Outline

Outline

Notation

- Data consists of p measurements (variables/attributes) on n examples (observations/cases)
- X is a $n \times p$-matrix with $X_{ij} :=$ the j-th measurement for the i-th example

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{np} \end{bmatrix}$$

Crabs Data ($n = 200$, $p = 5$)

Campbell (1974) studied rock crabs of the genus *leptograpsus*. One species, *L. variegatus*, had been split into two new species, previously grouped by colour, orange and blue. Preserved specimens lose their colour, so it was hoped that morphological differences would enable museum material to be classified.

Data are available on 50 specimens of each sex of each species, collected on sight at Fremantle, Western Australia. Each specimen has measurements on the width of the frontal lip `FL`, the rear width `RW`, and length along the midline `CL` and the maximum width `CW` of the carapace, and the body depth `BD` in mm.

Crabs Data

Looking at the crabs dataset, $n = 200$ measurements on $p = 5$ morphological features of crabs

- 'FL' frontal lip size (mm)
- 'RW' rear width (mm)
- 'CL' carapace length (mm)
- 'CW' carapace width (mm)
- 'BD' body depth (mm)

Also available, the colour ('B' or 'O') and sex ('M' or 'F').

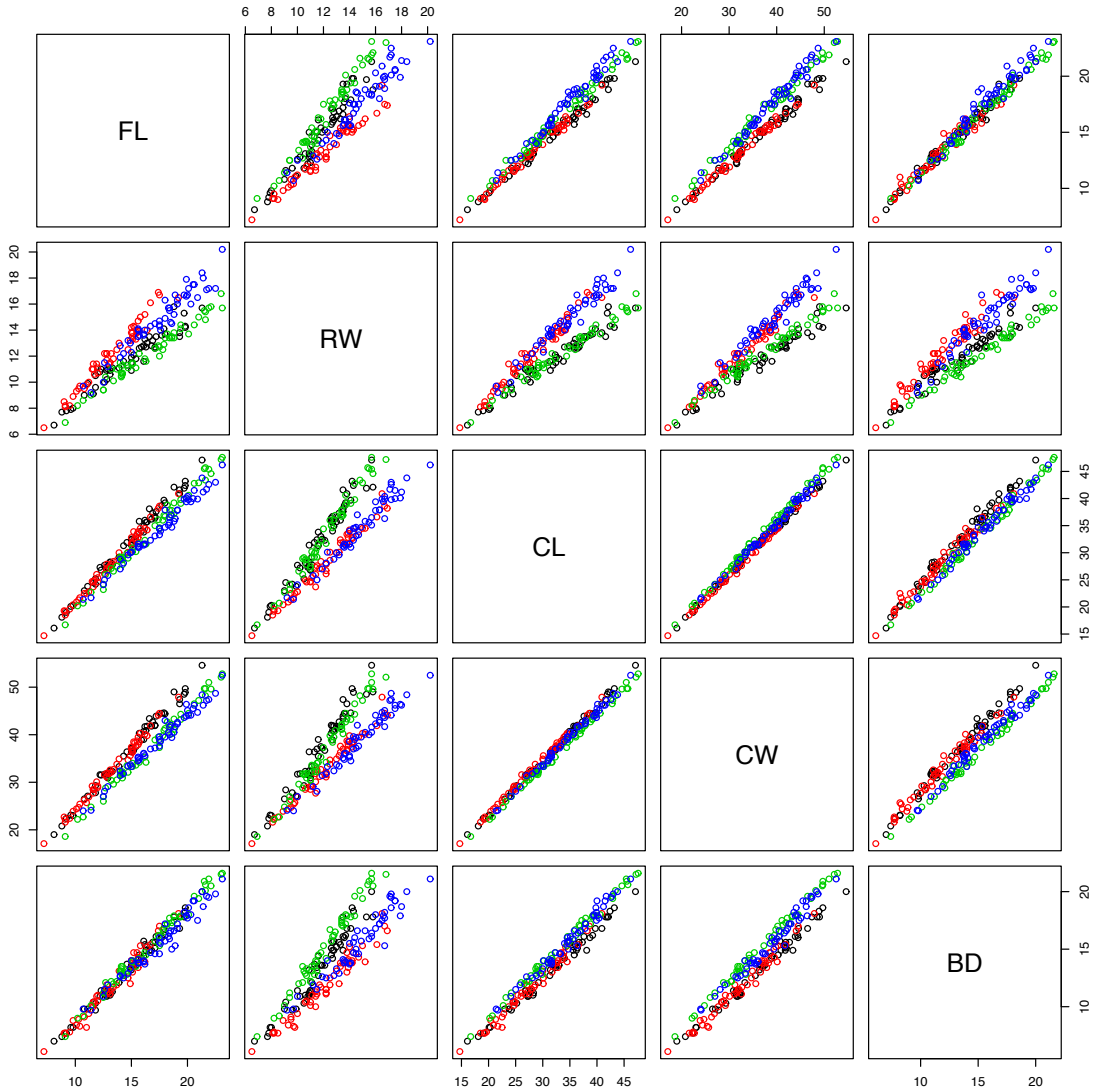
```
## load package MASS containing the data
library(MASS)
## look at data
crabs
```

```
     sp sex index   FL   RW   CL   CW   BD
1    B   M      1  8.1  6.7 16.1 19.0  7.0
2    B   M      2  8.8  7.7 18.1 20.8  7.4
...
```

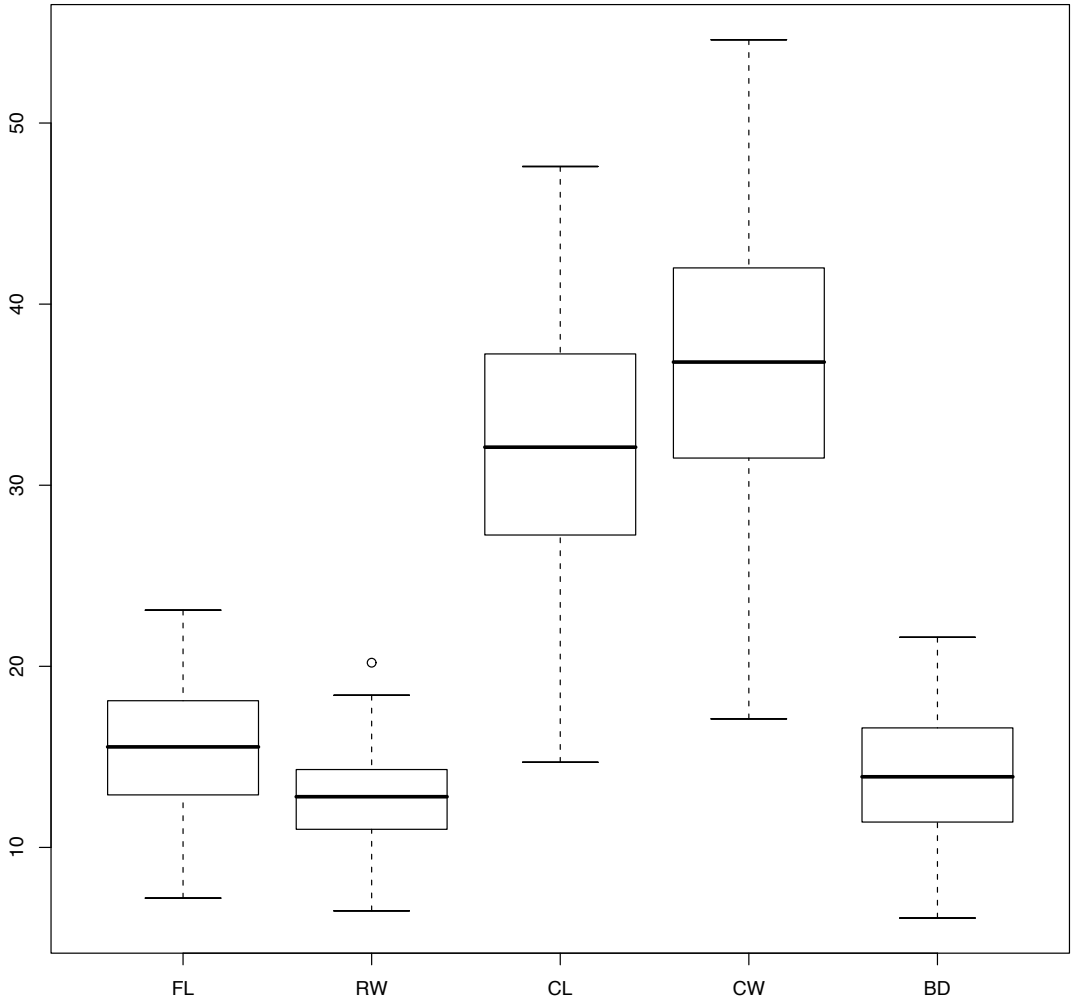
R code

```
## assign predictor and class variables
Crabs <- crabs[,4:8]
Crabs.class <- factor(paste(crabs[,1],crabs[,2],sep=""))

## plot data using pair plots
plot(Crabs,col=unclass(Crabs.class))

##boxplots
boxplot(Crabs)



## parallel coordinates
parcoord(Crabs)
```

