

Multivariate Statistics and Machine Learning

Korbinian Strimmer

23 October 2021

Contents

Welcome	5
License	5
Preface	7
About the author	7
About the module	7
Acknowledgements	8
1 Multivariate random variables	9
1.1 Why multivariate statistics?	9
1.2 Essentials in multivariate statistics	10
1.3 Multivariate normal distribution	14
1.4 Estimation in large sample and small sample settings	20
1.5 Categorical and multinomial distribution	32
1.6 Further multivariate distributions	42
2 Transformations and dimension reduction	49
2.1 Linear Transformations	49
2.2 Nonlinear transformations	52
2.3 Whitening transformations	54
2.4 Natural whitening procedures	60
2.5 Principal Component Analysis (PCA)	70
2.6 Correlation loadings plot to interpret PCA components	73
2.7 CCA whitening (Canonical Correlation Analysis)	75
3 Unsupervised learning and clustering	79
3.1 Challenges in supervised learning	79
3.2 Hierarchical clustering	82
3.3 K-means clustering	88
3.4 Mixture models	94
3.5 Fitting mixture models to data	101
4 Supervised learning and classification	107

4.1	Introduction	107
4.2	Bayesian discriminant rule or Bayes classifier	109
4.3	Normal Bayes classifier	110
4.4	The training step — learning QDA, LDA and DDA classifiers from data	112
4.5	Goodness of fit and variable ranking	115
4.6	Variable selection and cross-validation	119
5	Multivariate dependencies	123
5.1	Measuring the linear association between two sets of random variables	123
5.2	Mutual information as generalisation of correlation	125
5.3	Graphical models	130
6	Nonlinear and nonparametric models	137
6.1	Limits of linear models and correlation	138
6.2	Random forests	141
6.3	Gaussian processes	143
6.4	Neural networks	150
A	Brief refresher on matrices	153
A.1	Matrix basics	153
A.2	Simple matrix operations	155
A.3	Matrix summaries	155
A.4	Matrix inverse	157
A.5	Orthogonal matrices	158
A.6	Eigenvalues and eigenvectors	161
A.7	Matrix decompositions	164
A.8	Matrix summaries based on eigenvalues and singular values . . .	166
A.9	Functions of symmetric matrices	167
A.10	Matrix calculus	168
B	Further study	171
B.1	Recommended reading	171
B.2	Advanced reading	171
	Bibliography	173

Welcome

These are the lecture notes for MATH20802, a course in **Multivariate Statistics and Machine Learning** for third year mathematics students at the [Department of Mathematics of the University of Manchester](#).

The course text was written by [Korbinian Strimmer](#) from 2018–2021. This version is from 23 October 2021.

The notes will be updated from time to time. To view the current version visit the [online MATH38161 lecture notes](#). You may also [download the MATH38161 lecture notes as PDF](#).

License

These notes are licensed to you under [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](#).

Preface

About the author

Hello! My name is Korbinian Strimmer and I am a Professor in Statistics. I am part of the [Statistics group at the Department of Mathematics of the University of Manchester](#). You can find more information about me on [my home page](#).

I have first taught this module in the winter term 2018 at the University of Manchester, and subsequently also in 2019, 2020, and 2021.

I hope you enjoy the course! If you have any questions, comments, or corrections then please email me at korbinian.strimmer@manchester.ac.uk.

About the module

Topics covered

The MATH38161 module is designed to run over the course of 11 weeks. It has six parts, each covering a particular aspect of multivariate statistics and machine learning:

1. Multivariate random variables and estimation in large and small sample settings (W1 and W2)
2. Transformations and dimension reduction (W3 and W4)
3. Unsupervised learning/clustering (W5 and W6)
4. Supervised learning/classification (W7 and W8)
5. Measuring and modelling multivariate dependencies (W9)
6. Nonlinear and nonparametric models (W10, W11)

This module focuses on:

- *Concepts and methods (not on theory)*
- *Implementation and application in R*
- *Practical data analysis and interpretation (incl. report writing)*
- *Modern tools in data science and statistics (R markdown, R studio)*

Additional support material

Accompanying these notes are

- [lecture videos](#) (visualiser style).

Furthermore, there is also an [MATH38161 online reading list](#) hosted by the University of Manchester library.

If you are a University of Manchester student and enrolled in this module you will find on [Blackboard](#):

- a weekly learning plan for an 11 week study period,
- weekly worksheets with with examples (theory and application in R) and solutions in R Markdown, and
- exam papers of previous years.

Acknowledgements

Many thanks to [Beatriz Costa Gomes](#) for her help to compile the first draft of these course notes in the winter term 2018 while she was a graduate teaching assistant for this course. I also thank the many students who suggested corrections.

Chapter 1

Multivariate random variables

1.1 Why multivariate statistics?

Science uses experiments to verify hypotheses about the world. Statistics provides tools to quantify this procedure and offers methods to link data (experiments) with probabilistic models (hypotheses). Since the world is complex we need complex models and complex data, hence the need for multivariate statistics and machine learning.

Specifically, multivariate statistics (as opposed to univariate statistics) is concerned with methods and models for **random vectors** and **random matrices**, rather than just random univariate (scalar) variables. Therefore, in multivariate statistics we will frequently make use of matrix notation.

Closely related to multivariate statistics (traditionally a subfield of statistics) is machine learning (ML) which is traditionally a subfield of computer science. ML used to focus more on algorithms rather on probabilistic modelling but nowadays most machine learning methods are fully based on statistical multivariate approaches, so the two fields are converging.

Multivariate models provide a means to learn dependencies and interactions among the components of the random variables which in turn allow us to draw conclusion about underlying mechanisms of interest (e.g. biological or medical).

Two main tasks:

- unsupervised learning (finding structure, clustering)
- supervised learning (training from labeled data, followed by prediction)

Challenges:

- complexity of model needs to be appropriate for problem and available data,

- high dimensions make estimation and inference difficult
- computational issues.

1.2 Essentials in multivariate statistics

1.2.1 Univariate vs. multivariate random variables

Univariate random variable (dimension $d = 1$):

$$x \sim F$$

where x is a **scalar** and F is the distribution. $E(x) = \mu$ denotes the mean and $\text{Var}(x) = \sigma^2$ the variance of x .

Multivariate random **vector** of dimension d :

$$\mathbf{x} = (x_1, x_2, \dots, x_d)^T \sim F$$

\mathbf{x} is **vector** valued random variable.

The vector \mathbf{x} is column vector (=matrix of size $d \times 1$). Its components x_1, x_2, \dots, x_d are univariate random variables. The dimension d is also often denoted by p or q .

1.2.2 Mean of a random vector

The mean / expectation of a random vector with dimensions d is also a vector with dimensions d :

$$E(\mathbf{x}) = \boldsymbol{\mu} = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_d) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

1.2.3 Variance of a random vector

Recall the definition of mean and variance for a univariate random variable:

$$E(x) = \mu$$

$$\text{Var}(x) = \sigma^2 = E((x - \mu)^2) = E((x - \mu)(x - \mu)) = E(x^2) - \mu^2$$

Definition of **variance of a random vector**:

$$\text{Var}(\mathbf{x}) = \underbrace{\boldsymbol{\Sigma}}_{d \times d} = \mathbb{E} \left(\underbrace{(\mathbf{x} - \boldsymbol{\mu})}_{d \times 1} \underbrace{(\mathbf{x} - \boldsymbol{\mu})^T}_{1 \times d} \right) = \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

The variance of a random vector is, therefore, **not** a vector but a **matrix**!

$$\boldsymbol{\Sigma} = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \dots & \sigma_{dd} \end{pmatrix}$$

This matrix is called the **Covariance Matrix**, with off-diagonal elements $\sigma_{ij} = \text{Cov}(x_i, x_j)$ and the diagonal $\sigma_{ii} = \text{Var}(X_i) = \sigma_i^2$.

1.2.4 Properties of the covariance matrix

1. $\boldsymbol{\Sigma}$ is real valued: $\sigma_{ij} \in \mathbb{R}$
2. $\boldsymbol{\Sigma}$ is symmetric: $\sigma_{ij} = \sigma_{ji}$
3. The diagonal of $\boldsymbol{\Sigma}$ contains $\sigma_{ii} = \text{Var}(x_i) = \sigma_i^2$, i.e. the variances of the components of \mathbf{x} .
4. Off-diagonal elements $\sigma_{ij} = \text{Cov}(x_i, x_j)$ represent linear dependencies among the x_i . \implies linear regression, correlation

How many separate entries does $\boldsymbol{\Sigma}$ have?

$$\boldsymbol{\Sigma} = (\sigma_{ij}) = \underbrace{\begin{pmatrix} \sigma_{11} & \dots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \dots & \sigma_{dd} \end{pmatrix}}_{d \times d}$$

with $\sigma_{ij} = \sigma_{ji}$.

Number of separate entries: $\frac{d(d+1)}{2}$.

This numbers grows with the square of the dimension d , i.e. is of order $O(d^2)$:

d	# entries
1	1
10	55
100	5050
1000	500500
10000	50005000

For large dimension d the covariance matrix has many components!

→ computationally expensive (both for storage and in handling) → very challenging to estimate in high dimensions d .

Note: matrix inversion requires $O(d^3)$ operations using standard algorithms such as [Gauss Jordan elimination](#).¹ Hence, computing Σ^{-1} is computationally expensive for large d !

1.2.5 Eigenvalue decomposition of Σ

Recall the orthogonal eigendecomposition from matrix analysis and linear algebra: A symmetric matrix with real entries has real eigenvalues and a complete set of orthogonal eigenvectors.

Applying eigenvalue decomposition to the covariance matrix yields

$$\Sigma = U\Lambda U^T$$

where U is an orthogonal matrix containing the eigenvectors and

$$\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}$$

contains the eigenvalues λ_i .

Importantly, the eigenvalues of the covariance matrix are not only real valued but they are further constrained to be non-negative. This can be seen by computing the quadratic form $z^T \Sigma z$ for a non-zero non-random vector z which yields

$$\begin{aligned} z^T \Sigma z &= z^T E \left((x - \mu)(x - \mu)^T \right) z \\ &= E \left(z^T (x - \mu)(x - \mu)^T z \right) \\ &= E \left(\left(z^T (x - \mu) \right)^2 \right) \geq 0. \end{aligned}$$

Therefore the **covariance matrix Σ is always positive semi-definite**.

In fact, **unless there is collinearity** (i.e. a variable is a linear function the other variables) all eigenvalues will be positive and **Σ is positive definite**.

¹[Specialised matrix algorithms improve this](#) to about $O(d^{2.373})$. Matrices with special symmetries (e.g. diagonal and block diagonal matrices) or properties (e.g. orthogonal matrix) can also be inverted much easier (see Appendix A).

1.2.6 Quantities related to the covariance matrix

1.2.6.1 Correlation matrix P

The correlation matrix P (= upper case greek “rho”) is the standardised covariance matrix

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \text{Cor}(x_i, x_j)$$

$$\rho_{ii} = 1 = \text{Cor}(x_i, x_i)$$

$$P = (\rho_{ij}) = \begin{pmatrix} 1 & \dots & \rho_{1d} \\ \vdots & \ddots & \vdots \\ \rho_{d1} & \dots & 1 \end{pmatrix}$$

where P (“capital rho”) is a symmetric matrix ($\rho_{ij} = \rho_{ji}$).

Note the **variance-correlation decomposition**

$$\Sigma = V^{\frac{1}{2}} P V^{\frac{1}{2}}$$

where V is a diagonal matrix containing the variances:

$$V = \begin{pmatrix} \sigma_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{dd} \end{pmatrix}$$

$$P = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$$

This is the definition of correlation written in matrix notation.

1.2.6.2 Precision matrix or concentration matrix

$$\Omega = (\omega_{ij}) = \Sigma^{-1}$$

Ω (“Omega”) is the inverse of the covariance matrix.

