MS1b Statistical Data Mining

Yee Whye Teh
Department of Statistics
Oxford

Outline

Outline

Administrivia

Lectures

- Wednesdays 1100-1200, Weeks 1-8.
- Thursdays 1100-1200, Weeks 1,3,5,7.

Part C students

- Practical classes: Wednesdays 1100-1200, Weeks 2,4,6,8.
- Problem classes: Thursday ????-????, Weeks 2-8.

MSc students

- Practical class: Friday afternoon, Week 3.

Syllabus I

Part I: Dimensionality Reduction

- Principal Components Analysis
- Multidimensional Scaling
- Isomap
- Hierarchical clustering

Part II: Clustering

- K-means
- Vector Quantization
- Self Organizing Maps
- Mixture Models

Part II: Supervised Learning

- Empirical Risk Minimization
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes

Syllabus II

- Bayesian Methods
- Logistic Regression

Part III: Supervised Learning

- Nearest Neighbours, Prototype Based Methods
- Classification and Regression Trees
- Neural Networks (not in; lecture15)

Part IV: Ensemble Methods

- Bootstrap, Bagging
- Random Forests
- Boosting
- Dropout Neural Networks (not in)

R

- Learning how to use R for Data Mining

Outline

What is Data Mining?

Traditional Problems in Applied Statistics

Well formulated question that we would like to answer.
Expensive to gathering data and/or expensive to do computation.
Create specially designed experiments to collect high quality data.

Current Situation

Information Revolution

- improvements in data-storage devices (both larger and cheaper).
- powerful data capturing devices (microphones, cameras, satellites).

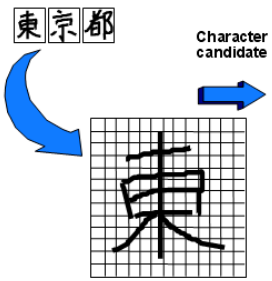
$\rightarrow$ lots of data with potentially valuable information available.
$\rightarrow$ Big Data....

What is Data Mining?

- To gain insight from secondary data possibly without a specific goal in mind.
- Often working with huge datasets.
  - Typically many variables (up to thousands or millions).
  - Often, but not always many observations (dozens to millions).

Applications of Data Mining

- Pattern Recognition



  - Sorting Cheques
  - Reading License Plates
  - Sorting Envelopes
  - Eye/ Face/ Fingerprint Recognition

Image data contain a lot of structure. Data mining usually refers to making sense of less structured data.

Applications of Data Mining

- Business applications
  - Help companies intelligently find information
  - Credit scoring
  - Predict which products people are going to buy
  - Recommender systems
  - Autonomous trading
- Scientific applications
  - Predict cancer occurence/type and health of patients
  - Make sense of complex physical models

...It is just a nice name for multivariate statistics (‘minus model checking’).

What Wal-Mart Knows About Customers' Habits

By CONSTANCE L. HAYS

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's computer network, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

The experts mined the data and found that the stores would indeed need certain products - and not just the usual flashlights. "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane," Ms. Dillman said in a recent interview. "And the pre-hurricane top-selling item was beer."

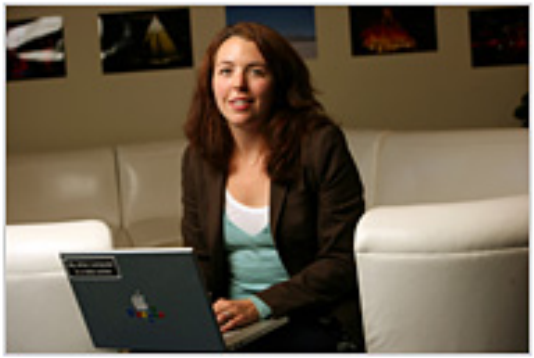
Thanks to those insights, trucks filled with toaster pastries and six-packs were soon speeding down Interstate 95 toward Wal-Marts in the path of Frances. Most of the products that were stocked for the storm sold quickly, the company said.

Full NY Times article on http://snipurl.com/ac5hc.

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.



Thor Swift for The New York Times

Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession

NYTimes article on use of R for data mining and data analysis:
snipurl.com/badgw

R, the Software, Finds Fans in Data Analysts - NYTimes.com

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

But R has also quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use.

"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon, a research scientist at Google, which uses the software widely. "It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

It is also free. R is an open-source program, and its popularity reflects a shift in the type of software used inside corporations. Open-source software is free for anyone to use and modify. I.B.M., Hewlett-Packard and Dell make billions of dollars a year selling servers that run the open-source Linux operating system, which competes with Windows from Microsoft. Most Web sites are displayed using an open-source application called Apache, and companies increasingly rely on the open-source MySQL database to store their critical information. Many people view the end results of all this technology via the Firefox Web browser, also open-source software.

Types of Data Mining

Unsupervised Learning

'Unclassified' data from which we would like to uncover hidden 'structure' or groupings

- Given detailed phone usage from many people, find interesting groups of people with similar behaviour.
- Shopping habits for people using loyalty cards: find groups of 'similar' shoppers.
- Given expression measurements of 1000s of genes for 100s of patients, find groups of functionally similar genes.

Goal: Hypothesis generation

Types of Data Mining

Supervised Learning

A database of 'classified' examples with predefined groupings

- Given detailed phone usage of many users *along with their historic churn*, predict when/if people are going to change contracts again.
- Given expression measurements of 1000s of genes for 100s of patients *along with a binary variable indicating absence or presence of a specific cancer*, predict if the cancer is present for a new patient.
- Given expression measurements of 1000s of genes for 100s of patients *along with survival length*, predict survival time.

Goal: Prediction.

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

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

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 lobe 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').