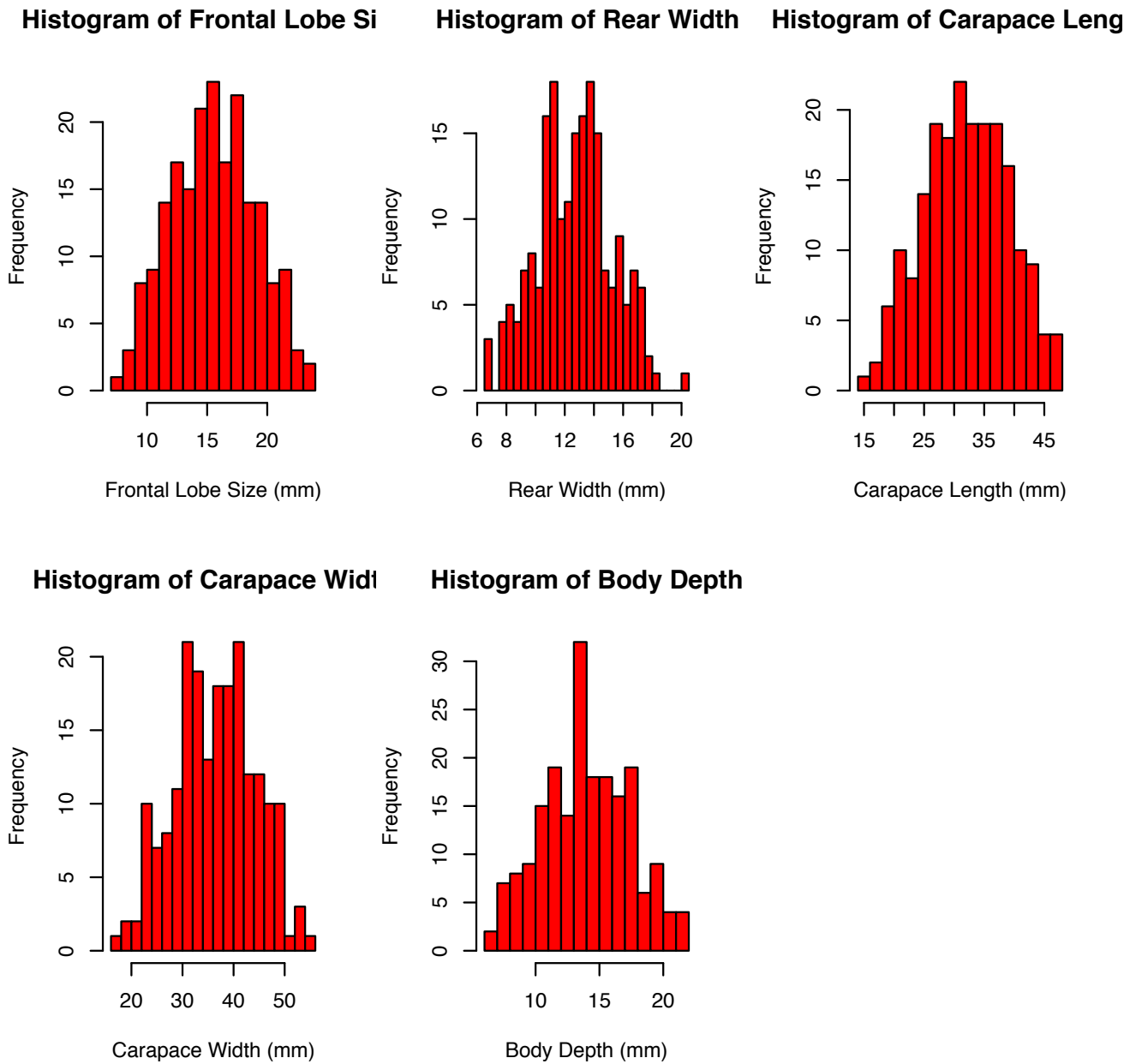
Simple Pairwise Scatterplots



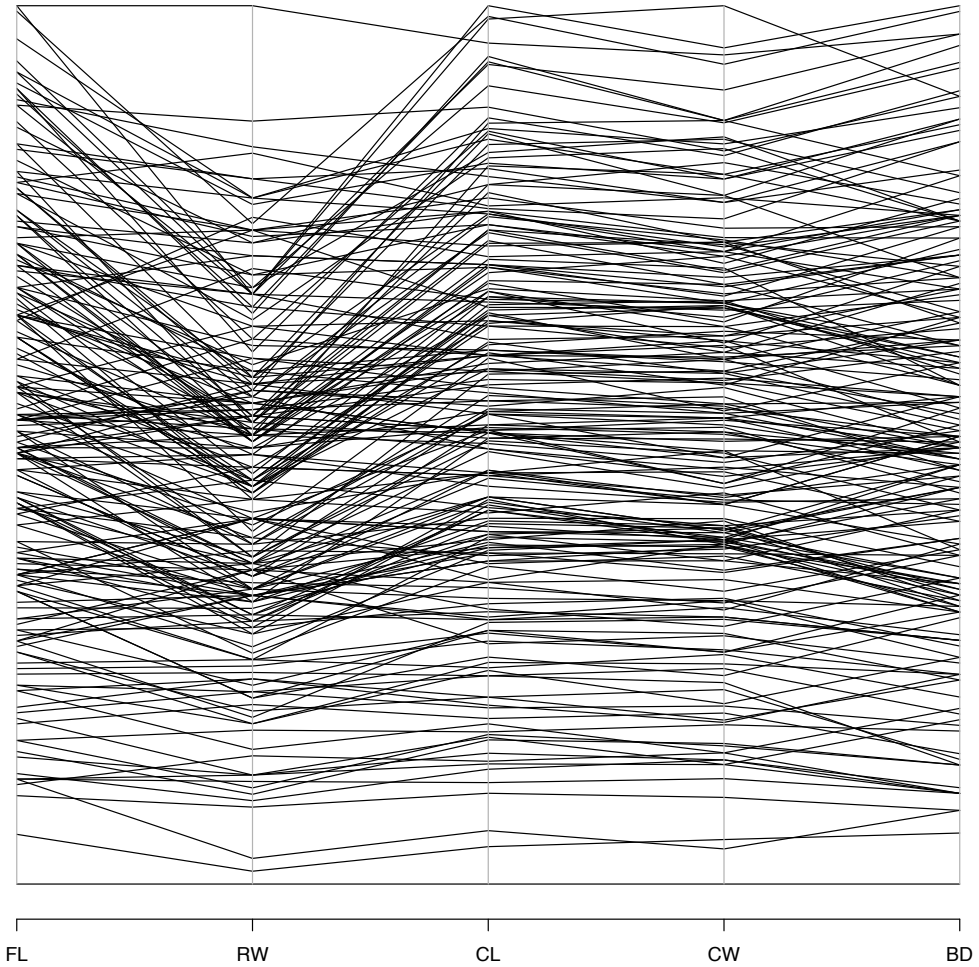
Univariate Boxplots



Univariate Histograms



Parallel Coordinate Plots



These summary plots are helpful, but do not really help very much if the dimensionality of the data is high (a few dozen or thousands).
Possible approaches for higher-dimensional problems.

