Al-sham Private University

Dr. George Karraz, Ph. D.

Principal Component Analysis

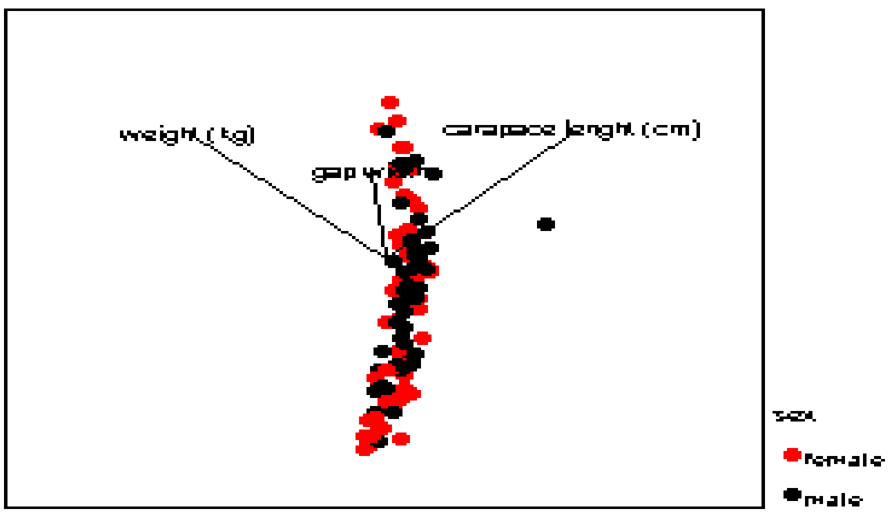
Dr. George Karraz, Ph. D.

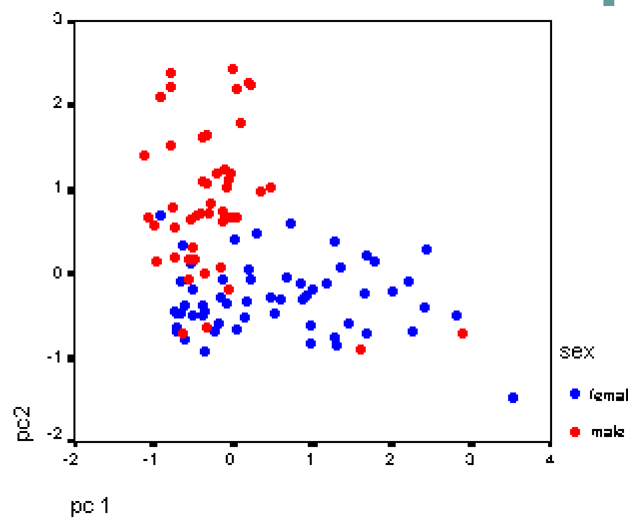
Philosophy of PCA

- Introduced by Pearson (1901) and Hotelling (1933) to describe the variation in a set of multivariate data in terms of a set of uncorrelated variables
- We typically have a data matrix of n observations on p correlated variables $x_1, x_2, \ldots x_p$
- PCA looks for a transformation of the x_i into p new variables y_i that are uncorrelated

The data matrix

case	ht (x_1)	wt(x_2)	age(x_3)	sbp(x_4)	heart rate (x_5)
1	175	1225	25	117	56
2	156	1050	31	122	63
n	202	1350	58	154	67





Reduce dimension

- The simplet way is to keep one variable and discard all others: not reasonable!
- Wheigt all variable equally: not reasonable (unless they have same variance)
- Wheigted average based on some citerion.
- Which criterion?