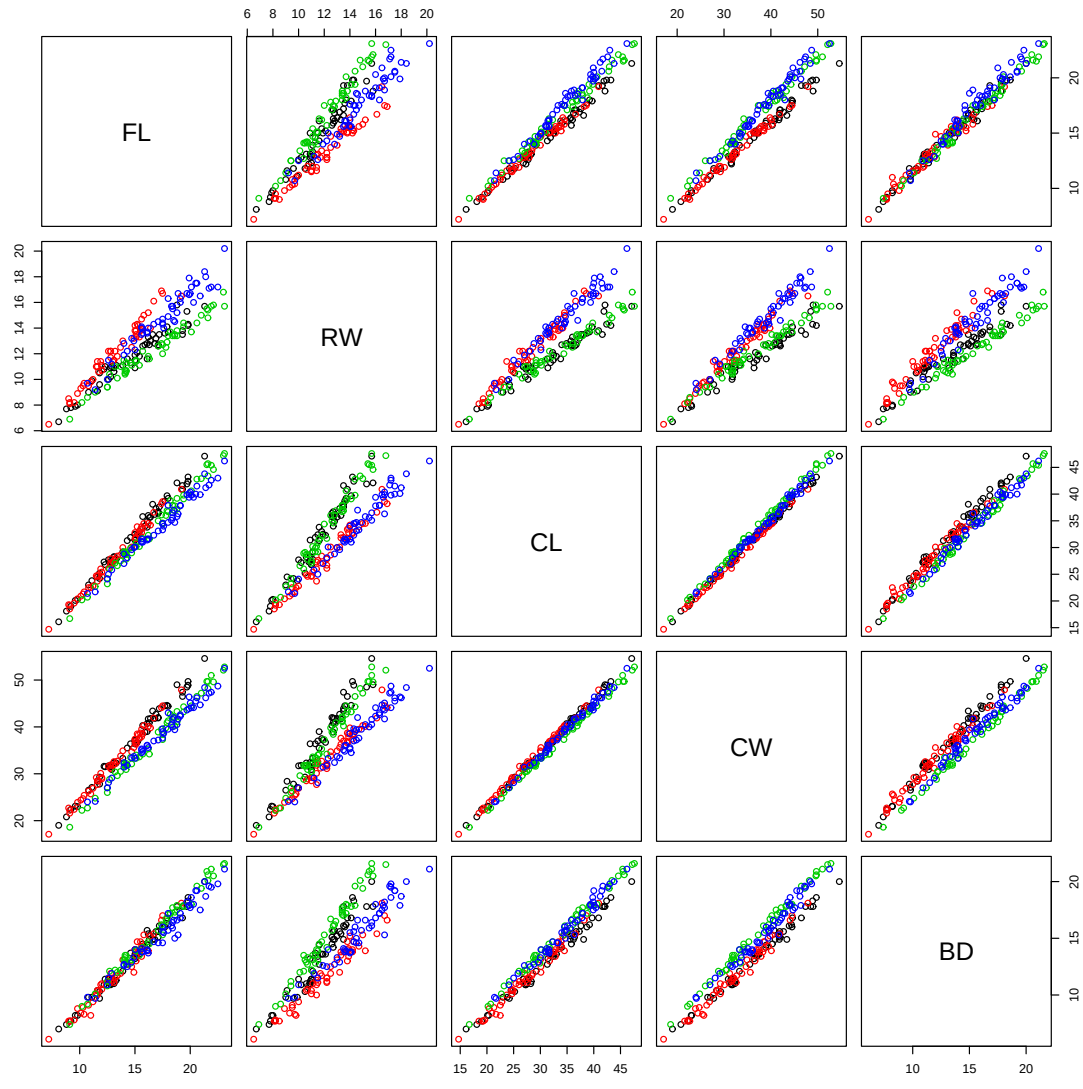
```
## load package MASS containing the data
library(MASS)
## look at data
crabs
```

```
     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
...
```

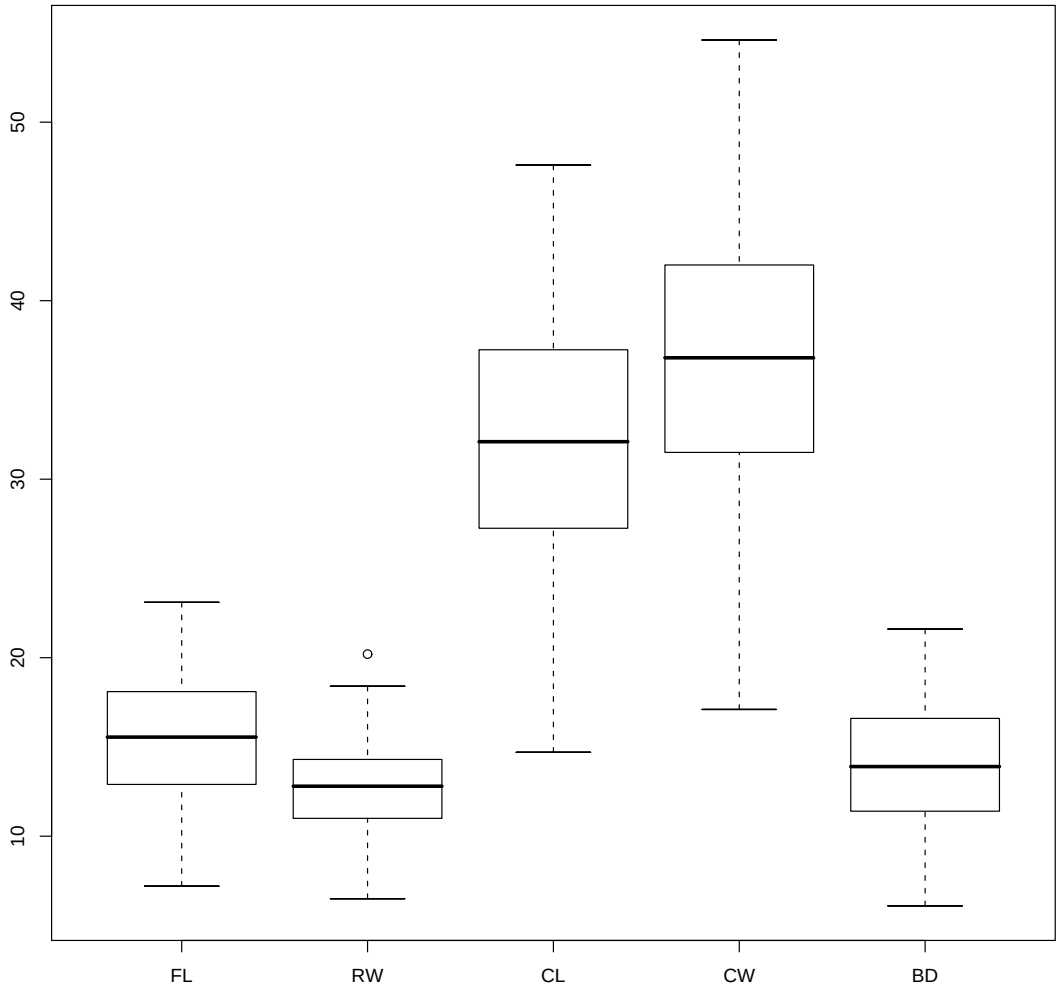
R code

```
## 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)
```