- We are constrained to view data in 2 or 3 dimensions
- Look for 'interesting' projections of X into lower dimensions
- Hope that for large p, considering only $k \ll p$ dimensions is just as informative

Outline

Principal Components Analysis (PCA)

- Seek to rotate data to a new basis that represents the data in a more 'interesting' way.
- PCA considers interesting to be directions with greatest *variance*.
- Builds up an orthogonal basis where new basis vectors are chosen to explain the greatest variance in data, the first few PCs should represent most of the variance-covariance structure in the data, i.e. the subspace spanned by first k PCs represents the 'best' k-dimensional view of the data.

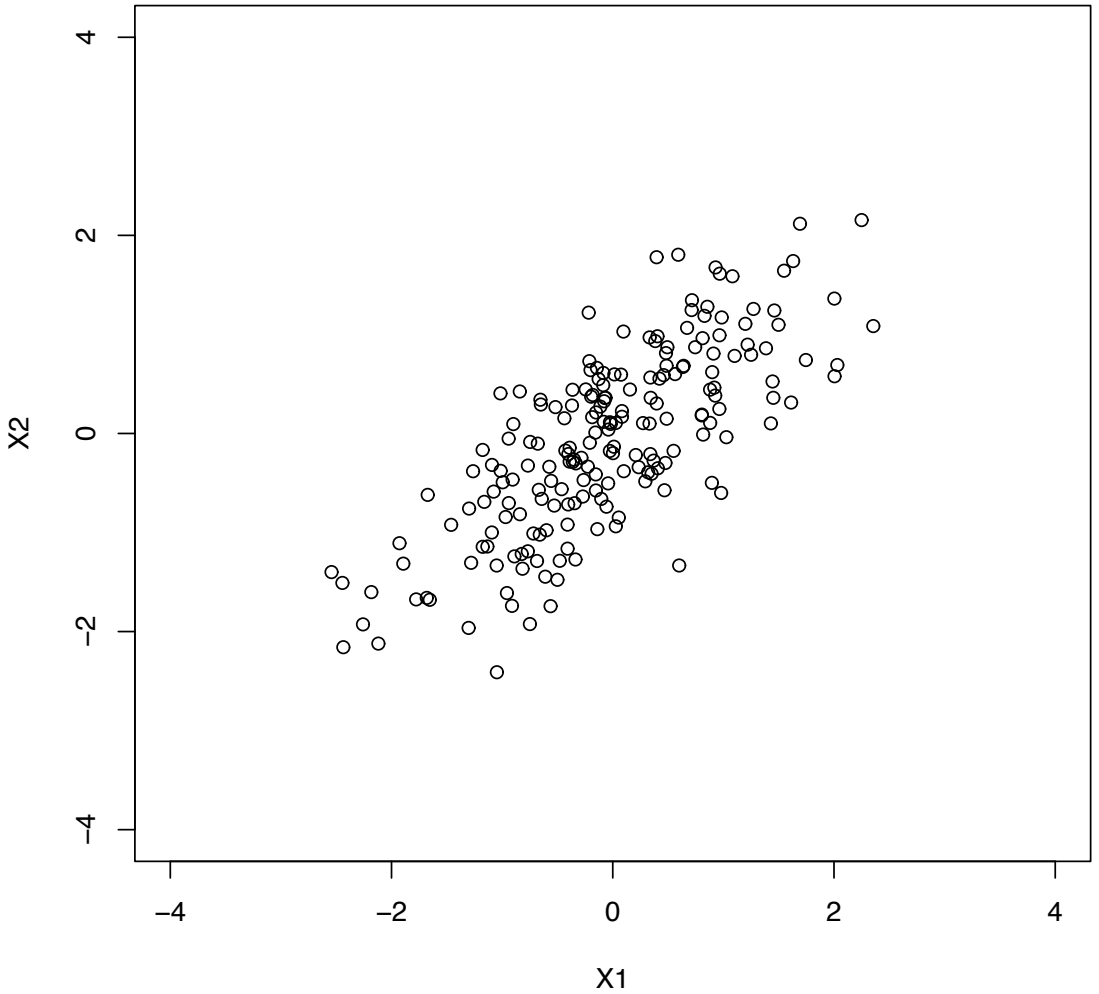
Principal Components Analysis (PCA)

- Consider a set of real-valued variables $X = (X_1 \dots X_p)^T$.
- For the 1^{st} PC, we seek a derived variable of the form

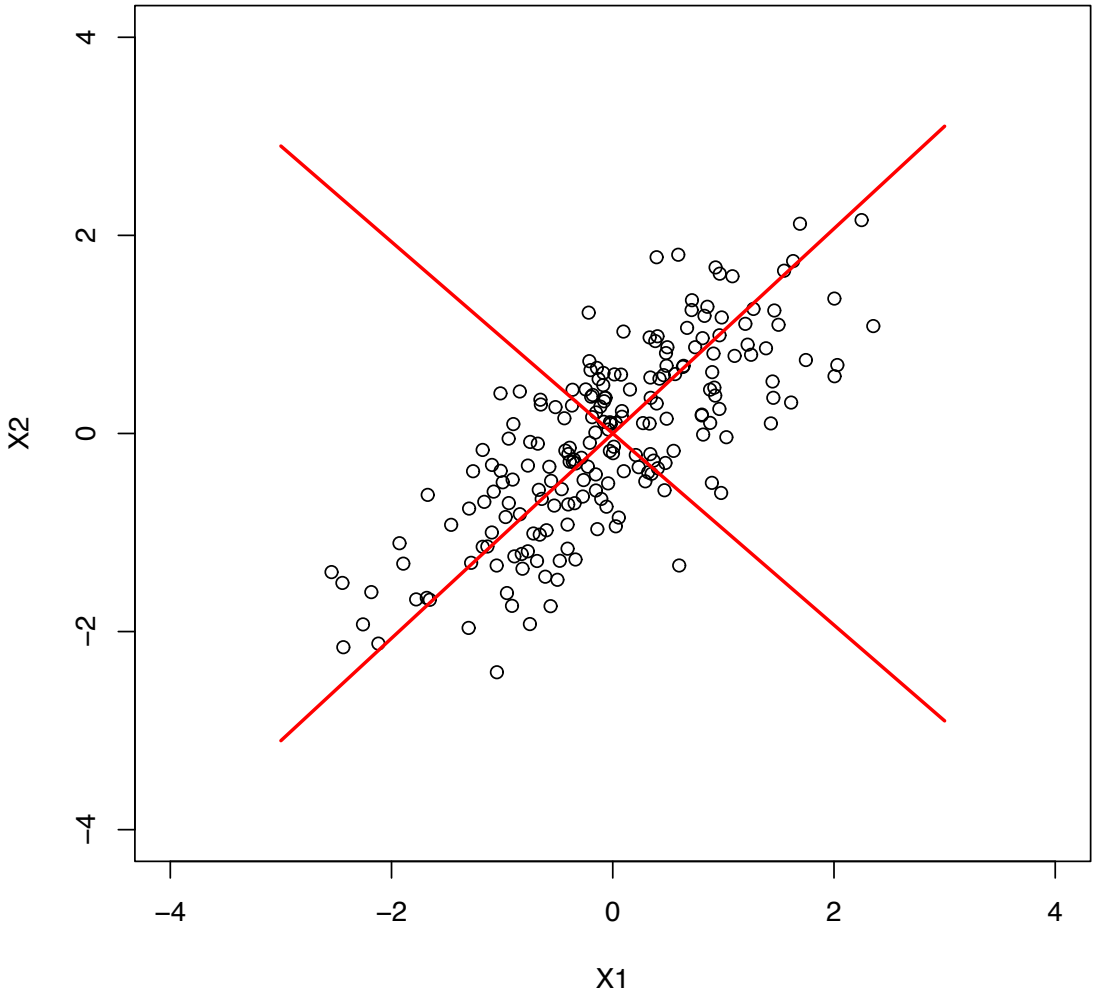
$$Z_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p = X^T \mathbf{a}_1$$

  where $a_{1i} \in \mathbb{R}$ are chosen to maximise var(Z_1).
- To get a well defined problem, we fix $\mathbf{a}_1^T \mathbf{a}_1 = 1$.
- The 1^{st} PC attempts to capture the common variation in all variables using a single derived variable.
- The 2^{nd} PC Z_2 is chosen to be orthogonal with the 1^{st} and is computed in a similar way. It will have the largest variance in the remaining $p - 1$ dimensions, etc.

