MS1b
Statistical Machine Learning and Data Mining

Yee Whye Teh
Department of Statistics
Oxford

http://www.stats.ox.ac.uk/~teh/smldm.html

Course Information

- Course webpage:
  http://www.stats.ox.ac.uk/~teh/smldm.html
- Lecturer: Yee Whye Teh
- TA for Part C: Thibaut Lienant
- TA for MSc: Balaji Lakshminarayanan and Maria Lomeli
- Please subscribe to Google Group:
  https://groups.google.com/forum/?hl=en-GB#!forum/smldm
- Sign up for course using sign up sheets.

Course Structure

Lectures

- 1400-1500 Mondays in Math Institute L4.
- 1000-1100 Wednesdays in Math Institute L3.

Part C:

- 6 problem sheets.
- Classes: 1600-1700 Tuesdays (Weeks 3-8) in 1 SPR Seminar Room.
- Due Fridays week before classes at noon in 1 SPR.

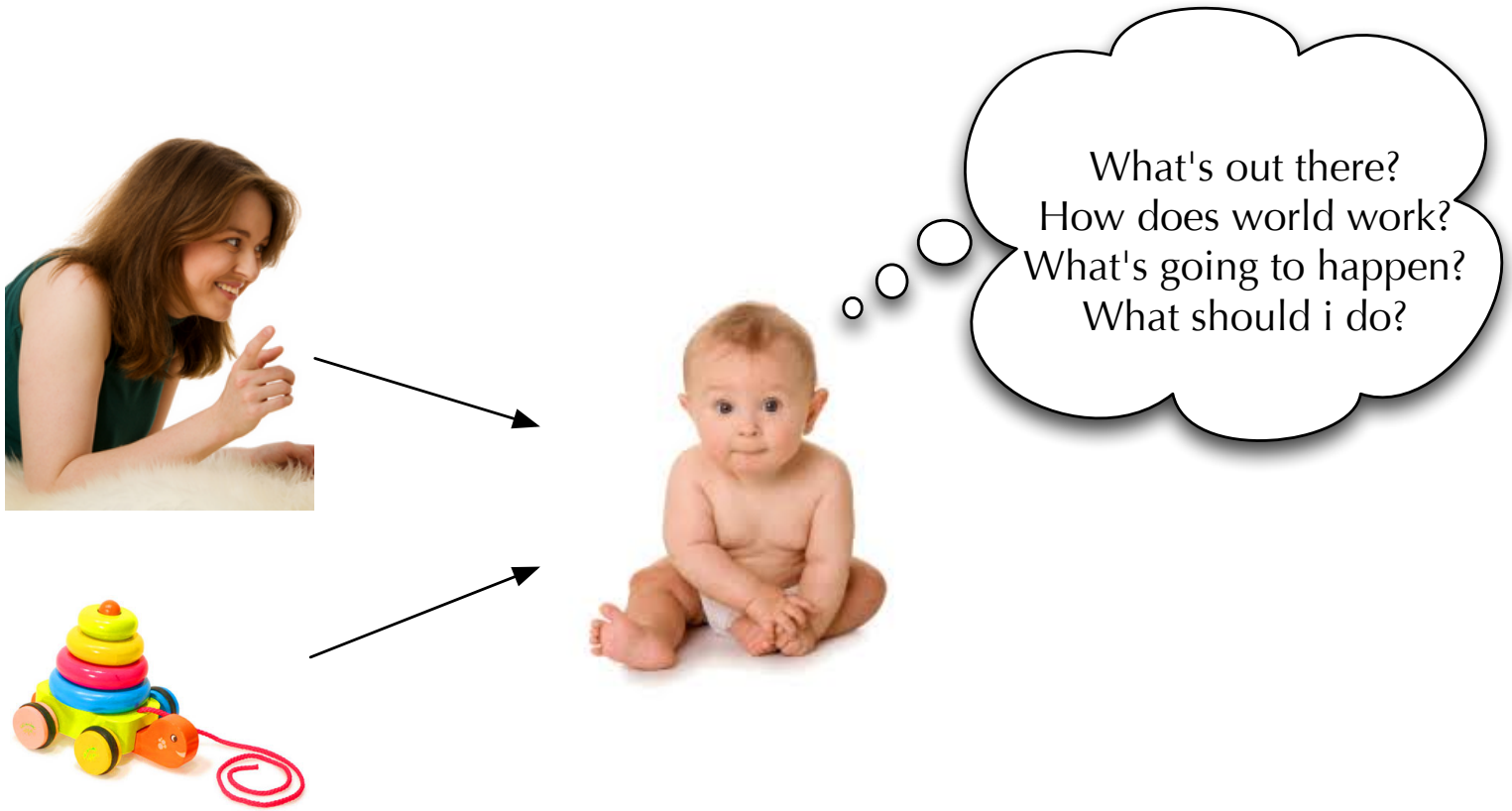
MSc:

- 4 problem sheets.
- Classes: Tuesdays (Weeks 3, 5, 7, 9) in 2 SPR Seminar Room.
- Group A: 1400-1500, Group B: 1500-1600.
- Due Fridays week before classes at noon in 1 SPR.
- Practical: Week 5 and 7 (assessed) in 1 SPR Computing Lab.
- Group A: 1400-1600, Group B: 1600-1800.

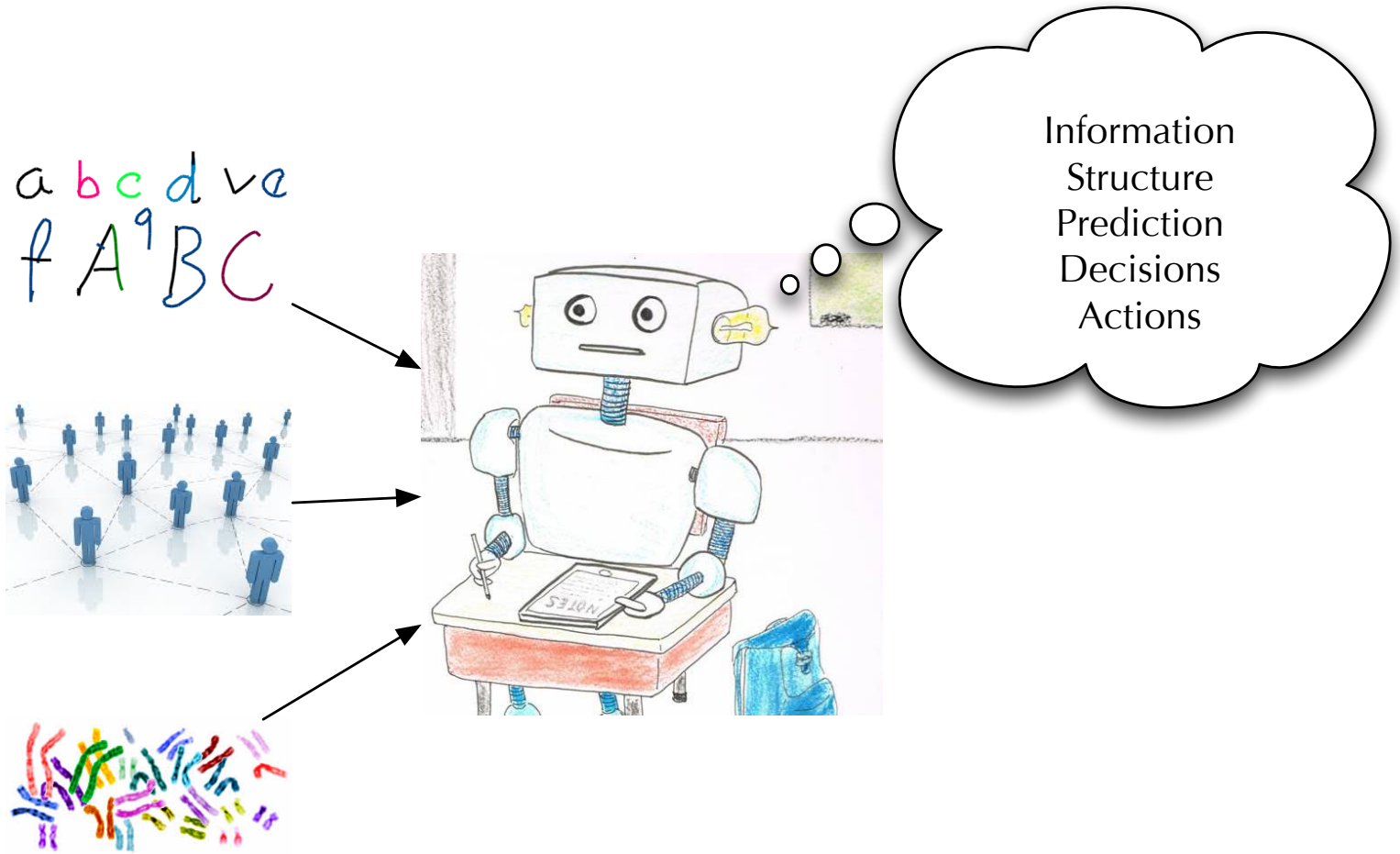
Course Aims

1. Have ability to use the relevant R packages to analyse data, interpret results, and evaluate methods.
2. Have ability to identify and use appropriate methods and models for given data and task.
3. Understand the statistical theory framing machine learning and data mining.
4. Able to construct appropriate models and derive learning algorithms for given data and task.