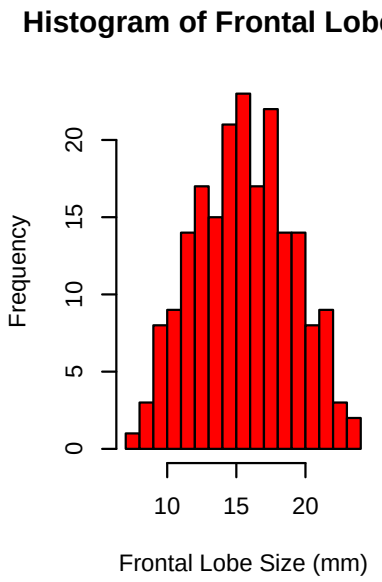
Simple Pairwise Scatterplots

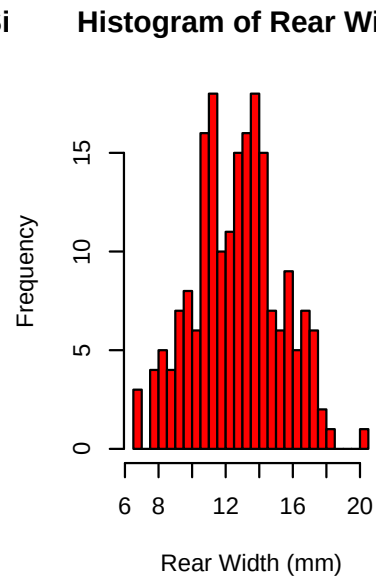


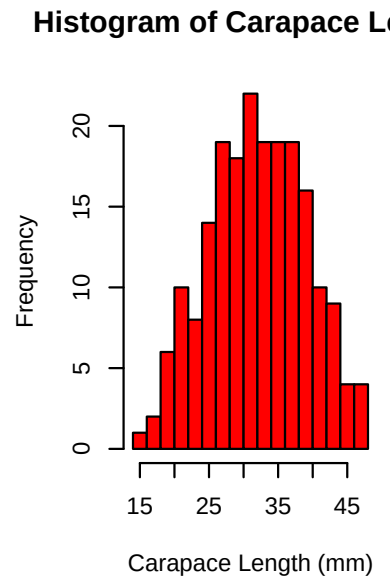
Univariate Boxplots

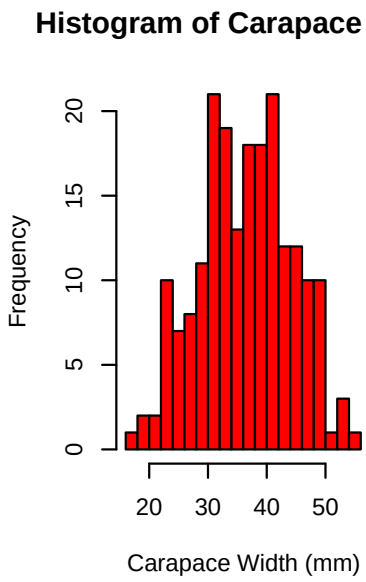


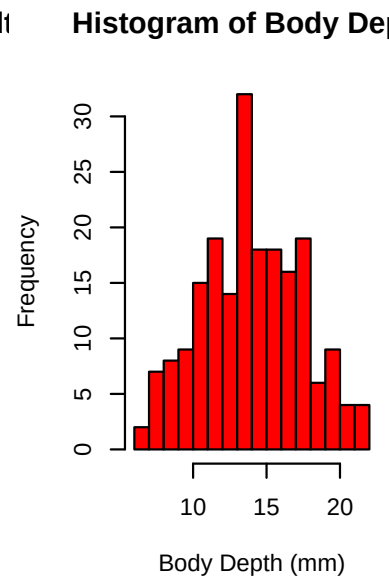
Univariate Histograms



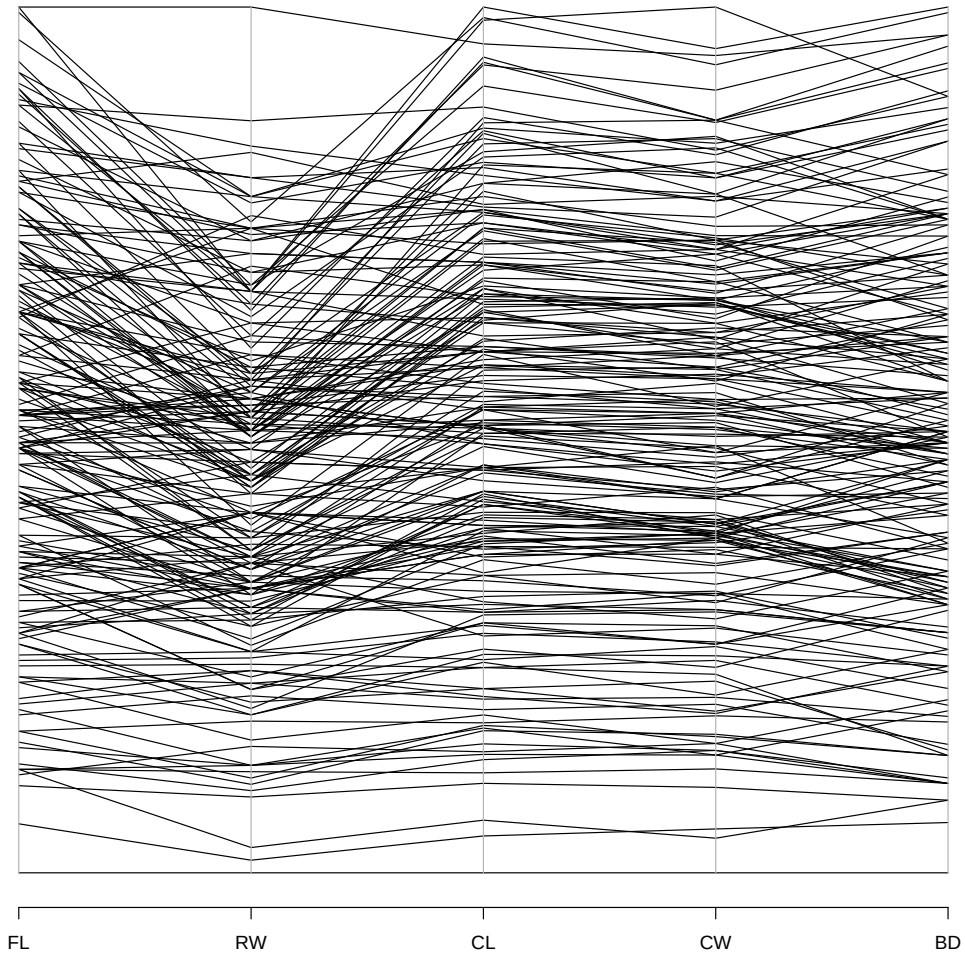








Parallel Coordinate Plots



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

Overview of 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.