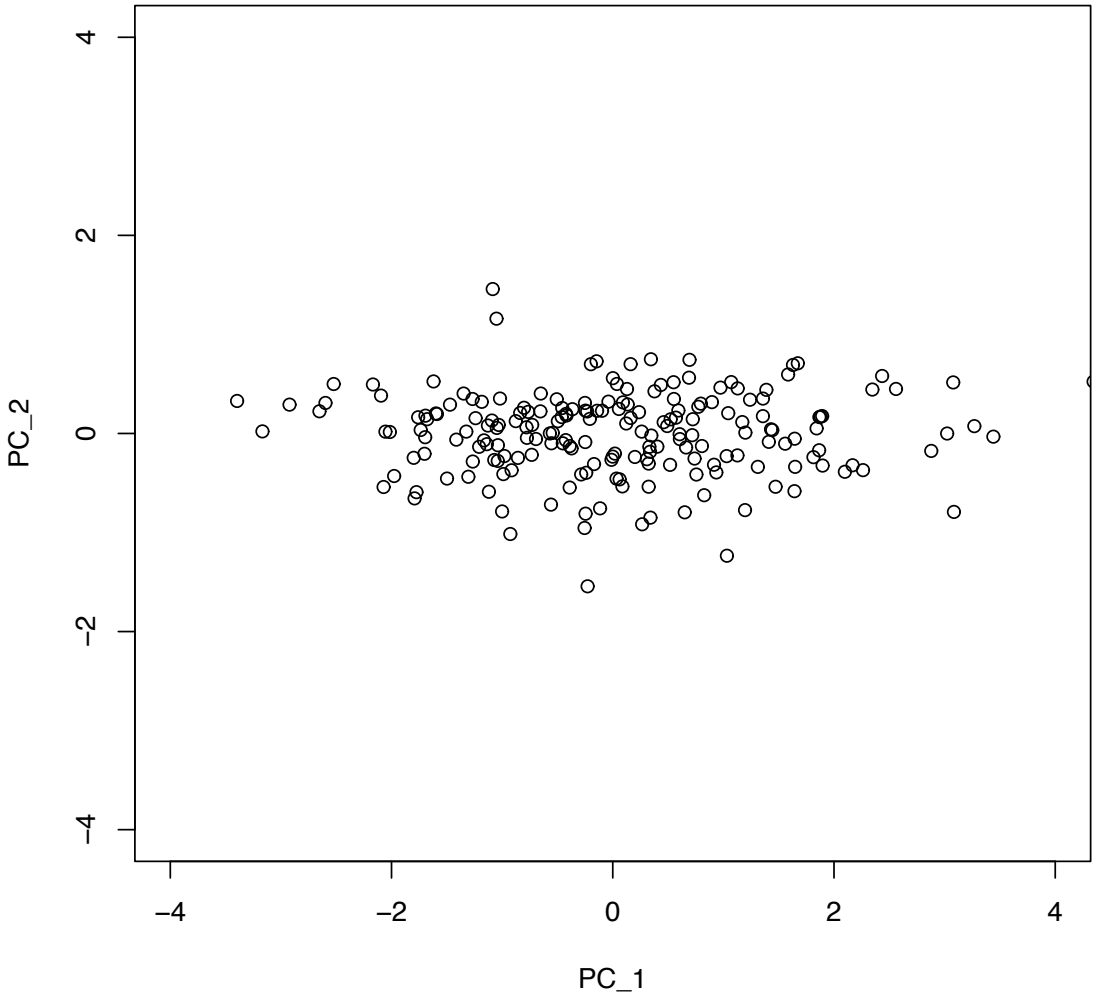
Principal Components Analysis (PCA)



Principal Components Analysis (PCA)



Principal Components Analysis (PCA)



How to Obtain the Coefficients?

To find the 1^{st} PC given by $Z_1 = X^T \mathbf{a}_1$

- Maximise $var(Z_1) = var(X\mathbf{a}_1) = \mathbf{a}_1^T cov(X)\mathbf{a}_1 \approx \mathbf{a}_1^T S\mathbf{a}_1$ subject to $\mathbf{a}_1^T\mathbf{a}_1 = 1$ where $S = n^{-1}X^TX$ is a $p \times p$ sample covariance matrix of the centred $n \times p$ data matrix X.
- Rewriting this as a constrained maximisation problem,

$$\text{Maximise } F(\mathbf{a}_1) = \mathbf{a}_1^T S\mathbf{a}_1 - \lambda_1\left(\mathbf{a}_1^T\mathbf{a}_1 - 1\right) \text{ w.r.t. } \mathbf{a}_1.$$

- The corresponding vector of partial derivatives yields

$$\frac{\partial F}{\partial \mathbf{a}_1} = 2S\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1.$$

- Setting this to zero reveals the eigenvector equation, i.e. $\mathbf{a}_1$ must be an eigenvector of S and λ_1 the corresponding eigenvector.
- Since $\mathbf{a}_1^T S\mathbf{a}_1 = \lambda_1\mathbf{a}_1^T\mathbf{a}_1 = \lambda_1$, the 1^{st} PC must be the eigenvector associated with the largest eigenvalue of S.

How to Obtain the Coefficients?

How about the 2^{nd} PC?

- Proceed as before but include the additional constraint that the 2^{nd} PC must be orthogonal to the 1^{st} PC

$$\text{Maximise } F(\mathbf{a}_2) = \mathbf{a}_2^T S \mathbf{a}_2 - \lambda_2 \left(\mathbf{a}_2^T \mathbf{a}_2 - 1\right) - \mu \left(\mathbf{a}_1^T \mathbf{a}_2\right) \text{ w.r.t. } \mathbf{a}_2$$

- Solving this shows that a_2 must be the eigenvector of S associated with the 2^{nd} largest eigenvalue, and so on
- The eigenvalue decomposition of S is given by $S = A \Lambda A^T$ where Λ is a diagonal matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and A is a $p \times p$ orthogonal matrix whose columns are the p eigenvectors of S.