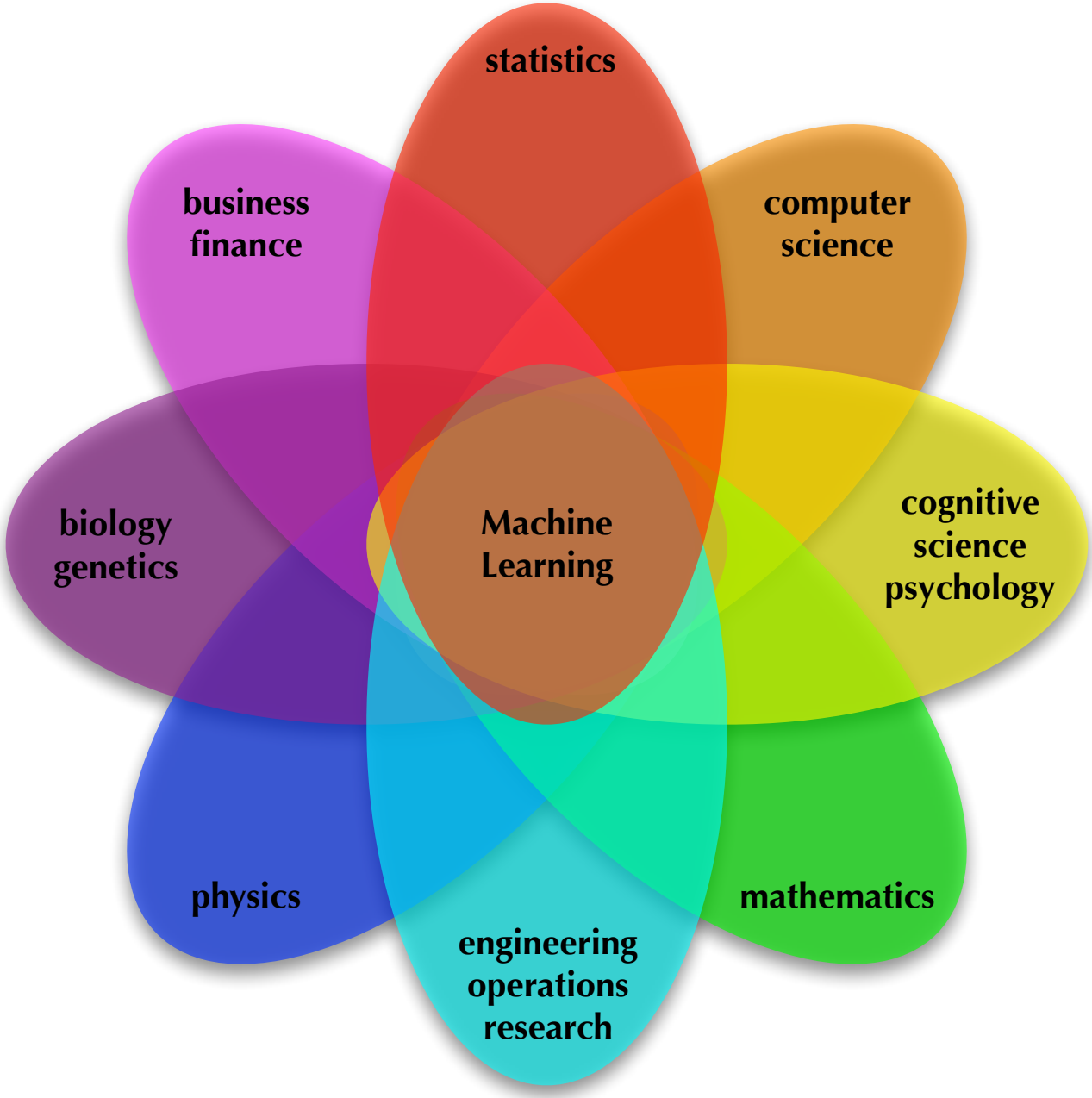
What is Machine Learning?



What is Machine Learning?



What is Machine Learning?



What is the Difference?

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 computers and data storage devices.
- Powerful data capturing devices.
- Lots of data with potentially valuable information available.

What is the Difference?

Data characteristics

- Size
- Dimensionality
- Complexity
- Messy
- Secondary sources

Focus on generalization performance

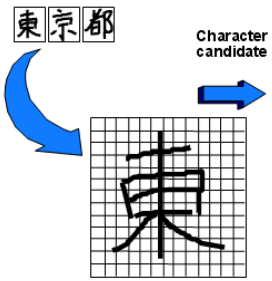
- Prediction on new data
- Action in new circumstances
- Complex models needed for good generalization.

Computational considerations

- Large scale and complex systems

Applications of Machine Learning

- Pattern Recognition



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

Applications of Machine Learning

- 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/personalized health
  - Make sense of complex physical, biological, ecological, sociological models

Further Readings, News and Applications

Links are clickable in pdf. More recent news posted on course webpage.

- Leo Breiman: Statistical Modeling: The Two Cultures
- NY Times: R
- NY Times: Career in Statistics
- NY Times: Data Mining in Walmart
- NY Times: Big Data's Impact In the World
- Economist: Data, Data Everywhere
- McKinsey: Big data: The Next Frontier for Competition
- NY Times: Scientists See Promise in Deep-Learning Programs
- New Yorker: Is "Deep Learning" a Revolution in Artificial Intelligence?

Types of Machine Learning

Unsupervised Learning

Uncover structure hidden in ‘unlabelled’ data.

- Given network of social interactions, find communities.
- Given shopping habits for people using loyalty cards: find groups of ‘similar’ shoppers.
- Given expression measurements of 1000s of genes for 1000s of patients, find groups of functionally similar genes.

Goal: Hypothesis generation, visualization.

Types of Machine Learning

Supervised Learning

A database of examples along with "labels" (task-specific).

- Given network of social interactions *along with their browsing habits*, predict what news might users find interesting.
- Given expression measurements of 1000s of genes for 1000s of patients *along with an indicator of 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 1000s of patients *along with survival length*, predict survival time.

Goal: Prediction *on new examples*.

Types of Machine Learning

Semi-supervised Learning

A database of examples, only a small subset of which are labelled.

Multi-task Learning

A database of examples, each of which has multiple labels corresponding to different prediction tasks.

Reinforcement Learning

An agent acting in an environment, given rewards for performing appropriate actions, learns to maximize its reward.

OxWaSP

Oxford-Warwick Centre for Doctoral Training in Statistics

- Programme aims to produce EuropeÕs future research leaders in statistical methodology and computational statistics for modern applications.
- 10 fully-funded (UK, EU) students a year (1 international).
- Website for prospective students.
- **Deadline: January 24, 2014**

Exploratory Data Analysis

Notation

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

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

- Denote the ith data item by $x_i \in \mathbb{R}^p$. (This is transpose of ith row of $\mathbf{X}$)
- Assume $x_1, \dots, x_n$ are *independently and identically distributed* samples of a *random vector* X over $\mathbb{R}^p$.

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 lobe `FL`,
- the rear width `RW`,
- the length along the carapace midline `CL`,
- the maximum width `CW` of the carapace, and
- the body depth `BD` in mm.

in addition to colour (species) and sex.