The inverse of the covariance matrix can be obtained via the spectral decomposition, followed by inverting the eigenvalues λ_i :

$$\Sigma^{-1} = U \Lambda^{-1} U^T = U \begin{pmatrix} \lambda_1^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d^{-1} \end{pmatrix} U^T$$

Note that **all eigenvalues λ_i need to be positive so that Σ can be inverted.** (i.e., Σ needs to be positive definite).

If any $\lambda_i = 0$ then Σ is singular and not invertible.

Importance of Σ^{-1} :

- Many expressions in multivariate statistics contain Σ^{-1} and not Σ .
- Σ^{-1} has close connection with graphical models (e.g. conditional independence graph, partial correlations).
- Σ^{-1} is a natural parameter from an exponential family perspective.

1.2.6.3 Partial correlation matrix

This is a standardised version of the precision matrix, see later chapter on graphical models.

1.2.6.4 Total variation and generalised variance

To summarise the covariance matrix Σ in a single scalar value there are two commonly used measures:

- **total variation:** $\text{Tr}(\Sigma) = \sum_{i=1}^d \lambda_i$
- **generalised variance:** $\det(\Sigma) = \prod_{i=1}^d \lambda_i$

The generalised variance $\det(\Sigma)$ is also known as the volume of Σ .

1.3 Multivariate normal distribution

The multivariate normal model is a generalisation of the univariate normal distribution from dimension 1 to dimension d .

1.3.1 Univariate normal distribution:

Dimension $d = 1$

$$x \sim N(\mu, \sigma^2)$$

$$E(x) = \mu, \text{Var}(x) = \sigma^2$$

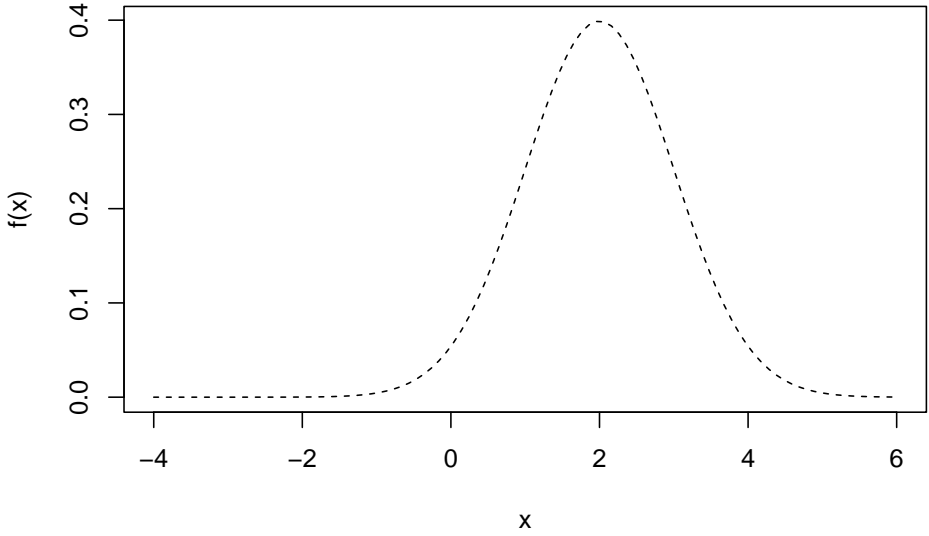
Density:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Plot of univariate normal density :

Unimodal with peak at μ , width determined by σ (in this plot: $\mu = 2, \sigma^2 = 1$)

Density Normal Distribution



Special case: **standard normal** with $\mu = 0$ and $\sigma^2 = 1$:

$$f(x|\mu = 0, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Differential entropy:

$$H(F) = \frac{1}{2}(\log(2\pi\sigma^2) + 1)$$

Cross-entropy:

$$H(F_{\text{ref}}, F) = \frac{1}{2} \left(\frac{(\mu - \mu_{\text{ref}})^2}{\sigma^2} + \frac{\sigma_{\text{ref}}^2}{\sigma^2} + \log(2\pi\sigma^2) \right)$$

KL divergence:

$$D_{\text{KL}}(F_{\text{ref}}, F) = H(F_{\text{ref}}, F) - H(F_{\text{ref}}) = \frac{1}{2} \left(\frac{(\mu - \mu_{\text{ref}})^2}{\sigma^2} + \frac{\sigma_{\text{ref}}^2}{\sigma^2} - \log\left(\frac{\sigma_{\text{ref}}^2}{\sigma^2}\right) - 1 \right)$$

Maximum entropy characterisation: the normal distribution is the unique distribution that has the highest (differential) entropy over all continuous distributions with support from minus infinity to plus infinity with a given mean and variance.

This is in fact one of the reasons why the normal distribution is so important (und useful) – if we only know that a random variable has a mean and variance,

and not much else, then using the normal distribution will be a reasonable and well justified working assumption!

1.3.2 Multivariate normal model

Dimension d

$$\mathbf{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}, \text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$$

Density:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \underbrace{(\mathbf{x} - \boldsymbol{\mu})^T}_{1 \times d} \underbrace{\boldsymbol{\Sigma}^{-1}}_{d \times d} \underbrace{(\mathbf{x} - \boldsymbol{\mu})}_{d \times 1} \right)$$

$1 \times 1 = \text{scalar!}$

- note that density contains precision matrix $\boldsymbol{\Sigma}^{-1}$
- inverting $\boldsymbol{\Sigma}$ implies inverting the eigenvalues λ_i of $\boldsymbol{\Sigma}$ (thus we need $\lambda_i > 0$)
- density also contains $\det(\boldsymbol{\Sigma}) = \prod_{i=1}^d \lambda_i \equiv$ product of eigenvalues of $\boldsymbol{\Sigma}$

Special case: **standard multivariate normal** with

$$\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$$

$$f(\mathbf{x}|\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I}) = (2\pi)^{-d/2} \exp \left(-\frac{1}{2} \mathbf{x}^T \mathbf{x} \right) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{x_i^2}{2} \right)$$

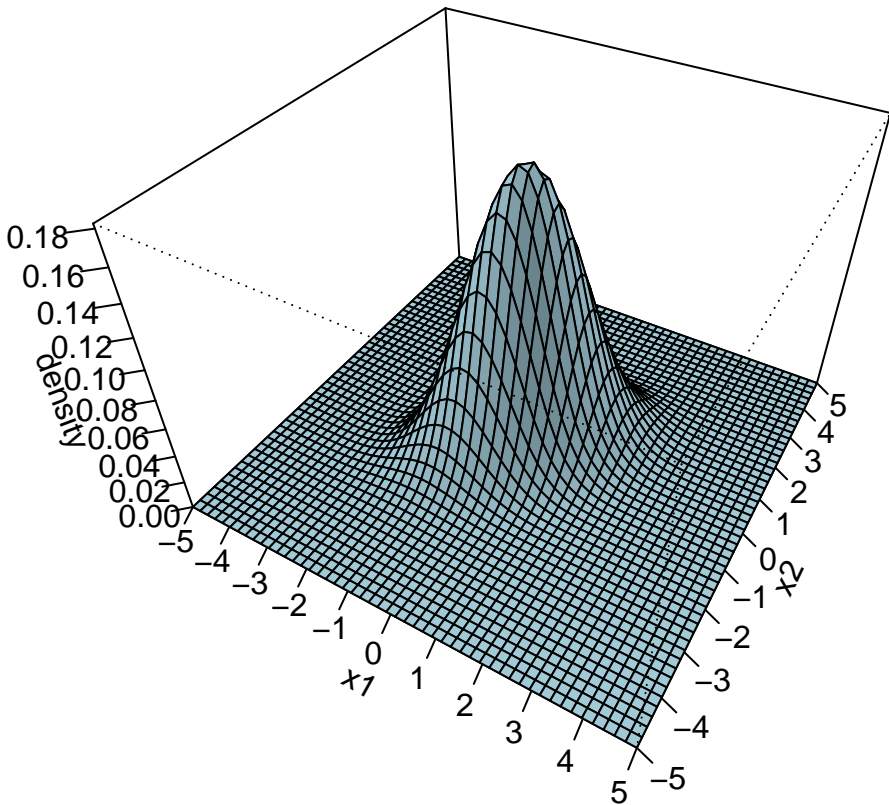
which is equivalent to the product of d univariate standard normals!

Misc:

- for $d = 1$, multivariate normal reduces to normal.
- for $\boldsymbol{\Sigma}$ diagonal (i.e. $\mathbf{P} = \mathbf{I}$, no correlation), MVN is the product of univariate normals (see Worksheet 2).

Plot of MVN density:

Bivariate Normal Density



- Location: μ
- Shape: Σ
- Unimodal: **one** peak
- Support from $-\infty$ to $+\infty$ in each dimension

An interactive [R Shiny web app](https://minerva.it.manchester.ac.uk/shiny/strimmer/bvn/) of the bivariate normal density plot is available online at <https://minerva.it.manchester.ac.uk/shiny/strimmer/bvn/>.

Differential entropy:

$$H = \frac{1}{2}(\log \det(2\pi\Sigma) + d)$$

Cross-entropy:

$$H(F_{\text{ref}}, F) = \frac{1}{2} \left\{ (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{ref}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{ref}}) + \text{Tr} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\text{ref}} \right) + \log \det \left(2\pi \boldsymbol{\Sigma} \right) \right\}$$

KL divergence:

$$\begin{aligned} D_{\text{KL}}(F_{\text{ref}}, F) &= H(F_{\text{ref}}, F) - H(F_{\text{ref}}) \\ &= \frac{1}{2} \left\{ (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{ref}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{ref}}) + \text{Tr} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\text{ref}} \right) - \log \det \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\text{ref}} \right) - d \right\} \end{aligned}$$

1.3.3 Shape of the multivariate normal density

Now we show that the contour lines of the multivariate normal density always take on the form of an ellipse, and that the radii of the ellipse is determined by the eigenvalues of $\boldsymbol{\Sigma}$.

We start by observing that a circle with radius r around the origin can be described as the set of points (x_1, x_2) satisfying $x_1^2 + x_2^2 = r^2$, or equivalently, $\frac{x_1^2}{r^2} + \frac{x_2^2}{r^2} = 1$. This is generalised to the shape of an ellipse by allowing (in two dimensions) for two radii r_1 and r_2 with $\frac{x_1^2}{r_1^2} + \frac{x_2^2}{r_2^2} = 1$, or in vector notation $\mathbf{x}^T \text{Diag}(r_1^2, r_2^2)^{-1} \mathbf{x} = 1$. In d dimensions and allowing for rotation of the axes and a shift of the origin from 0 to $\boldsymbol{\mu}$ the condition for an ellipse is

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q} \text{Diag}(r_1^2, \dots, r_d^2)^{-1} \mathbf{Q}^T (\mathbf{x} - \boldsymbol{\mu}) = 1$$

where \mathbf{Q} is an orthogonal matrix.

A contour line of a probability density function is a set of connected points where the density assumes the same constant value. In the case of the multivariate normal distribution keeping the density at some fixed value implies that $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$ where c is a constant. Using the eigenvalue decomposition of $\boldsymbol{\Sigma} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$ we can rewrite this condition as

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{U} \boldsymbol{\Lambda}^{-1} \mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}) = c.$$

This implies that the contour lines of the multivariate normal density are indeed ellipses and that the squared radii are proportional to the eigenvalues of $\boldsymbol{\Sigma}$. Equivalently, the positive square roots of the eigenvalues are proportional to the radii of the ellipse. Hence, for singular covariance matrix with one or more $\lambda_i = 0$ the corresponding radii are zero.

An interactive [R Shiny web app](https://minerva.it.manchester.ac.uk/shiny/striimmer/bvn/) to play with the contour lines of the bivariate normal distribution is online at <https://minerva.it.manchester.ac.uk/shiny/striimmer/bvn/>.