Properties of the Principal Components

- PCs are *uncorrelated*

$$cov(X^T \mathbf{a}_i, X^T \mathbf{a}_j) \approx \mathbf{a}_i^T S \mathbf{a}_j = 0 \text{ for } i \neq j.$$

- The total sample variance is given by

$$\text{Total sample variance} = \sum_{i=1}^{p} s_{ii} = \lambda_1 + \ldots + \lambda_p,$$

so the proportion of total variance explained by the k^{th} PC is

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad k = 1, 2, \ldots, p$$

R code

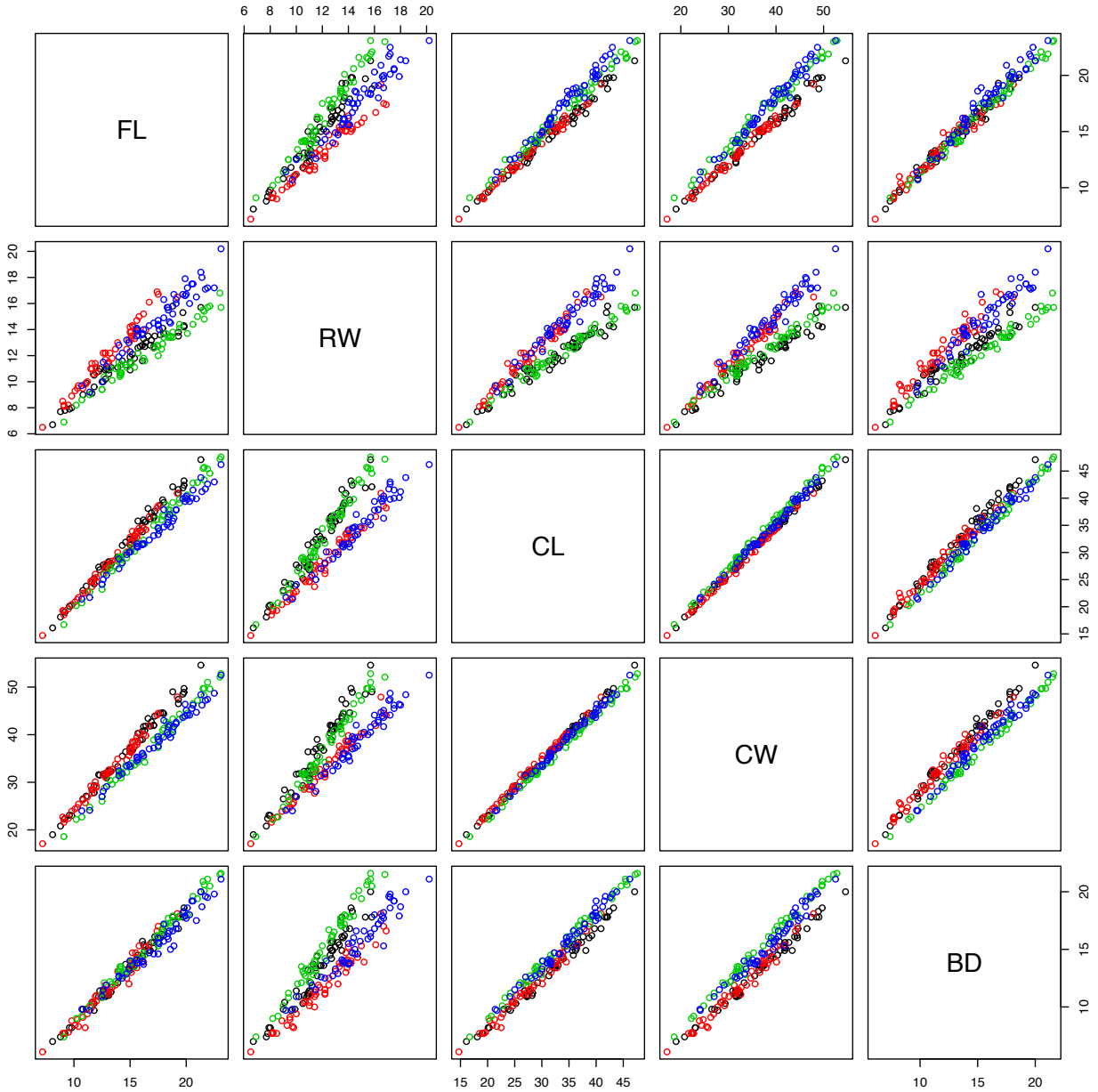
This is what we have had before:

```
library(MASS)
Crabs <- crabs[,4:8]
Crabs.class <- factor(paste(crabs[,1],crabs[,2],sep=""))
plot(Crabs,col=unclass(Crabs.class))
```

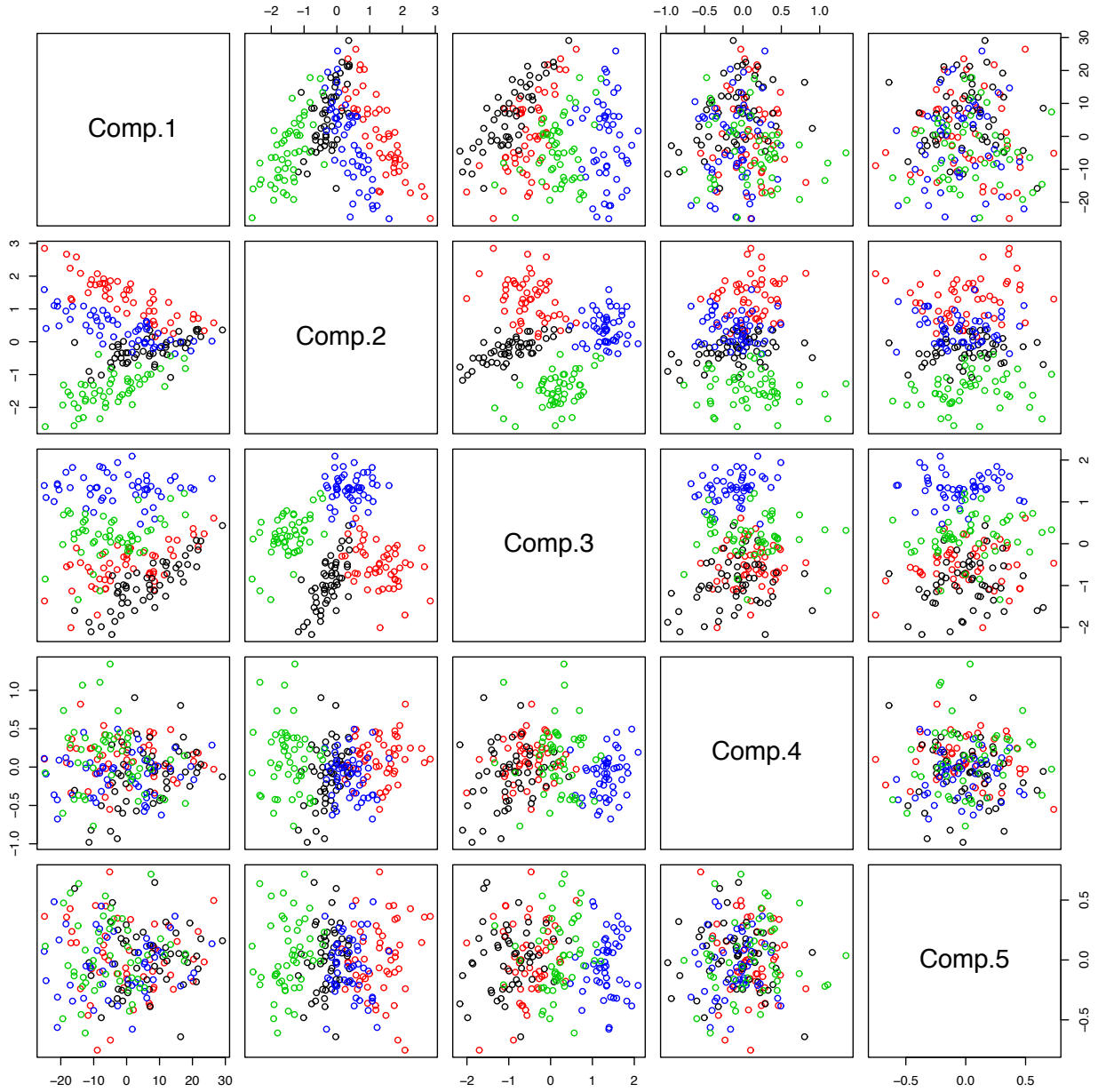
Now perform PCA analysis with function `princomp`.
Alternatively, use `eigen` or `svd` instead (later).

```
Crabs.pca <- princomp(Crabs,cor=FALSE)
plot(Crabs.pca)
pairs(predict(Crabs.pca),col=unclass(Crabs.class))
```

