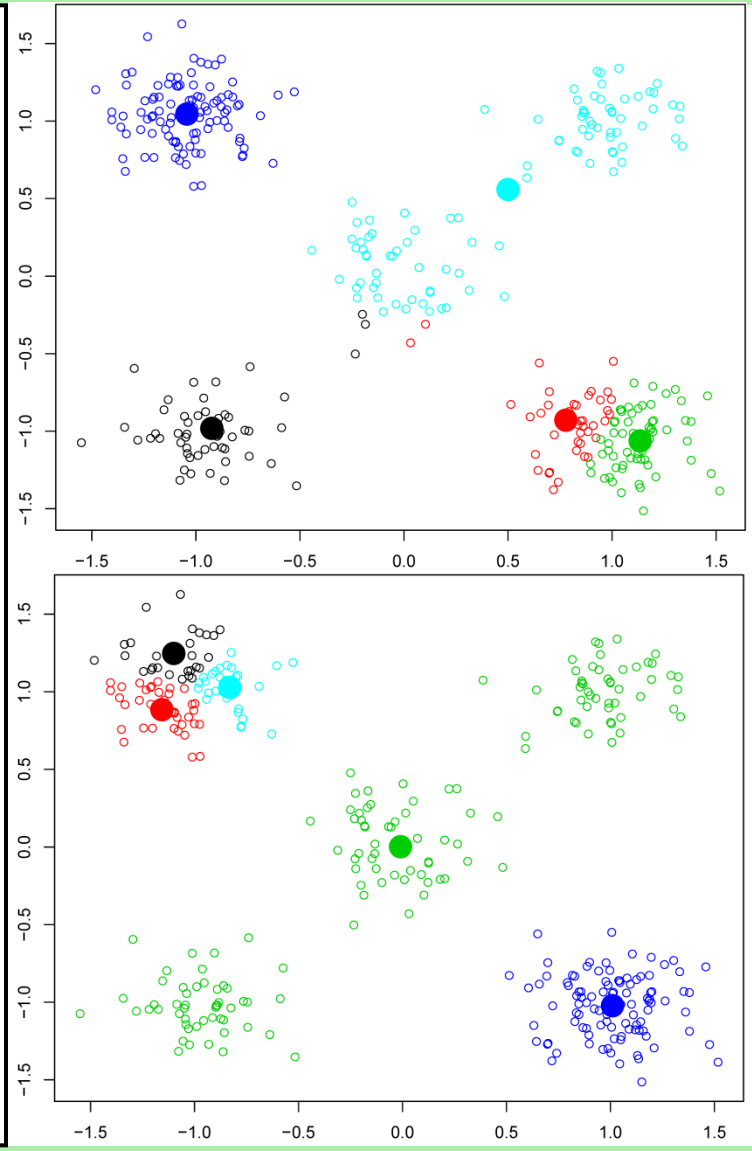
optimal solution with k=5

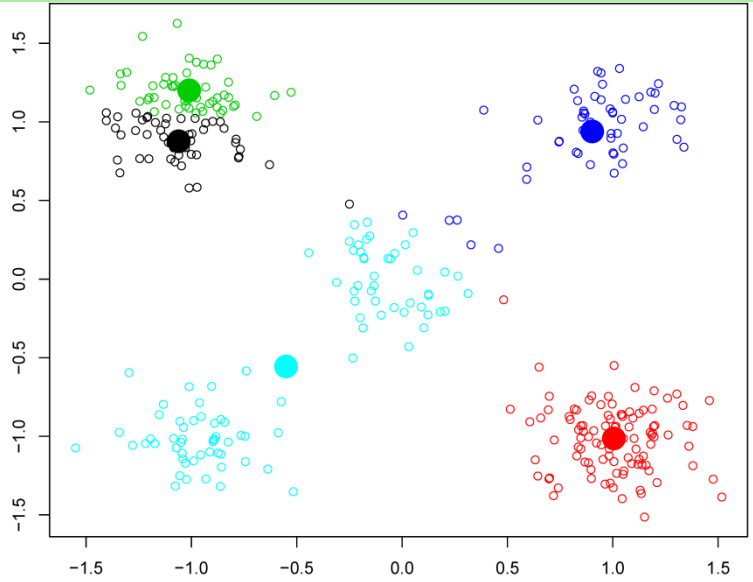
color indicate cluster membership

filled circles mark the cluster centers



Summarizing Multivariate Data

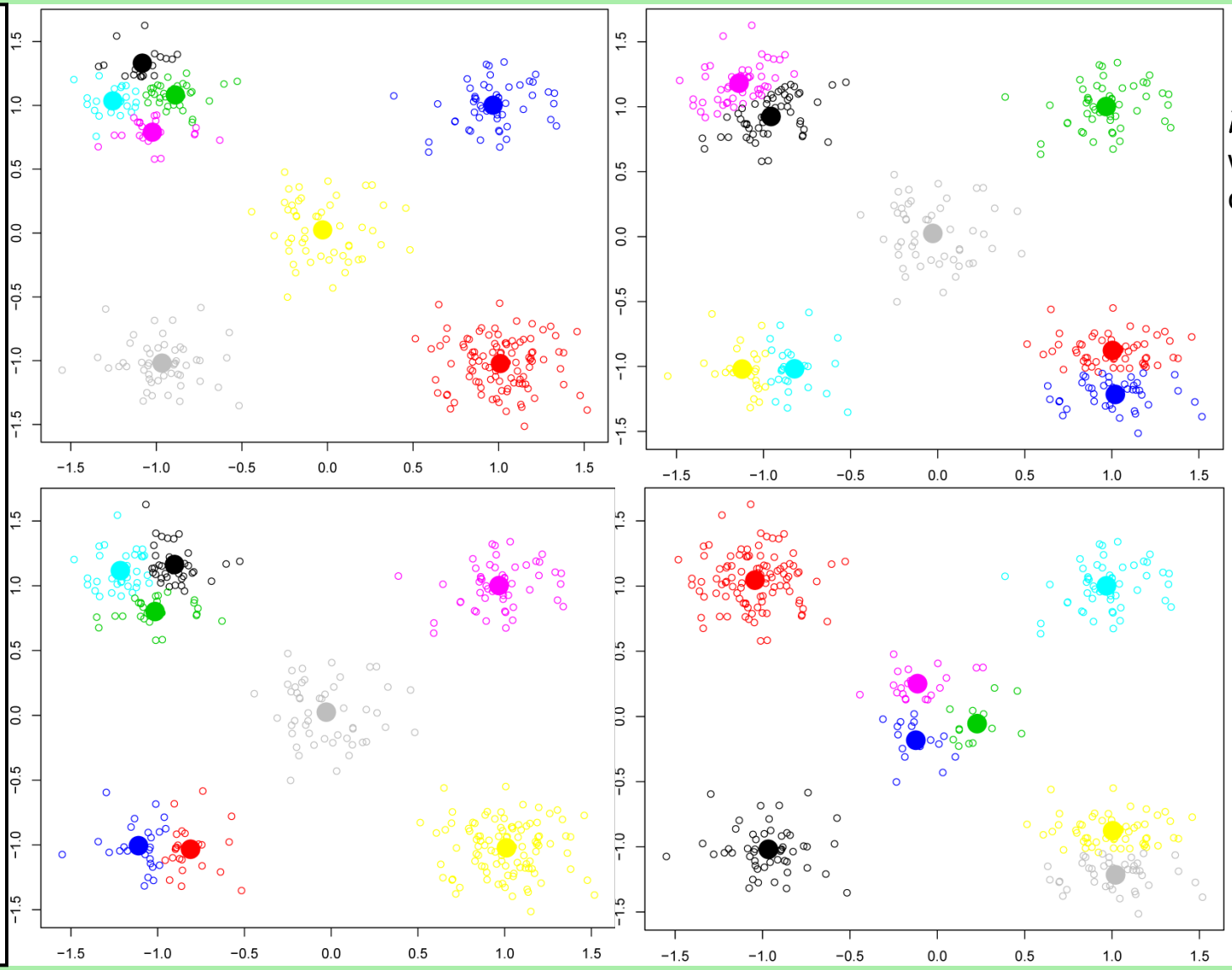




Local minima are shown:

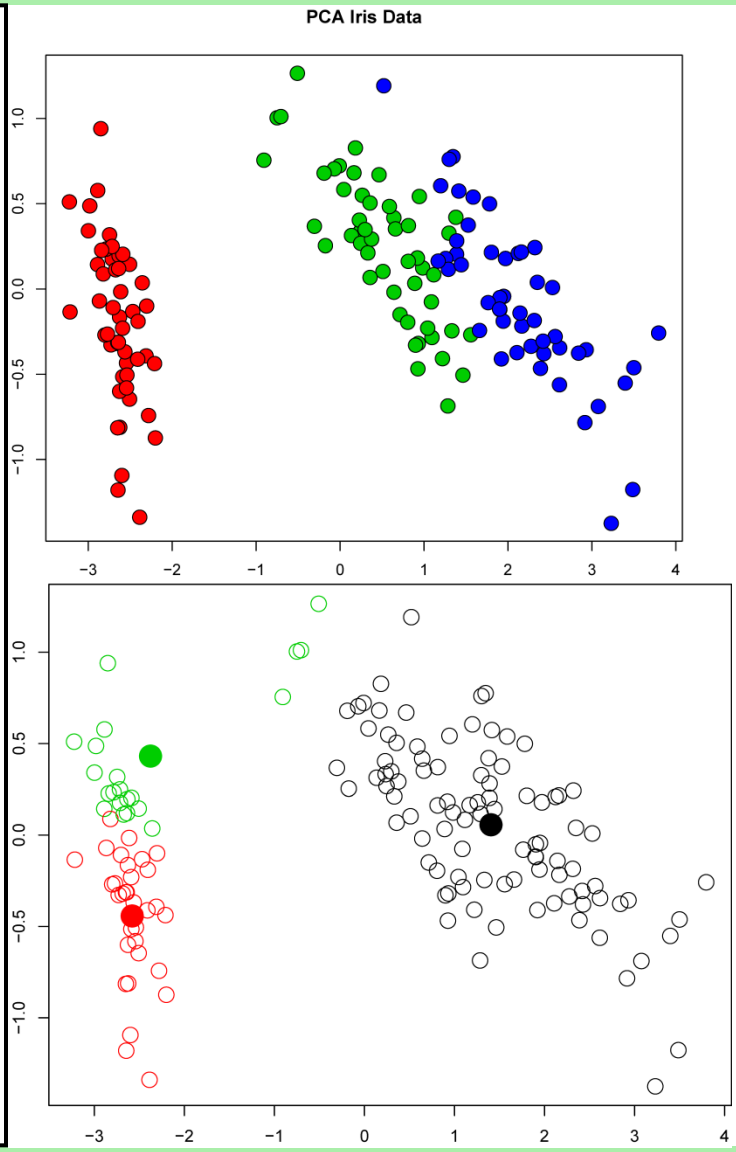
- top row one cluster explains two true clusters while one true cluster is divided into two model clusters
- lower row: three model clusters share one true cluster

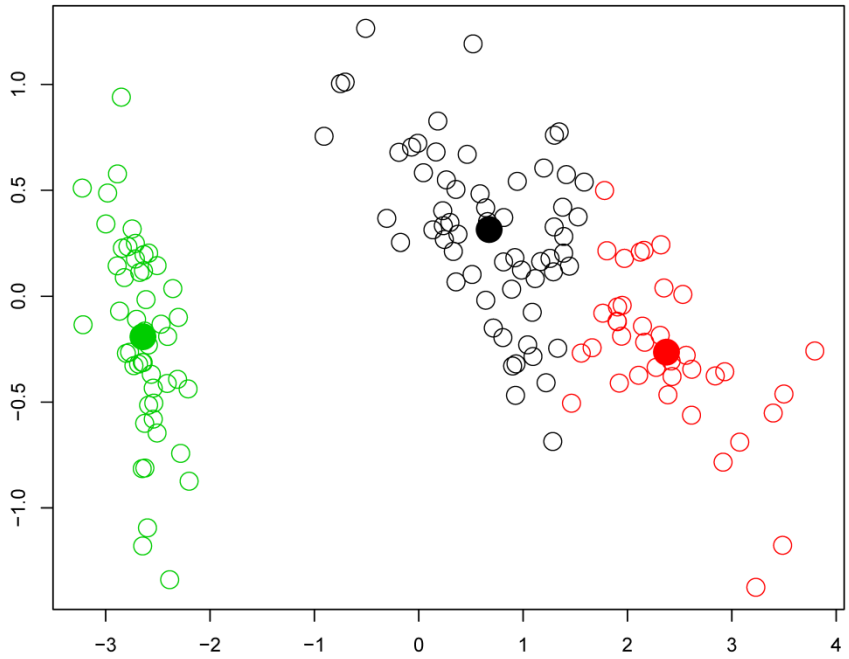
Summarizing Multivariate Data



k-means with 8 comp.

Summarizing Multivariate Data





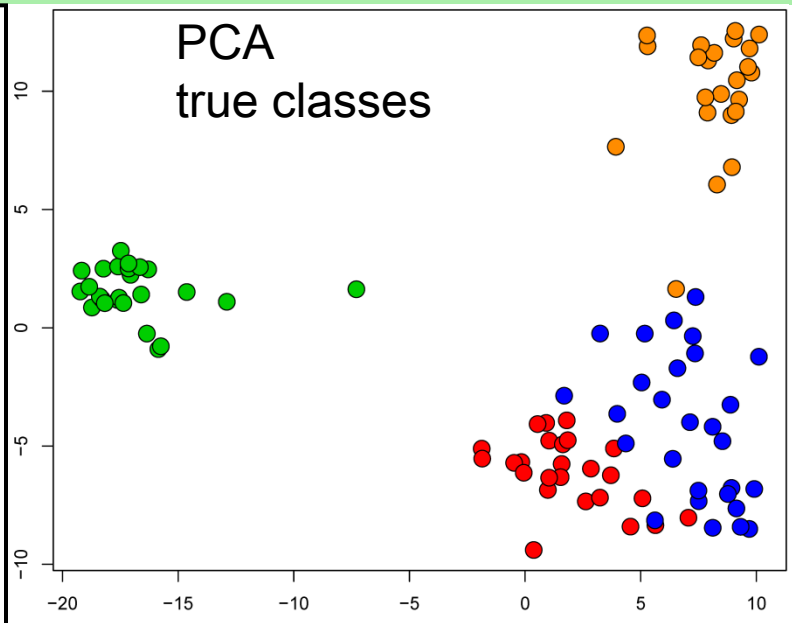
k-means applied to the Iris data

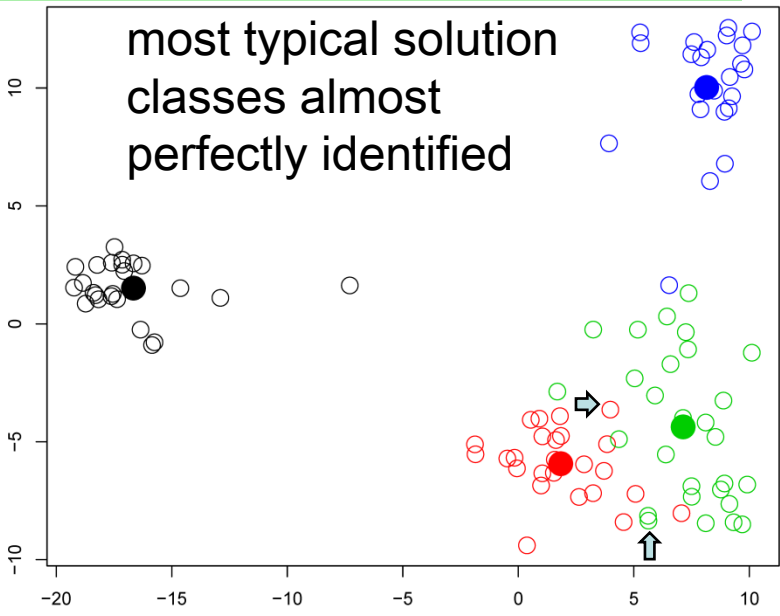
Upper right: typical quite good solution
only errors at class borders