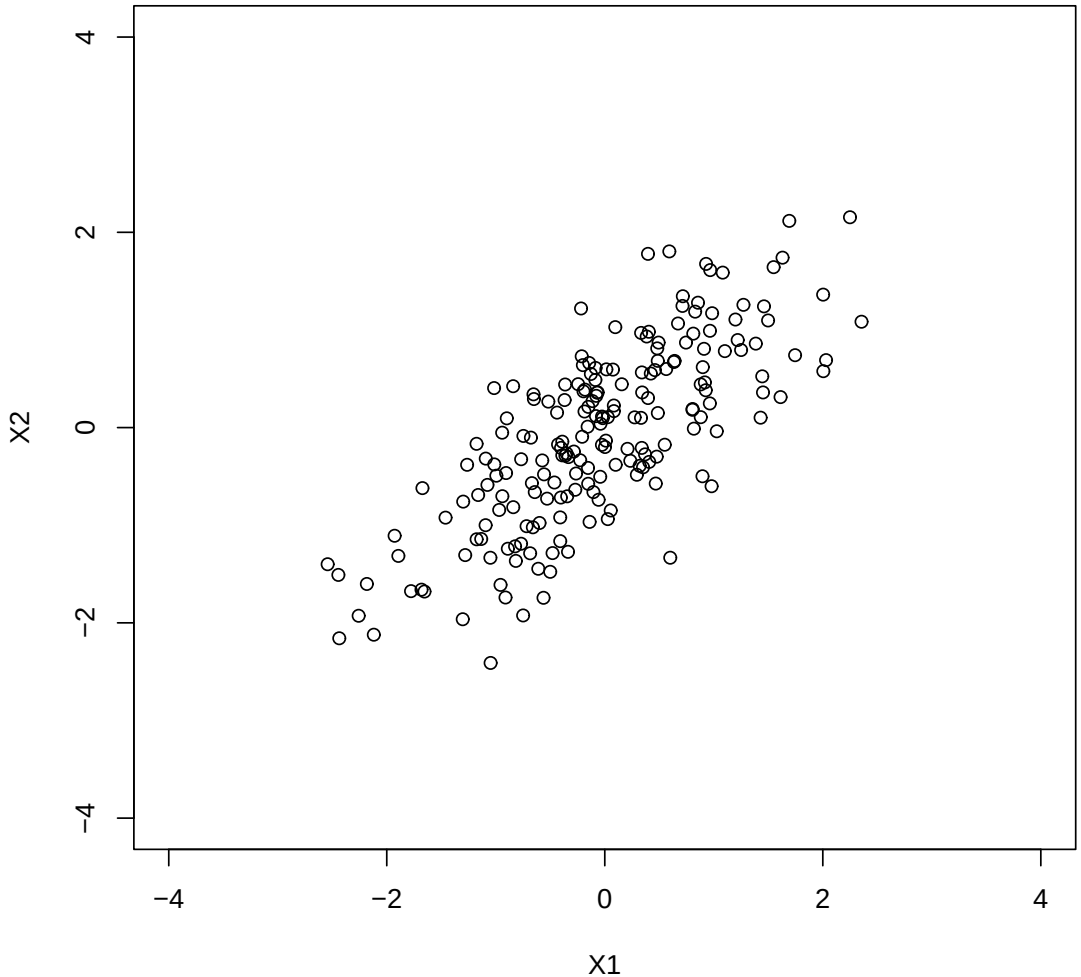
Overview of PCA

- 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

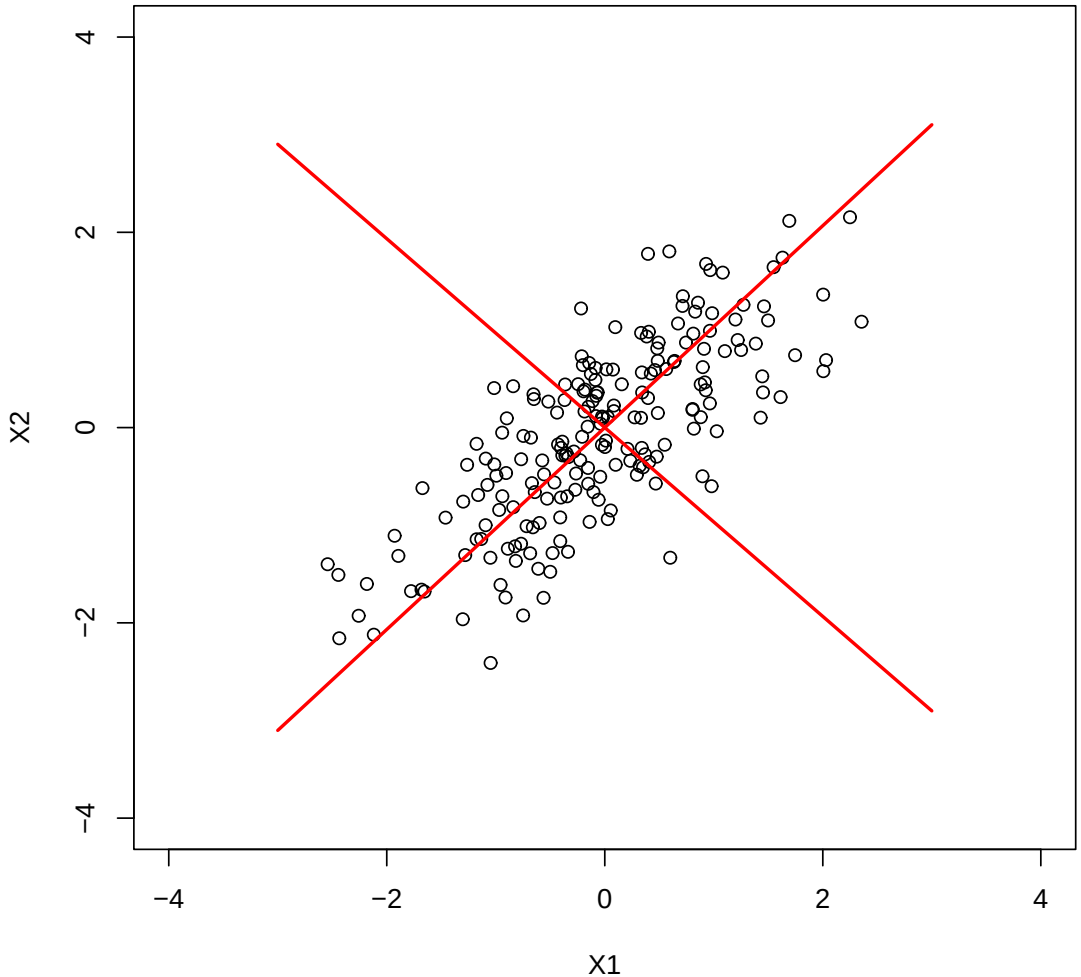
$$Z_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p = X^T \mathbf{a}_1$$

  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.

