Introduction to high-dimensional data

What are high-dimensional data?

Data in which the number of features, *p*, are equal or larger than the number of observations, *n*

Compare with low-dimensional data in which *n* is much greater than *p*

High-dimensional data have become more common as new automated data collection techniques have been developed

Subjects like genomics and medical sciences often use both large (in terms of *n*) and wide (in terms of *p*) datasets that can be difficult to analyse or visualise using standard statistical tools

What are high-dimensional data?

An example of a high-dimensional dataset might look like this:

	Blood pressure	Heart rate	Respiratory rate	Platelets	Lymphocytes	Red cells	Eosinophils	Neutrophils	BMI	Body fat	Cholesterol	Contact with infected person	Length of contact	Symptom severity	Number of symptoms	Length of stay	Survival
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																	

Challenge 1

Descriptions of three research questions and their datasets are given below. Which of these are considered to have high-dimensional data?

A. Predicting patient blood pressure using cholesterol level in blood, age, and BMI measurements collected from 100 patients
B. Predicting patient blood pressure using cholesterol level in blood, age, and BMI as well as information from 200,000 single nucleotide polymorphisms from 100 patients
C. Predicting length of time patients spend in hospital with pneumonia infection using measurements on age, BMI, length of time with symptoms, number of symptoms, and percentage of neutrophils in blood using data from 200 patients
D. Predicting probability of a patient's cancer progressing using gene expression data as well as data associated with general patient health (age, weight, BMI, blood pressure) and cancer growth (tumour size, localised spread, blood test results)

Challenges dealing with high-dimensional data

Datasets with large numbers of features are difficult to visualise

With low-dimensional data, we can easily plot the response variable against each explanatory variable. With high-dimensional data this is difficult due to large numbers of variables

In some high-dimensional datasets it can also be difficult to identify a single response variable

Let's have a look at a simple dataset with lots of features to understand some of the challenges we are facing when working with high-dimensional data

Challenge 2

Load the `Prostate` dataset from the `lasso2` package and examine the column names.

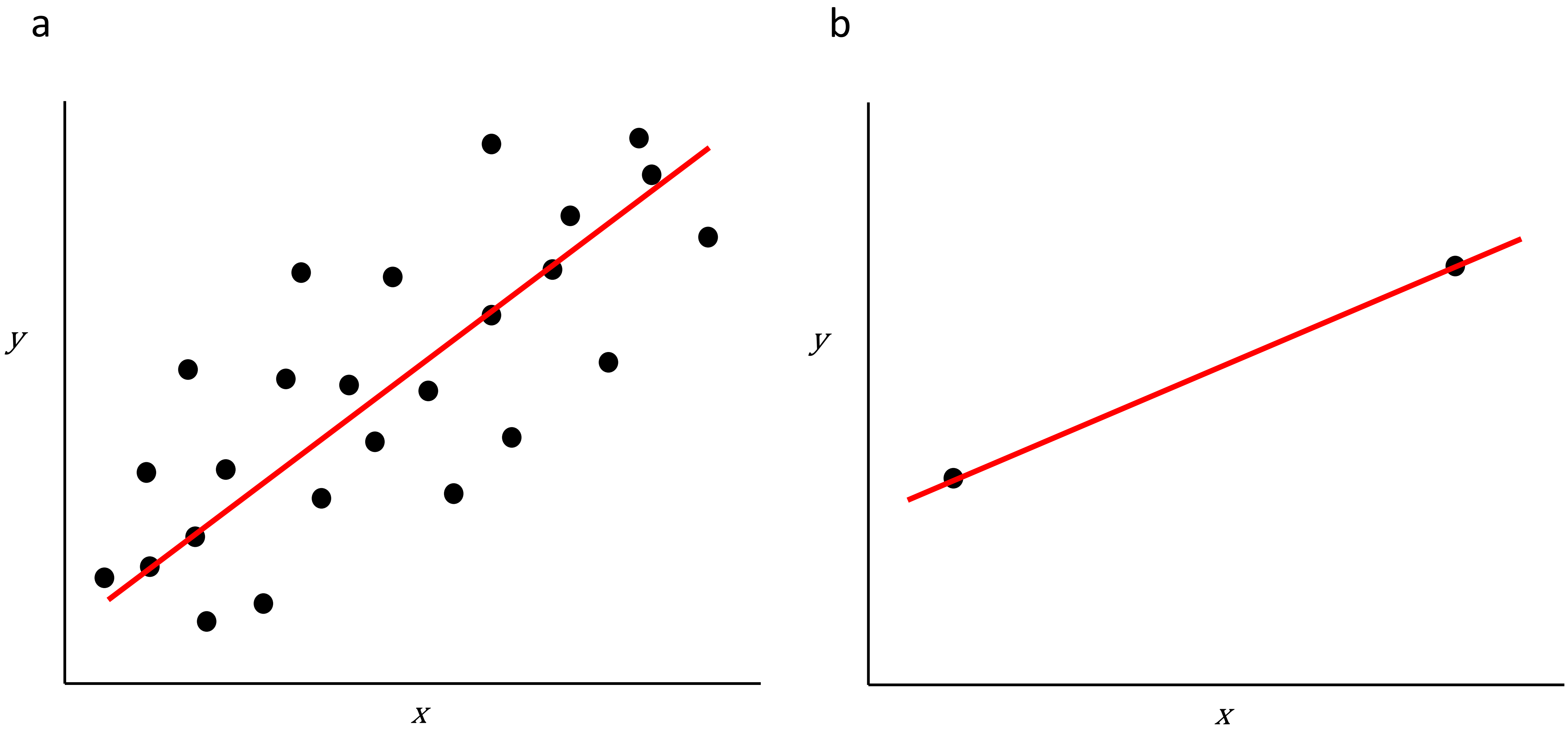
Examine the dataset (in which each row represents a single patient) and plot relationships between the variables using the `pairs` function. Why does it become more difficult to plot relationships between pairs of variables with increasing numbers of variables? Discuss in groups.