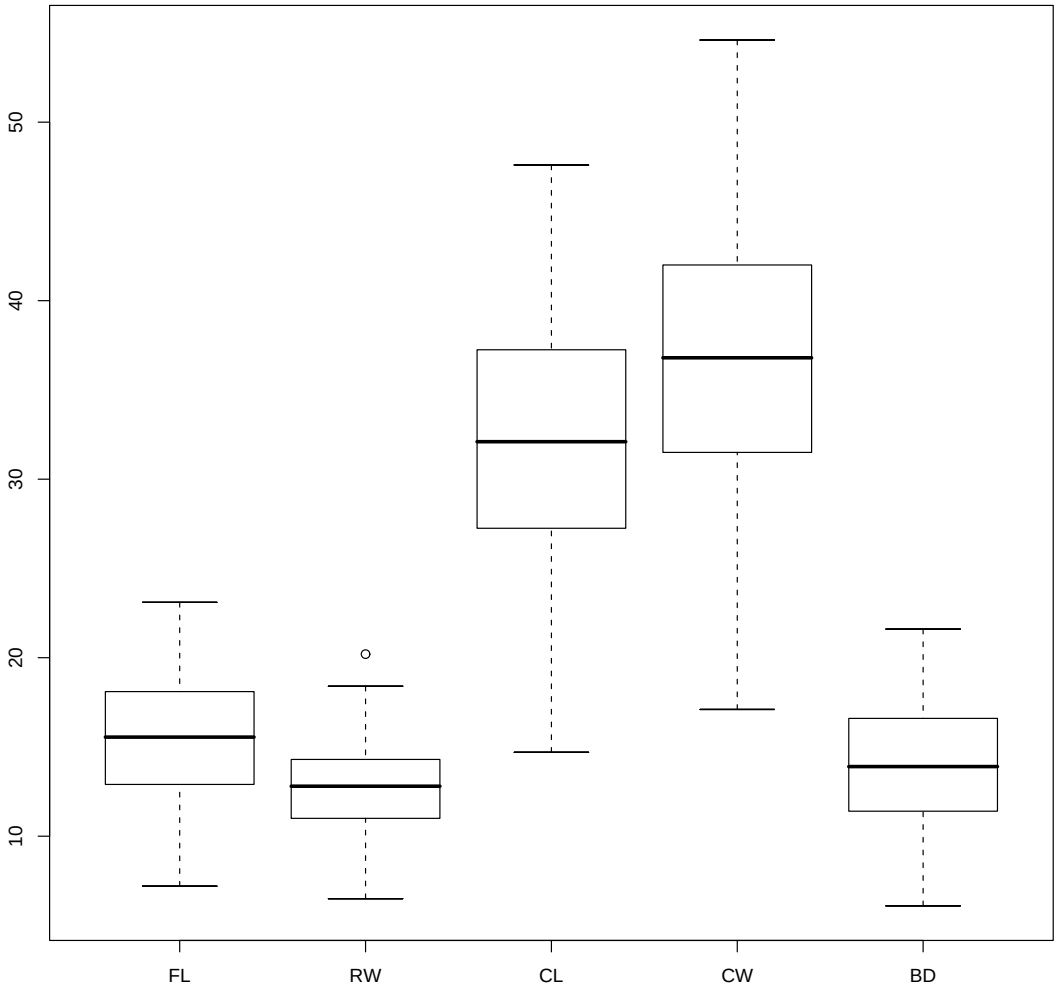
Crabs Data I

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

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

## various plots
boxplot(Crabs)
hist(Crabs$FL,col='red',breaks=20,xname='Frontal Lobe Size (mm)')
hist(Crabs$RW,col='red',breaks=20,xname='Rear Width (mm)')
hist(Crabs$CL,col='red',breaks=20,xname='Carapace Length (mm)')
hist(Crabs$CW,col='red',breaks=20,xname='Carapace Width (mm)')
hist(Crabs$BD,col='red',breaks=20,xname='Body Depth (mm)')
plot(Crabs,col=unclass(Crabs.class))
parcoord(Crabs)
```

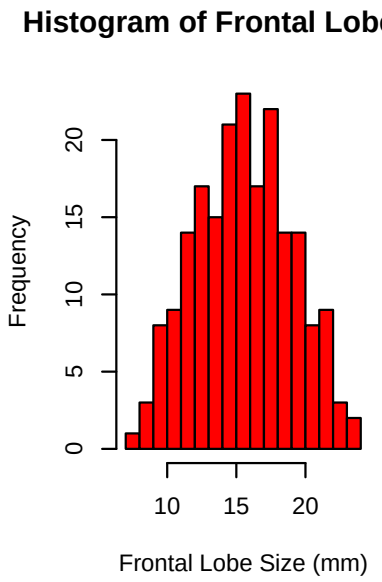
Crabs data

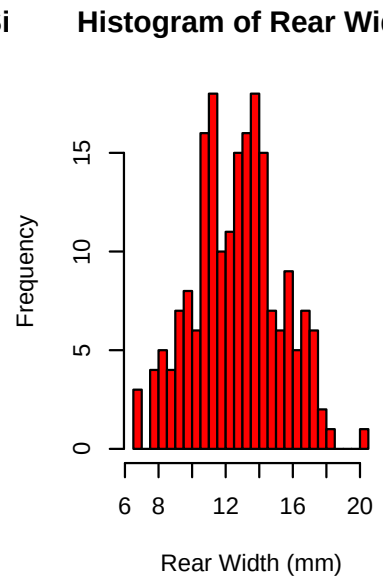
	sp	sex	index	FL	RW	CL	CW	BD
1	B	M	1	8.1	6.7	16.1	19.0	7.0
2	B	M	2	8.8	7.7	18.1	20.8	7.4
3	B	M	3	9.2	7.8	19.0	22.4	7.7
4	B	M	4	9.6	7.9	20.1	23.1	8.2
5	B	M	5	9.8	8.0	20.3	23.0	8.2
6	B	M	6	10.8	9.0	23.0	26.5	9.8
7	B	M	7	11.1	9.9	23.8	27.1	9.8
8	B	M	8	11.6	9.1	24.5	28.4	10.4
9	B	M	9	11.8	9.6	24.2	27.8	9.7
10	B	M	10	11.8	10.5	25.2	29.3	10.3
11	B	M	11	12.2	10.8	27.3	31.6	10.9
12	B	M	12	12.3	11.0	26.8	31.5	11.4
13	B	M	13	12.6	10.0	27.7	31.7	11.4
14	B	M	14	12.8	10.2	27.2	31.8	10.9
15	B	M	15	12.8	10.9	27.4	31.5	11.0
16	B	M	16	12.9	11.0	26.8	30.9	11.4
17	B	M	17	13.1	10.6	28.2	32.3	11.0
18	B	M	18	13.1	10.9	28.3	32.4	11.2
19	B	M	19	13.3	11.1	27.8	32.3	11.3
20	B	M	20	13.9	11.1	29.2	33.3	12.1

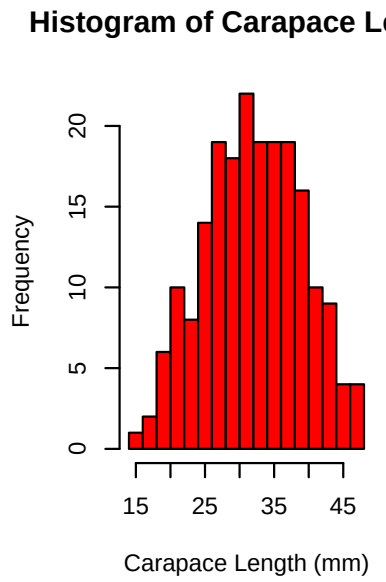
Univariate Boxplots

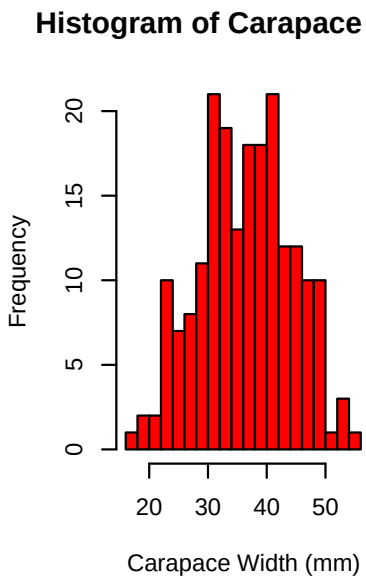


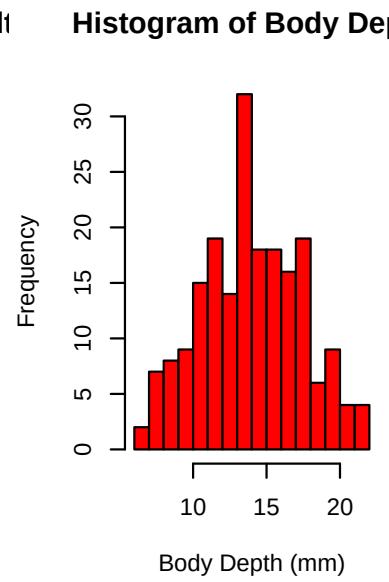
Univariate Histograms



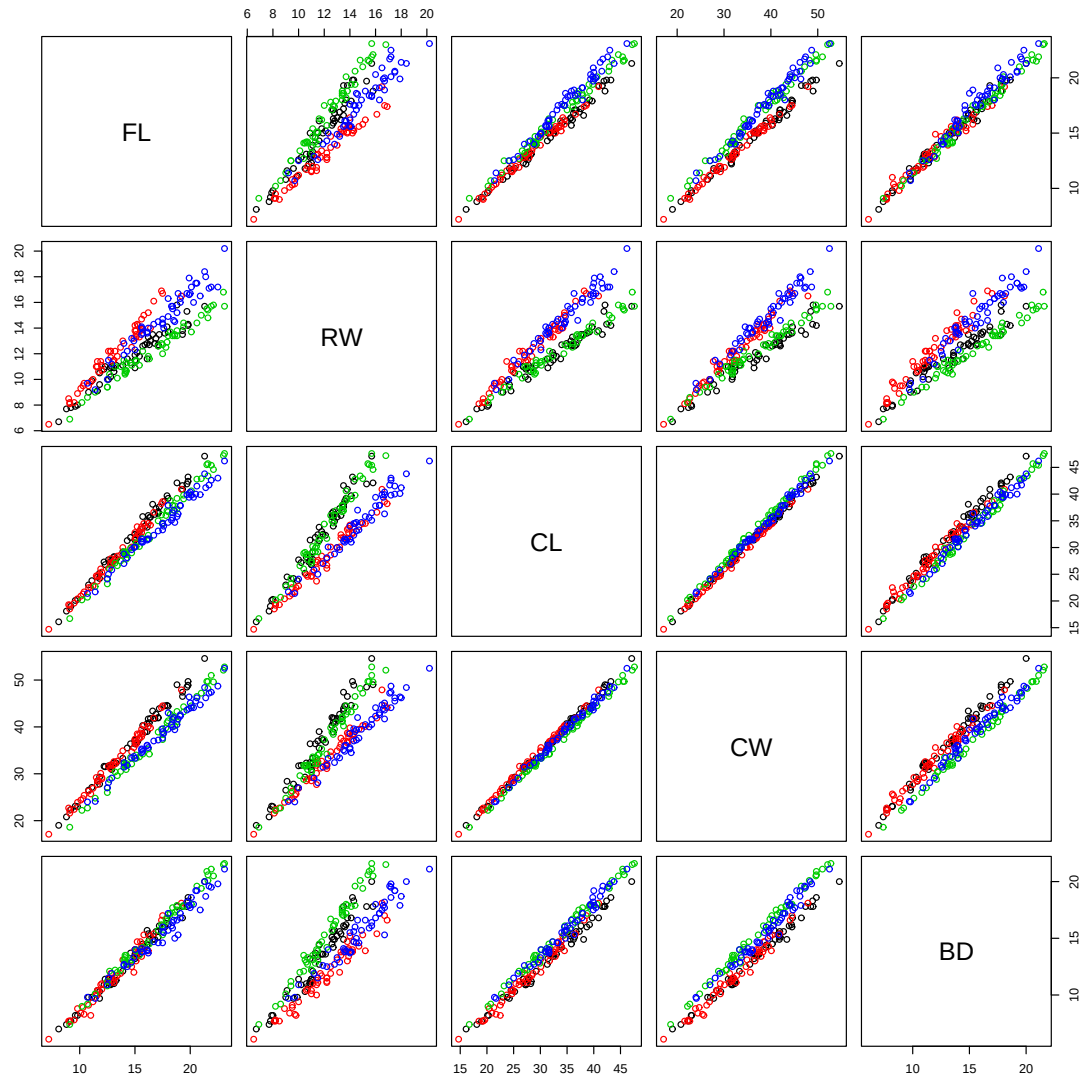




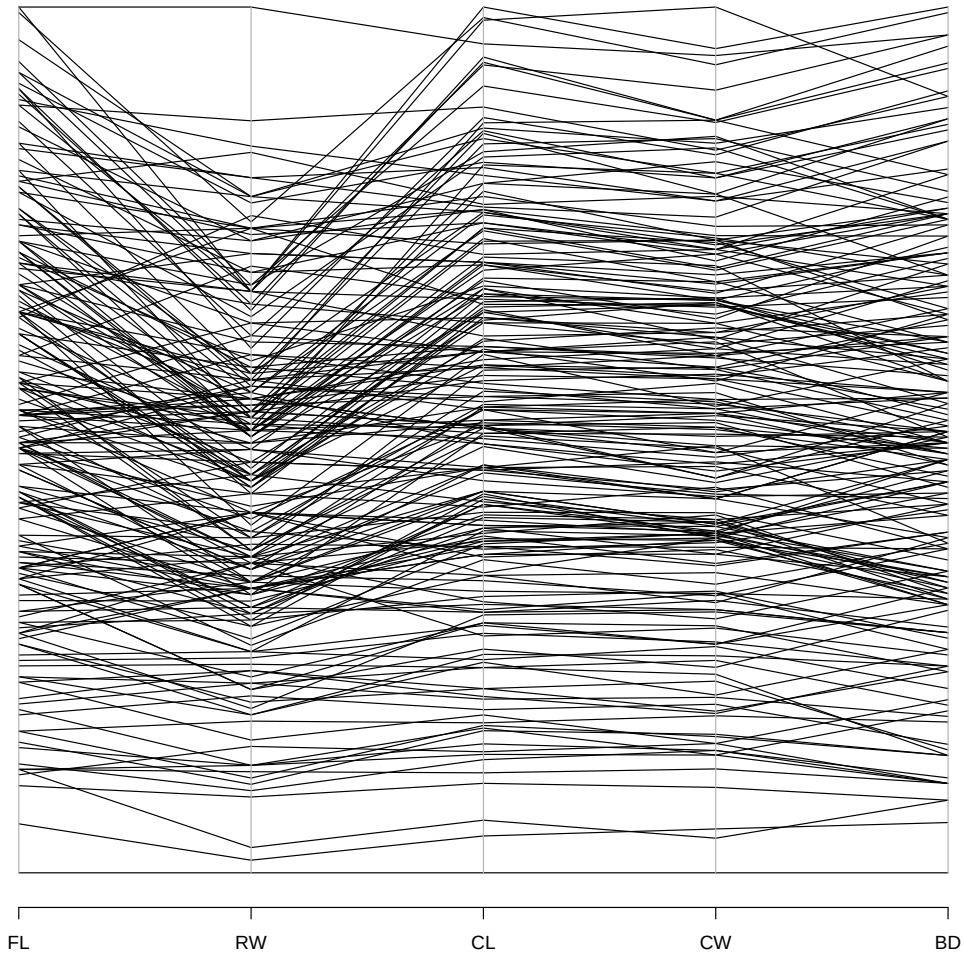




Simple Pairwise Scatterplots



Parallel Coordinate Plots



Visualization and Dimensionality Reduction

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).
Visualizing higher-dimensional problems:

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

Dimensionality reduction

- For each data item $x_i \in \mathbb{R}^p$, find a lower dimensional representation $y_i \in \mathbb{R}^k$ with $k \ll p$.
- Preserve as much as possible the interesting statistical properties/relationships of data items.

Principal Components Analysis (PCA)

- PCA considers interesting directions to be those with greatest *variance*.
- A *linear* dimensionality reduction technique:
- Finds an orthogonal basis $v_1, v_2, \ldots, v_p$ for the data space such that
  - The first principal component (PC) v_1 is the direction of greatest variance of data.
  - The second PC v_2 is the direction orthogonal to v_1 of greatest variance, etc.
  - The subspace spanned by the first k PCs represents the 'best' k-dimensional representation of the data.
  - The k-dimensional representation of x_i is:

$$z_i = Vx_i = \sum_{\ell=1}^{k} v_\ell^\top x_i$$

- For simplicity, we will assume from now on that our dataset is centred, i.e. we subtract the average $\bar{x}$ from each x_i.

Principal Components Analysis (PCA)

- Our data set is an iid sample of a random vector $X = [X_1 \dots X_p]^\top$.
- For the 1^{st} PC, we seek a derived variable of the form

$$Z_1 = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p = v_1^\top X$$

where $v_1 = [v_{11}, \dots, v_{1p}]^\top \in \mathbb{R}^p$ are chosen to maximise

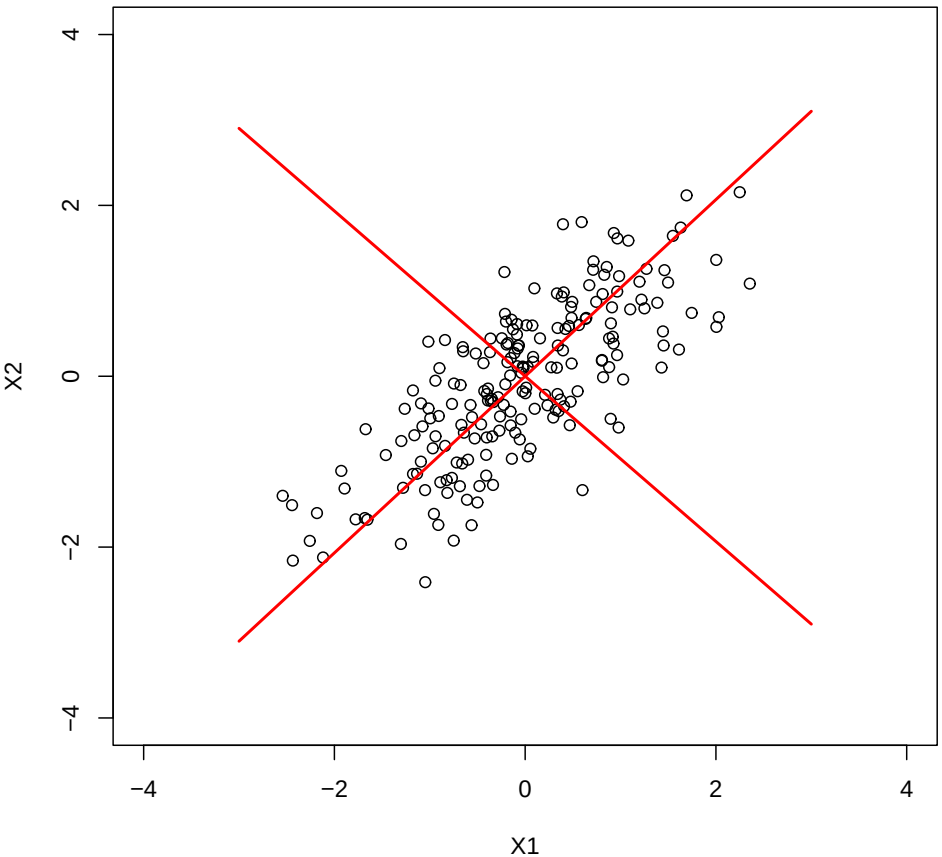
$$\mathrm{Var}(Z_1).$$

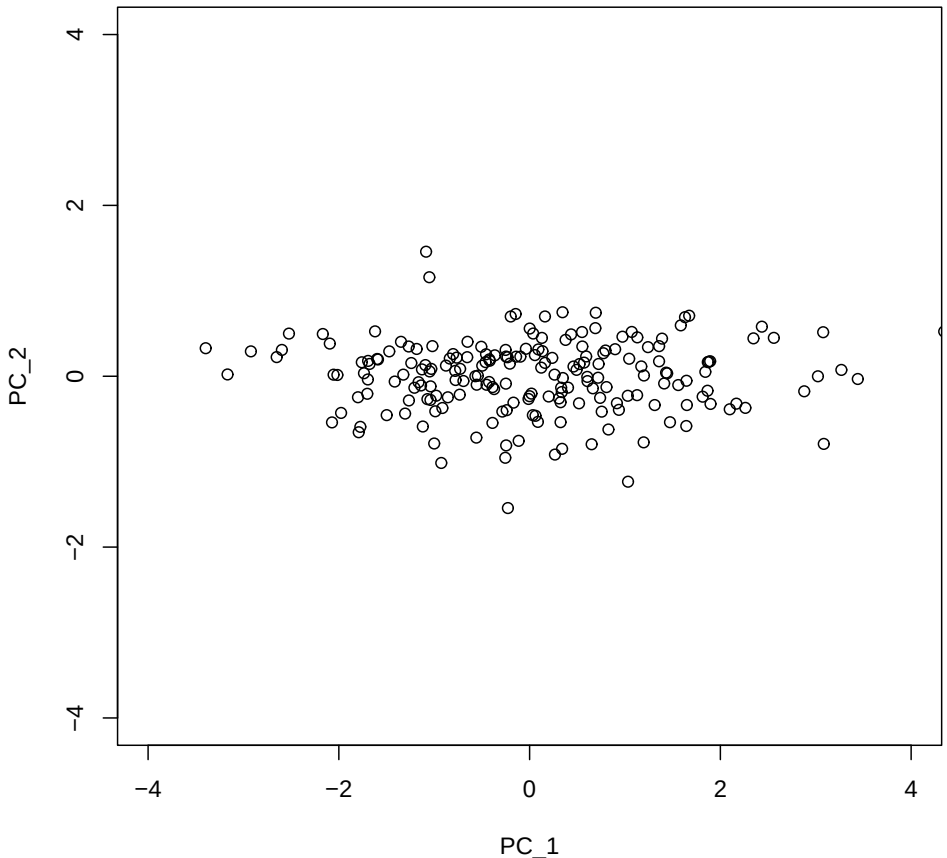
To get a well defined problem, we fix

$$v_1^\top v_1 = 1.$$

- The 2^{nd} PC 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)





Deriving the First Principal Component

- Maximise, subject to $v_1^\top v_1 = 1$:

$$\mathrm{Var}(Z_1) = \mathrm{Var}(v_1^\top X) = v_1^\top \mathrm{Cov}(X) v_1 \approx v_1^\top S v_1$$

where $S \in \mathbb{R}^{p \times p}$ is the sample covariance matrix, i.e.

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{n-1} \sum_{i=1}^{n} x_i x_i^\top = \frac{1}{n-1} \mathbf{X}^\top \mathbf{X}.$$

- Rewriting this as a constrained maximisation problem,

$$\mathcal{L}(v_1, \lambda_1) = v_1^\top S v_1 - \lambda_1 \left(v_1^\top v_1 - 1 \right).$$

- The corresponding vector of partial derivatives yields

$$\frac{\partial \mathcal{L}(v_1, \lambda_1)}{\partial v_1} = 2 S v_1 - 2 \lambda_1 v_1.$$

- Setting this to zero reveals the eigenvector equation, i.e. v_1 must be an eigenvector of S and λ_1 the corresponding eigenvalue.
- Since $v_1^\top S v_1 = \lambda_1 v_1^\top v_1 = \lambda_1$, the 1^{st} PC must be the eigenvector associated with the largest eigenvalue of S.

Deriving Subsequent Principal Components

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

