Principal component analysis

Problem Statement

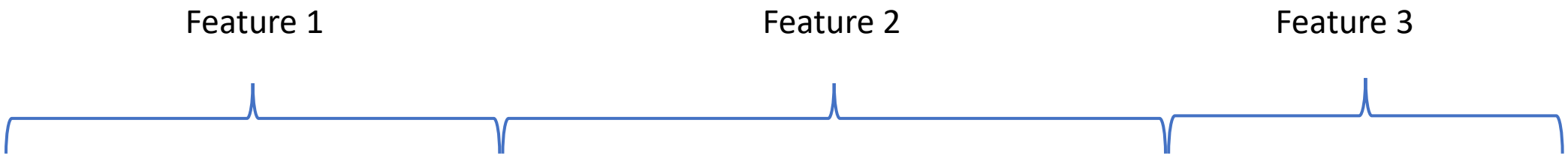
High dimensional data is difficult to analyse:

- many variables (p)
- sometimes no clear response variable

Key challenge is to reduce dimensionality in such datasets while retaining (most of) the original information

PCA is common unsupervised method for reducing dimensionality in high dimensional datasets

Problem Statement



	Blood pressure	Heart rate	Respiratory rate	Platelets	Lymphocytes	Red cells	Eosinophils	Neutrophils	BMI	Body fat	Cholesterol
Patient 1											
Patient 2											
Patient 3											
Patient 4											
Patient 5											
Patient 6											
Patient 7											
Patient 8											
Patient 9											
Patient 10											
Patient 11											
Patient 12											
Patient 13											
Patient 14											
Patient 15											