1.3.4 Three types of covariances

Following the above we can parameterise a covariance matrix in terms of its i) volume, ii) shape, and iii) orientation by writing

$$\Sigma = \lambda \mathbf{U} \mathbf{A} \mathbf{U}^T$$

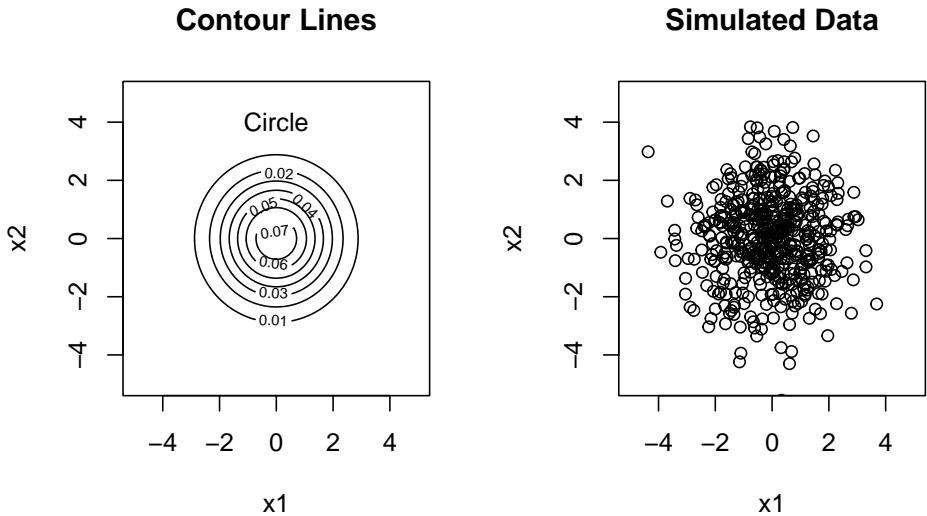
with $\mathbf{A} = \text{Diag}(a_1, \dots, a_d)$ and $\prod_{i=1}^d a_i = 1$ (and $\mathbf{\Lambda} = \lambda \mathbf{A}$).

- i) The **volume** is $\det(\Sigma) = \lambda^d$, determined by a single parameter λ .
- ii) The **shape** is determined by \mathbf{A} , with $d - 1$ free parameters.
- iii) The **orientation** is given by the orthogonal matrix \mathbf{U} , with $d(d - 1)/2$ free parameters.

This leads to classification of covariances into three varieties:

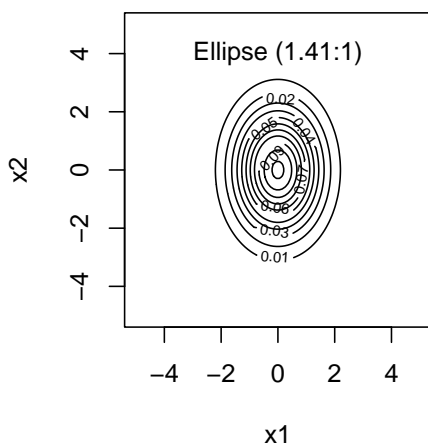
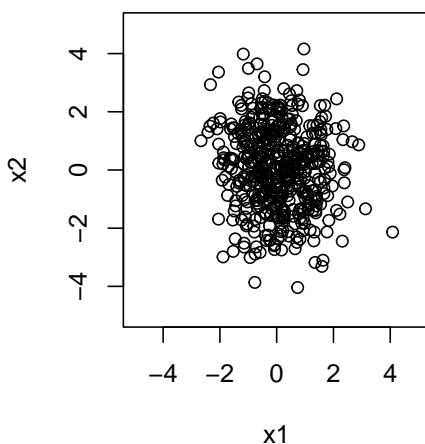
Type 1: spherical covariance $\Sigma = \lambda \mathbf{I}$, with spherical contour lines, 1 free parameter ($\mathbf{A} = \mathbf{I}$, $\mathbf{U} = \mathbf{I}$).

Example: $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ with $\sqrt{\lambda_1/\lambda_2} = 1$:



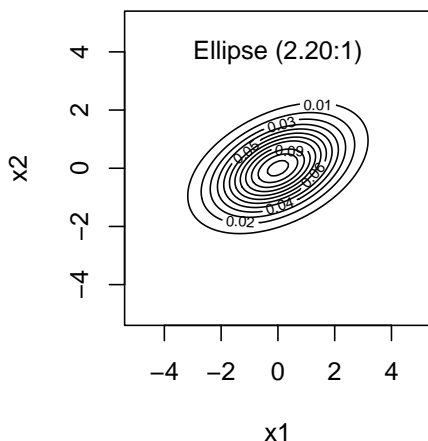
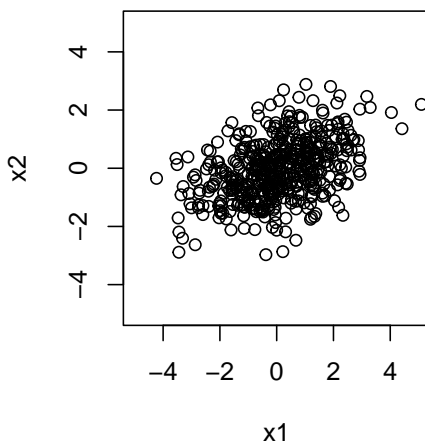
Type 2: diagonal covariance $\Sigma = \lambda \mathbf{A}$, with elliptical contour lines oriented along the coordinate axes, d free parameters ($\mathbf{U} = \mathbf{I}$).

Example: $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ with $\sqrt{\lambda_1/\lambda_2} \approx 1.41$:

Contour Lines**Simulated Data**

Type 3: general unrestricted covariance Σ , with elliptical contour lines oriented in any direction, $d(d+1)/2$ free parameters.

Example: $\Sigma = \begin{pmatrix} 2 & 0.6 \\ 0.6 & 1 \end{pmatrix}$ with $\sqrt{\lambda_1/\lambda_2} \approx 2.20$:

Contour Lines**Simulated Data**

1.4 Estimation in large sample and small sample settings

In practical application of multivariate normal model we need to learn its parameters from data. We first consider the case when there are many measurements

available, and then second the case when the number of data points is small compared to the number of parameters.

In a previous course in year 2 (see [MATH20802 Statistical Methods](#)) the method of maximum likelihood as well as essential of Bayesian statistics were introduced. Here we apply these approaches in the setting of the multivariate normal distribution.

1.4.1 Multivariate data

Vector notation:

Samples from a multivariate normal distribution are *vectors* (not scalars as for univariate normal):

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{iid}}{\sim} N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Matrix and component notation:

All the data points are commonly collected into a matrix \mathbf{X} .

In statistics the convention is to store each data vector in the rows of \mathbf{X} :

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & & & \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix}$$

Therefore,

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1d} \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} x_{21} \\ \vdots \\ x_{2d} \end{pmatrix}, \dots, \mathbf{x}_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{nd} \end{pmatrix}$$

Thus, in statistics the first index runs over $(1, \dots, n)$ and denotes the samples while the second index runs over $(1, \dots, d)$ and refers to the variables.

The statistics convention on data matrices is *not* universal! In fact, in most of the machine learning literature in engineering and computer science the data samples are stored in the columns so that the variables appear in the rows (thus in the engineering convention the data matrix is transposed compared to the statistics convention).

In order to avoid confusion it is recommended to use vector notation for data instead of matrix notation because this is not ambiguous.

1.4.2 Strategies for large sample estimation

1.4.2.1 Empirical estimators (outline)

For large n we have thanks to the law of large numbers:

$$\underbrace{F}_{\text{true}} \approx \underbrace{\hat{F}}_{\text{empirical}}$$

We now would like to estimate A which is a functional of F , i.e. $A = m(F)$. For example the mean, the median or some other quantity.

The *empirical estimate* is obtained by replacing the unknown true distribution F with the observed empirical distribution: $\hat{A} = m(\hat{F})$.

For example, the expectation of a random variable is approximated/estimated as the average over the observation:

$$E_F(x) \approx E_{\hat{F}}(x) = \frac{1}{n} \sum_{k=1}^n x_k$$

$$E_F(g(x)) \approx E_{\hat{F}}(g(x)) = \frac{1}{n} \sum_{k=1}^n g(x_k)$$

Simple recipe to obtain an empirical estimator: simply replace the expectation operator by the sample average in the quantity of interest.

What does this work: the empirical distribution \hat{F} is actually the nonparametric maximum likelihood estimate of F (see below for likelihood estimation).

Note: the approximation of F by \hat{F} also the basis other approaches such as Efron's bootstrap method.

1.4.2.2 Maximum likelihood estimation (outline)

R.A. Fisher (1922): model-based estimators using the density or probability mass function

log-likelihood function:

$$\log L(\theta) = \sum_{k=1}^n \underbrace{\log f}_{\text{log-density}} \left(\underbrace{x_i}_{\text{data}} \mid \underbrace{\theta}_{\text{parameters}} \right)$$

likelihood = probability to observe data given the model parameters

Maximum likelihood estimate:

$$\hat{\theta}^{\text{ML}} = \arg \max_{\theta} \log L(\theta)$$

Maximum likelihood (ML) finds the parameters that make the observed data most likely (it does *not* find the most probable model!)

Recall from [MATH20802 Statistical Methods](#) that maximum likelihood is closely linked to minimising the relative entropy (KL divergence) $D_{\text{KL}}(F, F_{\theta})$ between the unknown true model F to the specified model F_{θ} . Specifically, for large sample size n the model $F_{\hat{\theta}}$ fit by maximum likelihood is indeed the model that is closest to F .

Correspondingly, the great appeal of **maximum likelihood estimates** (MLEs) is that they **are optimal for large n** , i.e. so that **for large sample size no estimator can be constructed that outperforms the MLE** (note the emphasis on “for large n ”). A further advantage of the method of maximum likelihood is that it does not only provide a point estimate but also the asymptotic error (via the Fisher information which is related to the curvature of the log-likelihood function).

1.4.3 Large sample estimates of mean μ and covariance Σ

1.4.3.1 Empirical estimates:

Recall the definitions:

$$\mu = E(x)$$

and

$$\Sigma = E\left((x - \mu)(x - \mu)^T\right)$$

For the empirical estimate we replace the expectations by the corresponding sample averages.

These resulting estimators can be written in three different ways:

Vector notation:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})(x_k - \hat{\mu})^T = \frac{1}{n} \sum_{k=1}^n x_k x_k^T - \hat{\mu} \hat{\mu}^T$$

Data matrix notation:

The empirical mean and covariance can also be written in terms of the data matrix X (using statistics convention):

$$\hat{\mu} = \frac{1}{n} X^T \mathbf{1}_n$$

$$\widehat{\Sigma} = \frac{1}{n} \mathbf{X}^T \mathbf{X} - \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T$$

See Worksheet 2 for details.

Component notation:

The corresponding component notation with $\mathbf{X} = (x_{ki})$ is:

$$\hat{\mu}_i = \frac{1}{n} \sum_{k=1}^n x_{ki}$$

$$\hat{\sigma}_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - \hat{\mu}_i)(x_{kj} - \hat{\mu}_j)$$

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_d \end{pmatrix}, \widehat{\Sigma} = (\hat{\sigma}_{ij})$$

Variance estimate:

$$\hat{\sigma}_{ii} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - \hat{\mu}_i)^2$$

Note the factor $\frac{1}{n}$ (not $\frac{1}{n-1}$)

Engineering and machine learning convention:

Using the engineering and machine learning convention for the data matrix \mathbf{X} the estimators are written as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \mathbf{X} \mathbf{1}_n$$

$$\widehat{\Sigma} = \frac{1}{n} \mathbf{X} \mathbf{X}^T - \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T$$

In the corresponding component notation the two indices for the columns and rows are interchanged.