$$\mathcal{L}(v_2, \lambda_2, \mu) = v_2^\top S v_2 - \lambda_2 \left(v_2^\top v_2 - 1\right) - \mu \left(v_1^\top v_2\right).$$

- Solving this shows that v_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

$$S = V \Lambda V^\top$$

where Λ is a diagonal matrix with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

and V is a $p \times p$ orthogonal matrix whose columns are the p eigenvectors of S, i.e. the principal components $v_1, \ldots, v_p$.

Properties of the Principal Components

- PCs are *uncorrelated*

$$\mathrm{Cov}(X^\top v_i, X^\top v_j) \approx v_i^\top S v_j = 0 \text{ for } i \neq j.$$

- The *total sample variance* is given by

$$\sum_{i=1}^{p} S_{ii} = \lambda_1 + \ldots + \lambda_p,$$

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

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

- S is a real symmetric matrix, so eigenvectors (principal components) are orthogonal.
- Derived variables $Z_1, \ldots, Z_p$ have variances $\lambda_1, \ldots, \lambda_p$.

R code

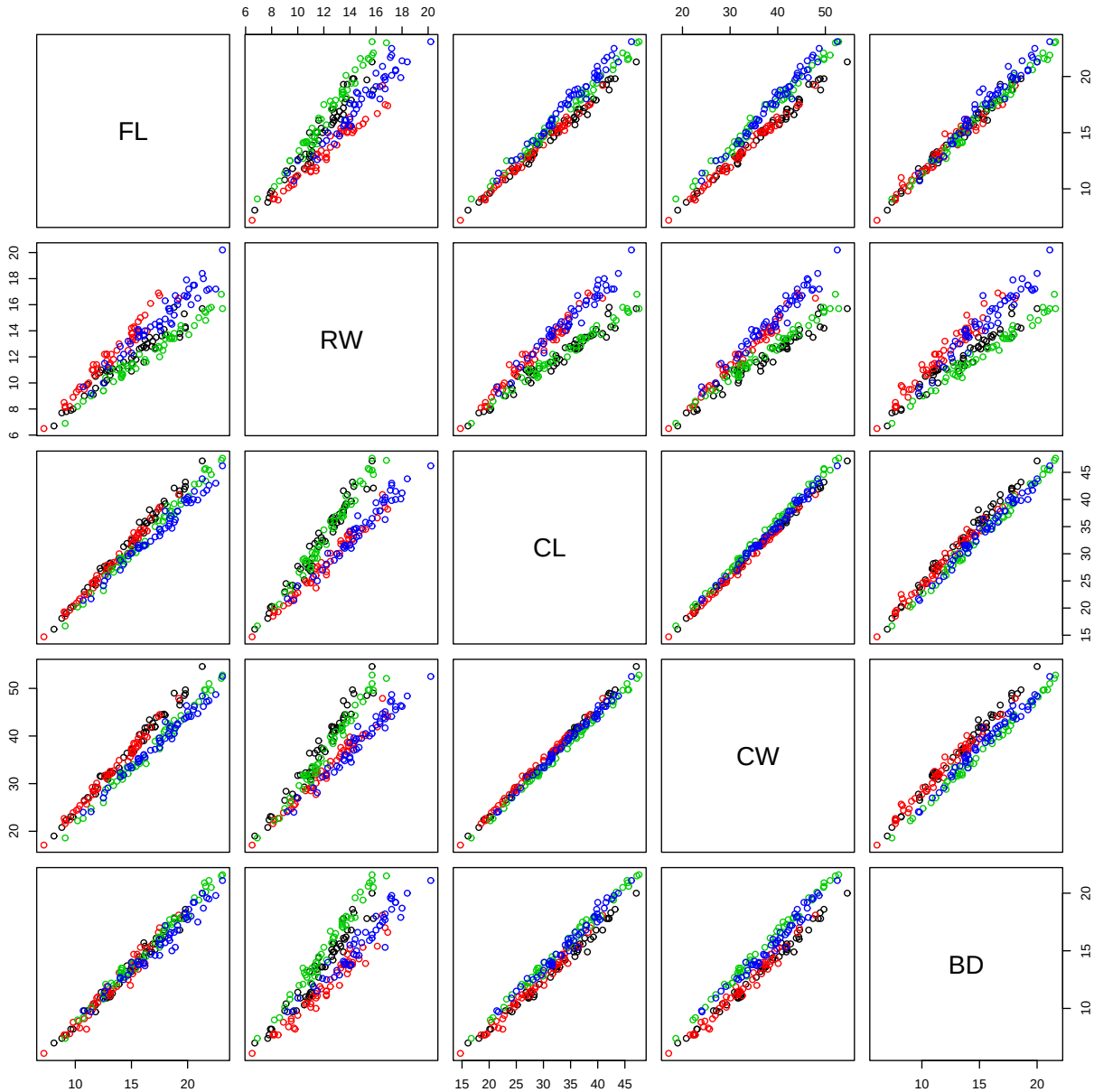
This is what we have had before:

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

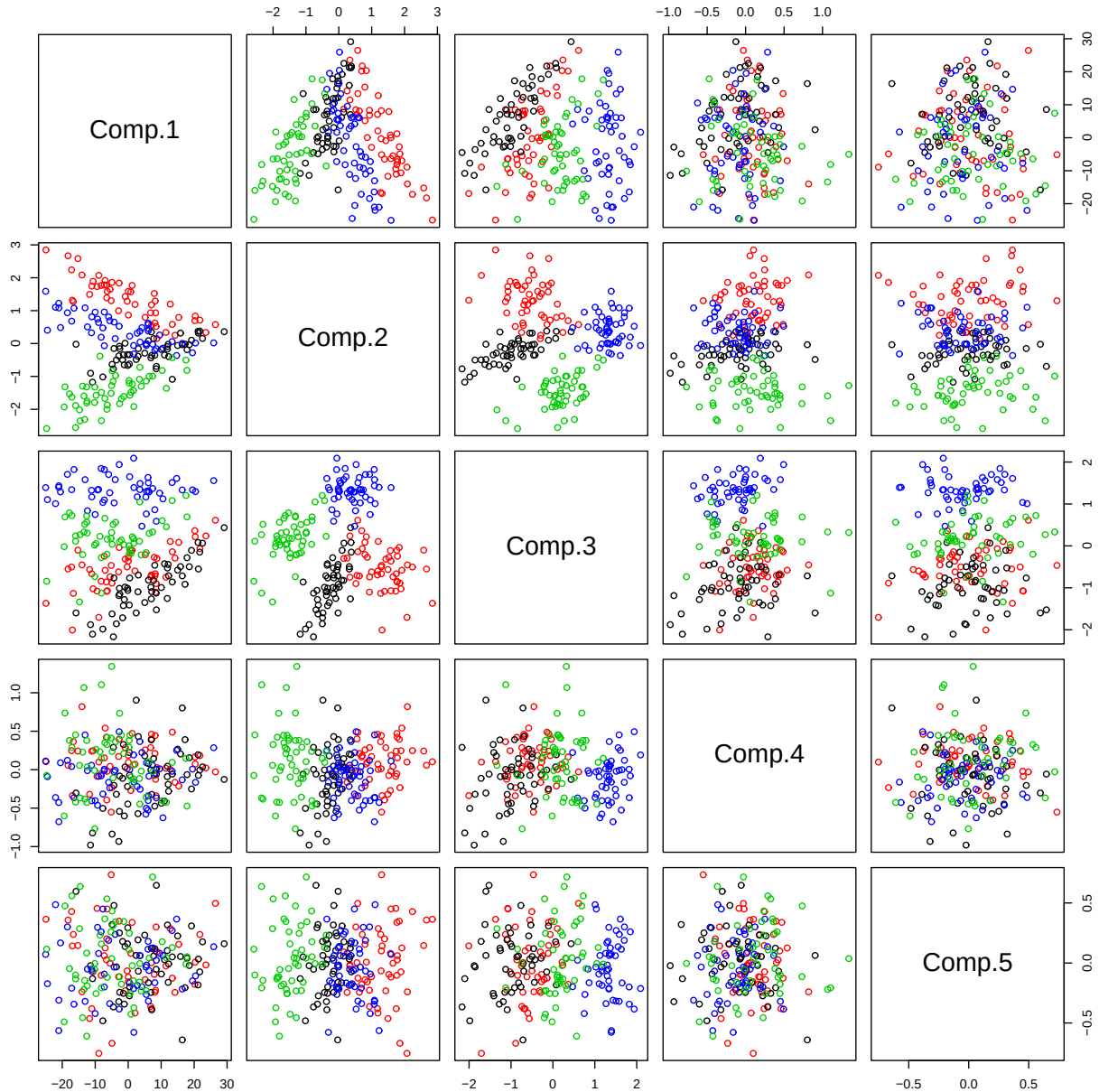
Now perform PCA with function `princomp`. (Alternatively, solve for the PCs yourself using `eigen` or `svd`).

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

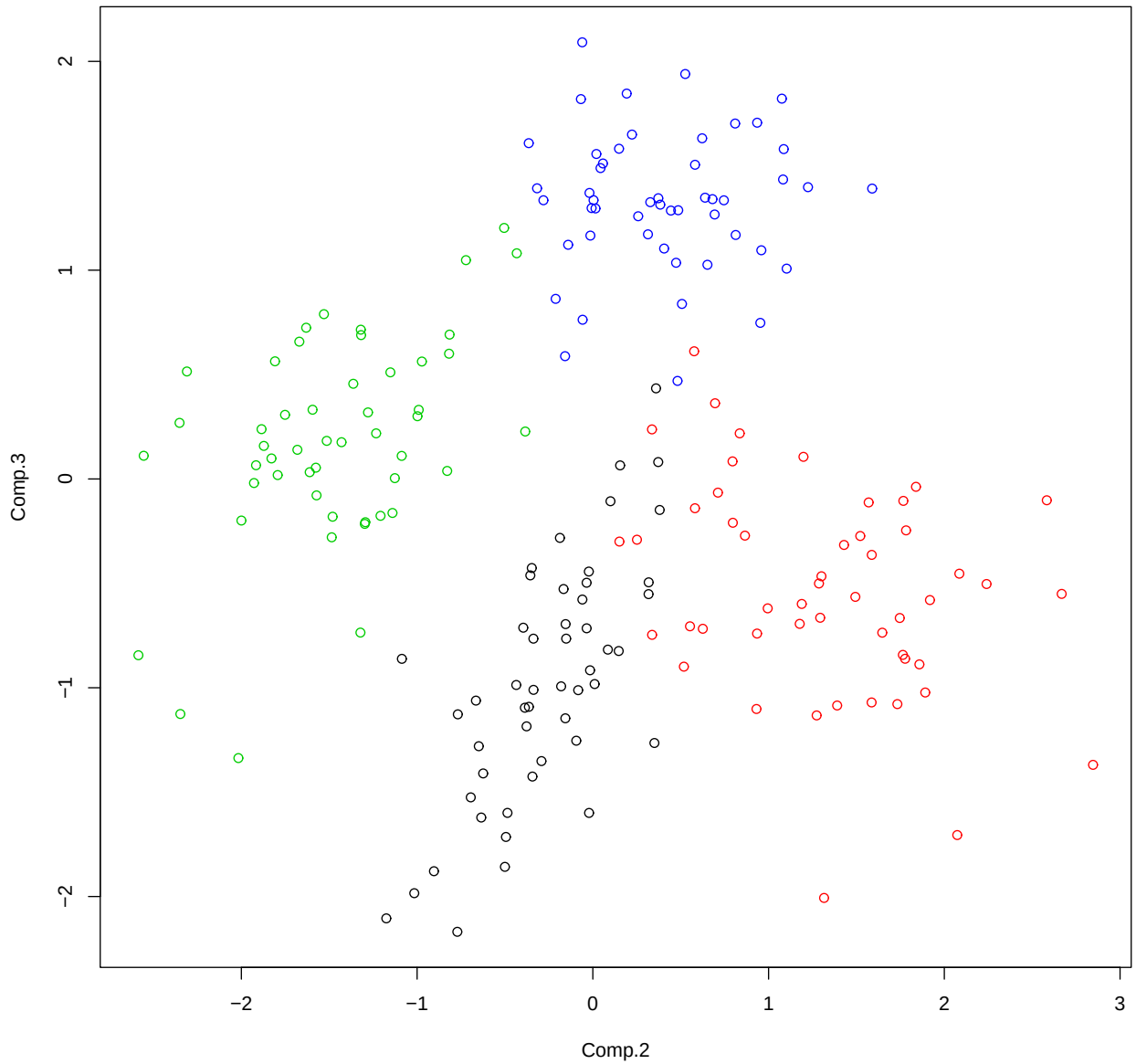
Original crabs data



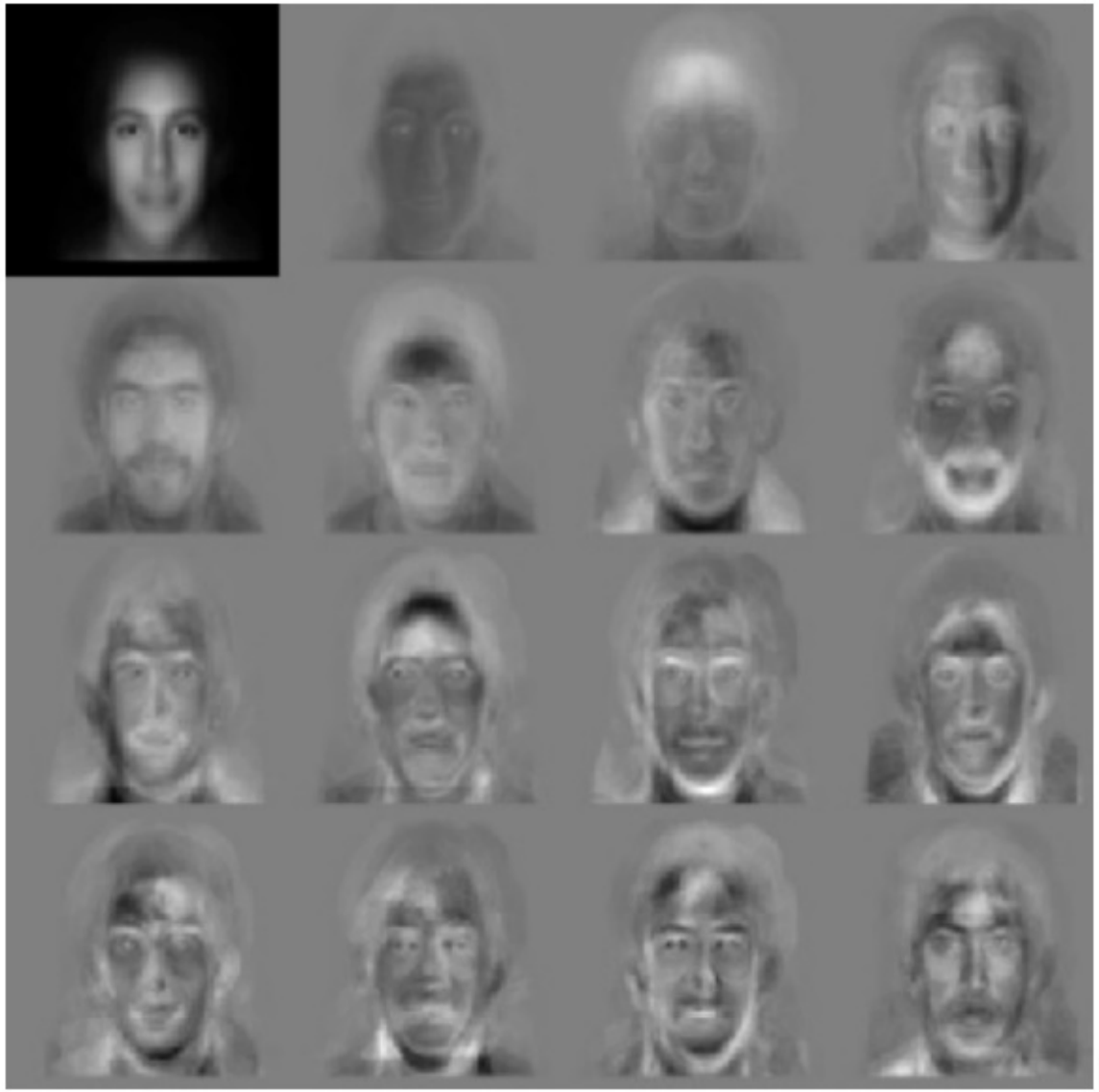
PCs of crabs data



PC 2 vs PC 3

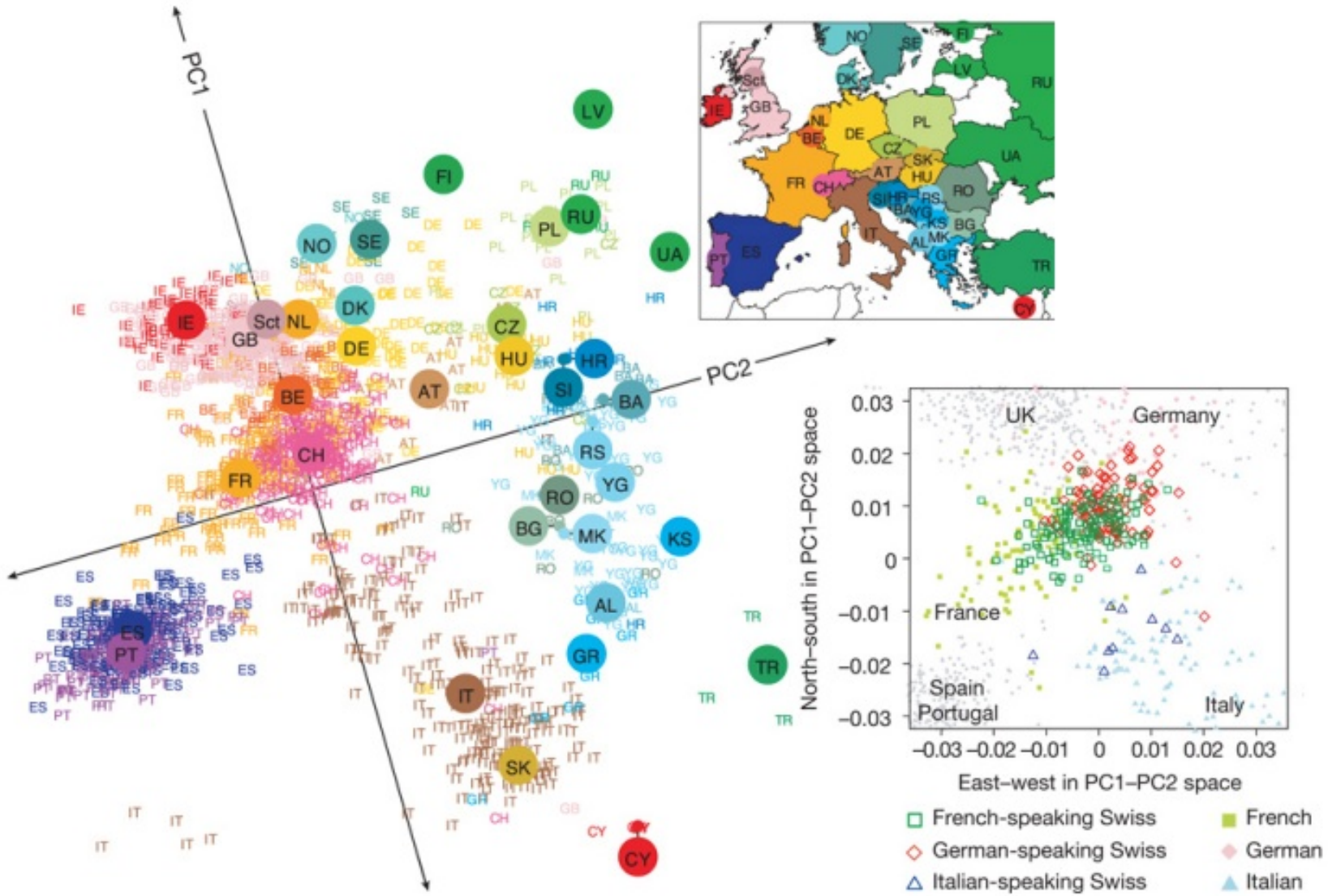


PCA on Face Images



http://vismod.media.mit.edu/vismod/demos/facerec/basic.html

PCA on European Genetic Variation



Comments on the use of PCA

- PCA commonly used to project data X onto the first k PCs giving the k-dimensional view of the data that best preserves the first two moments.
- Although PCs are uncorrelated, scatterplots sometimes reveal structures in the data other than linear correlation.
- 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 $\mathrm{Corr}(X)$ instead of $\mathrm{Cov}(X)$.
  - Robustness to outliers is also an issue. Variance is affected by outliers therefore so are PCs.

Eigenvalue Decomposition (EVD)