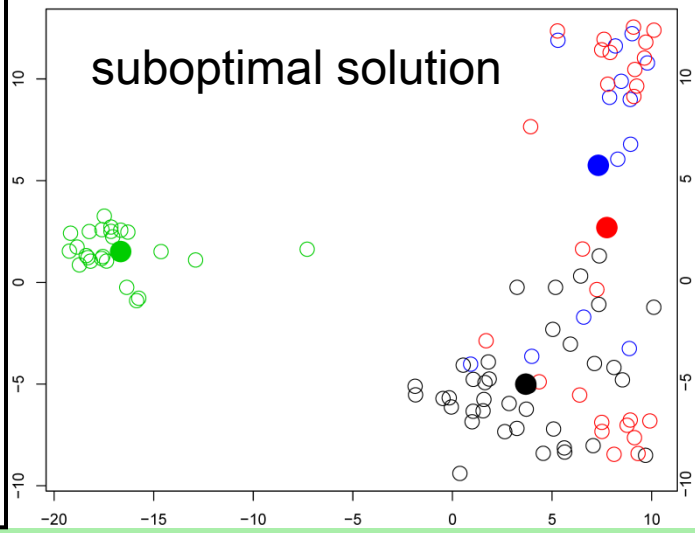
Lower left: another typical solution
which is not as good

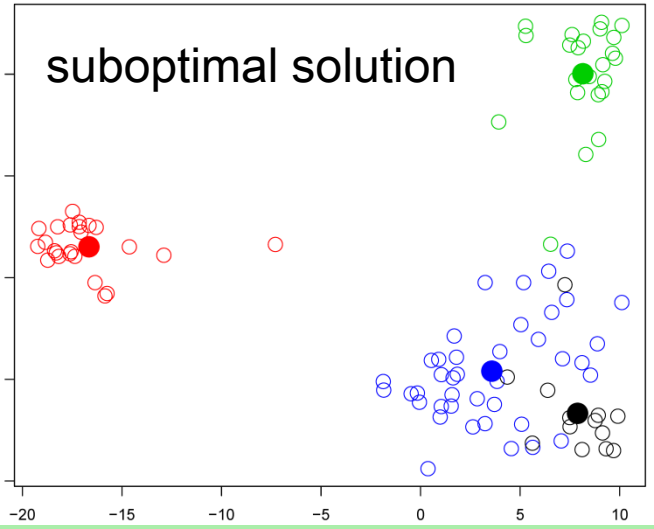
Can these solutions be distinguished?

Summarizing Multivariate Data





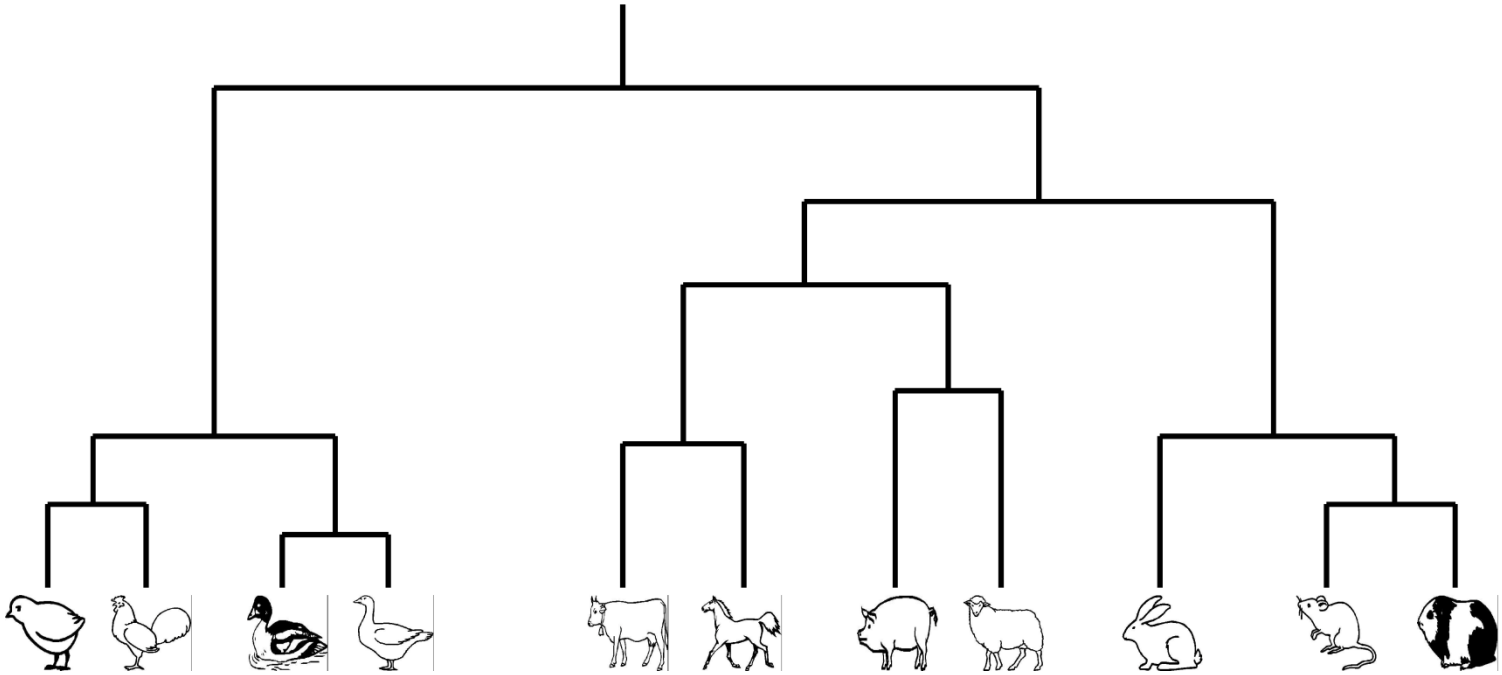




k-means applied to multiple tissues

Summarizing Multivariate Data

Hierarchical clustering supplies distances between clusters which are captured in a dendrogram. These distances allow to merge or cut clusters. Clustering is done agglomerative (bottom up) or divisive (top down)



Summarizing Multivariate Data

agglomerative hierarchical clustering (bottom up) merges the closest clusters to new clusters.
Starts with clusters single observations and iteratively merges clusters.

different distance measures between clusters A and B are used:

$$
\begin{aligned}
d_{\min}(A,B) &= \min_{\boldsymbol{a}\in A,\boldsymbol{b}\in B} \|\boldsymbol{a}-\boldsymbol{b}\| && \text{(single linkage)}\\
d_{\max}(A,B) &= \max_{\boldsymbol{a}\in A,\boldsymbol{b}\in B} \|\boldsymbol{a}-\boldsymbol{b}\| && \text{(complete linkage)}\\
d_{\text{avg}}(A,B) &= \frac{1}{n_A\ n_B} \sum_{\boldsymbol{a}\in A}\sum_{\boldsymbol{b}\in B} \|\boldsymbol{a}-\boldsymbol{b}\| && \text{(average linkage)}\\
d_{\text{mean}}(A,B) &= \|\bar{\boldsymbol{a}}-\bar{\boldsymbol{b}}\| && \text{(average linkage)}
\end{aligned}
$$

where n_A (n_B) is the number of elements in A (B) and $\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{b}}$) is the mean of cluster A (B).
For the element distance $\|.\|$ any distance measure is possible like the Euclidean distance, the Manhattan distance, or the Mahalanobis distance.