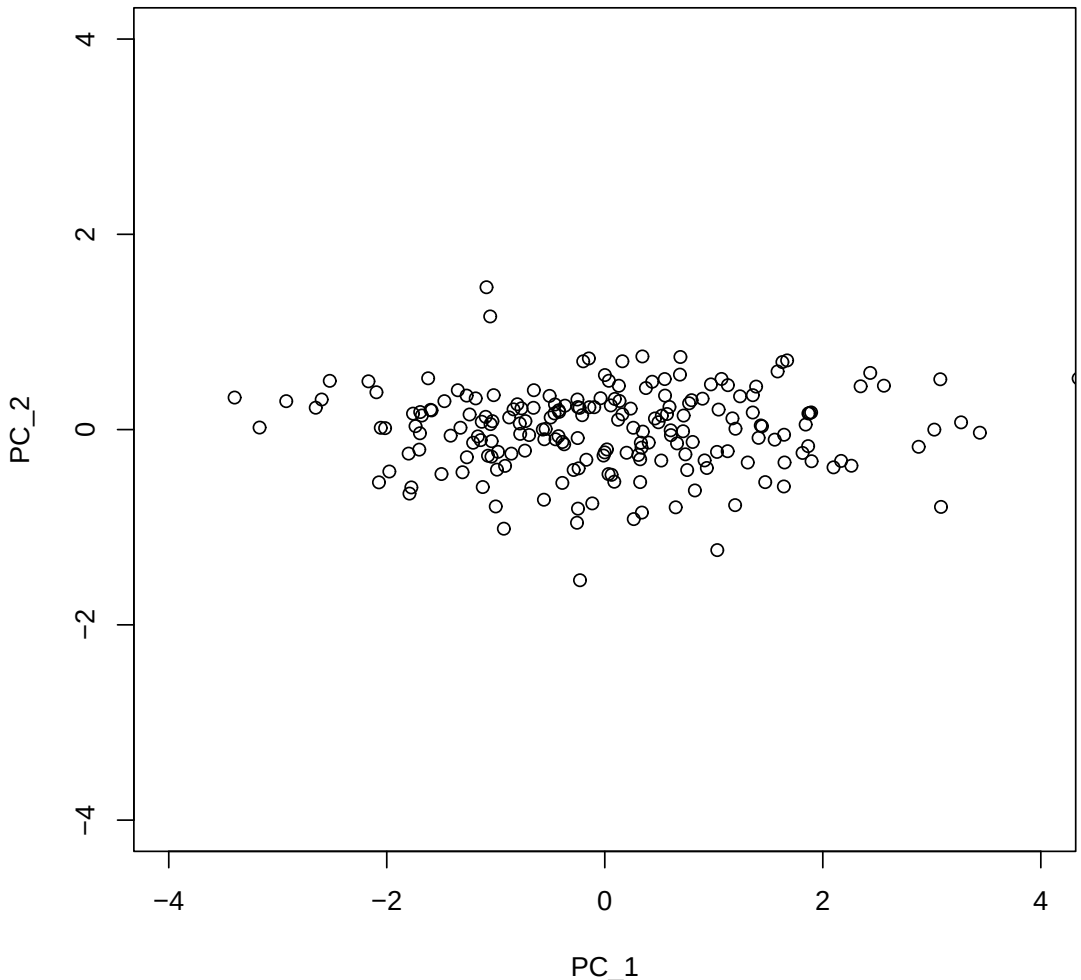
Principal Components Analysis



Principal Components Analysis



Principal Components Analysis



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 $S = n^{-1}\mathbf{X}^T\mathbf{X}$ 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

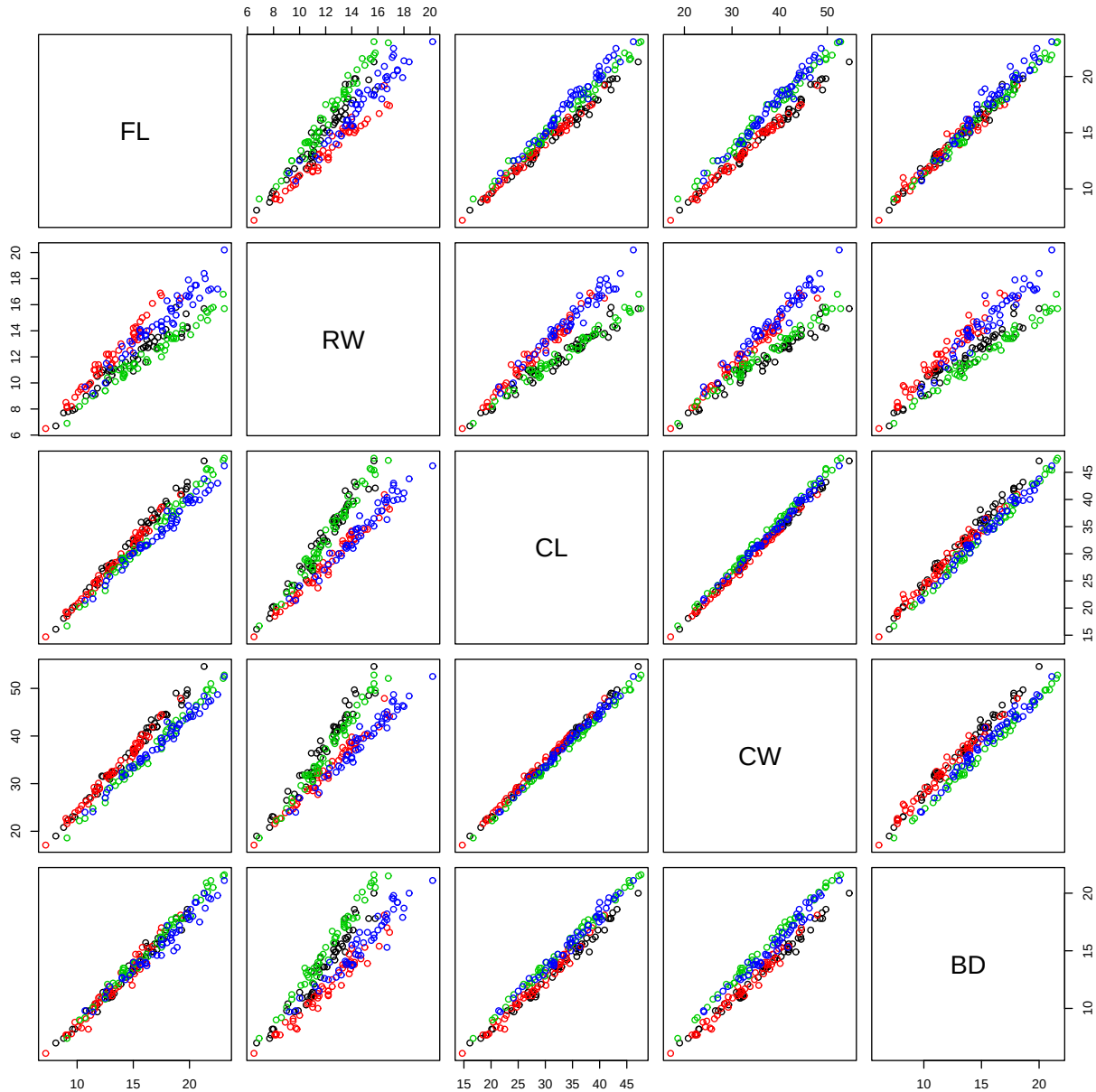
This is what we have had before:

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

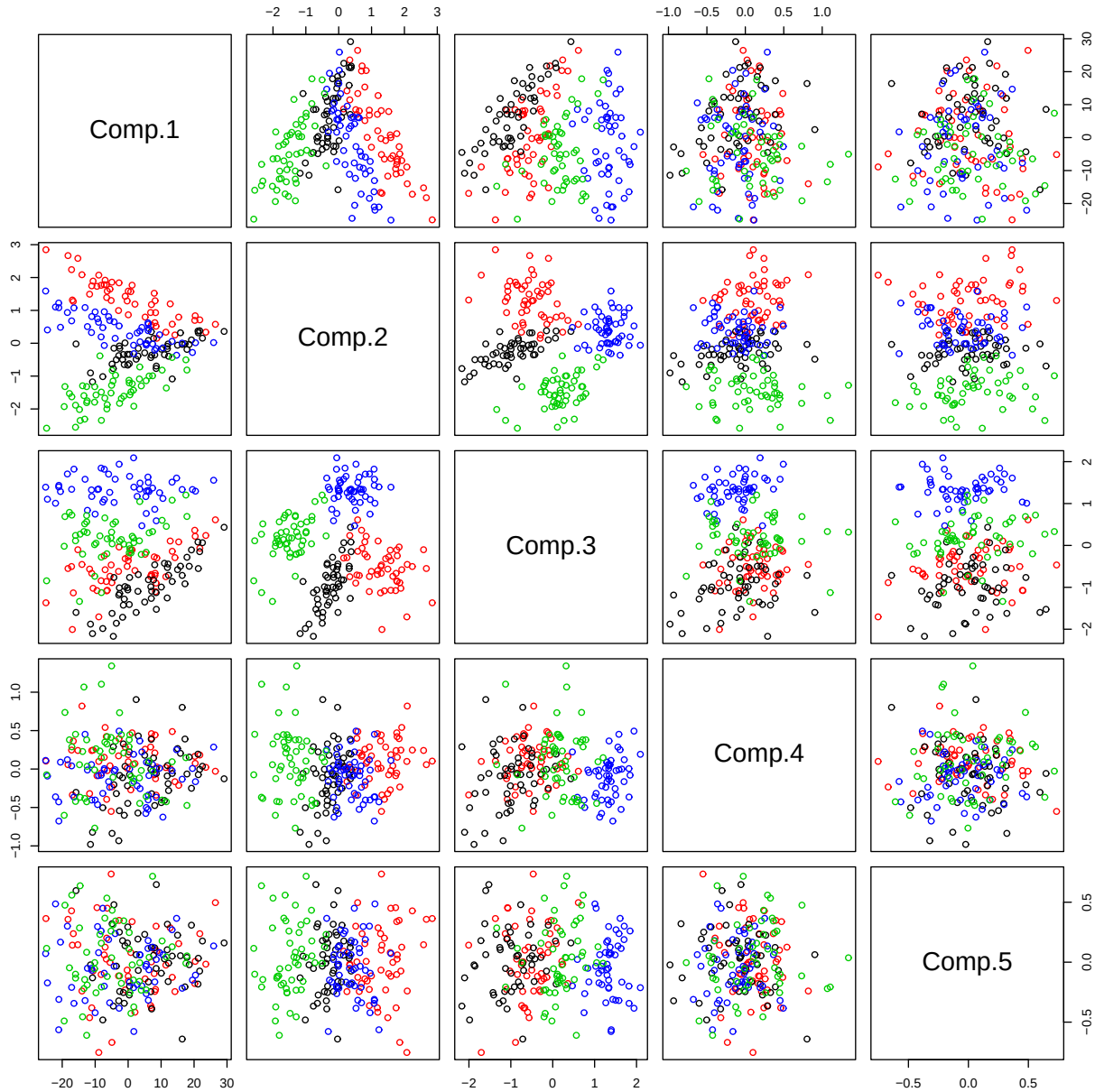
Now perform PCA analysis with function `princomp`.
Alternatively, use `eigen` or `svd` instead.