To avoid confusion when using matrix or component notation you need to always state which convention is used! In these notes we strictly follow the statistics convention.

1.4.3.2 Maximum likelihood estimates

We now derive the MLE of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the multivariate normal distribution. The corresponding log-likelihood function is

$$\begin{aligned}\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{k=1}^n \log f(\mathbf{x}_k | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log \det(\boldsymbol{\Sigma}) - \frac{1}{2} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}).\end{aligned}$$

Written in terms of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ this becomes

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Omega}) = -\frac{nd}{2} \log(2\pi) + \frac{n}{2} \log \det(\boldsymbol{\Omega}) - \frac{1}{2} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\mathbf{x}_k - \boldsymbol{\mu}).$$

Exploiting identities for the trace and log det (see Appendix) we can rewrite $(\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\mathbf{x}_k - \boldsymbol{\mu}) = \text{Tr}((\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Omega})$ and $\log \det(\boldsymbol{\Omega}) = \text{Tr}(\log \boldsymbol{\Omega})$. This leads to the log-likelihood

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Omega}) = -\frac{nd}{2} \log(2\pi) + \frac{n}{2} \text{Tr}(\log \boldsymbol{\Omega}) - \frac{1}{2} \sum_{k=1}^n \text{Tr}((\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Omega}).$$

First, to find the MLE for $\boldsymbol{\mu}$ we compute (see Appendix for some rules in vector and matrix calculus)

$$\frac{\partial \log L(\boldsymbol{\mu}, \boldsymbol{\Omega})}{\partial \boldsymbol{\mu}} = \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Omega}.$$

Setting this equal to zero we get $\sum_{k=1}^n \mathbf{x}_k = n\hat{\boldsymbol{\mu}}_{ML}$ and thus

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k.$$

Next, to obtain the MLE for $\boldsymbol{\Omega}$ we compute

$$\frac{\partial \log L(\boldsymbol{\mu}, \boldsymbol{\Omega})}{\partial \boldsymbol{\Omega}} = \frac{n}{2} \boldsymbol{\Omega}^{-1} - \frac{1}{2} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T.$$

Setting this equal to zero and substituting the MLE for $\boldsymbol{\mu}$ we get

$$\hat{\boldsymbol{\Omega}}_{ML}^{-1} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T = \hat{\boldsymbol{\Sigma}}_{ML}.$$

Therefore, the MLEs are identical to the empirical estimates.

Note the factor $\frac{1}{n}$ in the MLE of the covariance matrix.

1.4.3.3 Distribution of the empirical / maximum likelihood estimates

With $x_1, \dots, x_n \sim N_d(\mu, \Sigma)$ one can find the exact distributions of the estimators.

1. Distribution of the estimate of the mean:

$$\hat{\mu}_{ML} \sim N_d\left(\mu, \frac{\Sigma}{n}\right)$$

Since $E(\hat{\mu}_{ML}) = \mu \implies \hat{\mu}_{ML}$ is unbiased.

2. Distribution of the covariance estimate:

$$\hat{\Sigma}_{ML} \sim \text{Wishart}\left(\frac{\Sigma}{n}, n-1\right)$$

Since $E(\hat{\Sigma}_{ML}) = \frac{n-1}{n}\Sigma \implies \hat{\Sigma}_{ML}$ is biased, with $\text{Bias}(\hat{\Sigma}_{ML}) = \Sigma - E(\hat{\Sigma}_{ML}) = -\frac{\Sigma}{n}$.

Easy to make unbiased: $\hat{\Sigma}_{UB} = \frac{n}{n-1}\hat{\Sigma} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \hat{\mu})(x_k - \hat{\mu})^T$ is unbiased.

But unbiasedness of an estimator is **not** a very relevant criterion in multivariate statistics as we will see in the next section.

1.4.4 Problems with maximum likelihood in small sample settings and high dimensions

Modern data is high dimensional!

Data sets with $n < d$, i.e. high dimension d and small sample size n are now common in many fields, e.g., medicine, biology but also finance and business analytics.

$$n = 100 \text{ (e.g, patients/samples)}$$

$$d = 20000 \text{ (e.g., genes/SNPs/proteins/variables)}$$

Reasons:

- the number of measured variables is increasing quickly with technological advances (e.g. genomics)
- but the number of samples cannot be similarly increased (for cost and ethical reasons)

General problems of MLEs:

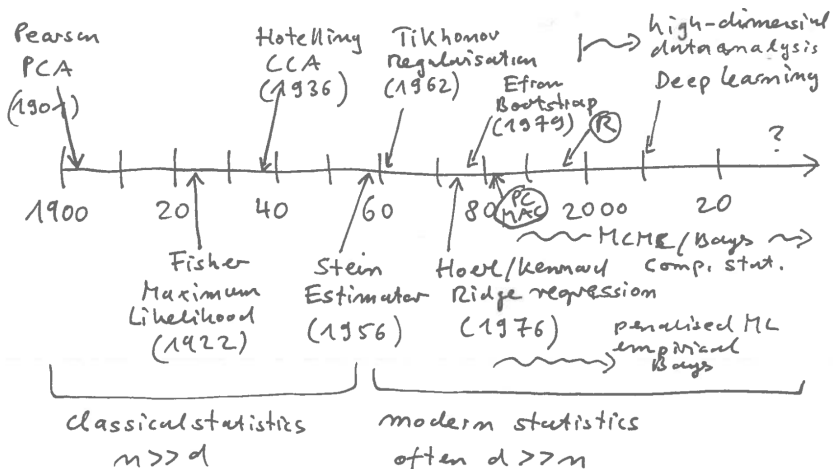
1. ML estimators are optimal only if **sample size is large** compared to the number of parameters. However, this optimality is not any more valid if sample size is moderate or smaller than the number of parameters.

2. If there is not enough data the **ML estimate overfits**. This means ML fits the current data perfectly but the resulting model does not generalise well (i.e. model will perform poorly in prediction)
3. If there is a choice between different models with different complexity **ML will always select the model with the largest number of parameters**.

-> for high-dimensional data with small sample size maximum likelihood estimation does not work!!!

History of Statistics:

Much of modern statistics (from 1960 onwards) is devoted to the development of inference and estimation techniques that work with complex, high-dimensional data.



- Maximum likelihood is a method from classical statistics (time up to about 1960).
- From 1960 modern (computational) statistics emerges, starting with “**Stein Paradox**” (1956): Charles Stein showed that in a **multivariate setting** ML estimators are **dominated by** (= are always worse than) shrinkage estimators!
- For example, there is a shrinkage estimator for the mean that is better (in terms of MSE) than the average (which is the MLE)!

Modern statistics has developed many different (but related) methods for use in high-dimensional, small sample settings:

- regularised estimators
- shrinkage estimators
- penalised maximum likelihood estimators
- Bayesian estimators
- Empirical Bayes estimators
- KL / entropy-based estimators

Most of this is out of scope for our class, but will be covered in advanced statistical courses.

Next, we describe a **simple regularised estimator for the estimation of the covariance** that we will use later (i.e. in classification).

1.4.5 Estimation of covariance matrix in small sample settings

Problems with ML estimate of Σ

1. Σ has $O(d^2)$ number of parameters! $\implies \hat{\Sigma}^{\text{MLE}}$ requires *a lot* of data!
 $n \gg d$ or d^2
2. if $n < d$ then $\hat{\Sigma}$ is positive **semi**-definite (even if Σ is p.d.f.!)
 $\implies \hat{\Sigma}$ will have **vanishing eigenvalues** (some $\lambda_i = 0$) and thus **cannot be inverted** and is singular!

Simple regularised estimate of Σ

Regularised estimator S^* = convex combination of $S = \hat{\Sigma}^{\text{MLE}}$ and I_d (identity matrix) to get

Regularisation:

$$\underbrace{S^*}_{\text{regularised estimate}} = \underbrace{\lambda}_{\text{shrinkage intensity}} \underbrace{I_d}_{\text{target}} + (1 - \lambda) \underbrace{S}_{\text{ML estimate}}$$

Next, choose $\lambda \in [0, 1]$ such that S^* is better (in terms of MSE) than both S and I_d .

Bias-variance trade-off

MSE is the Mean Squared Error, composed of squared bias and variance.

$$\text{MSE}(\theta) = E((\hat{\theta} - \theta)^2) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

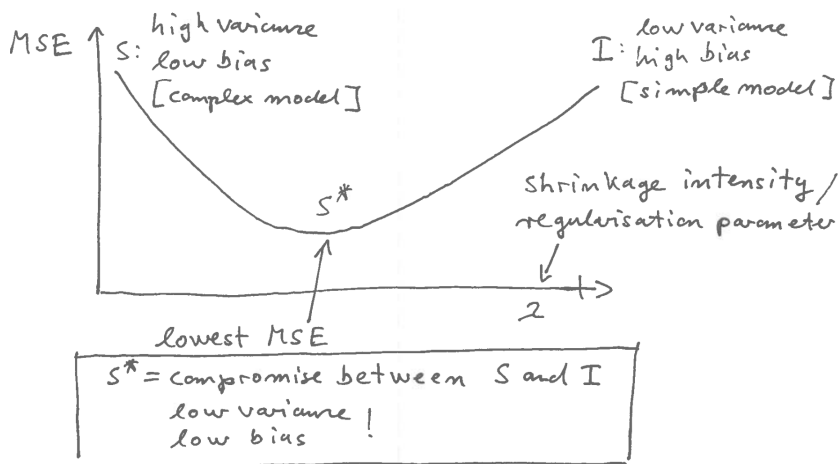
with $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

S : ML estimate, many parameters, low bias, high variance

I_d : “target”, no parameters, high bias, low variance

\implies **reduce high variance of S by introducing a bit of bias through I_d !**

\implies overall, MSE is decreased



How to find optimal shrinkage / regularisation parameter λ ? Minimise MSE!

Challenge: since we don't know the true Σ we cannot actually compute the MSE directly but have to estimate it! How is this done in practise?

- by cross-validation (=resampling procedure)
- by using some analytic approximation (e.g. Stein's formula)

Why does regularisation of $\hat{\Sigma}$ work?

1. S^* is **positive definite**:

Matrix Theory:

$$\underbrace{M_1}_{\text{symmetric positive definite, } \lambda I} + \underbrace{M_2}_{\text{symmetric positive semi-definite, } (1-\lambda)S} = \underbrace{M_3}_{\text{symmetric positive definite, } S^*}$$

$$\implies S^* \text{ can be inverted even if } n < d$$

(see Appendix A for details).

2. It's **Bayesian** in disguise!

$$\underbrace{S^*}_{\text{posterior mean}} = \underbrace{\lambda I_d}_{\text{prior information}} + (1-\lambda) \underbrace{S}_{\text{data summarised by maximum likelihood}}$$

- Prior information helps to infer Σ even in small samples
- Since λ is chosen from data, it is actually an empirical Bayes.
- also called shrinkage estimator since the off-diagonal entries are shrunk towards zero
- this type of linear shrinkage/regularisation is natural for exponential family models (Diaconis and Ylvisaker, 1979)

In Worksheet 2 the empirical estimator of covariance is compared with the covariance estimator implemented in the R package "corpcor". This uses a

regularisation similar as above (but for the correlation rather than covariance matrix) and it employs an analytic data-adaptive estimate of the shrinkage intensity λ . This estimator is a variant of an empirical Bayes / James-Stein estimator (see [MATH20802 Statistical Methods](#)). .

Summary

- In multivariate statistics, it is useful (and often necessary) to utilise prior information!
- Regularisation introduces bias and reduces variance, minimising overall MSE
- Unbiased estimation (a highly valued property in classical statistics!) is not a good idea in multivariate settings and often leads to poor estimators.

1.5 Categorical and multinomial distribution

1.5.1 Categorical distribution

The **categorical distribution** is a generalisation of the Bernoulli distribution and is correspondingly also known as **Multinoulli** distribution.