Eigenvalue decomposition plays a significant role in PCA. PCs are eigenvectors of $S = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}$ and PCA properties are derived from those of eigenvectors and eigenvalues.

- For any $p \times p$ *symmetric* matrix S, there exists p eigenvectors $v_1, \ldots, v_p$ that are pairwise orthogonal and p associated eigenvalues $\lambda_1, \ldots, \lambda_p$ which satisfy the eigenvalue equation $Sv_i = \lambda_i v_i \ \forall i$.
- S can be written as $S = V\Lambda V^\top$ where
  - $V = [v_1, \ldots, v_p]$ is a $p \times p$ orthogonal matrix
  - $\Lambda = diag\{\lambda_1, \ldots, \lambda_p\}$
  - If S is a real-valued matrix, then the eigenvalues are real-valued as well, $\lambda_i \in \mathbb{R} \ \forall i$
- To compute the PCA of a dataset $\mathbf{X}$, we can:
  - First estimate the covariance matrix using the sample covariance S.
  - Compute the EVD of S using the R command `eigen`.

Singular Value Decomposition (SVD)

Though the EVD does not always exist, the singular value decomposition is another matrix factorization technique that *always* exist, even for non-square matrices.

- $\mathbf{X}$ can be written as $X = UDV^\top$ where
  - U is an $n \times n$ matrix with orthogonal columns.
  - D is a $n \times p$ matrix with decreasing non-negative elements on the diagonal (the singular values) and zero off-diagonal elements.
  - V is a $p \times p$ matrix with orthogonal columns.
- SVD can be computed using very fast and numerically stable algorithms. The relevant R command is `svd`.

Some Properties of the SVD

- Let $\mathbf{X} = UDV^\top$ be the SVD of the $n \times p$ data matrix $\mathbf{X}$.
- Note that

$$(n-1)S = \mathbf{X}^\top\mathbf{X} = (UDV^\top)^\top(UDV^\top) = VD^\top U^\top UDV^\top = VD^\top DV^\top,$$

  using orthogonality ($U^\top U = I_n$) of U.
- The eigenvalues of S are thus the diagonal entries of $\frac{1}{n-1}D^2$ and the columns of the orthogonal matrix V are the eigenvectors of S.
- We also have

$$\mathbf{X}\mathbf{X}^\top = (UDV^\top)(UDV^\top)^\top = UDV^\top VD^\top U^\top = UDD^\top U^\top,$$

  using orthogonality ($V^\top V = I_p$) of V.
- SVD also gives the optimal low-rank approximations of $\mathbf{X}$:

$$\min_{\tilde{X}} \|\tilde{X} - \mathbf{X}\|^2 \qquad \text{s.t. } \tilde{X} \text{ has maximum rank } r < n, p.$$