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

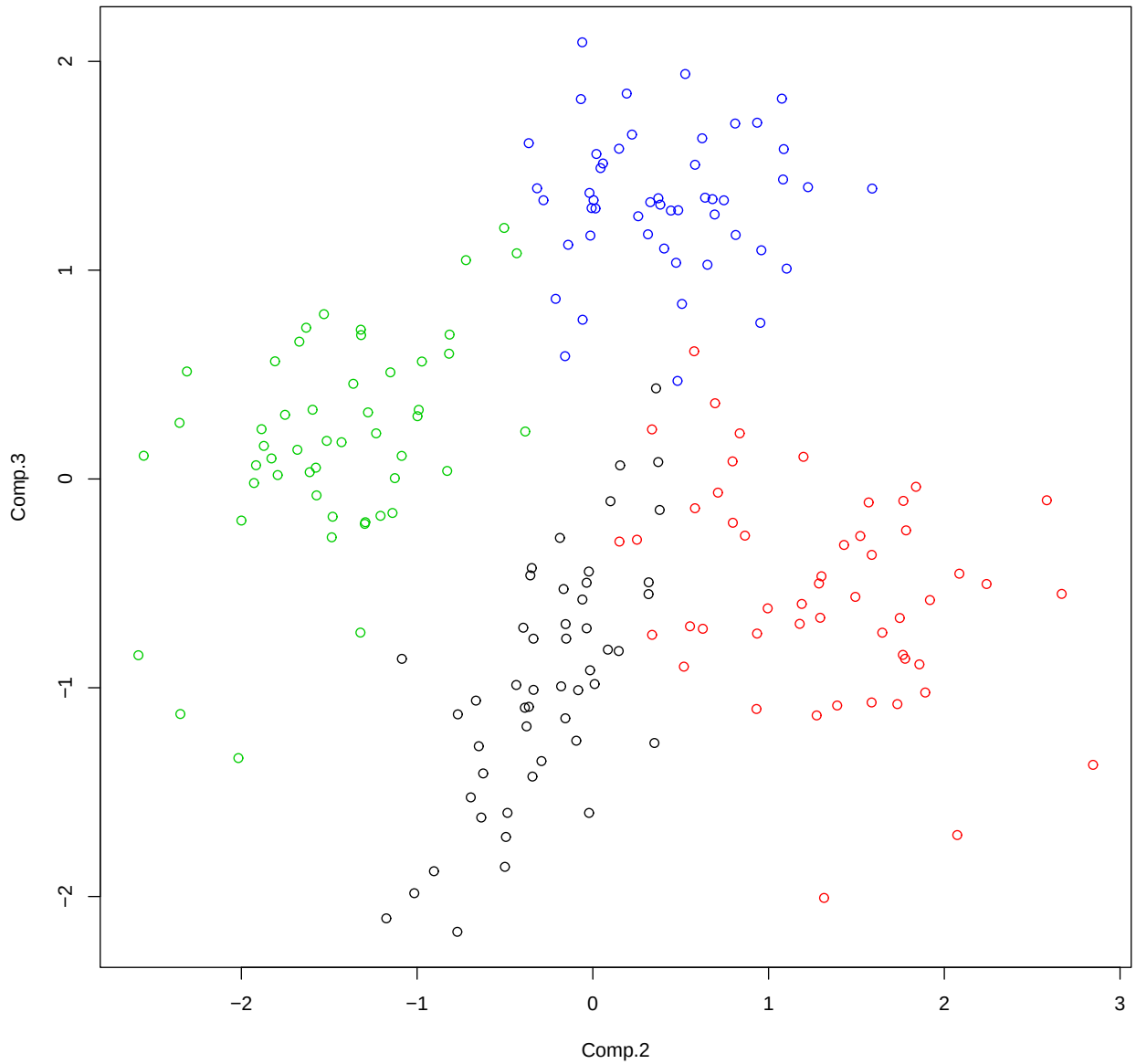
PCA Example 1: Original crabs data



PCA Example 1: Rotated crabs data



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

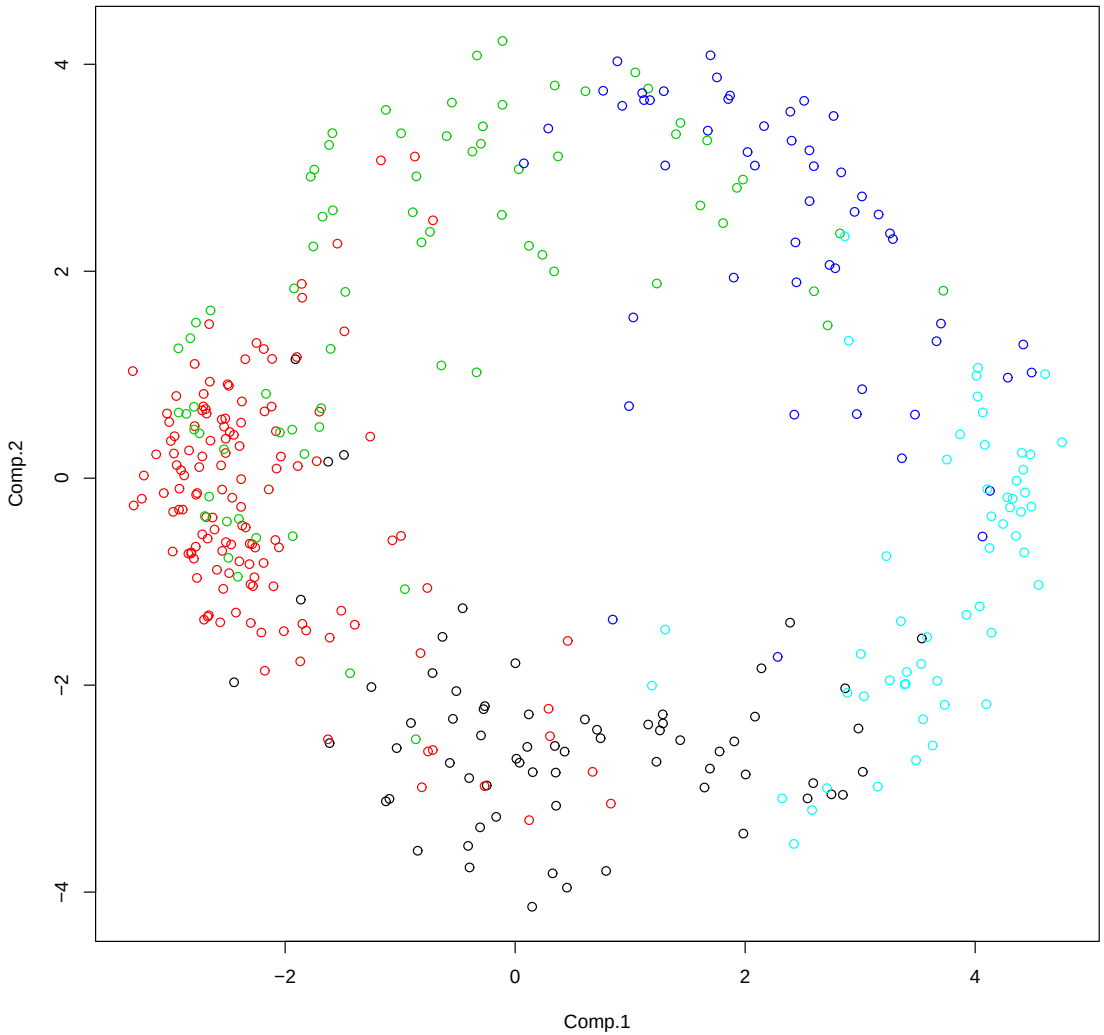


PCA Example 2: Yeast Cell Cycle Data ($n = 384$, $p = 17$)

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

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.
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.