Assume we have K classes labeled “class 1”, “class 2”, ..., “class K ”. A *discrete* random variable with a state space consisting of these K classes has a categorical distribution $\text{Cat}(\boldsymbol{\pi})$. The parameter vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$ specifies the probabilities of each of the K classes with $\Pr(\text{“class } k\text{”}) = \pi_k$. The parameters satisfy $\pi_k \in [0, 1]$ and $\sum_{k=1}^K \pi_k = 1$, hence there are $K - 1$ independent parameters in a categorical distribution (and not K).

Sampling from a categorical distributions $\text{Cat}(\boldsymbol{\pi})$ yields one of K classes. There are several ways to numerically represent “class k ”, for example simply by the corresponding number k . However, instead of this “integer encoding” it is often more convenient to use the so-called “one hot encoding” where the class is represented by an indicator vector $\mathbf{x} = (x_1, \dots, x_K)^T = (0, 0, \dots, 1, \dots, 0)^T$ containing zeros everywhere except for the element $x_k = 1$ at position k . Thus all $x_k \in \{0, 1\}$ and $\sum_{k=1}^K x_k = 1$.

The expectation of $\mathbf{x} \sim \text{Cat}(\boldsymbol{\pi})$ is $E(\mathbf{x}) = \boldsymbol{\pi}$, with $E(x_k) = \pi_k$. The covariance matrix is $\text{Var}(\mathbf{x}) = \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$. In component notation $\text{Var}(x_i) = \pi_i(1 - \pi_i)$ and $\text{Cov}(x_i, x_j) = -\pi_i\pi_j$. This follows directly from the definition of the variance $\text{Var}(\mathbf{x}) = E(\mathbf{x}\mathbf{x}^T) - E(\mathbf{x})E(\mathbf{x})^T$ and noting that $x_i^2 = x_i$ and $x_i x_j = 0$ if $i \neq j$. Note that the variance matrix $\text{Var}(\mathbf{x})$ is singular by construction, as the K random variables x_1, \dots, x_K are dependent through the constraint $\sum_{k=1}^K x_k = 1$.

The corresponding probability mass function (pmf) can be written conveniently in terms of x_k as

$$f(\mathbf{x}) = \prod_{k=1}^K \pi_k^{x_k} = \begin{cases} \pi_k & \text{if } x_k = 1 \end{cases}$$

and the log pmf as

$$\log f(\mathbf{x}) = \sum_{k=1}^K x_k \log \pi_k = \begin{cases} \log \pi_k & \text{if } x_k = 1 \end{cases}$$

In order to be more explicit that the categorical distribution has $K - 1$ and not K parameters we rewrite the log-density with $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ and $x_K = 1 - \sum_{k=1}^{K-1} x_k$ as

$$\begin{aligned} \log f(\mathbf{x}) &= \sum_{k=1}^{K-1} x_k \log \pi_k + x_K \log \pi_K \\ &= \sum_{k=1}^{K-1} x_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k\right) \end{aligned}$$

Note that there is no particular reason to choose π_K as derived, in its place any other of the π_k may be selected.

For $K = 2$ the categorical distribution reduces to the Bernoulli $\text{Ber}(p)$ distribution, with $\pi_1 = p$ and $\pi_2 = 1 - p$.

1.5.2 Multinomial distribution

The multinomial distribution arises from repeated categorical sampling, just like the Binomial distribution arises from repeated Bernoulli sampling.

1.5.2.1 Univariate case

Binomial Distribution:

Repeat Bernoulli $\text{Ber}(\pi)$ experiment n times:

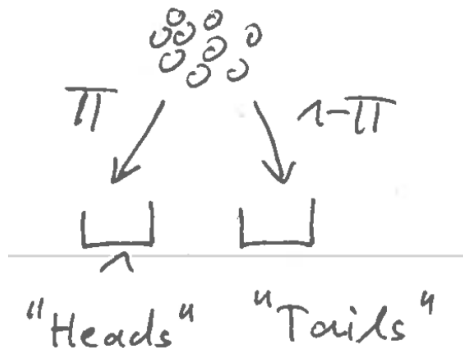
$$\begin{aligned}x &\sim \text{Bin}(n, \pi) \\x &\in \{0, \dots, n\} \\E(x) &= n \pi \\ \text{Var}(x) &= n \pi(1 - \pi)\end{aligned}$$

Standardised to unit interval:

$$\begin{aligned}\frac{x}{n} &\in \left\{0, \frac{1}{n}, \dots, 1\right\} \\E\left(\frac{x}{n}\right) &= \pi \\ \text{Var}\left(\frac{x}{n}\right) &= \frac{\pi(1 - \pi)}{n}\end{aligned}$$

Urn model:

distribute n balls into two bins



1.5.2.2 Multivariate case

Multinomial distribution:

Draw n times from categorical distribution $\text{Cat}(\pi)$:

$$\mathbf{x} \sim \text{Mult}(n, \pi)$$

$$x_i \in \{0, 1, \dots, n\}; \sum_{i=1}^K x_i = n$$

$$\mathbb{E}(\mathbf{x}) = n \pi$$

$$\text{Var}(x_i) = n \pi_i (1 - \pi_i)$$

$$\text{Cov}(x_i, x_j) = -n \pi_i \pi_j$$

Standardised to unit interval:

$$\frac{x_i}{n} \in \left\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$$

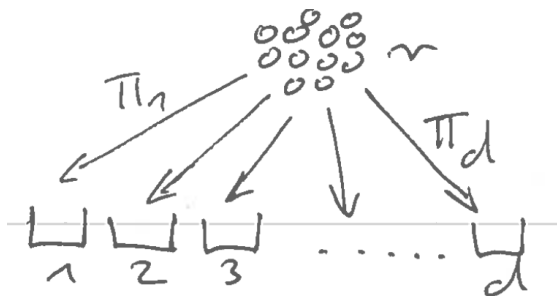
$$\mathbb{E}\left(\frac{\mathbf{x}}{n}\right) = \pi$$

$$\text{Var}\left(\frac{x_i}{n}\right) = \frac{\pi_i(1 - \pi_i)}{n}$$

$$\text{Cov}\left(\frac{x_i}{n}, \frac{x_j}{n}\right) = -\frac{\pi_i \pi_j}{n}$$

Urn model:

distribute n balls into K bins:



1.5.3 Entropy and maximum likelihood analysis for the categorical distribution

In the following examples we compute the KL divergence and MLE as well related quantities for the categorical distribution.

These generalise the same calculations for the Bernoulli distribution discussed in the earlier module [MATH20802 Statistical Methods](#).

Example 1.1. KL divergence between two categorical distributions with K classes:

With $P = \text{Cat}(p)$ and $Q = \text{Cat}(q)$ and corresponding probabilities p_1, \dots, p_K and q_1, \dots, q_K satisfying $\sum_{i=1}^K p_i = 1$ and $\sum_{i=1}^K q_i = 1$ we get:

$$D_{\text{KL}}(P, Q) = \sum_{i=1}^K p_i \log \left(\frac{p_i}{q_i} \right)$$

To be explicit that there are only $K - 1$ parameters in a categorical distribution we can also write

$$D_{\text{KL}}(P, Q) = \sum_{i=1}^{K-1} p_i \log \left(\frac{p_i}{q_i} \right) + p_K \log \left(\frac{p_K}{q_K} \right)$$

with $p_K = \left(1 - \sum_{i=1}^{K-1} p_i\right)$ and $q_K = \left(1 - \sum_{i=1}^{K-1} q_i\right)$.

Example 1.2. Expected Fisher information of the categorical distribution:

We first compute the Hessian matrix $\nabla^T \nabla \log f(\mathbf{x})$ of the log-probability mass function, where the differentiation is with regard to π_1, \dots, π_{K-1} .

The diagonal entries of the Hessian matrix (with $i = 1, \dots, K - 1$) are

$$\frac{\partial^2}{\partial \pi_i^2} \log f(\mathbf{x}) = -\frac{x_i}{\pi_i^2} - \frac{x_K}{\pi_K^2}$$

and its off-diagonal entries are (with $j = 1, \dots, K - 1$)

$$\frac{\partial^2}{\partial \pi_i \partial \pi_j} \log f(\mathbf{x}) = -\frac{x_K}{\pi_K^2}$$

Recalling that $E(x_i) = \pi_i$ we can compute the expected Fisher information

matrix for a categorical distribution as

$$\begin{aligned}
 \mathbf{I}^{\text{Fisher}}(\pi_1, \dots, \pi_{K-1}) &= -\mathbb{E} \left(\nabla^T \nabla \log f(\mathbf{x}) \right) \\
 &= \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_K} & \cdots & \frac{1}{\pi_K} \\ \vdots & \ddots & \vdots \\ \frac{1}{\pi_K} & \cdots & \frac{1}{\pi_{K-1}} + \frac{1}{\pi_K} \end{pmatrix} \\
 &= \text{Diag} \left(\frac{1}{\pi_1}, \dots, \frac{1}{\pi_{K-1}} \right) + \frac{1}{\pi_K} \mathbf{1}
 \end{aligned}$$

For $K = 2$ and $\pi_1 = p$ this reduces to the expected Fisher information of a Bernoulli variable

$$\begin{aligned}
 I^{\text{Fisher}}(p) &= \left(\frac{1}{p} + \frac{1}{1-p} \right) \\
 &= \frac{1}{p(1-p)}
 \end{aligned}$$

Example 1.3. Quadratic approximation of KL divergence of the categorical distribution:

The expected Fisher information arises from a local quadratic approximation of the KL divergence:

$$D_{\text{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) \approx \frac{1}{2} \boldsymbol{\varepsilon}^T \mathbf{I}^{\text{Fisher}}(\boldsymbol{\theta}) \boldsymbol{\varepsilon}$$

and

$$D_{\text{KL}}(F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}, F_{\boldsymbol{\theta}}) \approx \frac{1}{2} \boldsymbol{\varepsilon}^T \mathbf{I}^{\text{Fisher}}(\boldsymbol{\theta}) \boldsymbol{\varepsilon}$$

We now consider the KL divergence $D_{\text{KL}}(P, Q)$ between the categorical distribution $P = \text{Cat}(\mathbf{p})$ with probabilities $\mathbf{p} = (p_1, \dots, p_K)^T$ with the categorical distribution $Q = \text{Cat}(\mathbf{q})$ with probabilities $\mathbf{q} = (q_1, \dots, q_K)^T$.

First, we keep P fixed and assume that Q is a perturbed version of P with $\mathbf{q} = \mathbf{p} + \boldsymbol{\varepsilon}$. Note that the perturbations $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_K)^T$ satisfy $\sum_{k=1}^K \varepsilon_k = 0$

because $\sum_{k=1}^K p_i = 1$ and $\sum_{k=1}^K q_i = 1$. Thus $\varepsilon_K = -\sum_{k=1}^{K-1} \varepsilon_k$. Then

$$\begin{aligned}
 D_{\text{KL}}(P, Q = P + \varepsilon) &= D_{\text{KL}}(\text{Cat}(\mathbf{p}), \text{Cat}(\mathbf{p} + \varepsilon)) \\
 &\approx \frac{1}{2}(\varepsilon_1, \dots, \varepsilon_{K-1}) \mathbf{I}^{\text{Fisher}}(p_1, \dots, p_{K-1}) \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{K-1} \end{pmatrix} \\
 &= \frac{1}{2} \left(\sum_{k=1}^{K-1} \frac{\varepsilon_k^2}{p_k} + \frac{\left(\sum_{k=1}^{K-1} \varepsilon_k\right)^2}{p_K} \right) \\
 &= \frac{1}{2} \sum_{k=1}^K \frac{\varepsilon_k^2}{p_k} \\
 &= \frac{1}{2} \sum_{k=1}^K \frac{(p_k - q_k)^2}{p_k} \\
 &= \frac{1}{2} D_{\text{Neyman}}(P, Q)
 \end{aligned}$$

Similarly, if we keep Q fixed and consider P as a disturbed version of Q we get

$$\begin{aligned}
 D_{\text{KL}}(P = Q + \varepsilon, Q) &= D_{\text{KL}}(\text{Cat}(\mathbf{q} + \varepsilon), \text{Cat}(\mathbf{q})) \\
 &\approx \frac{1}{2} \sum_{k=1}^K \frac{(p_k - q_k)^2}{q_k} \\
 &= \frac{1}{2} D_{\text{Pearson}}(P, Q)
 \end{aligned}$$

Note that in both approximations we divide by the probabilities of the distribution that is kept fixed.