Let us write it first

- Looking for a transformation of the data matrix **X** (nxp) such that
-

$$Y = \boldsymbol{\delta}^T \mathbf{X} = \delta_1 X_1 + \delta_2 X_2 + .. + \delta_p X_p$$

- Where $\boldsymbol{\delta} = (\delta_1, \delta_2, .., \delta_p)^T$ is a column vector of wheights with

$$\delta_1^2 + \delta_2^2 + .. + \delta_p^2 = 1$$

One good criterion

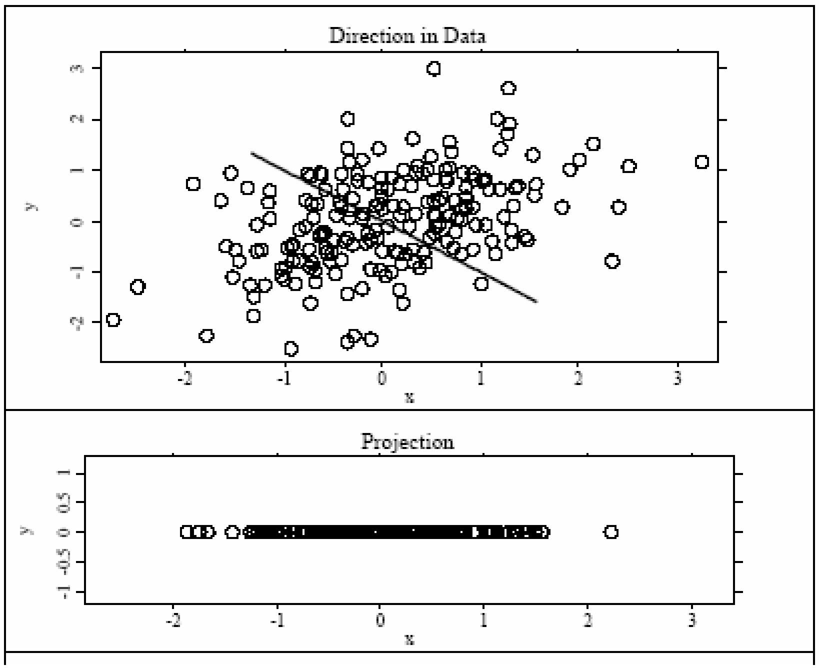
- Maximize the variance of the projection of the observations on the Y variables
- Find δ *so that*

$$Var(\delta^T \mathbf{X}) = \delta^T \mathbf{Var(X)}\, \delta \quad \textit{is maximal}$$

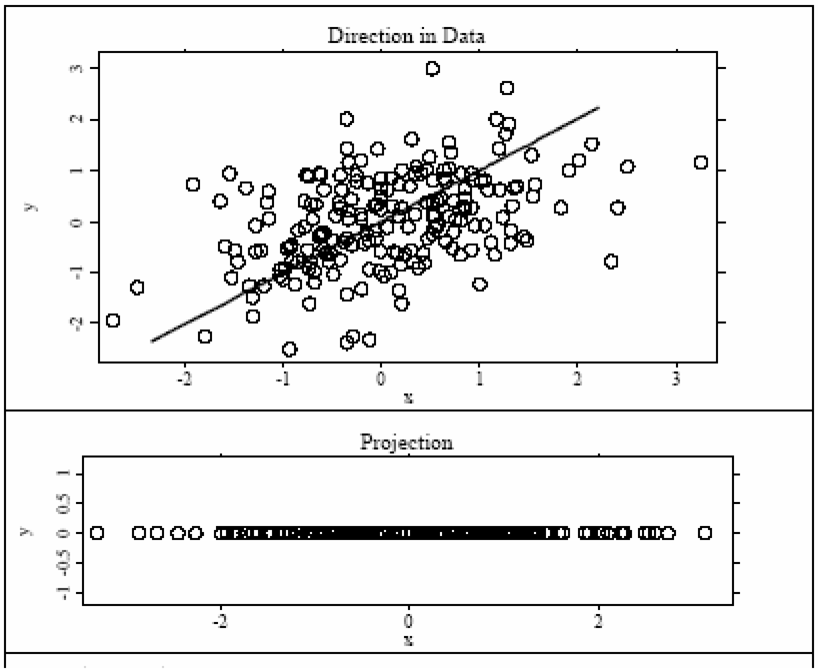
- The matrix **C=Var(X)** is the covariance matrix of the X_i variables