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,

- 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,

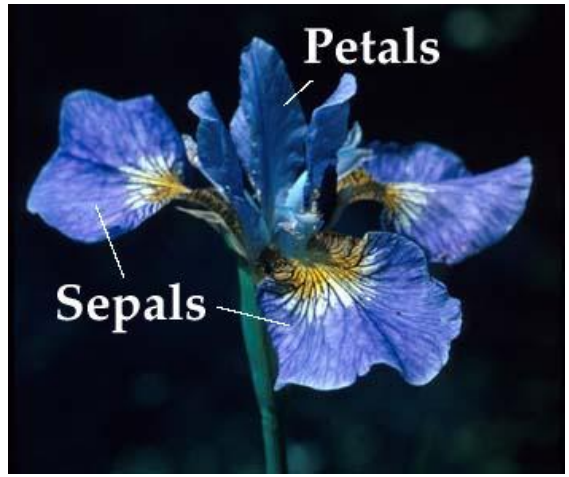
However, the quality of this image depends on the proportion of variance explained by the PCs used

Biplot Example 1: Fisher's Iris Data

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

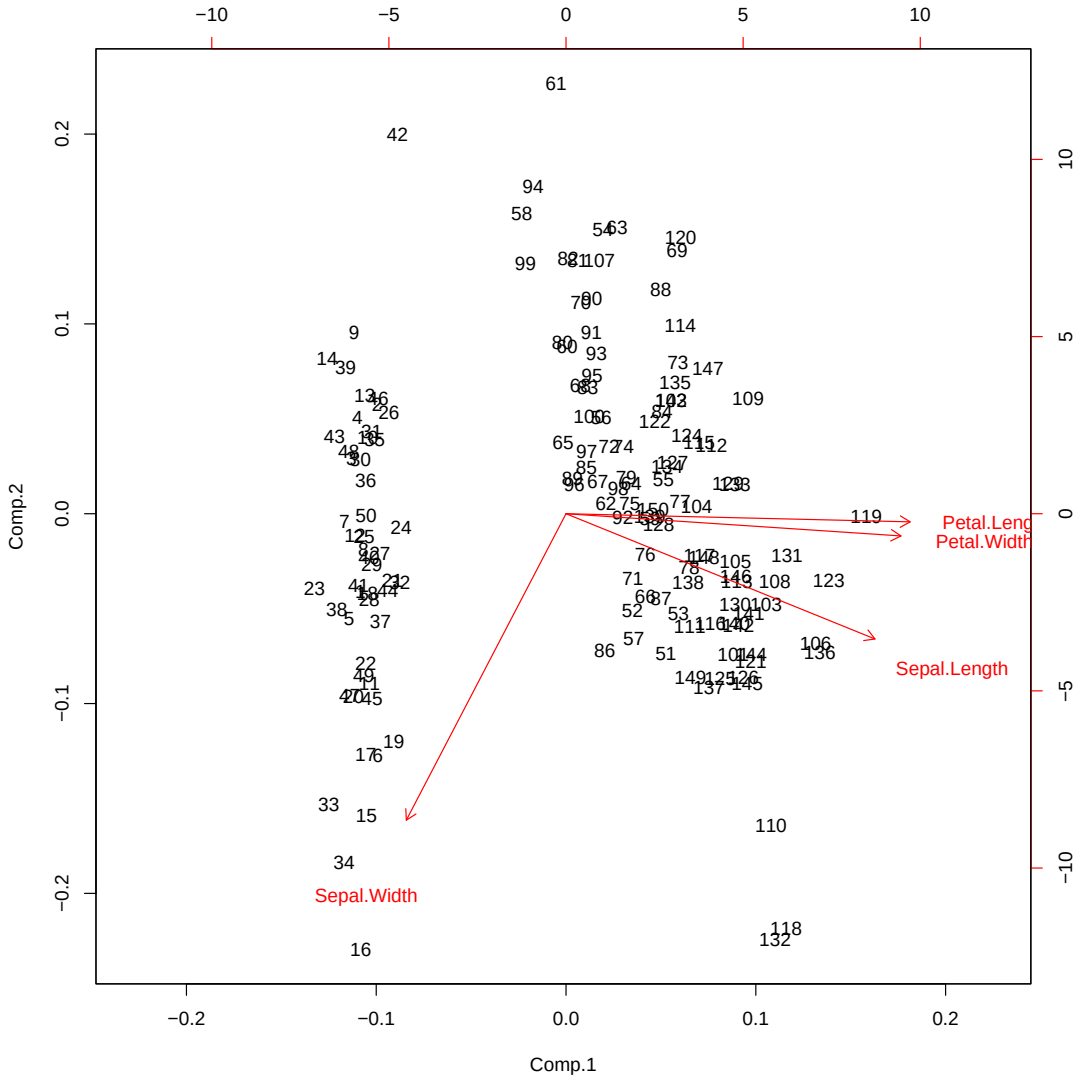
Each measuring the length and widths of both sepal and petals