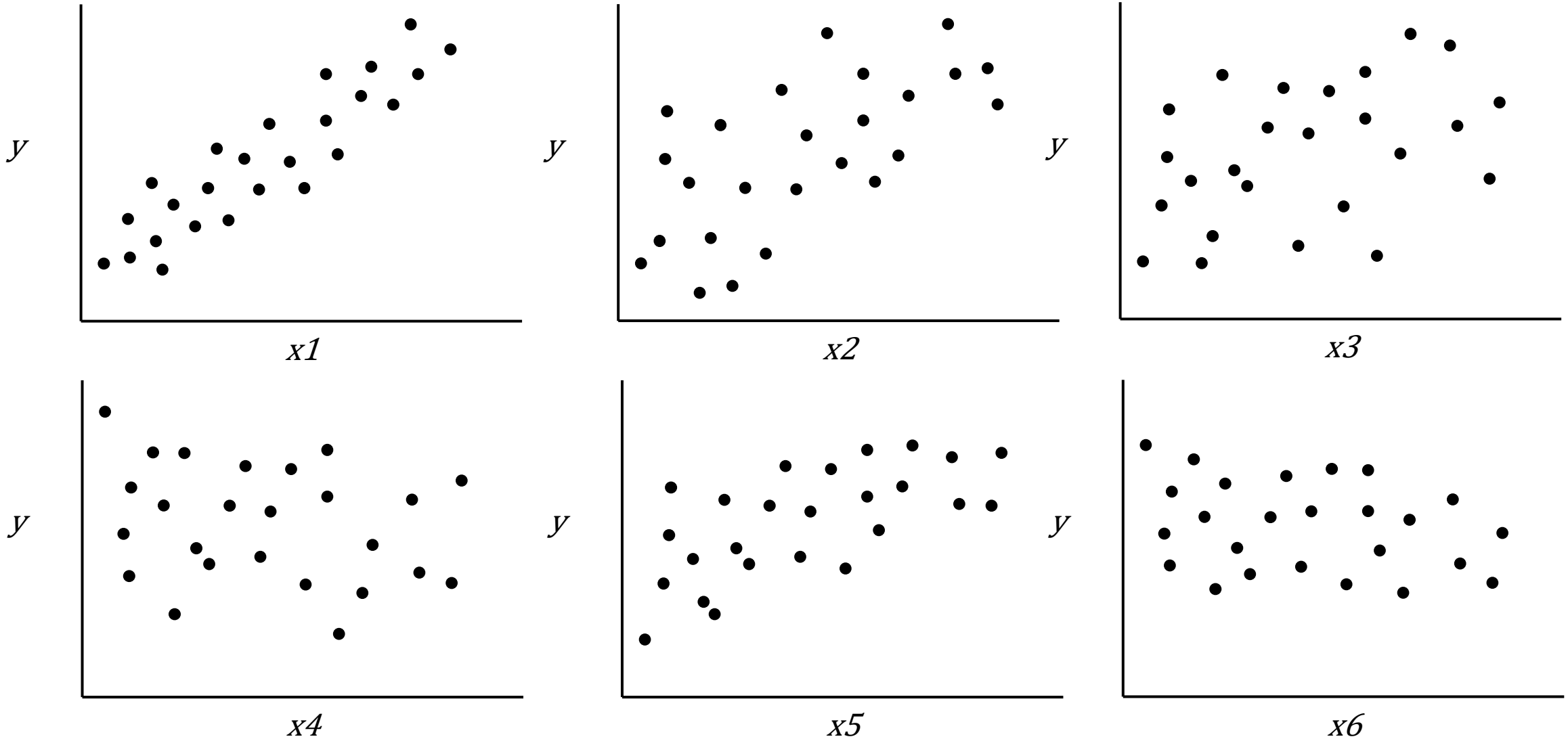
PCA and high dimensional data

PCA allows **large numbers** of correlated variables to be summarised into **smaller numbers** of uncorrelated variables (**principal components**)

Useful data exploration tool as it allows relationships between variables to be observed. Principal components calculated from PCA can be used in further analysis, e.g. linear regression

Scores for each principal component are calculated for each data point in the original dataset. The principal components are single variables which are calculated using a linear combination of several variables

PCA and high dimensional data



PCA and high dimensional data

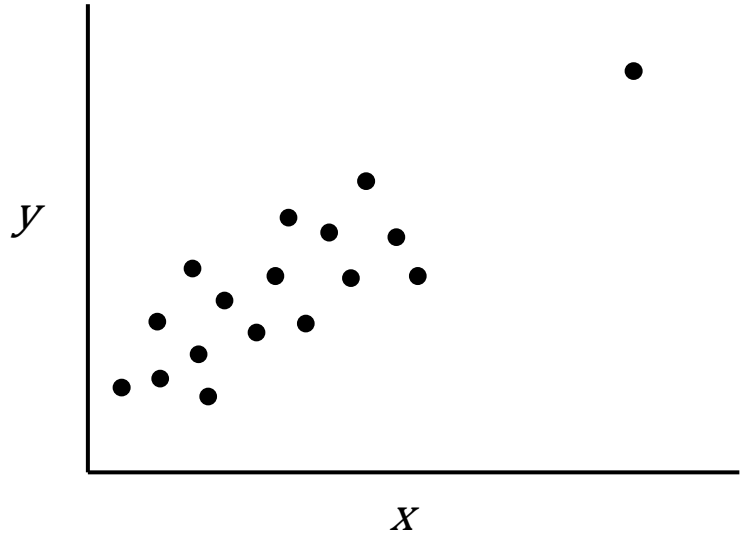
Advantages:

- popular and well understood method
- easy to implement with many softwares/packages available
- basic calculations used for PCs and PC scores are easy to understand

$$Z_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$$

Limitations:

- assumes that variables in original dataset are correlated
- scale sensitive
- outliers
- assumes linear relationship between variables
- difficult to interpret principal components



Challenge 1

Descriptions of three datasets and research questions are given below. For which of these might PCA be considered a useful tool for analysing data so that the research questions may be addressed?

A. An epidemiologist has data collected from different patients admitted to hospital with infectious respiratory disease. They would like to determine whether length of stay in hospital differs in patients with different respiratory diseases.
B. An online retailer has collected data on user interactions with its online app and has information on the number of times each user interacted with the app, what products they viewed per interaction, and the type and cost of these products. The retailer would like to use this information to predict whether or not a user will be interested in a new product.
C. A scientist has assayed gene expression levels in 1000 cancer patients and has data from probes targeting different genes in tumour samples from patients. She would like to create new variables representing relative abundance of different groups of genes to i) find out if genes form subgroups based on biological function and ii) use these new variables in a linear regression examining how gene expression varies with disease severity.
D. All of the above