Summarizing Multivariate Data

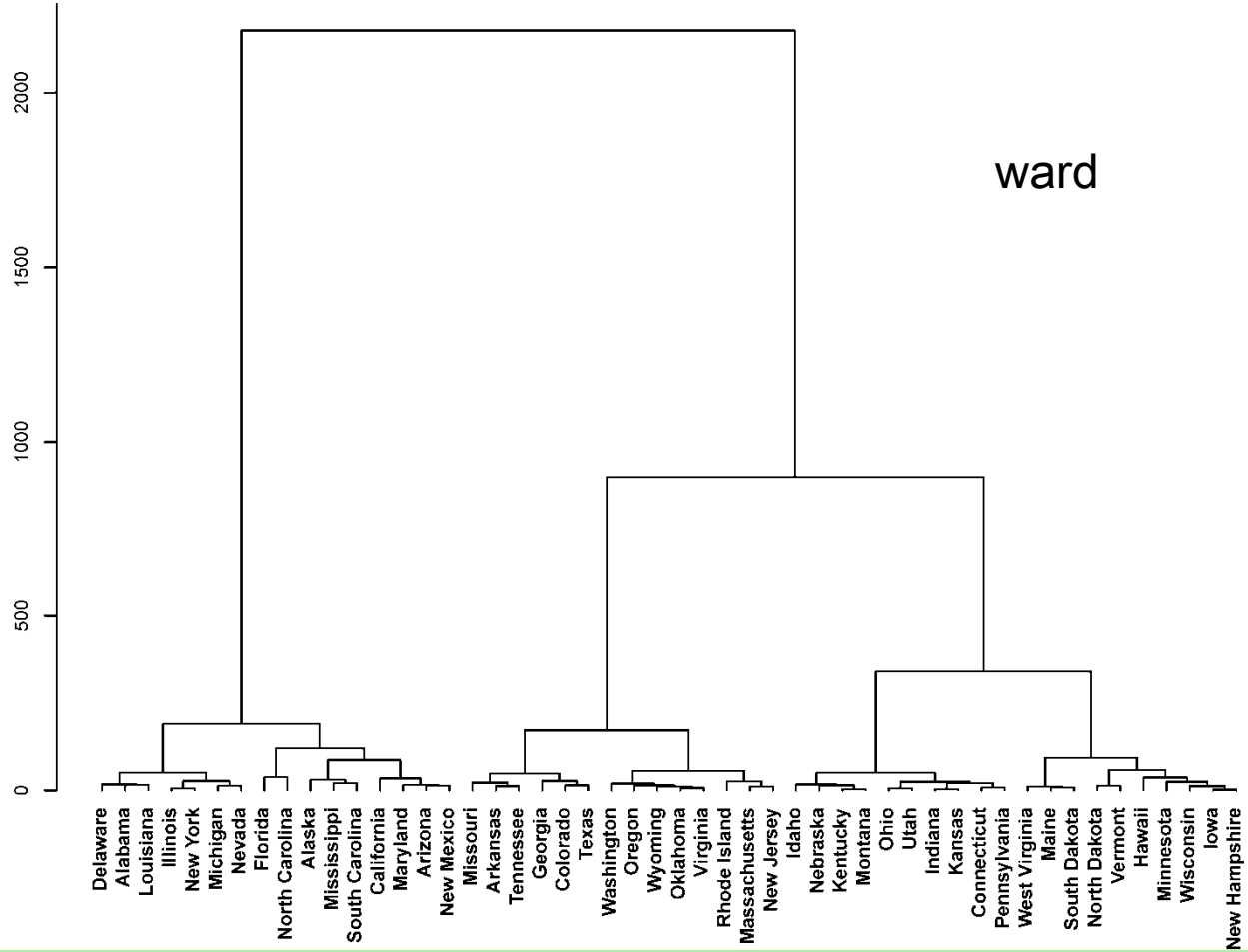
single element clusters: distance measures are equivalent
For more elements in cluster:

- complete linkage $d_{\max}$ avoids that clusters are elongated in some direction (smallest distance between points remains small).
  → cluster may not be well separated.
- single linkage $d_{\min}$ ensures that each pair of elements from different clusters has a minimal distance. Single linkage clustering is relevant for leave-one-cluster-out cross-validation, which assumes that a whole new group of objects is unknown and left out.
- average linkage d_{avg} is "Unweighted Pair Group Method using arithmetic Averages" (UPGMA)

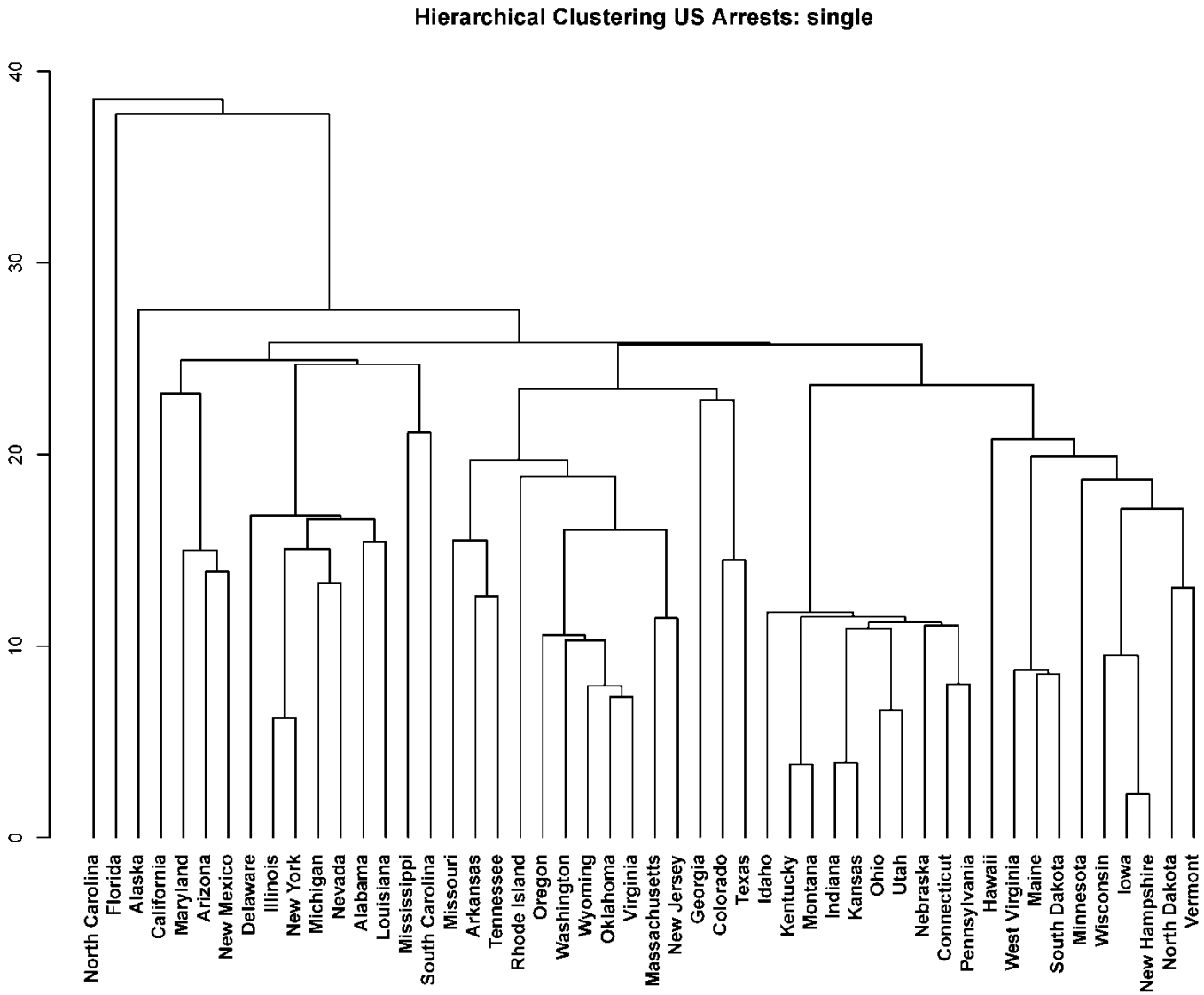
Divisive or top down clustering is often based on graph theoretic considerations. First the **minimal spanning tree** is built.
Then the **largest edge is removed** which gives two clusters.
Now the second largest edge can be removed and so on.
It might be more appropriate to compute the average edge length within a cluster and find the edge which is considerably larger than other edges in the cluster.