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))
```

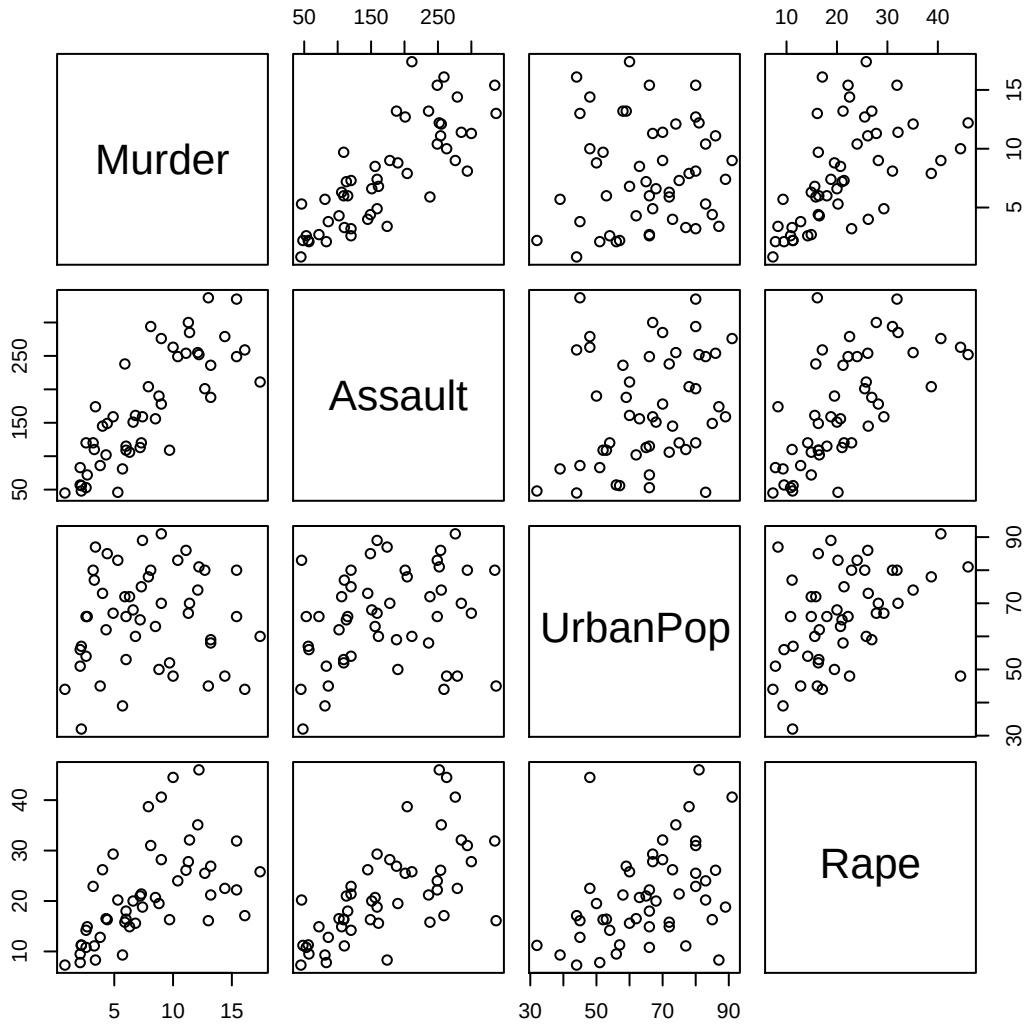

Comp.1
Comp.2
Petal.Leng
Petal.Width
Sepal.Length
Sepal.Width

Biplot Example 2: US Arrests

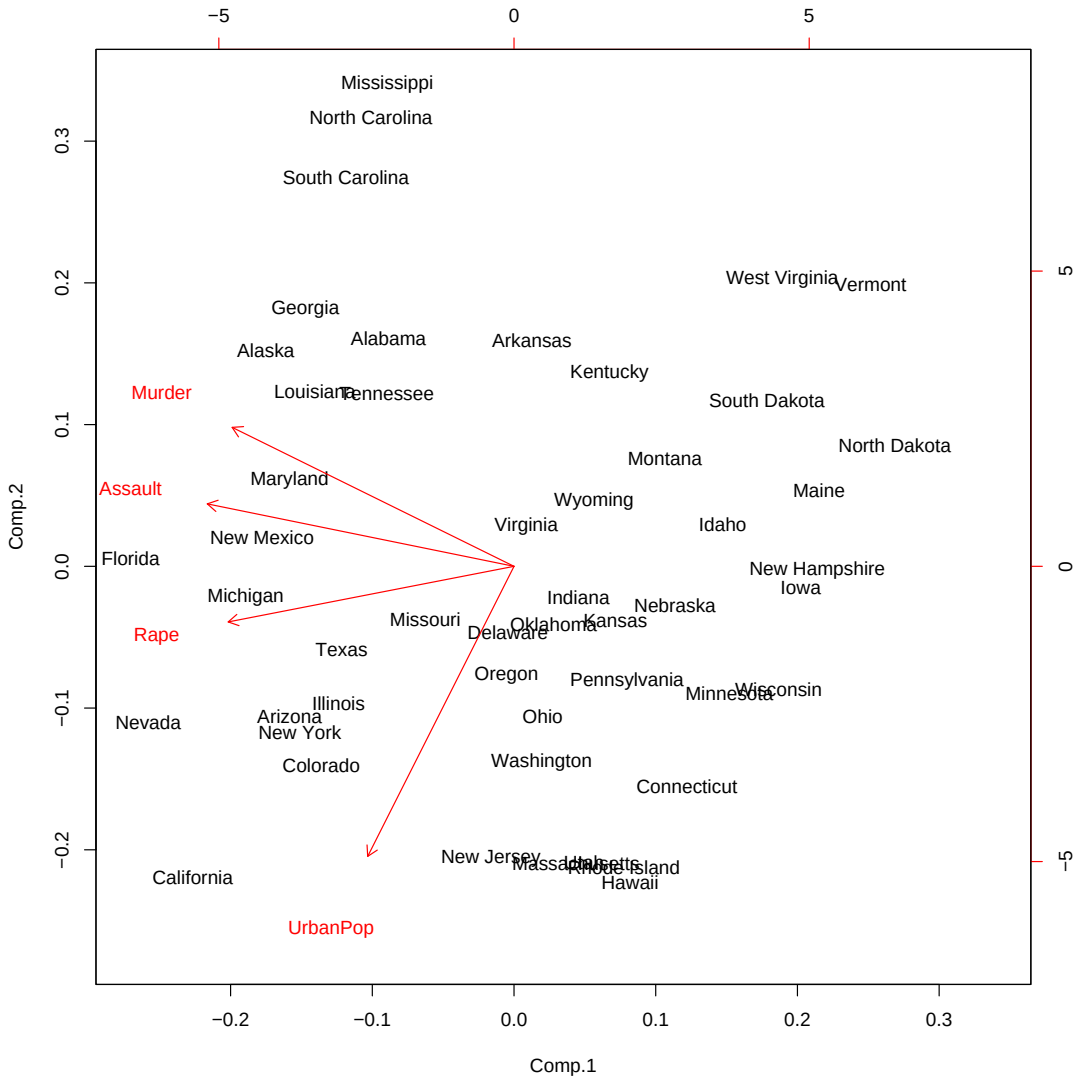
This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

```
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)
```

Pairs Plot: US States