Let us see it on a figure

Good



Better



Covariance matrix

C= $$\begin{pmatrix} v(x_1)\ c(x_1,x_2)\ ....c(x_1,x_p) \\ c(x_1,x_2)\ v(x_2)\ ....c(x_2,x_p) \\ \\ c(x_1,x_p)\ c(x_2,x_p).........v(x_p) \end{pmatrix}$$

And so.. We find that

- The direction of $\boldsymbol{\delta}$ is given by the eigenvector $\boldsymbol{\gamma_1}$ correponding to the largest eigenvalue of matrix **C**
- The second vector that is orthogonal (uncorrelated) to the first is the one that has the second highest variance which comes to be the eignevector corresponding to the second eigenvalue
- And so on …

So PCA gives

- New variables Y_i that are linear combination of the original variables (x_i):
- $Y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots a_{ip}x_p \;\; ; \; i=1..p$
- The new variables Y_i are derived in decreasing order of importance;
- they are called 'principal components'

Calculating eignevalues and eigenvectors

- The eigenvalues λ_i are found by solving the equation

$$det(C-\lambda I)=0$$

- Eigenvectors are columns of the matrix A such that

$$C=A\ D\ A^T$$

- Where

$$D=\begin{pmatrix} \lambda 1\ 0 \ldots\ldots\ 0 \\ 0 \quad \lambda_2 \ldots\ldots\ 0 \\ 0 \\ 0 \ldots\ldots\ \ldots\lambda_p \end{pmatrix}$$

An example

- Let us take two variables with covariance $c>0$
- $\mathbf{C}=$ $\begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$ $\mathbf{C}-\lambda\mathbf{I}=$ $\begin{pmatrix} 1-\lambda & c \\ c & 1-\lambda \end{pmatrix}$

$$det(\mathbf{C}-\lambda\mathbf{I})=(1-\lambda)^2-c^2$$

- Solving this we find $\lambda_1 = 1+c$

  $\lambda_2 = 1-c < \lambda_1$

and eigenvectors

- Any eigenvector A satisfies the condition

$$\mathbf{C}A=\lambda A$$

A= $\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ CA= $\begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_1 + ca_2 \\ ca_1 + a_2 \end{pmatrix} = \begin{pmatrix} \lambda a_1 \\ \lambda a_2 \end{pmatrix}$

Solving we find $A_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, $A_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$

PCA is sensitive to scale

- If you multiply one variable by a scalar you get different results

(can you show it?)

- This is because it uses covariance matrix (and not correlation)
- PCA should be applied on data that have approximately the same scale in each variable