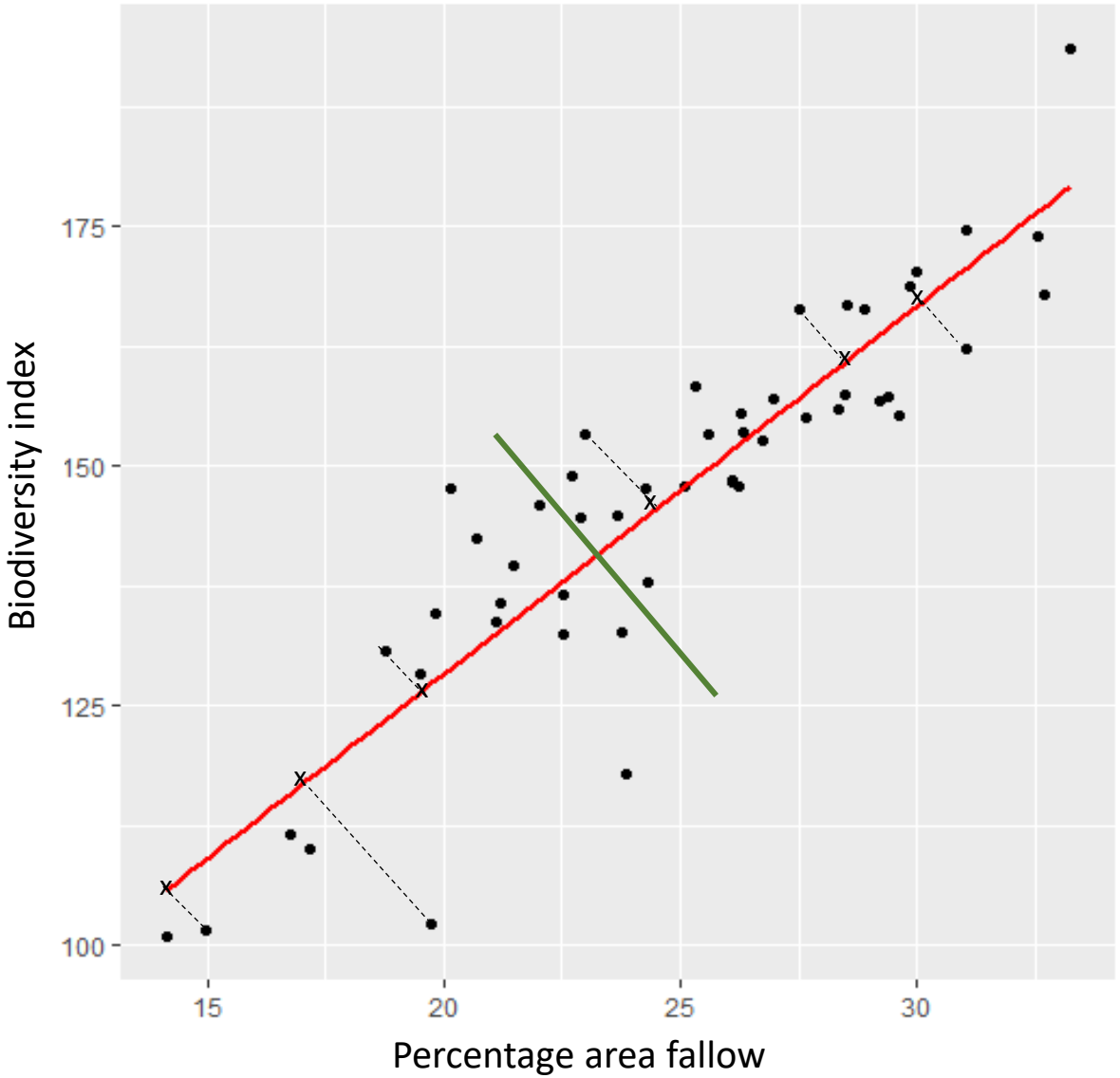
Carrying out PCA

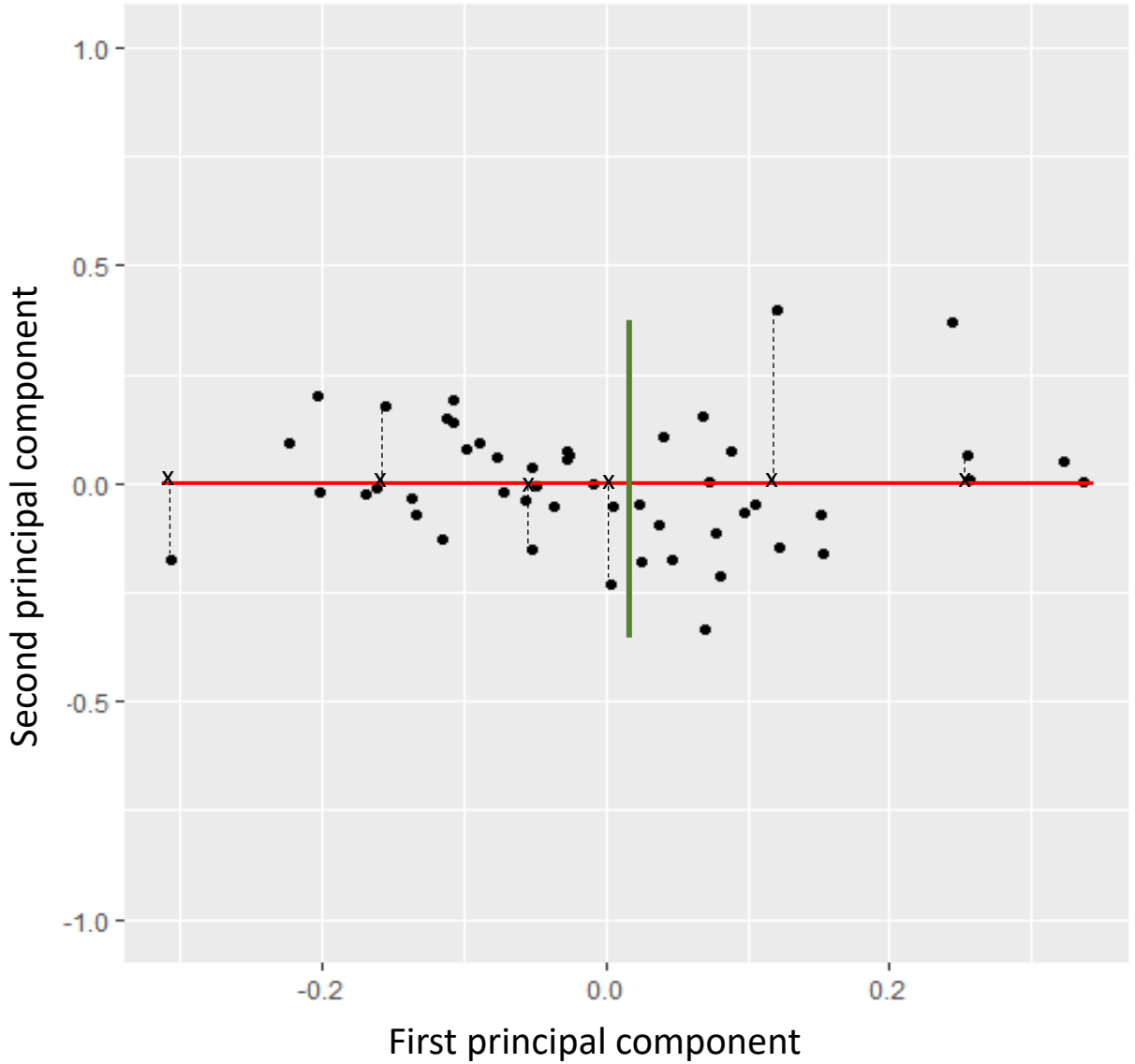
PCA is carried out by calculating a matrix from the original dataset which shows how each of the variables in the dataset relate to each other

This matrix can then be broken down so that the direction and magnitude of the data can be observed (i.e. how strongly variables are related and in what direction)

The first principal component is the direction along which the data show the **most variation**

The second principal component is the direction along which the data show the **next highest amount of variation**, unrelated to the first principal component


Biodiversity index
175
150
125
100
15
20
25
30
Percentage area fallow


Second principal component
1.0
0.5
0.0
-0.5
-1.0
-0.2
0.0
0.2
First principal component

Carrying out PCA

The principal component score for the first principal component in the previous example is calculated using:

$$Z_{i1} = a_1 \times (fallow_i - \overline{fallow}) + a_2 \times (bio\ index_i - \overline{bio\ index})$$

a_1 and a_2 are the **principal component loadings** which represent the direction along which the variation in the data is maximised. A loading value is calculated for each variable added to this linear equation

Loadings describe how much each variable contributes to a specific principal component

There are as many principal components as there are variables in your dataset, but as we'll see, some are more useful at explaining your data than others

Challenge 2

Why might it be necessary to standardise variables before performing a PCA?