Summarizing Multivariate Data

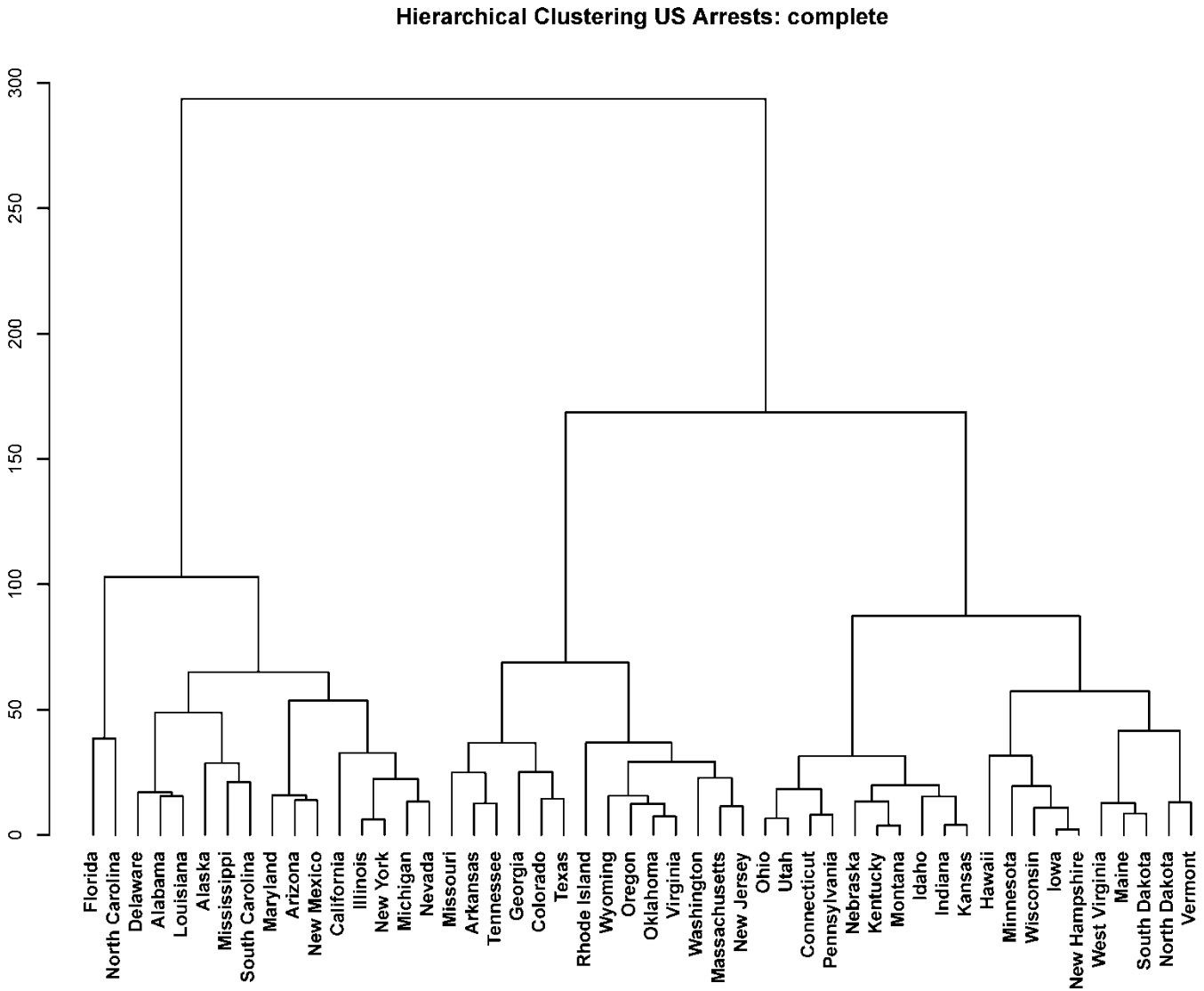
hierarchical clustering of the US Arrest data distance measures “ward”, “single”, “complete”, “average”, “mcquitty”, “median”, and “centroid”:
R function `hclust`: `hc <- hclust(dist(USArrests), method="ward")`