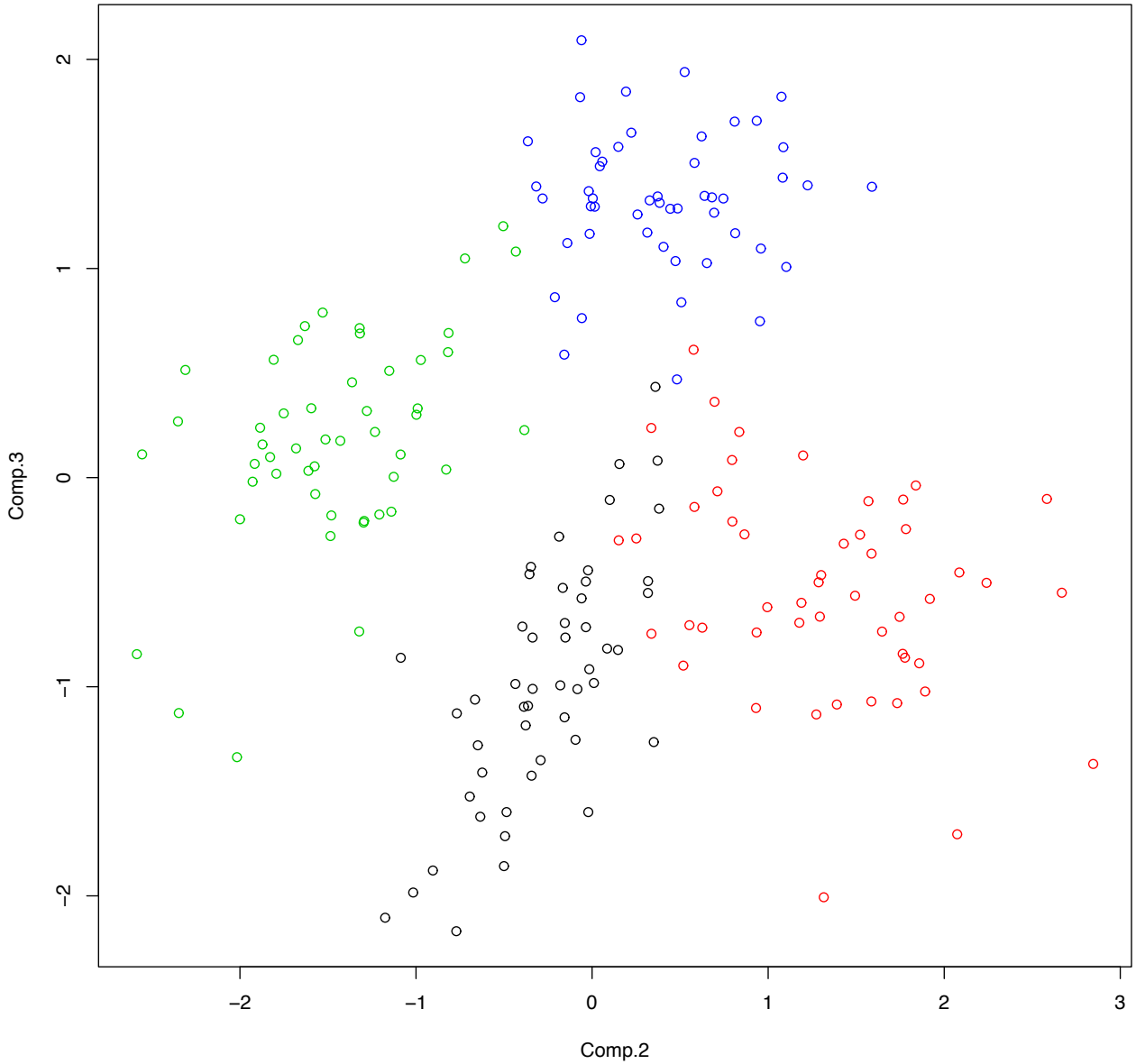
PCA Example 1: Original crabs data



PCA Example 1: Rotated crabs data



PCA Example 1: Crabs Data ($n = 200$, $p = 5$)



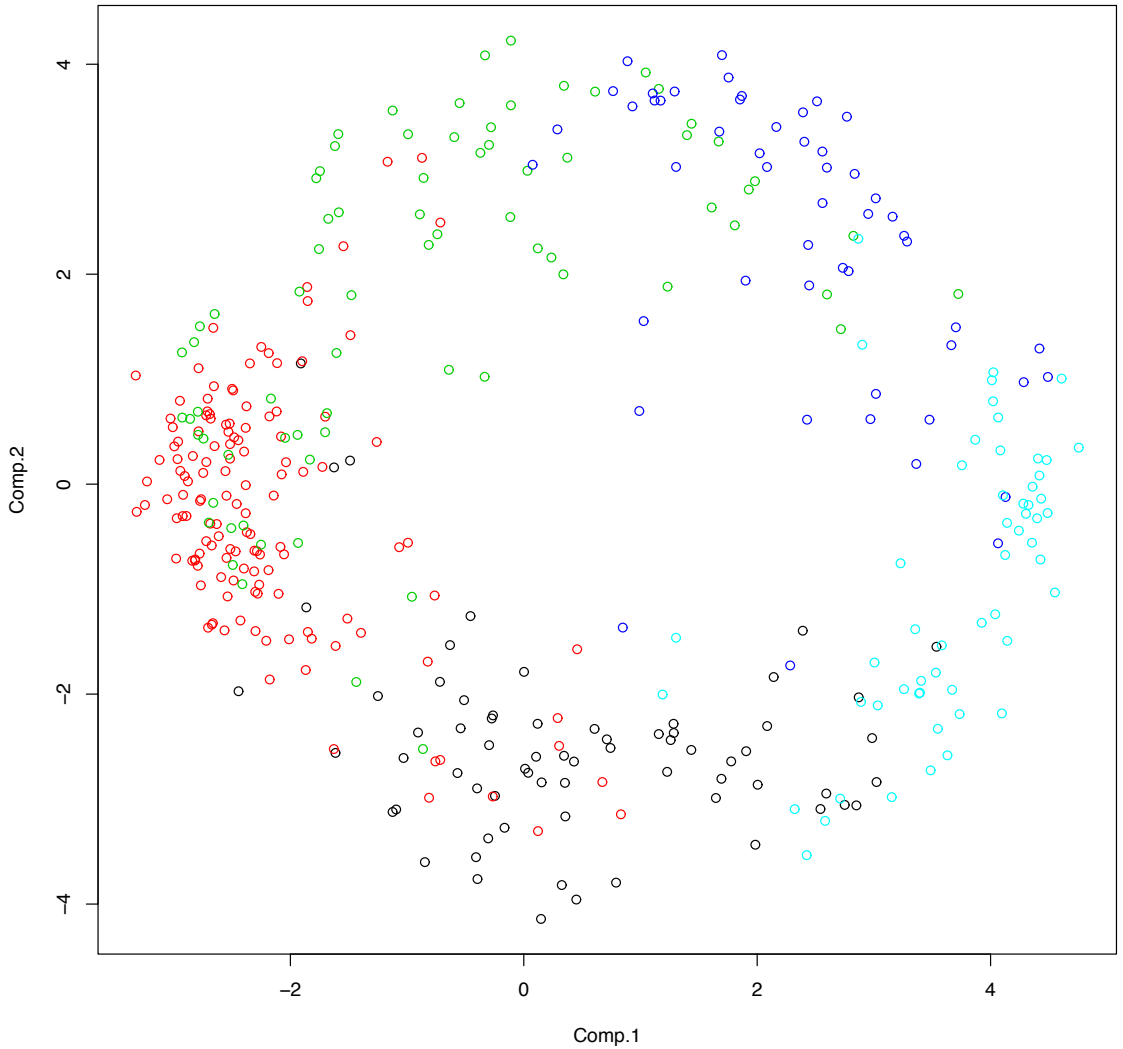
PCA Example 2: Yeast Cell Cycle Data ($n = 384$, $p = 17$)

Cho *et al* (1998) present gene expression data on the cell cycle of yeast. They identify a subset of genes that can be categorised into five different phases of the cell-cycle. Changes in expression for the genes are measured over two cell cycles (17 time points).