Note the appearance of the *Pearson χ^2 divergence* and the *Neyman χ^2 divergence* in the above. Both are, like the KL divergence, part of the family of *f-divergences*. The Neyman χ^2 divergence is also known as the reverse Pearson divergence as $D_{\text{Neyman}}(P, Q) = D_{\text{Pearson}}(Q, P)$.

Example 1.4. Maximum likelihood estimation of the parameters of the categorical distribution:

Maximum likelihood estimation seems trivial at first sight but it is in fact a bit more complicated since there are only $K - 1$ free parameters, and not K . So we either need to optimise with regard to a specific set of $K - 1$ parameters (which is what we do below) or use a constrained optimisation procedure to enforce that $\sum_{k=1}^K \pi_k = 1$ (for example by using a Lagrange multiplier).

- The data: We observe n samples x_1, \dots, x_n . The data matrix of dimension $n \times K$ is $\mathbf{X} = (x_1, \dots, x_n)^T = (x_{ik})$. It contains each $x_i = (x_{i1}, \dots, x_{iK})^T$. The corresponding summary (minimal sufficient)

statistics are $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = (\bar{x}_1, \dots, \bar{x}_K)^T$ with $\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}$. We can also write $\bar{x}_K = 1 - \sum_{k=1}^{K-1} \bar{x}_k$. The number of samples for class k is $n_k = n\bar{x}_k$ with $\sum_{k=1}^K n_k = n$.

- The log-likelihood is

$$\begin{aligned} l_n(\pi_1, \dots, \pi_{K-1}) &= \sum_{i=1}^n \log f(\mathbf{x}_i) \\ &= \sum_{i=1}^n \left(\sum_{k=1}^{K-1} x_{ik} \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_{ik} \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \\ &= n \left(\sum_{k=1}^{K-1} \bar{x}_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} \bar{x}_k \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \right) \\ &= n \left(\sum_{k=1}^{K-1} \bar{x}_k \log \pi_k + \bar{x}_K \log \pi_K \right) \end{aligned}$$

- Score function (gradient, row vector)

$$\begin{aligned} S_n(\pi_1, \dots, \pi_{K-1}) &= \nabla l_n(\pi_1, \dots, \pi_{K-1}) \\ &= \begin{pmatrix} \frac{\partial}{\partial \pi_1} l_n(\pi_1, \dots, \pi_{K-1}) \\ \vdots \\ \frac{\partial}{\partial \pi_{K-1}} l_n(\pi_1, \dots, \pi_{K-1}) \end{pmatrix}^T \\ &= n \begin{pmatrix} \frac{\bar{x}_1}{\pi_1} - \frac{\bar{x}_K}{\pi_K} \\ \vdots \\ \frac{\bar{x}_{K-1}}{\pi_{K-1}} - \frac{\bar{x}_K}{\pi_K} \end{pmatrix}^T \end{aligned}$$

Note in particular the need for the second term that arises because π_K depends on all the π_1, \dots, π_{K-1} .

- Maximum likelihood estimate: Setting $S_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) = 0$ yields $K-1$ equations

$$\bar{x}_i \left(1 - \sum_{k=1}^{K-1} \hat{\pi}_k^{ML} \right) = \hat{\pi}_i^{ML} \left(1 - \sum_{k=1}^{K-1} \bar{x}_k \right)$$

for $i = 1, \dots, K-1$ and with solution

$$\hat{\pi}_i^{ML} = \bar{x}_i$$

It also follows that

$$\hat{\pi}_K^{ML} = 1 - \sum_{k=1}^{K-1} \hat{\pi}_k^{ML} = 1 - \sum_{k=1}^{K-1} \bar{x}_k = \bar{x}_K$$

The maximum likelihood estimator is therefore the frequency of the occurrence of a class among the n samples.

Example 1.5. Observed Fisher information of the categorical distribution:

We first need to compute the negative Hessian matrix of the log likelihood function $-\nabla^T \nabla l_n(\pi_1, \dots, \pi_{K-1})$ and then evaluate it at the MLEs $\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}$.

The diagonal entries of the Hessian matrix (with $i = 1, \dots, K-1$) are

$$\frac{\partial^2}{\partial \pi_i^2} l_n(\pi_1, \dots, \pi_{K-1}) = -n \left(\frac{\bar{x}_i}{\pi_i^2} + \frac{\bar{x}_K}{\pi_K^2} \right)$$

and its off-diagonal entries are (with $j = 1, \dots, K-1$)

$$\frac{\partial^2}{\partial \pi_i \partial \pi_j} l_n(\pi_1, \dots, \pi_{K-1}) = -\frac{n \bar{x}_K}{\pi_K^2}$$

Thus, the observed Fisher information matrix at the MLE for a categorical distribution is

$$J_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) = n \begin{pmatrix} \frac{1}{\hat{\pi}_1^{ML}} + \frac{1}{\hat{\pi}_K^{ML}} & \cdots & \frac{1}{\hat{\pi}_K^{ML}} \\ \vdots & \ddots & \vdots \\ \frac{1}{\hat{\pi}_K^{ML}} & \cdots & \frac{1}{\hat{\pi}_{K-1}^{ML}} + \frac{1}{\hat{\pi}_K^{ML}} \end{pmatrix}$$

For $K = 2$ this reduces to the observed Fisher information of a Bernoulli variable

$$\begin{aligned} J_n(\hat{p}_{ML}) &= n \left(\frac{1}{\hat{p}_{ML}} + \frac{1}{1 - \hat{p}_{ML}} \right) \\ &= \frac{n}{\hat{p}_{ML}(1 - \hat{p}_{ML})} \end{aligned}$$

The inverse of the observed Fisher information is:

$$J_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML})^{-1} = \frac{1}{n} \begin{pmatrix} \hat{\pi}_1^{ML}(1 - \hat{\pi}_1^{ML}) & \cdots & -\hat{\pi}_1^{ML} \hat{\pi}_{K-1}^{ML} \\ \vdots & \ddots & \vdots \\ -\hat{\pi}_{K-1}^{ML} \hat{\pi}_1^{ML} & \cdots & \hat{\pi}_{K-1}^{ML}(1 - \hat{\pi}_{K-1}^{ML}) \end{pmatrix}$$

To show that this is indeed the inverse we use the Woodbury matrix identity (see Appendix), with $B=1$, $\mathbf{u} = (\pi_1, \dots, \pi_{K-1})^T$, $\mathbf{v} = -\mathbf{u}^T$, $A = \text{Diag}(\mathbf{u})$ and its inverse $A^{-1} = \text{Diag}(\pi_1^{-1}, \dots, \pi_{K-1}^{-1})$. Then $A^{-1}\mathbf{u} = \mathbf{1}_{K-1}$ and $1 - \mathbf{u}^T A^{-1} \mathbf{u} = \pi_K$. With this $J_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML})^{-1} = \frac{1}{n} (A - \mathbf{u} \mathbf{u}^T)$ and $J_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) = n \left(A^{-1} + \frac{1}{\pi_K} \mathbf{1}_{K-1 \times K-1} \right)$.

For $K = 2$ the inverse observed Fisher information of the categorical distribution reduces to that of the Bernoulli distribution

$$J_n(\hat{p}_{ML})^{-1} = \frac{\hat{p}_{ML}(1 - \hat{p}_{ML})}{n}$$

The inverse observed Fisher information is useful, e.g., as the asymptotic variance of the maximum likelihood estimates.

Example 1.6. Wald statistic for the categorical distribution:

The squared Wald statistic is

$$\begin{aligned} t(\mathbf{p}_0)^2 &= (\hat{\pi}_1^{ML} - p_1^0, \dots, \hat{\pi}_{K-1}^{ML} - p_{K-1}^0) J_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) \begin{pmatrix} \hat{\pi}_1^{ML} - p_1^0 \\ \vdots \\ \hat{\pi}_{K-1}^{ML} - p_{K-1}^0 \end{pmatrix} \\ &= n \left(\sum_{k=1}^{K-1} \frac{(\hat{\pi}_k^{ML} - p_k^0)^2}{\hat{\pi}_k^{ML}} + \frac{(\sum_{k=1}^{K-1} (\hat{\pi}_k^{ML} - p_k^0))^2}{\hat{\pi}_K^{ML}} \right) \\ &= n \left(\sum_{k=1}^K \frac{(\hat{\pi}_k^{ML} - p_k^0)^2}{\hat{\pi}_k^{ML}} \right) \\ &= n D_{\text{Neyman}}(\text{Cat}(\hat{\pi}_{ML}), \text{Cat}(\mathbf{p}_0)) \end{aligned}$$

With n_1, \dots, n_K the observed counts with $n = \sum_{k=1}^K n_k$ and $\hat{\pi}_k^{ML} = \frac{n_k}{n} = \bar{x}_k$, and $n_1^{\text{expect}}, \dots, n_K^{\text{expect}}$ the expected counts $n_k^{\text{expect}} = n p_k^0$ under \mathbf{p}_0 we can write the squared Wald statistic as follows:

$$t(\mathbf{p}_0)^2 = \sum_{k=1}^K \frac{(n_k - n_k^{\text{expect}})^2}{n_k} = \chi_{\text{Neyman}}^2$$

This is known as the Neyman chi-squared statistic (note the *observed* counts in its denominator) and it is asymptotically distributed as χ_{K-1}^2 because there are $K - 1$ free parameters in \mathbf{p}_0 .

Example 1.7. Wilks log-likelihood ratio statistic for the categorical distribution:

The Wilks log-likelihood ratio is

$$W(\mathbf{p}_0) = 2(l_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) - l_n(p_1^0, \dots, p_{K-1}^0))$$

with $\mathbf{p}_0 = c(p_1^0, \dots, p_{K-1}^0)^T$. As the probabilities sum up to 1 there are only $K - 1$ free parameters.

The log-likelihood at the MLE is

$$l_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) = n \sum_{k=1}^K \bar{x}_k \log \hat{\pi}_k^{ML} = n \sum_{k=1}^K \bar{x}_k \log \bar{x}_k$$

with $\hat{\pi}_k^{ML} = \frac{n_k}{n} = \bar{x}_k$. Note that here and in the following the sums run from 1 to K where the K -th component is always computed from the components 1 to $K - 1$, as in the previous section. The log-likelihood at \mathbf{p}_0 is

$$l_n(p_1^0, \dots, p_{K-1}^0) = n \sum_{k=1}^K \bar{x}_k \log p_k^0$$

so that the Wilks statistic becomes

$$W(\mathbf{p}_0) = 2n \sum_{k=1}^K \bar{x}_k \log \left(\frac{\bar{x}_k}{p_k^0} \right)$$

It is asymptotically chi-squared distributed with $K - 1$ degrees of freedom.

Note that for this model the Wilks statistic is equal to the KL Divergence

$$W(\mathbf{p}_0) = 2n D_{\text{KL}}(\text{Cat}(\hat{\pi}_{ML}), \text{Cat}(\mathbf{p}_0))$$

The Wilks log-likelihood ratio statistic for the categorical distribution is also known as the **G test statistic** where $\hat{\pi}_{ML}$ corresponds to the observed frequencies (as observed in data) and \mathbf{p}_0 are the expected frequencies (i.e. hypothesised to be the true frequencies).