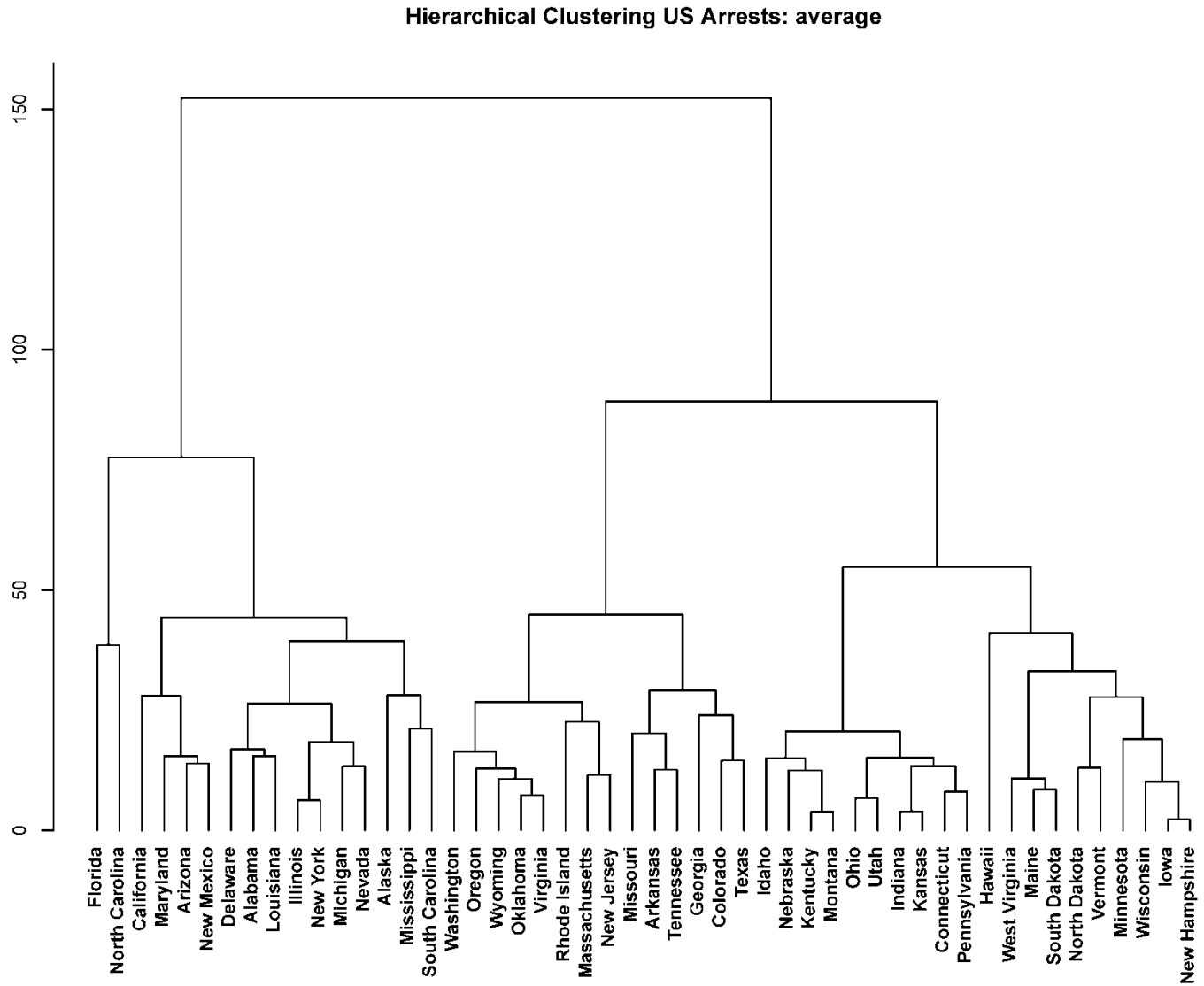
Summarizing Multivariate Data



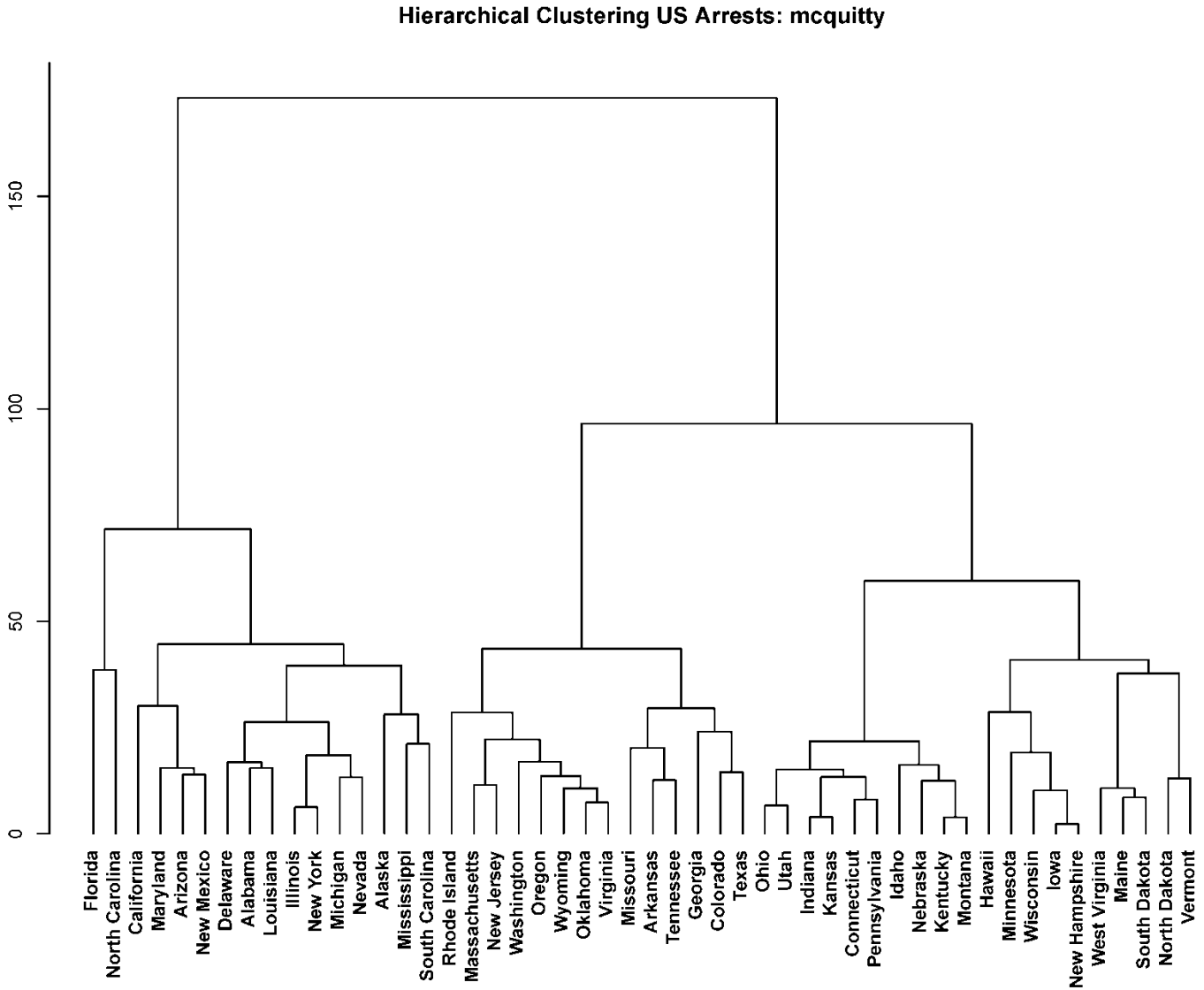
Summarizing Multivariate Data



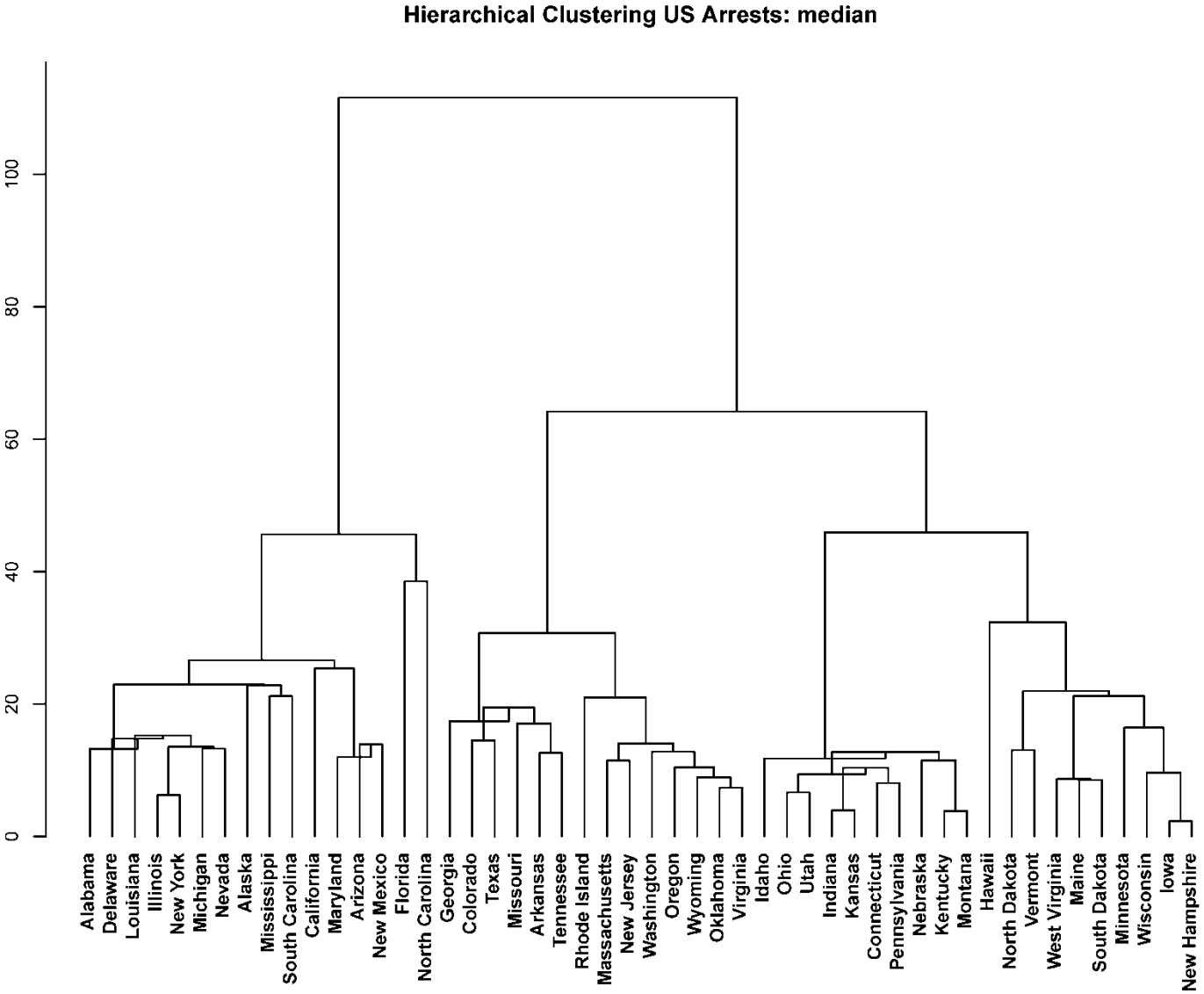
Summarizing Multivariate Data



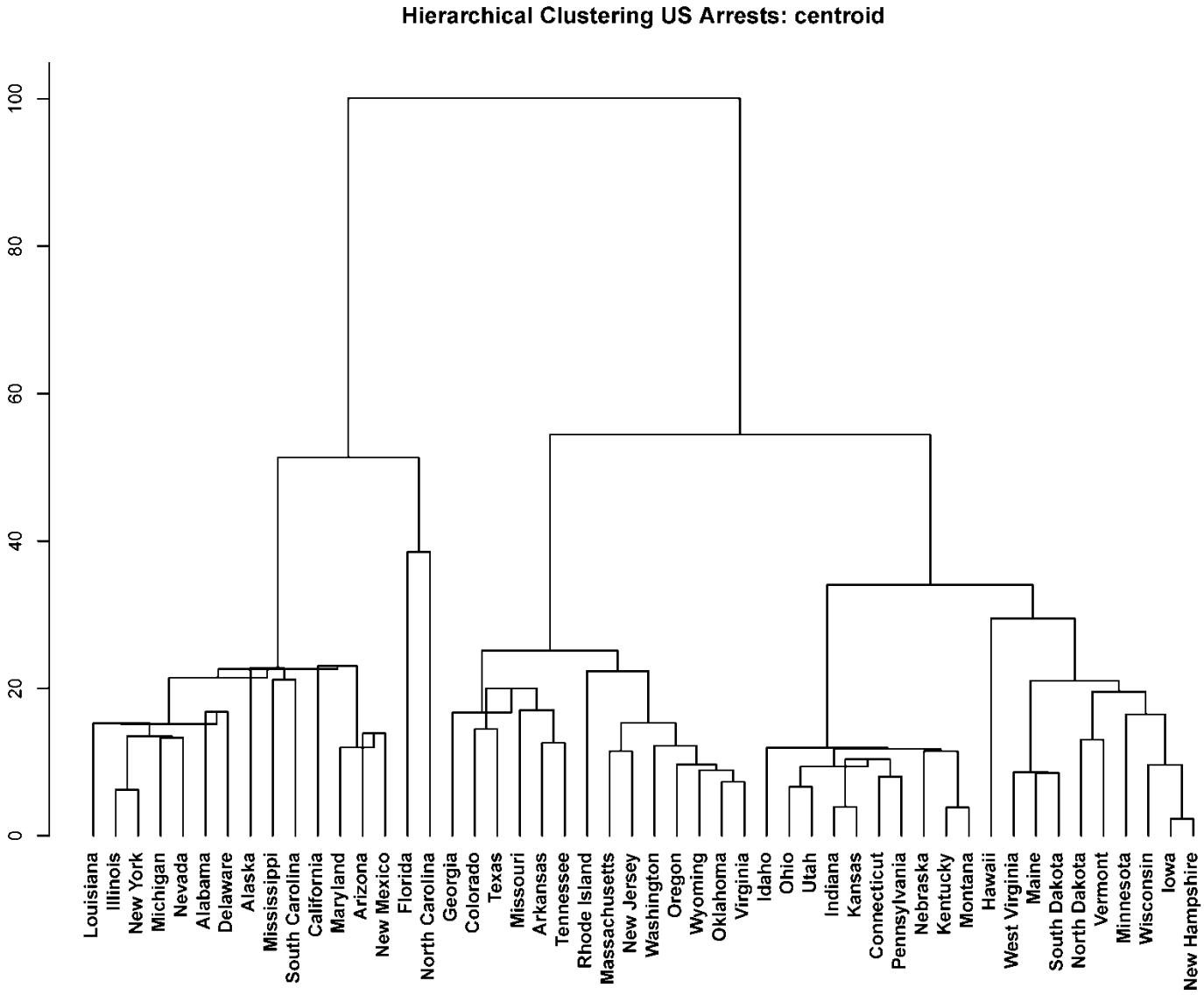
Summarizing Multivariate Data



Summarizing Multivariate Data



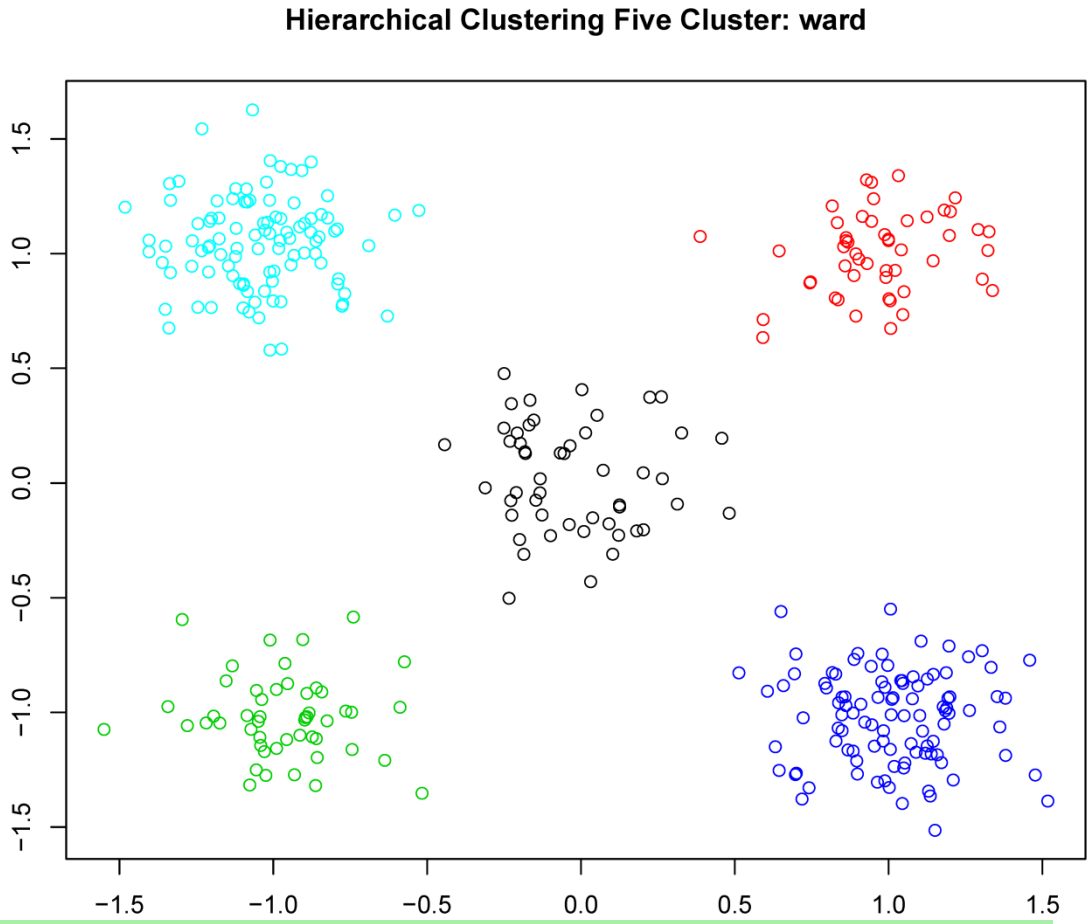