The data were normalised so that the expression values for each gene has mean zero and unit variance across the cell cycles.

We visualise the 384 genes in the space of the first two prinicipal components.

PCA Example 2: Yeast Cell Cycle Data ($n = 384$, $p = 17$)



Comments on the use of PCA

- PCA commonly used to project data X onto the first k PCs giving the 'best' k-dimensional view of the data.
- PCA commonly used for lossy compression of high dimensional data.
- Emphasis on variance is where the weaknesses of PCA stem from:
  - The PCs depend heavily on the units measurement. Where the data matrix contains measurements of vastly differing orders of magnitude, the PC will be greatly biased in the direction of larger measurement. It is therefore recommended to calculate PCs from $cor(X)$ instead of $cov(X)$.
  - Robustness to outliers is also an issue. Variance is affected by outliers therefore so are PCs.
- Although PCs are uncorrelated, scatterplots sometimes reveal structures in the data other than linear correlation.

Biplots

- When viewing projections of data matrix X into its PC space, it is instructive to view the contribution from the original variables to the PCs that are plot.
- Biplots overlay projection of *unit vectors* of the original variables into the PC space
- As PCs are linear combinations of the original variables, it is straightforward to invert this relationship to yield the contributions of the original variables to the PCs

Biplots

Biplots show us an *image* of the data and unit vectors of the original axes into the projected space.

- The distance of projected points away from the projected original axes tell us its original location.
- Unit vectors of the original variables give us a common denominator to compare how much weighting each PC gives to the original variables.
- It can be shown that $\cos\theta$ (where θ is the angle that subtends two projected original axes) approximates the correlation between these variables.

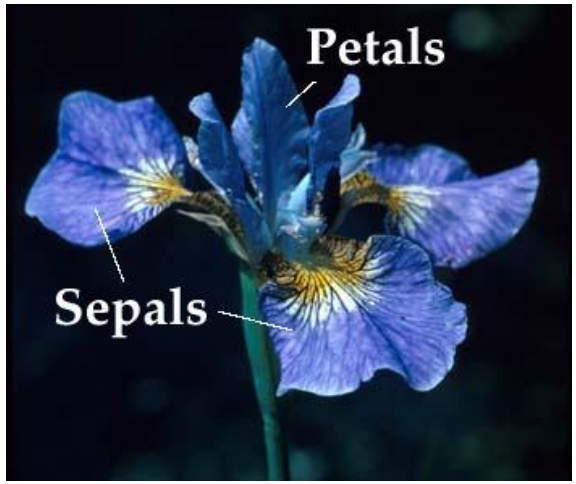
However, the quality of this image depends on the proportion of variance explained by the PCs used.

Biplot Example 1: Fisher's Iris Data

50 sample from 3 species of iris: *iris setosa*, *versicolor*, and *virginica*

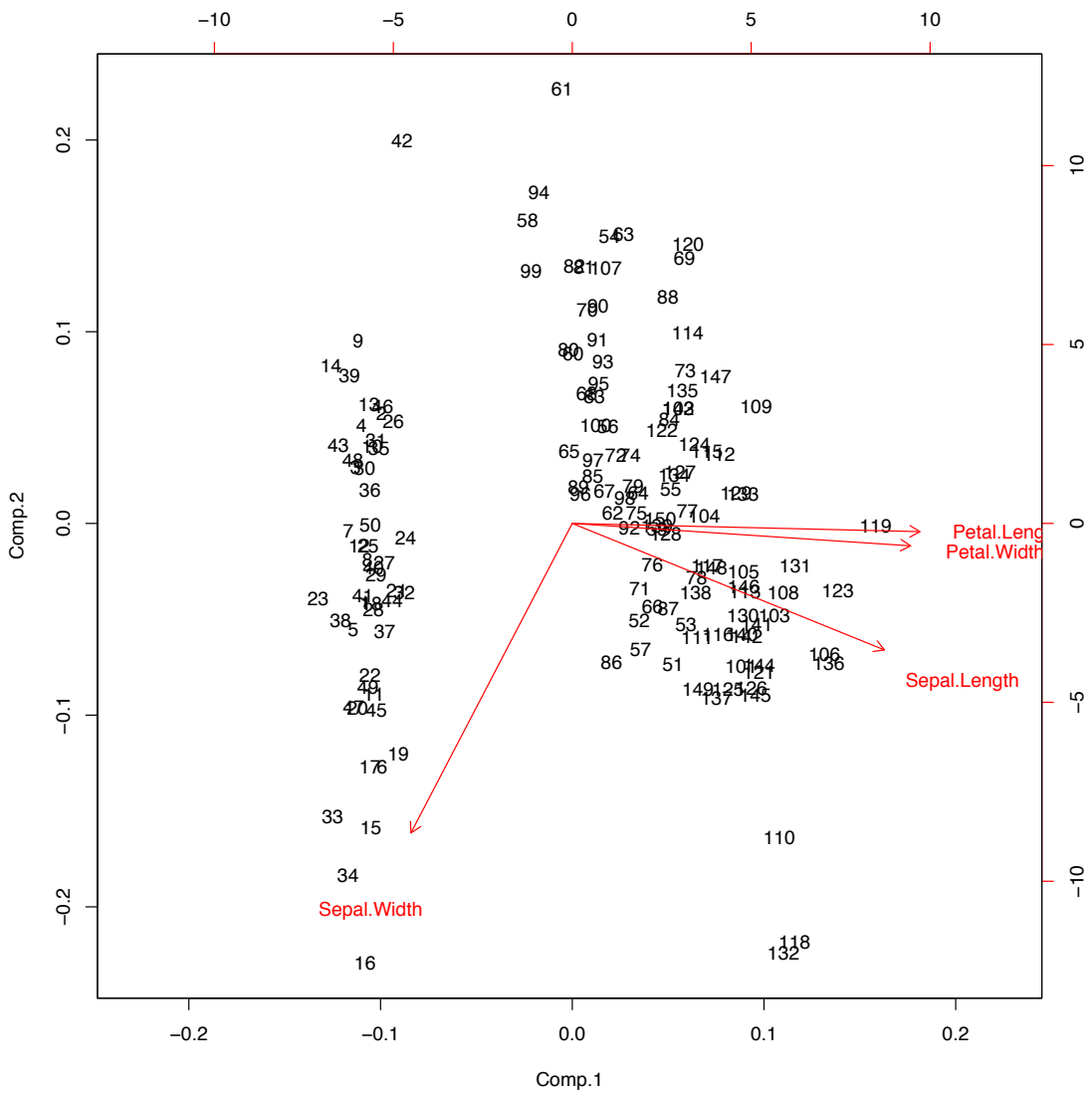
Each measuring the length and widths of both sepal and petals

Collected by E. Anderson (1935) and analysed by R.A. Fisher (1936)



Using again function `princomp` and `biplot`.

```
iris1 <- iris
iris1 <- iris1[,-5]
biplot(princomp(iris1,cor=T))
```