A. To make the results of the PCA interesting
B. To ensure that variables with different ranges of values contributes equally to analysis
C. To allow the feature matrix to be calculated faster, especially in cases where there are a lot of input variables
D. To allow both continuous and categorical variables to be included in the PCA
E. All of the above

Can you think of datasets where it might not be necessary to standardise variables? Discuss in groups.

Example 1

The prostate data have **97** rows and **9** columns.

Columns (red = clinical continuous variables):

lcavol (log-transformed cancer volume)
lweight (log-transformed prostate weight)
lbph (log-transformed amount of benign prostate enlargement)
svi (seminal vesicle invasion)
lcp (log capsular penetration; amount of spread of cancer in outer walls of prostate)
gleason (Gleason score; grade of cancer cells)
pgg45 (percentage Gleason scores 4 or 5)
lpsa (log prostate specific antigen; level of PSA in blood)
age (patient age in years)

Example 1

We will calculate 97 principal component scores for each of the 97 rows in this dataset, using five principal components. We will include five clinical variables in our PCA, each of the continuous variables in the prostate dataset, so that we can create fewer variables representing clinical markers of cancer progression.

We will create a new dataset only including the variables lcavol, lweight, lbph, lcp and lpsa. We will standardise these variables before carrying out the PCA so that they have a mean of 0 and a standard deviation of 1.

Example 2

PCAtools provides functions for data exploration using PCA. Functions to apply different methods for choosing appropriate numbers of principal components are available

We will use PCAtools to explore some gene expression microarray data downloaded from the Gene Expression Omnibus.

A microarray is used to detect the expression of multiple different genes at the same time. Microarrays used to detect DNA sequences are microscope slides containing thousands of tiny spots in defined positions, in which each spot contains a DNA sequence encoding a particular gene. These DNA molecules can be thought of as probes which detect whether or not a particular gene is expressed in an input sample.

Example 2

The dataset we will be analysing using **PCAtools** includes two subsets of data:

- a matrix of gene expression data showing microarray results for different probes used to examine gene expression profiles in 91 different breast cancer patient samples
- metadata associated with the gene expression results with information from patients

To start our analysis we will download the **BiocManager** and **PCAtools** packages from BioConductor. The BiocManager package is used to install packages from Bioconductor and PCAtools provides functions that can be used to explore data via PCA and produce useful figures and analysis tools.

Example 2

Microarray data are difficult to analyse because:

- they are high dimensional
- formulating a research question using microarray data can be difficult, especially if not much is known about which genes code for phenotypes of interest
- exploratory analysis is difficult due to the number of potentially interesting response variables (i.e. expression data from probes targeting different genes).

Hypothesis: groups of genes are associated with different phenotypic characteristics of cancers (e.g. histologic grade, tumour size)

Challenge 3

Apply a PCA to the cancer gene expression data using the '**pca**' function from **PCAtools**. You can use the help files in PCAtools to find out about the 'pca' function (type ?pca in R). Remove the lower 20% of principal components from your PCA using the removeVar argument in the pca function. As in the example using prostate data above, examine the first 5 rows and columns of rotated data and loadings from your PCA

Challenge 4

Using the **screeplot** function in **PCAtools**, create a screeplot to show proportion of variance explained by each principal component. Explain the output of the screeplot in terms of proportion of variance in data explained by each principal component

Challenge 5

Create a biplot of the first two principal components from your PCA (using **biplot** function in PCAtools - see helpfile for arguments) and determine whether samples cluster based on variables in metadata. Explain you results

Challenge 6

Use '**colby**' and '**lab**' arguments in **biplot** to explore whether these two groups may cluster by Age or by whether or not the sample expresses the Estrogen Receptor gene (ER+ or ER-)

Further reading

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R. Chapter 6.3 (Dimension Reduction Methods), Chapter 10 (Unsupervised Learning)

Jolliffe, I.T. & Cadima, J. (2016) Principal component analysis: a review and recent developments. Phil. Trans. R. Soc A 374. http://dx.doi.org/10.1098/rsta.2015.0202

Johnstone, I.M. & Titterington, D.M. (2009) Statistical challenges of high-dimensional data. Phil. Trans. R. Soc A 367. doi:10.1098/rsta.2009.0159

PCA: A Practical Guide to Principal Component Analysis
https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

A One-Stop Shop for Principal Component Analysis
https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c