Summarizing Multivariate Data



Summarizing Multivariate Data

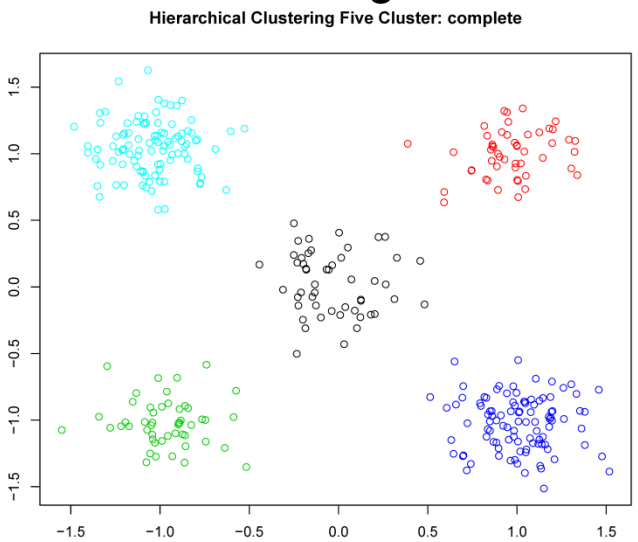
hierarchical clustering of the five cluster data set: Ward's distance is perfect. To determine the clusters, the dendrogram has to be cut which we do by the R function `cutree()`

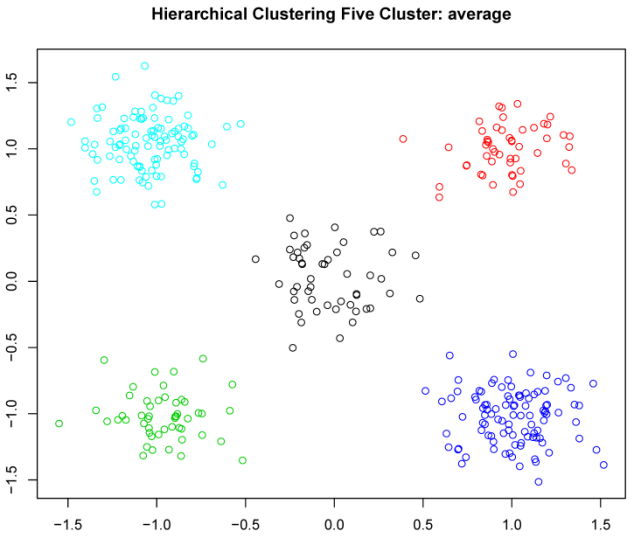
```
hc <- hclust(dist(x), method="ward")
cl <- cutree(hc,k=5)
```