Using observed counts n_k and expected counts $n_k^{\text{expect}} = np_k^0$ we can write the Wilks statistic respectively the G-statistic as follows:

$$W(\mathbf{p}_0) = 2 \sum_{k=1}^K n_k \log \left(\frac{n_k}{n_k^{\text{expect}}} \right)$$

Example 1.8. Quadratic approximation of the Wilks log-likelihood ratio statistic for the categorical distribution:

Developing the Wilks statistic $W(\mathbf{p}_0)$ around the MLE $\hat{\pi}_{ML}$ yields the squared Wald statistic which for the categorical distribution is the Neyman chi-squared statistic:

$$\begin{aligned} W(\mathbf{p}_0) &= 2n D_{\text{KL}}(\text{Cat}(\hat{\pi}_{ML}), \text{Cat}(\mathbf{p}_0)) \\ &\approx n D_{\text{Neyman}}(\text{Cat}(\hat{\pi}_{ML}), \text{Cat}(\mathbf{p}_0)) \\ &= \sum_{k=1}^K \frac{(n_k - n_k^{\text{expect}})^2}{n_k} \\ &= \chi_{\text{Neyman}}^2 \end{aligned}$$

If instead we approximate the KL divergence assuming \mathbf{p}_0 as fixed we arrive at

$$\begin{aligned} 2n D_{\text{KL}}(\text{Cat}(\hat{\pi}_{ML}), \text{Cat}(\mathbf{p}_0)) &\approx n D_{\text{Pearson}}(\text{Cat}(\hat{\pi}_{ML}), \text{Cat}(\mathbf{p}_0)) \\ &= \sum_{k=1}^K \frac{(n_k - n_k^{\text{expect}})^2}{n_k^{\text{expect}}} \\ &= \chi_{\text{Pearson}}^2 \end{aligned}$$

which is the well-known Pearson chi-squared statistic (note the *expected* counts in its denominator).

1.6 Further multivariate distributions

For most univariate distributions there are multivariate versions.

In the following we describe the multivariate versions of the Beta distribution, the Gamma distribution (also known as scaled χ^2 distribution) and of the inverse Gamma distribution.

1.6.1 Dirichlet distribution

1.6.1.1 Univariate case

Beta distribution

$$x \sim \text{Beta}(\alpha, \beta)$$

$$x \in [0, 1]$$

$$\alpha > 0; \beta > 0$$

$$m = \alpha + \beta$$

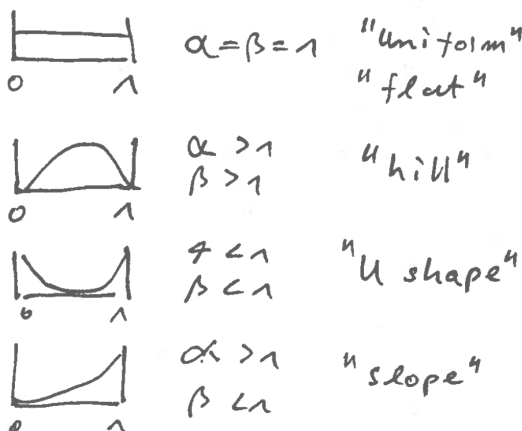
$$\mu = \frac{\alpha}{m} \in [0, 1]$$

$$E(x) = \mu$$

$$\text{Var}(x) = \frac{\mu(1-\mu)}{m+1}$$

compare with unit standardised binomial!

Different shapes



Useful as distribution for a proportion π

Bayesian Model:

Beta prior: $\pi \sim \text{Beta}(\alpha, \beta)$

Binomial likelihood: $x|\pi \sim \text{Bin}(n, \pi)$

1.6.1.2 Multivariate case

Dirichlet distribution

$$\mathbf{x} \sim \text{Dir}(\boldsymbol{\alpha})$$

$$x_i \in [0, 1]; \sum_{i=1}^d x_i = 1$$

$$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)^T > 0$$

$$m = \sum_{i=1}^d \alpha_i$$

$$\mu_i = \frac{\alpha_i}{m} \in [0, 1]$$

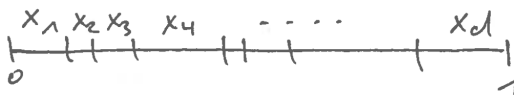
$$\mathbb{E}(x_i) = \mu_i$$

$$\text{Var}(x_i) = \frac{\mu_i(1 - \mu_i)}{m + 1}$$

$$\text{Cov}(x_i, x_j) = -\frac{\mu_i \mu_j}{m + 1}$$

compare with unit standardised multinomial!

Stick breaking" model



Useful as distribution for a proportion π

Bayesian Model:

Dirichlet prior: $\pi \sim \text{Dir}(\alpha)$

Multinomial likelihood: $x|\pi \sim \text{Mult}(n, \pi)$

1.6.2 Wishart distribution

This multivariate distribution generalises the univariate scaled χ^2 distribution (also known as Gamma distribution).

1.6.2.1 Univariate case

Scaled χ^2 distribution (=Gamma distribution, see below)

$$z_1, z_2, \dots, z_m \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$x = \sum_{i=1}^m z_i^2$$

$$x \sim \sigma^2 \chi_m^2 = W_1(\sigma^2, m)$$

$$E(x) = m \sigma^2$$

$$\text{Var}(x) = m 2\sigma^4$$

Useful as the distribution of sample variance:

$$y_1, \dots, y_n \sim N(\mu, \sigma^2)$$

Known mean μ :

$$\frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2 \sim W_1\left(\frac{\sigma^2}{n}, n\right)$$

Unknown mean μ (estimated by \bar{y}):

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \sim W_1\left(\frac{\sigma^2}{n}, n-1\right)$$

$$\hat{\sigma}_{UB}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \sim W_1\left(\frac{\sigma^2}{n-1}, n-1\right)$$

1.6.2.2 Multivariate case

Wishart distribution

$$\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m \stackrel{\text{iid}}{\sim} N_d(\mathbf{0}, \mathbf{\Sigma})$$

$$\underbrace{\mathbf{X}}_{d \times d} = \sum_{i=1}^m \underbrace{\mathbf{z}_i \mathbf{z}_i^T}_{d \times d}$$

Note that \mathbf{X} is a *matrix*!

$$\mathbf{X} \sim W_d(\mathbf{\Sigma}, m)$$

$$\mathbb{E}(\mathbf{X}) = m\mathbf{\Sigma}$$

$$\text{Var}(x_{ij}) = m \left(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj} \right)$$

Useful as distribution of sample covariance:

$$\mathbf{y}_1, \dots, \mathbf{y}_n \sim N_d(\boldsymbol{\mu}, \mathbf{\Sigma})$$

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T \sim W_d(\mathbf{\Sigma}/n, n)$$

$$\hat{\mathbf{\Sigma}}_{ML} = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T \sim W_d(\mathbf{\Sigma}/n, n-1)$$

$$\hat{\mathbf{\Sigma}}_{UB} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T \sim W_d(\mathbf{\Sigma}/(n-1), n-1)$$

1.6.2.3 Relationship to Gamma distribution

The scaled χ^2 distribution (=one-dimensional Wishart distribution) with parameters σ^2 and m is a reparameterised **Gamma distribution** with shape parameter α and scale parameter β :

$$\text{Gam} \left(\underbrace{\frac{m}{2}}_{\text{shape}}, \underbrace{2\sigma^2}_{\text{scale}} \right) = \sigma^2 \chi_m^2 = W_1(\sigma^2, m)$$

or, equivalently ($m = 2\alpha$, $\sigma^2 = \beta/2$)

$$\text{Gam} \left(\underbrace{\alpha}_{\text{shape}}, \underbrace{\beta}_{\text{scale}} \right) = \frac{\beta}{2} \chi_{2\alpha}^2 = W_1\left(\frac{\beta}{2}, 2\alpha\right)$$

The mean of the Gamma distribution is $E(x) = \alpha\beta$ and the variance is $\text{Var}(x) = \alpha\beta^2$.

The **exponential distribution** with scale parameter β is a special case of the Gamma distribution with $\alpha = 1$:

$$\text{Exp}(\beta) = \text{Gam}(1, \beta) = \frac{\beta}{2} \chi_2^2 = W_1\left(\frac{\beta}{2}, 2\right)$$

The corresponding mean is β and the variance is β^2 .

The density expressed in terms of scale parameter β is

$$f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}$$

Instead of the scale parameter β the exponential distribution is often also parameterised in terms of a rate parameter $\lambda = \frac{1}{\beta}$.

1.6.3 Inverse Wishart distribution

1.6.3.1 Univariate case

Inverse χ^2 Distribution:

$$x \sim W_1^{-1}(\psi, k+2) = \psi \text{Inv-}\chi_{k+2}^2$$

$$E(x) = \frac{\psi}{k}$$

$$\text{Var}(x) = \frac{2\psi^2}{k^2(k-2)}$$

Relationship to scaled χ^2 :

$$\frac{1}{x} \sim W_1(\psi^{-1}, k+2) = \psi^{-1} \chi_{k+2}^2$$

1.6.3.2 Multivariate case

Inverse Wishart distribution:

$$\underbrace{\mathbf{X}}_{d \times d} \sim W_d^{-1}\left(\underbrace{\mathbf{\Psi}}_{d \times d}, k+d+1\right)$$

$$E(\mathbf{X}) = \mathbf{\Psi}/k$$

$$\text{Var}(x_{ij}) = \frac{2}{k^2(k-2)} \frac{(k+2)\psi_{ij} + k\psi_{ii}\psi_{jj}}{2k+2}$$

Relationship to Wishart:

$$X^{-1} \sim W_d(\Psi^{-1}, k + d + 1)$$

1.6.3.3 Relationship to inverse Gamma distribution

Another way to express the univariate inverse Wishart distribution is via the **inverse Gamma distribution**:

$$\text{Inv-Gam}\left(\underbrace{1 + \frac{k}{2}}_{\text{shape } \alpha}, \underbrace{\frac{\psi}{2}}_{\text{scale } \beta}\right) = \psi \text{Inv-}\chi_{k+2}^2 = W_1^{-1}(\psi, k + 2)$$

or equivalently ($k = 2(\alpha - 1)$ and $\psi = 2\beta$)

$$\text{Inv-Gam}(\alpha, \beta) = 2\beta \text{Inv-}\chi_{2\alpha}^2 = W_1^{-1}(2\beta, 2\alpha)$$

The mean of the inverse Gamma distribution is $E(x) = \frac{\beta}{\alpha - 1} = \mu$ the variance $\text{Var}(x) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)} = \frac{2\mu^2}{k - 2}$.

The inverse of x is Gamma distributed:

$$\frac{1}{x} \sim \text{Gam}\left(1 + \frac{k}{2}, 2\psi^{-1}\right) = \text{Gam}(\alpha, \beta^{-1})$$

The inverse Wishart distribution is useful as conjugate distribution for Bayesian modelling of the variance, with k the sample size parameter and $\Psi = k\Sigma$ (or $\psi = k\sigma^2$).

1.6.4 Further distributions

https://en.wikipedia.org/wiki/List_of_probability_distributions

Wikipedia is a quite good source for information on distributions!

Chapter 2

Transformations and dimension reduction

Motivation: In the following we study transformations of random vectors and their distributions. These transformation are very important since they either transform simple distributions into more complex distributions or allow to simplify complex models. In machine learning invertible mappings of transformations for probability distributions are known as “normalising flows”.