Challenges dealing with high-dimensional data

Fitting regression models to datasets with large numbers of variables is difficult due to the potential for overfitting and in identifying a response variable

Imagine we are carrying out least squares regression on a dataset with 25 observations. Fitting a best fit line through these data produces a plot shown in Figure 2a

Imagine the effective number of observations per features is low. The result of fitting a best fit line through few observations can be seen in Figure 2b


a
y
x
b
y
x

Challenges dealing with high-dimensional data

In the first plot, the least squares regression lines does not fit the data perfectly and there is some error around the regression line

When there are only two observations the regression line will fit through the points exactly, resulting in overfitting of the data

Another problem in carrying out regression on high-dimensional data is dealing with correlations between explanatory variables

Challenge 3

Use the `cor` function to examine correlations between all variables in the Prostate dataset. Are some variables highly correlated (i.e. correlation coefficiants >0.6)? Carry out a linear regression predicting patient age using all variables in the Prostate dataset.

Statistical methods for high-dimensional data

As we found out in the challenges, carrying out linear regression on datasets with large numbers of features is difficult due to: high correlation between variables; difficulty in identifying clear a response variable; risk of overfitting

While linear regression cannot be used in datasets with many features, high-dimensional regression methods are available with methods to deal with overfitting

In situations where the response variable is difficult to identify or where explanatory variables are highly correlated, dimensionality reduction may be used to create fewer variables that represent variation in the original dataset

Statistical methods (such as hierarchical clustering and k-means clustering) are often used to identify clusters within complex datasets

Challenge 4

Change the value of `sd` in the above example. What happens to the data when `sd` is increased?

High-dimensional data in the biosciences

In this workshop, we will look at statistical methods that can be used to visualise and analyse high-dimensional biological data using packages available from Bioconductor (https://www.bioconductor.org/)

Bioconductor packages can be installed and used in `R` using the `BiocManager` package. We can explore these packages by browsing the vignettes provided in Bioconductor

We load the `methylation` dataset which represents data collected using Illumina Infinium methylation arrays which are used to examine methylation across the human genome

The size of the dataset makes it difficult to examine in full, a common challenge in analysing high-dimensional genomics data

Further reading

Buhlman, P. & van de Geer, S. (2011) Statistics for High-Dimensional Data. Springer, London.

Buhlman, P., Kalisch, M. & Meier, L. (2014) High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application https://doi.org/10.1146/annurev-statistics-022513-115545

Johnstone, I.M. & Titterington, D.M. (2009) Statistical challenges of high-dimensional data. Philosophical Transactions of the Royal Society A 367:4237-4253.

https://www.bioconductor.org/packages/release/workflows/vignettes/methylationArrayAnalysis/inst/doc/methylationArrayAnalysis.html