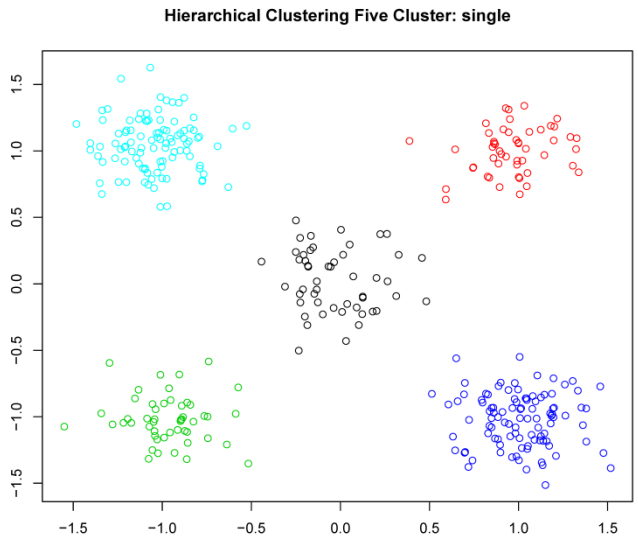


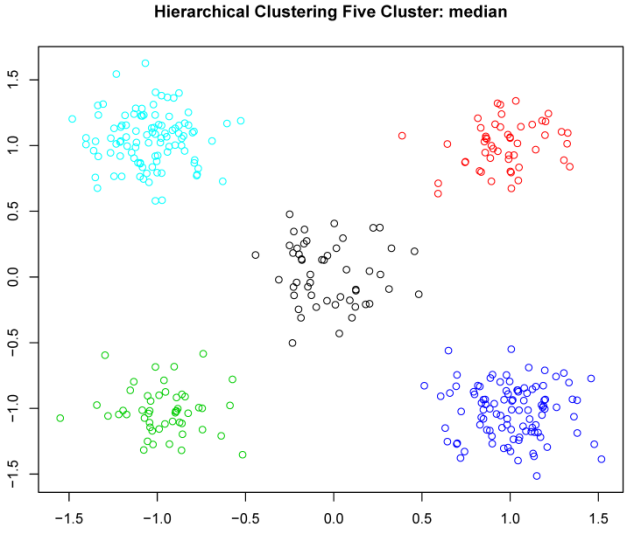
Summarizing Multivariate Data

hierarchical clustering of the five cluster data set

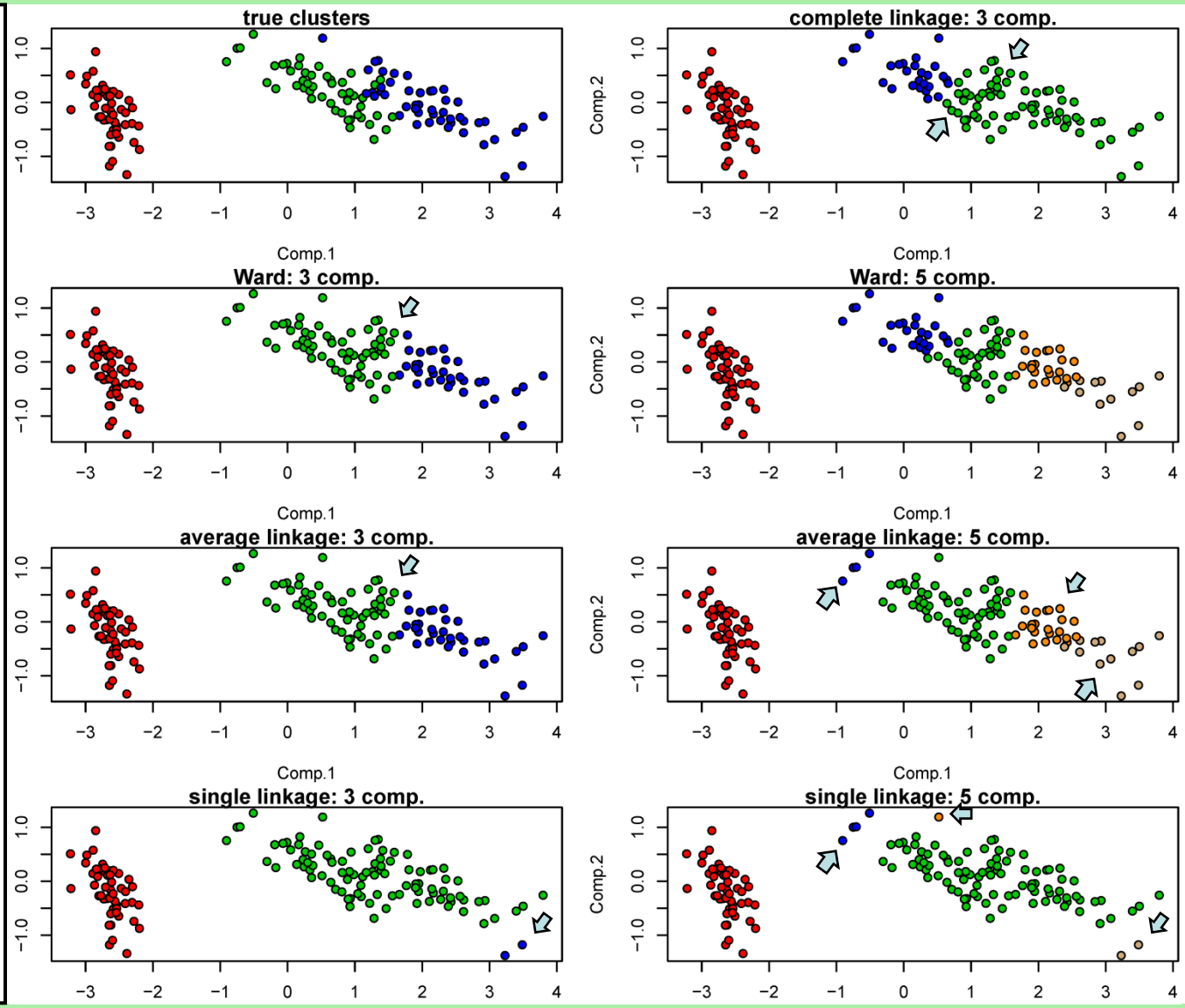






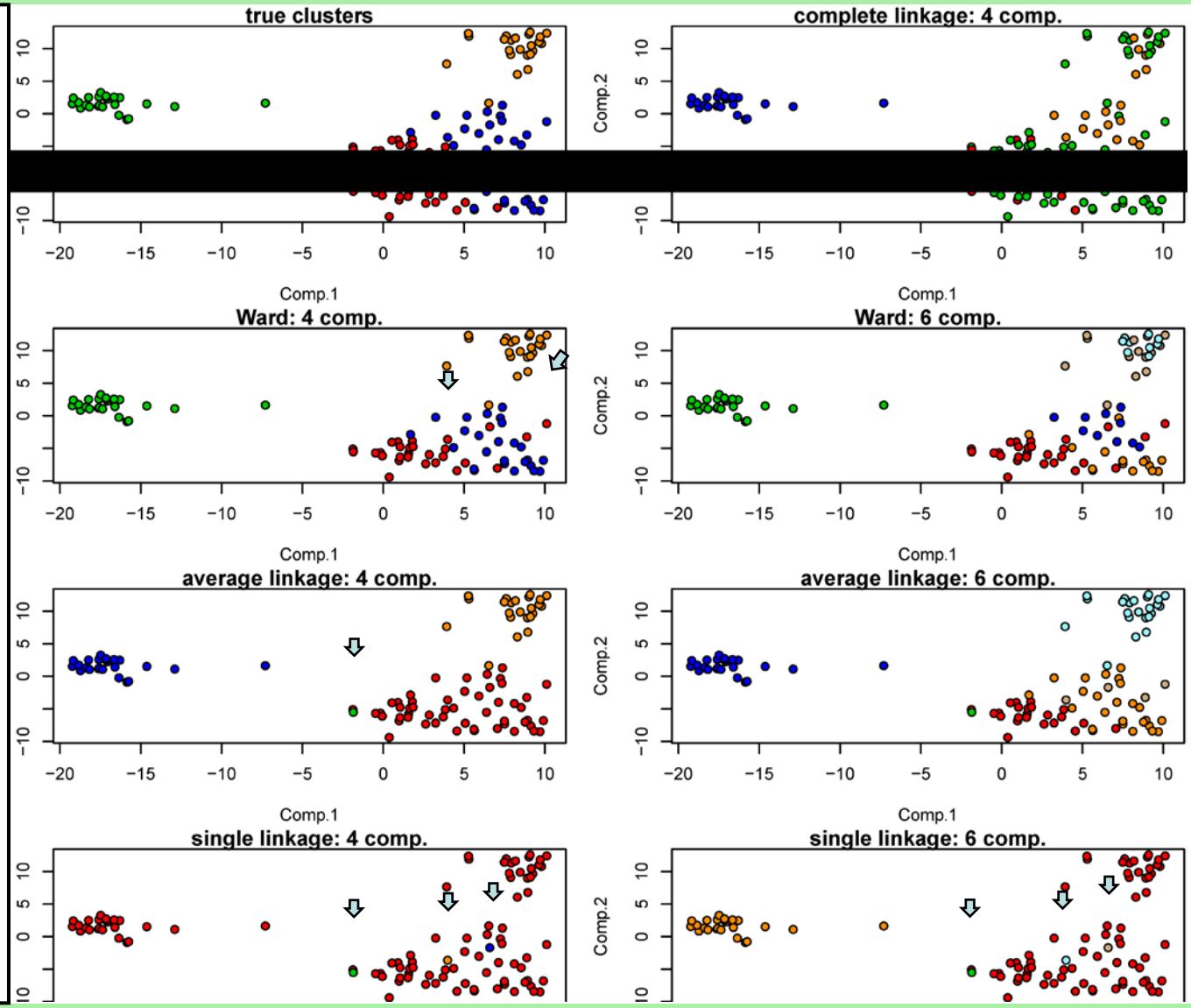


Summarizing Multivariate Data



hierarchical clustering of the iris data set

Summarizing Multivariate Data



hierarchical clustering of the multiple tissue data