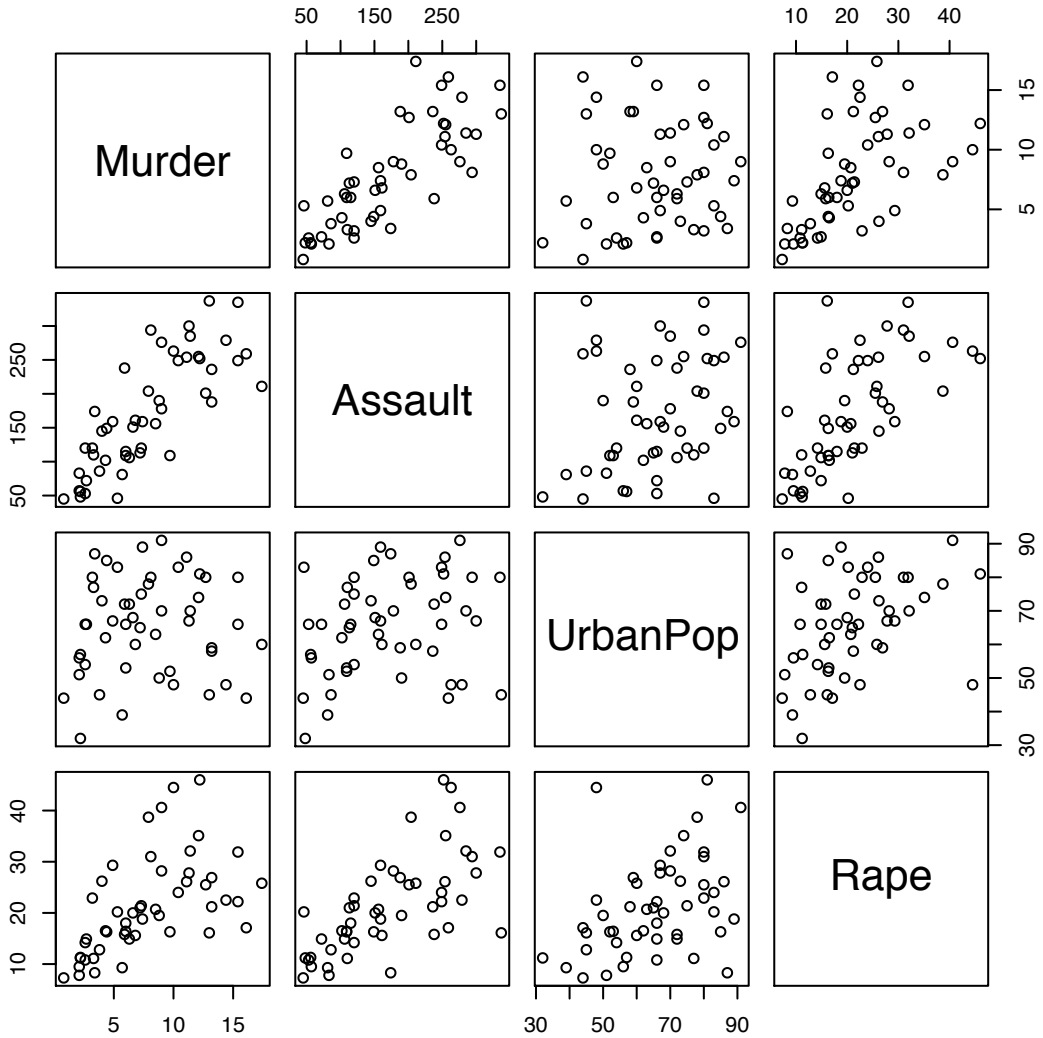

Comp.1
Comp.2
Petal.Leng
Petal.Width
Sepal.Length
Sepal.Width

Biplot Example 2: US Arrests

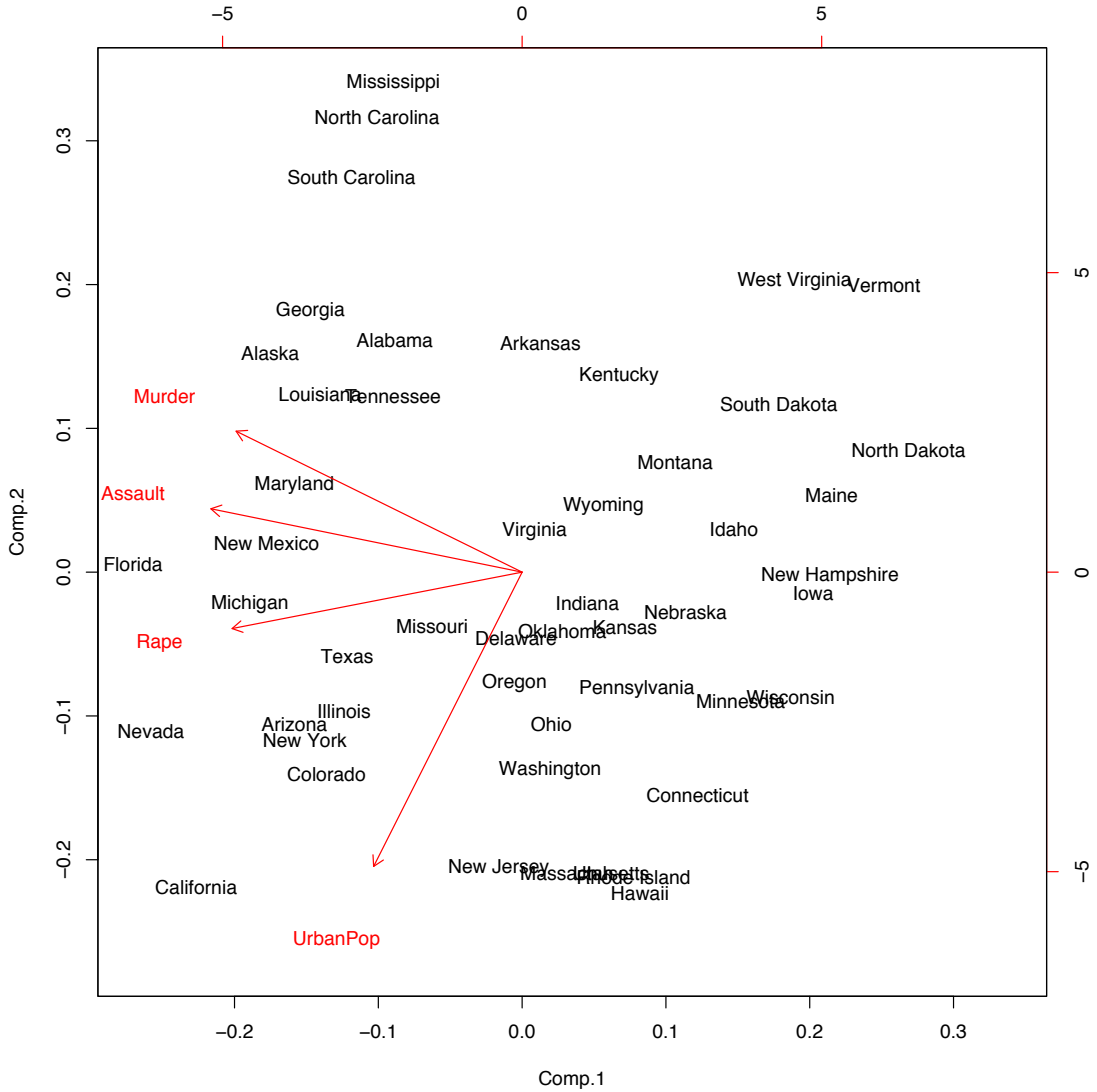
This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

```
pairs(USArrests)
usarrests.pca <- princomp(USArrests,cor=T)
plot(usarrests.pca)

pairs(predict(usarrests.pca))
biplot(usarrests.pca)
```

Pairs Plot: US Arrests



Biplot Example 2: US Arrests



Biplot Example 3: US State data

This data set contains statistics like illiteracy and life expectancy on 50 US states.

```
data(state)                          ## load state data
state <- state.x77[, 2:7]            ## extract useful info
row.names(state)<-state.abb
state[1:5,]                          ## lets have a look
```

```
   Income Illiteracy Life Exp Murder HS Grad Frost
AL   3624        2.1    69.05   15.1    41.3    20
AK   6315        1.5    69.31   11.3    66.7   152
AZ   4530        1.8    70.55    7.8    58.1    15
AR   3378        1.9    70.66   10.1    39.9    65
CA   5114        1.1    71.71   10.3    62.6    20
```

```
## calculate the pc's of the data and show biplot
state.pca <- princomp(state,cor=TRUE)
biplot(state.pca,pc.biplot=TRUE,cex=0.8,font=2,expand=0.9)
```

Biplot: US States