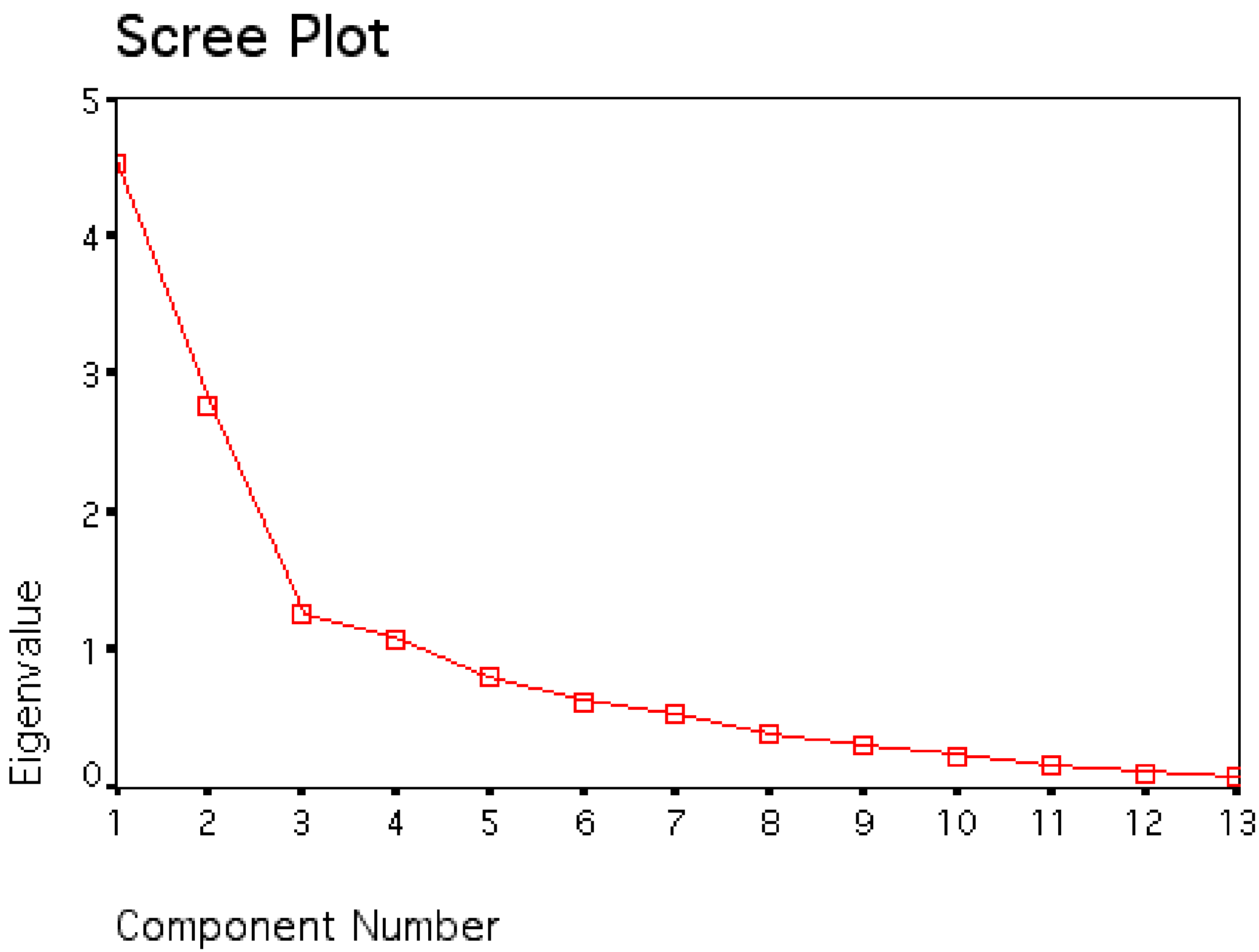
Interpretation of PCA

- The new variables (PCs) have a variance equal to their corresponding eigenvalue