  This problem can be solved by keeping only the r largest singular values of X, zeroing out the smaller singular values in the SVD.

Biplots

- PCA plots show the data items (as rows of $\mathbf{X}$) in the PC space.
- *Biplots* allow us to visualize the *original variables* (as columns $\mathbf{X}$) in the same plot.
- As for PCA, we would like the geometry of the plot to preserve as much of the covariance structure as possible.

Biplots

Recall that $X = [X_1, \ldots, X_p]^\top$ and $\mathbf{X} = UDV^\top$ is the SVD of the data matrix.

- The PC projection of x_i is:

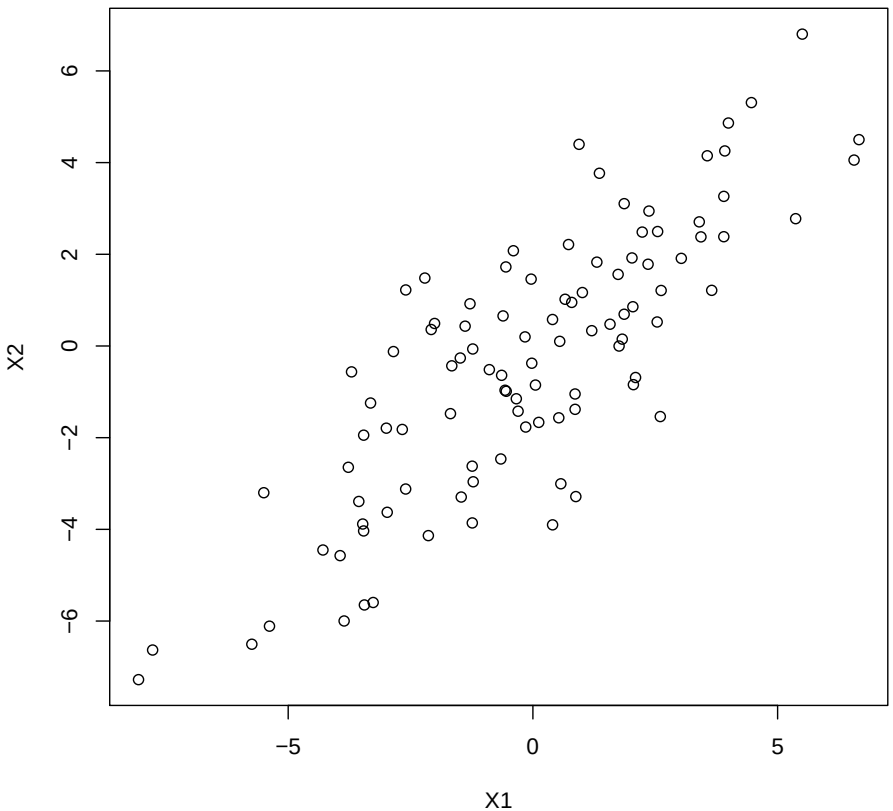
$$z_i = V^\top x_i = DU_i^\top = [D_{11}U_{i1}, \ldots, D_{kk}U_{ik}]^\top.$$

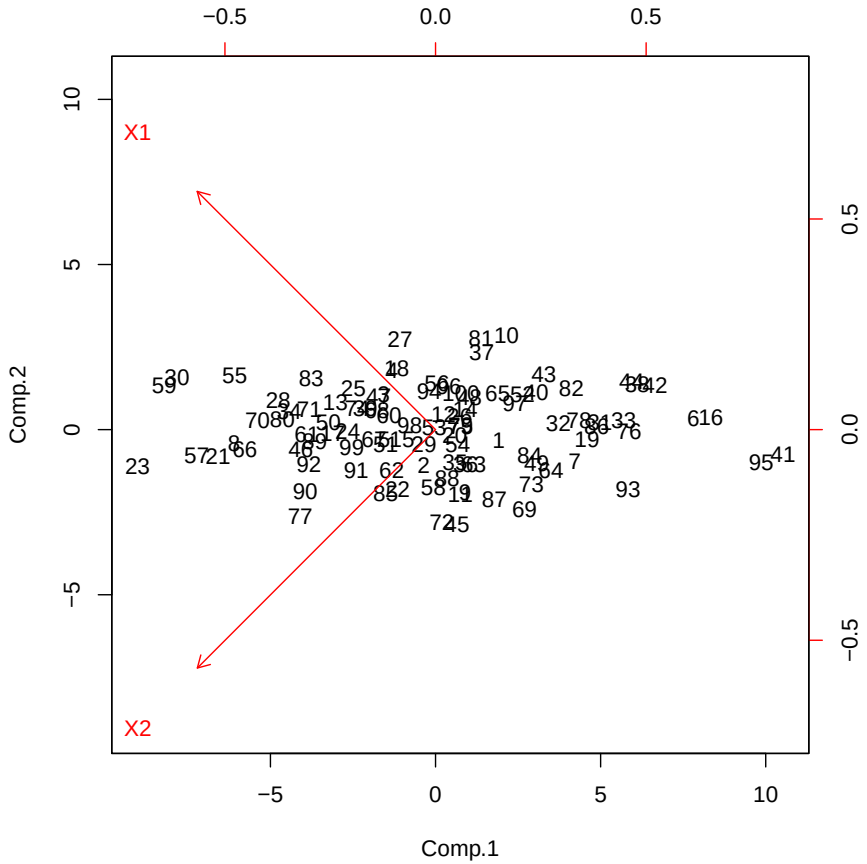
- The jth unit vector $\mathbf{e}_j \in \mathbb{R}^p$ points in the direction of X_j. Its PC projection is $V_j^\top = V^\top \mathbf{e}_j$, the jth row of V.
- The projection of the variable indicates the weighting each PC gives to the original variables.
- Dot products between the projections gives entries of the data matrix:

$$x_{ij} = \sum_{k=1}^{p} U_{ik}D_{kk}V_{jk} = \langle DU_i^\top, V_j^\top \rangle.$$

- Distance of projected points from projected variables gives original location.
- These relationships can be plotted in 2D by focussing on first two PCs.

Biplots





Biplots

- There are other projections we can consider for biplots:

$$x_{ij} = \sum_{k=1}^{p} U_{ik} D_{kk} V_{jk} = \langle D U_i^\top, V_j^\top \rangle = \langle D^{1-\alpha} U_i^\top, D^\alpha V_j^\top \rangle.$$

  where $0 \leq \alpha \leq 1$. The $\alpha = 1$ case has some nice properties.
- Covariance of the projected points is:

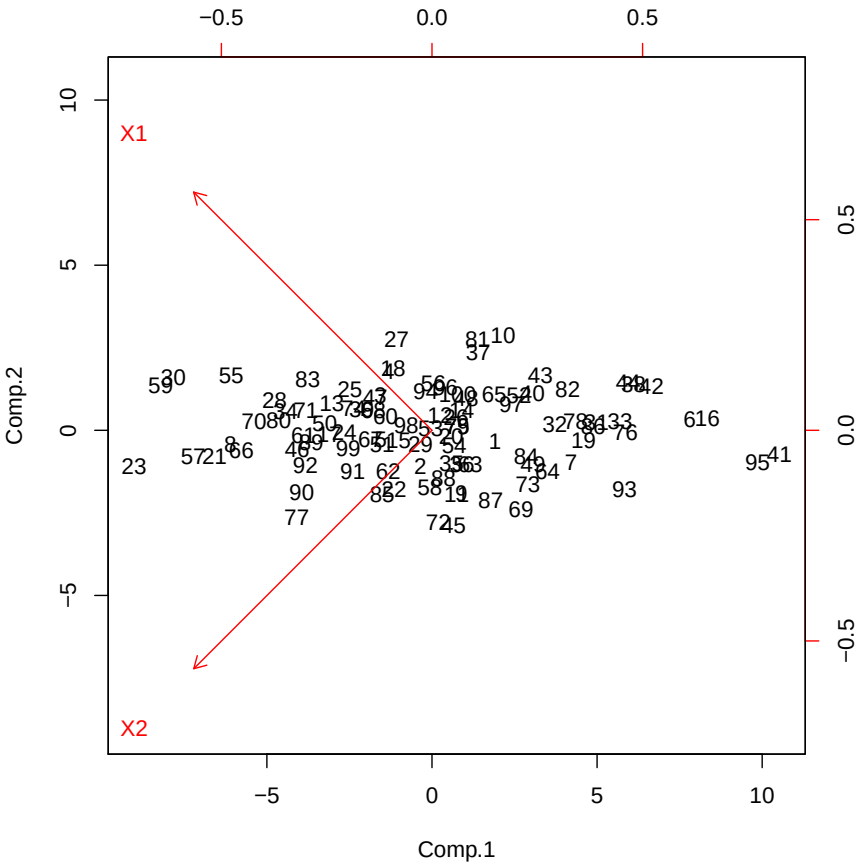
$$\frac{1}{n-1} \sum_{i=1}^{n} U_i^\top U_i = \frac{1}{n-1} I.$$

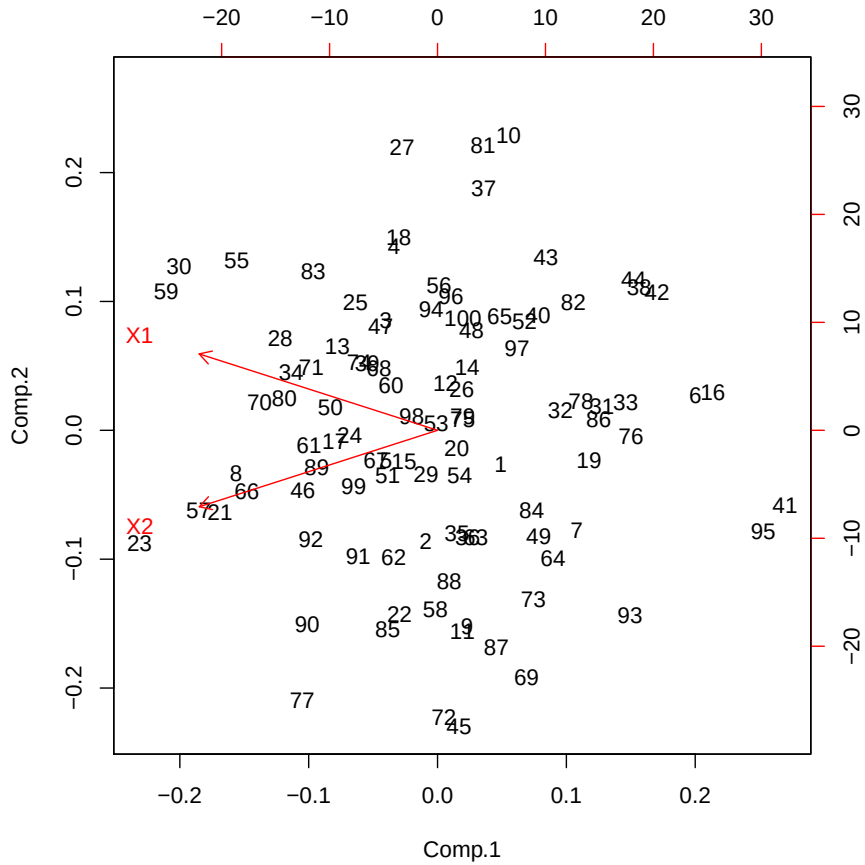
  Projected points are uncorrelated and dimensions are equi-variance.
- The covariance between X_j and X_ℓ is:

$$\mathrm{Var}(X_j X_\ell) = \frac{1}{n-1} \langle D V_j^\top, D V_\ell^\top \rangle$$

  So the angle between the projected variables gives the correlation.
- When using $k < p$ PCs, quality depends on the proportion of variance explained by the PCs.

Biplots





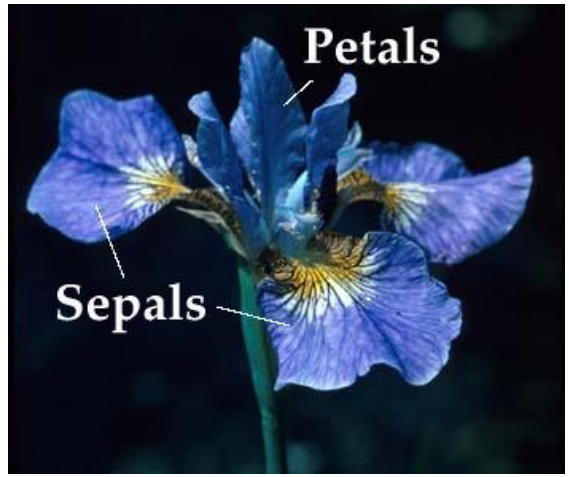