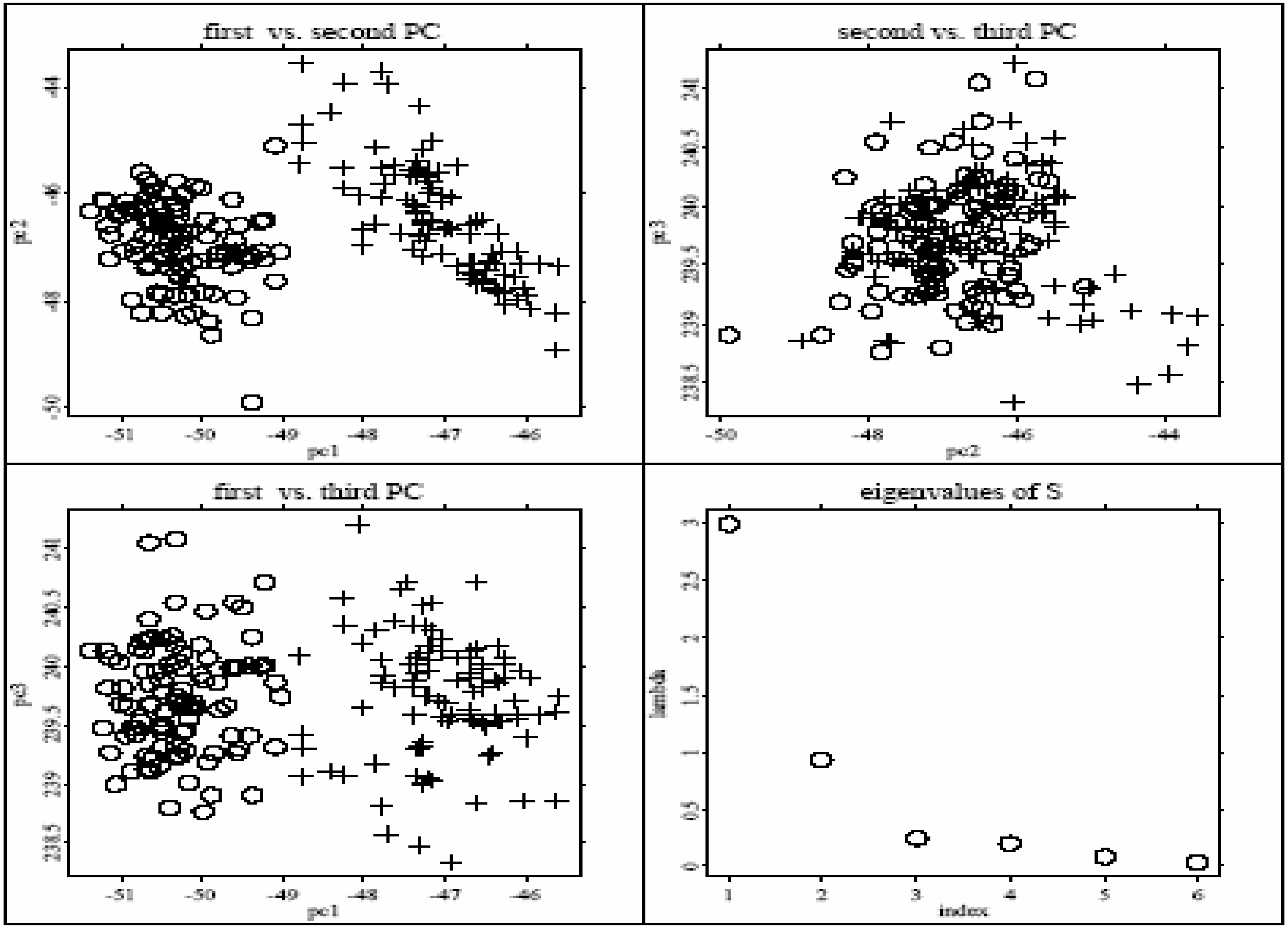
  $Var(Y_i) = \lambda_i$ for all $i = 1 \ldots p$
- Small λ_i ⇔ small variance ⇔ data change little in the direction of component Y_i
- The relative variance explained by each PC is given by $\lambda_i / \Sigma \lambda_i$

How many components to keep?

- Enough PCs to have a cumulative variance explained by the PCs that is >50-70%
- Kaiser criterion: keep PCs with eigenvalues >1
- Scree plot: represents the ability of PCs to explain de variation in data


Scree Plot
Eigenvalue
Component Number
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13

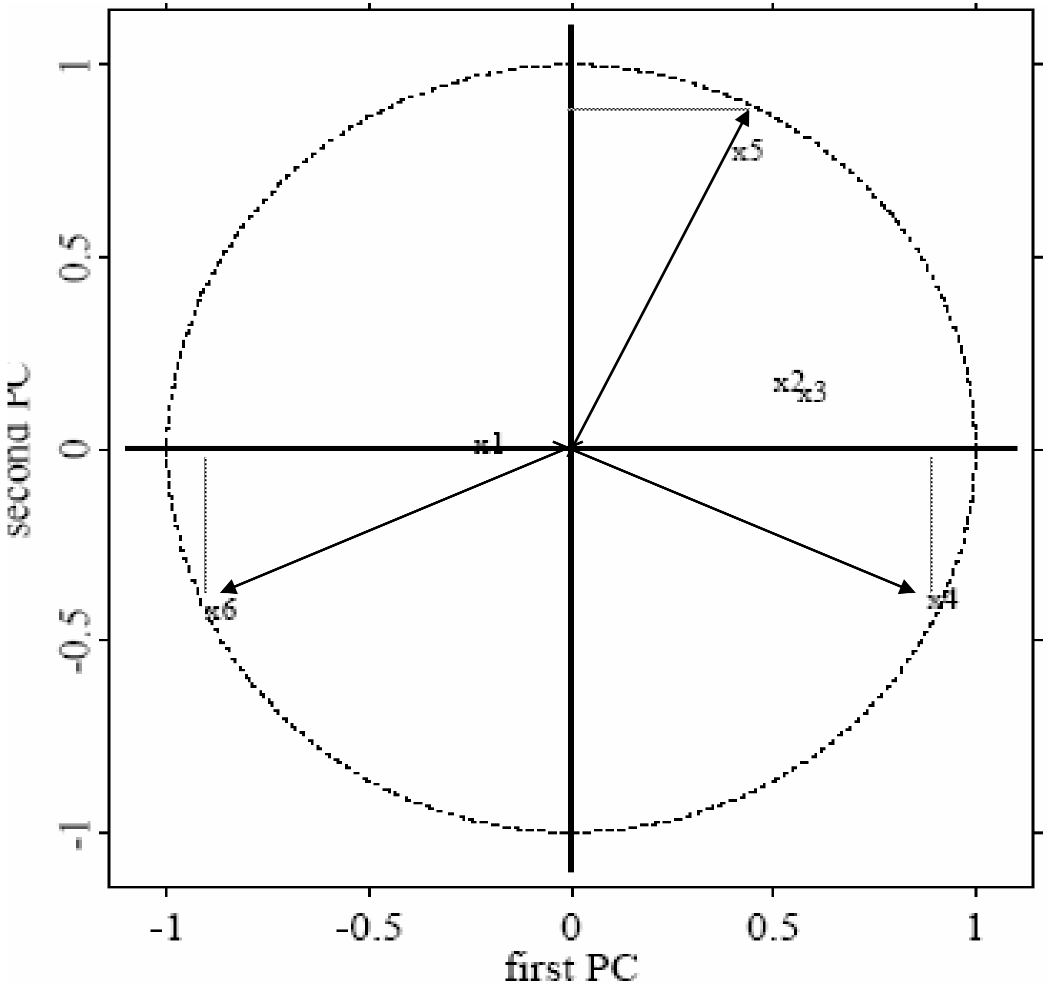
Do it graphically



Interpretation of components

- See the wheights of variables in each component
- If $Y_1 = 0.89\, X_1 + 0.15 X_2 - 0.77 X_3 + 0.51 X_4$
- Then X_1 *and* X_3 have the highest wheights and so are the most important variable in the first PC
- See the correlation between variables X_i and PCs: circle of correlation

Circle of correlation



Normalized (standardized) PCA

- If variables have very heterogenous variances we standardize them
- The standardized variables X_i*

$$X_i^* = (X_i - mean) / \sqrt{variance}$$

- The new variables all have the same variance (1), so each variable have the same wheight.

Application of PCA in Genomics

- PCA is useful for finding new, more informative, uncorrelated features; it reduces dimensionality by rejecting low variance features
- Analysis of expression data
- Analysis of metabolomics data (Ward et al., 2003)

However

- PCA is only powerful if the biological question is related to the highest variance in the dataset
- If not other techniques are more useful : Independent Component Analysis
- Introduced by Jutten in 1987