```
pc <- princomp(x)
biplot(pc,scale=0)
biplot(pc,scale=1)
```

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

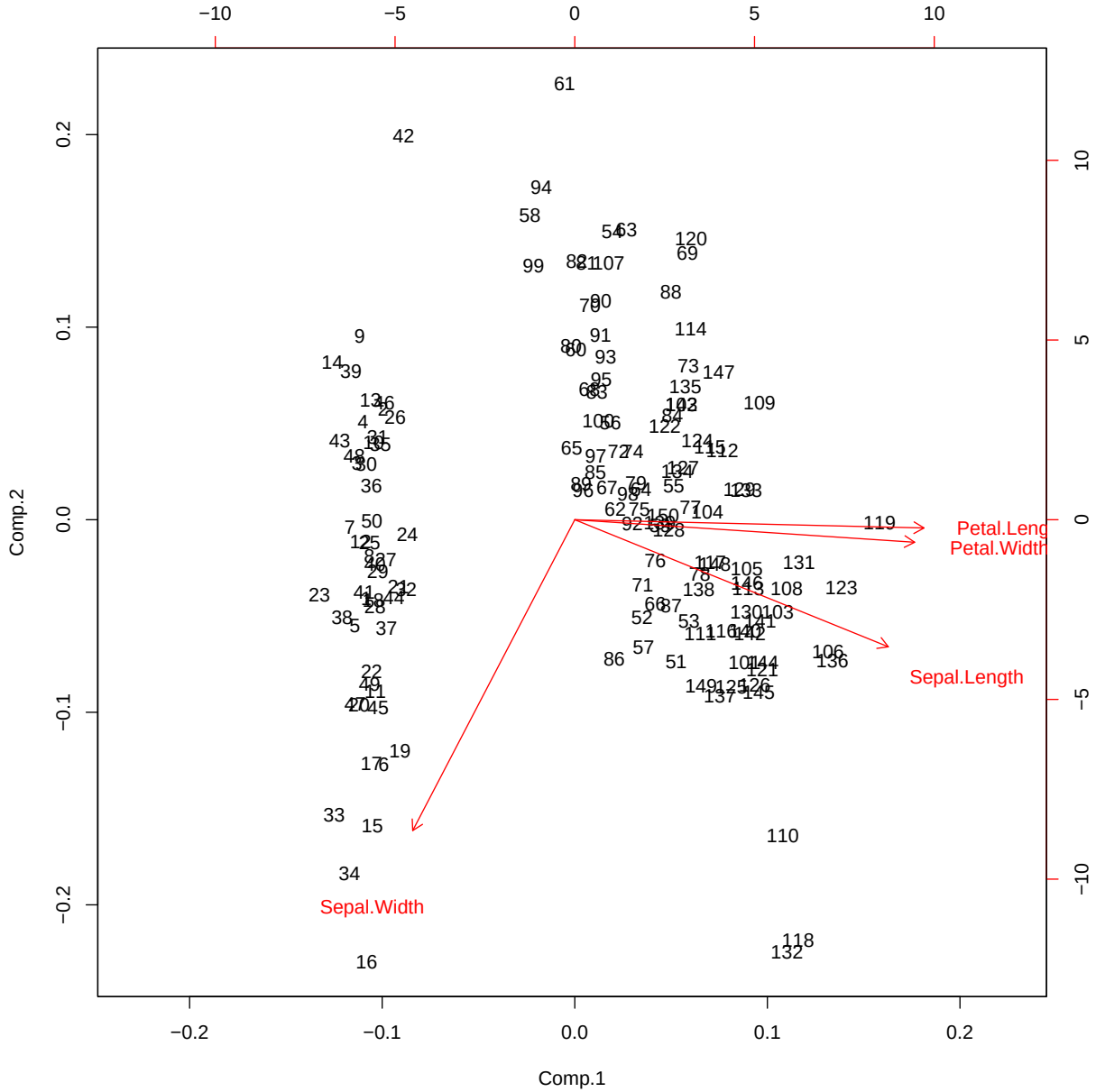
Collected by E. Anderson (1935) and analysed by R.A. Fisher (1936)



Using again function `princomp` and `biplot`.

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

Iris Data

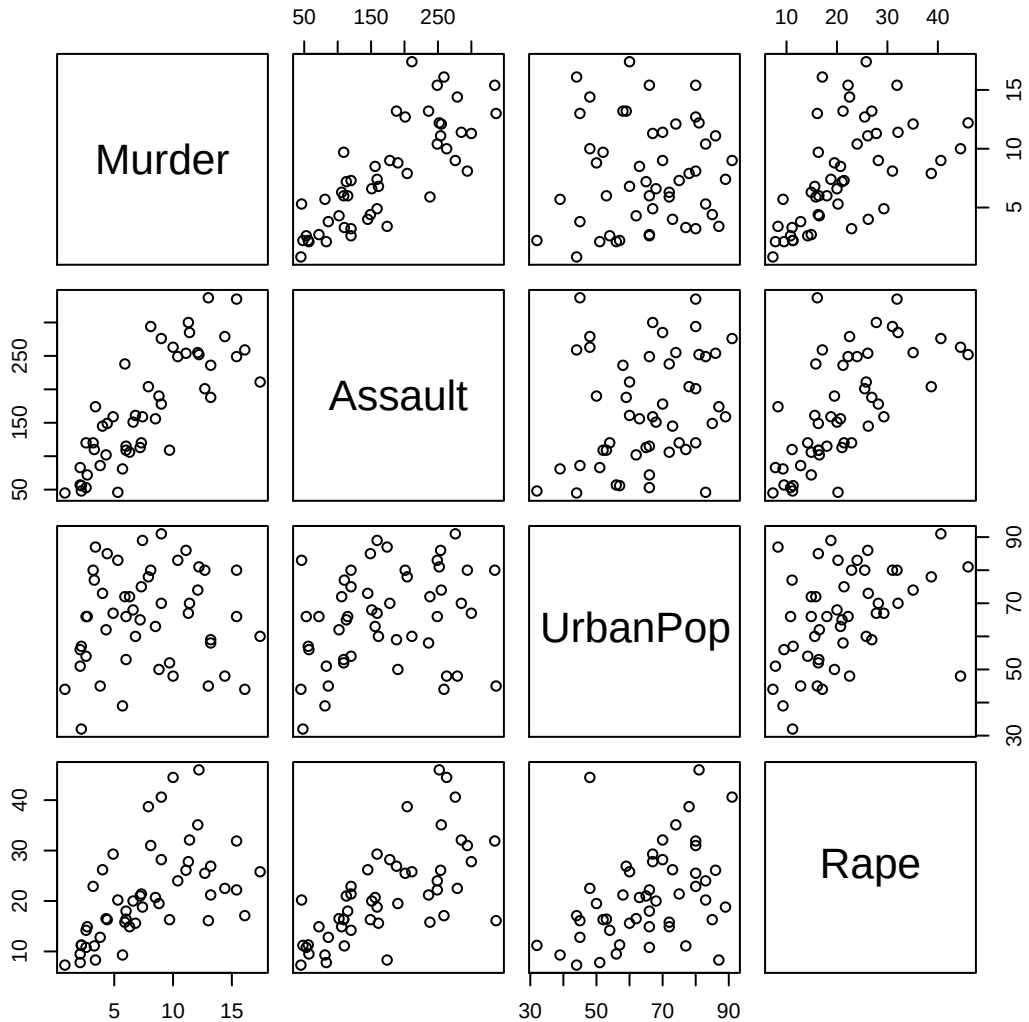

Comp.1
Comp.2
Petal.Leng
Petal.Width
Sepal.Length
Sepal.Width

US Arrests Data

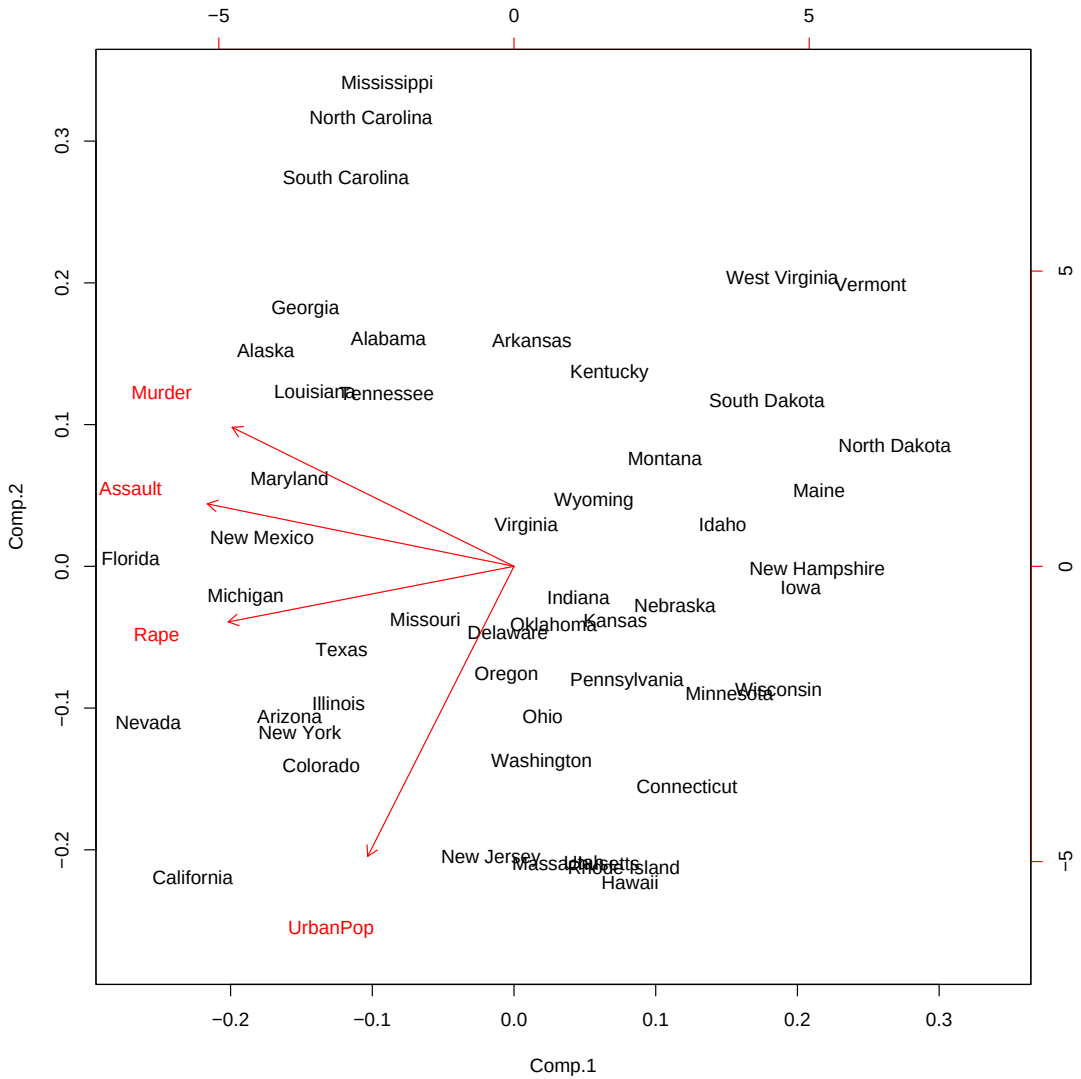
This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

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

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

US Arrests Data Pairs Plot



US Arrests Data Biplot



Further Readings

- Venables and Ripley, Chapter 11.
- Hastie et al, Chapter 14.
- James et al, Chapter 10.
- Tukey, John W. (1980). We need both exploratory and confirmatory. The American Statistician 34 (1): 23-25.