2.1 Linear Transformations

2.1.1 Location-scale transformation

Also known as affine transformation.

$$y = \underbrace{a}_{\text{location parameter}} + \underbrace{B}_{\text{scale parameter}} x$$

$y : m \times 1$ random vector
 $a : m \times 1$ vector, location parameter
 $B : m \times d$ matrix, scale parameter, $m \geq 1$
 $x : d \times 1$ random vector

$$\begin{matrix} E(x) = \mu \\ \text{Var}(x) = \Sigma \end{matrix} \implies \begin{matrix} E(y) = a + B\mu \\ \text{Var}(y) = B\Sigma B^T \end{matrix}$$

Special cases/examples:

Example 2.1. Univariate case ($d = 1, m = 1$):

- $E(y) = a + b\mu$
- $\text{Var}(y) = b^2\sigma^2$

Example 2.2. Sum of two random univariate variables: $y = x_1 + x_2$, i.e. $a = 0$ and $B = (1, 1)$

- $E(x_1 + x_2) = \mu_1 + \mu_2$
- $\text{Var}(x_1 + x_2) = (1, 1) \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} = \text{Var}(x_1) + \text{Var}(x_2) + 2\text{Cov}(x_1, x_2)$

Example 2.3. $y_1 = a_1 + b_1x_1$ and $y_2 = a_2 + b_2x_2$, i.e. $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ and

$$B = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$$

- $E(y) = \begin{pmatrix} a_1 + b_1\mu_1 \\ a_2 + b_2\mu_2 \end{pmatrix}$
- $\text{Var}(y) = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix} = \begin{pmatrix} b_1^2\sigma_1^2 & b_1b_2\sigma_{12} \\ b_1b_2\sigma_{21} & b_2^2\sigma_2^2 \end{pmatrix}$
note that $\text{Cov}(y_1, y_2) = b_1b_2\text{Cov}(x_1, x_2)$

2.1.2 Invertible location-scale transformation

If $m = d$ and $\det(B) \neq 0$ then we get an **invertible** transformation:

$$y = a + Bx$$

$$x = B^{-1}(y - a)$$

Transformation of density: $x \sim F_x$ with density $f_x(x)$

$\implies y \sim F_y$ with density

$$f_y(y) = |\det(B)|^{-1} f_x(B^{-1}(y - a))$$

Example 2.4. Mahalanobis transform $y = \Sigma^{-1/2}(x - \mu)$

We assume a positive definite and thus invertible Σ , so that the inverse principal matrix square root $\Sigma^{-1/2}$ can be computed, and the transformation itself is invertible.

$$a = -\Sigma^{-1/2}\mu$$

$$B = \Sigma^{-1/2}$$

$$\begin{aligned} E(\mathbf{x}) &= \boldsymbol{\mu} \text{ and } \text{Var}(\mathbf{x}) = \boldsymbol{\Sigma} \\ \implies E(\mathbf{y}) &= \mathbf{0} \text{ and } \text{Var}(\mathbf{y}) = \mathbf{I}_d \end{aligned}$$

Mahalanobis transformation performs three functions:

1. Centering $(-\boldsymbol{\mu})$
2. Standardisation $\text{Var}(y_i) = 1$
3. Decorrelation $\text{Cor}(y_i, y_j) = 0$ for $i \neq j$

Univariate case ($d = 1$)

$$y = \frac{x - \mu}{\sigma}$$

= centering + standardisation

The **Mahalanobis transformation** appears implicitly in many places in multivariate statistics, e.g. in the multivariate normal density.

It is a particular example of a whitening transformation (of which there are infinitely many, see later chapters).

Example 2.5. Inverse Mahalanobis transformation $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{x}$

This transformation is the **inverse** of the Mahalanobis transformation. As the Mahalanobis transform is a particular whitening transform the inverse transform is sometimes called the Mahalanobis colouring transformation.

$$\begin{aligned} \mathbf{a} &= \boldsymbol{\mu} \\ \mathbf{B} &= \boldsymbol{\Sigma}^{1/2} \end{aligned}$$

$$\begin{aligned} E(\mathbf{x}) &= \mathbf{0} \text{ and } \text{Var}(\mathbf{x}) = \mathbf{I}_d \\ \implies E(\mathbf{y}) &= \boldsymbol{\mu} \text{ and } \text{Var}(\mathbf{y}) = \boldsymbol{\Sigma} \end{aligned}$$

Assume \mathbf{x} is multivariate standard normal $\mathbf{x} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ with density

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right)$$

Then the density after applying this inverse Mahalanobis transform is

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y}) &= |\det(\boldsymbol{\Sigma}^{1/2})|^{-1} (2\pi)^{-d/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right) \end{aligned}$$

$\Rightarrow \mathbf{y}$ has multivariate normal density!!

Application: e.g. random number generation: draw from $N_d(\mathbf{0}, \mathbf{I}_d)$ (easy!) then convert to multivariate normal by transformation (see Worksheet 3).

2.2 Nonlinear transformations

2.2.1 General transformation

$$\mathbf{y} = \mathbf{h}(\mathbf{x})$$

with \mathbf{h} an arbitrary vector-valued function

- linear case: $\mathbf{h}(\mathbf{x}) = \mathbf{a} + \mathbf{B}\mathbf{x}$

2.2.2 Delta method

Assume that we know the mean $E(\mathbf{x}) = \boldsymbol{\mu}$ and variance $\text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$ of \mathbf{x} . Is it possible to say something about the mean and variance of the transformed random variable \mathbf{y} ?

$$E(\mathbf{y}) = E(\mathbf{h}(\mathbf{x})) = ?$$

$$\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{h}(\mathbf{x})) = ?$$

In general, for a transformation $\mathbf{h}(\mathbf{x})$ the exact mean and variance of the transformed variable cannot be obtained analytically.

However, we can find a **linear approximation** and then compute its mean and variance. This approximation is called the “Delta Method”, or the “law of propagation of errors”, and is credited to Gauss.¹

Linearisation of $\mathbf{h}(\mathbf{x})$ is achieved by a Taylor series approximation of first order of $\mathbf{h}(\mathbf{x})$ around \mathbf{x}_0 :

$$\mathbf{h}(\mathbf{x}) \approx \mathbf{h}(\mathbf{x}_0) + \underbrace{D\mathbf{h}(\mathbf{x}_0)}_{\text{Jacobian matrix}} (\mathbf{x} - \mathbf{x}_0) = \underbrace{\mathbf{h}(\mathbf{x}_0) - D\mathbf{h}(\mathbf{x}_0)\mathbf{x}_0}_{\mathbf{a}} + \underbrace{D\mathbf{h}(\mathbf{x}_0)}_{\mathbf{B}} \mathbf{x}$$

∇ , the nabla operator, is the row vector $(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_d})$, which when applied to univariate h gives the gradient:

$$\nabla h(\mathbf{x}) = \left(\frac{\partial h}{\partial x_1}, \dots, \frac{\partial h}{\partial x_d} \right)$$

The **Jacobian matrix** is the **generalisation of the gradient** if \mathbf{h} is vector-valued:

¹Gorroochurn, P. 2020. Who Invented the Delta Method, Really? The Mathematical Intelligencer 42:46–49. <https://doi.org/10.1007/s00283-020-09982-0>

$$Dh(\mathbf{x}) = \begin{pmatrix} \nabla h_1(\mathbf{x}) \\ \nabla h_2(\mathbf{x}) \\ \vdots \\ \nabla h_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_m}{\partial x_1} & \cdots & \frac{\partial h_m}{\partial x_d} \end{pmatrix}$$

First order approximation of $h(\mathbf{x})$ around $\mathbf{x}_0 = \boldsymbol{\mu}$ yields $\mathbf{a} = h(\boldsymbol{\mu}) - Dh(\boldsymbol{\mu})\boldsymbol{\mu}$ and $\mathbf{B} = Dh(\boldsymbol{\mu})$ and leads directly to the **multivariate Delta method**:

$$\begin{aligned} E(\mathbf{y}) &\approx h(\boldsymbol{\mu}) \\ \text{Var}(\mathbf{y}) &\approx Dh(\boldsymbol{\mu}) \boldsymbol{\Sigma} (Dh(\boldsymbol{\mu}))^T \end{aligned}$$

The **univariate Delta method** is a special case:

$$\begin{aligned} E(y) &\approx h(\mu) \\ \text{Var}(y) &\approx \sigma^2 h'(\mu)^2 \end{aligned}$$

Note that the Delta approximation breaks down if $\text{Var}(\mathbf{y})$ is singular, for example if the first derivative (or gradient or Jacobian matrix) at $\boldsymbol{\mu}$ is zero.

Example 2.6. Variance of the odds ratio

The proportion $\hat{p} = \frac{n_1}{n}$ resulting from n repeats of a Bernoulli experiment has expectation $E(\hat{p}) = p$ and variance $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$. What are the (approximate) mean and the variance of the corresponding odds ratio $\widehat{OR} = \frac{\hat{p}}{1-\hat{p}}$?

With $h(x) = \frac{x}{1-x}$, $\widehat{OR} = h(\hat{p})$ and $h'(x) = \frac{1}{(1-x)^2}$ we get using the Delta method $E(\widehat{OR}) \approx h(p) = \frac{p}{1-p}$ and $\text{Var}(\widehat{OR}) \approx h'(p)^2 \text{Var}(\hat{p}) = \frac{p}{n(1-p)^3}$.

Example 2.7. Log-transform as variance stabilisation

Assume x has some mean $E(x) = \mu$ and variance $\text{Var}(x) = \sigma^2 \mu^2$, i.e. the standard deviation $\text{SD}(x)$ is proportional to the mean μ . What are the (approximate) mean and the variance of the log-transformed variable $\log(x)$?

With $h(x) = \log(x)$ and $h'(x) = \frac{1}{x}$ we get using the Delta method $E(\log(x)) \approx h(\mu) = \log(\mu)$ and $\text{Var}(\log(x)) \approx h'(\mu)^2 \text{Var}(x) = \left(\frac{1}{\mu}\right)^2 \sigma^2 \mu^2 = \sigma^2$. Thus, after applying the log-transform the variance does not depend any more on the mean!

2.2.3 Transformation of a probability density function under a general invertible transformation

Assume $h(\mathbf{x}) = \mathbf{y}(\mathbf{x})$ is invertible: $h^{-1}(\mathbf{y}) = \mathbf{x}(\mathbf{y})$

$\mathbf{x} \sim F_{\mathbf{x}}$ with probability density function $f_{\mathbf{x}}(\mathbf{x})$

The density $f_{\mathbf{y}}(\mathbf{y})$ of the transformed random vector \mathbf{y} is then given by

$$f_{\mathbf{y}}(\mathbf{y}) = |\det(D\mathbf{x}(\mathbf{y}))| f_{\mathbf{x}}(\mathbf{x}(\mathbf{y}))$$

where $D\mathbf{x}(\mathbf{y})$ is the Jacobian matrix of the inverse transformation.

Special cases:

- Univariate version: $f_y(y) = \left| \frac{dx(y)}{dy} \right| f_x(x(y))$
- Linear transformation $\mathbf{h}(\mathbf{x}) = \mathbf{a} + \mathbf{B}\mathbf{x}$, with $\mathbf{x}(\mathbf{y}) = \mathbf{B}^{-1}(\mathbf{y} - \mathbf{a})$ and $D\mathbf{x}(\mathbf{y}) = \mathbf{B}^{-1}$:

$$f_{\mathbf{y}}(\mathbf{y}) = |\det(\mathbf{B})|^{-1} f_{\mathbf{x}}(\mathbf{B}^{-1}(\mathbf{y} - \mathbf{a}))$$

2.2.4 Normalising flows

In this module we will focus mostly on linear transformations as these underpin much of classical multivariate statistics, but it is important to keep in mind for later study the importance of nonlinear transformations

In machine learning (sequences of) invertible nonlinear transformations are known as “normalising flows”. They are used both in a generative way (building complex models from simple models) and for simplification and dimension reduction.

If you are interested in normalising flows then a good start to learn more are the review papers by Kobyzev et al (2021)² and Papamakarios et al. (2021).³

2.3 Whitening transformations

2.3.1 Overview

The *Mahalanobis* transform (also know as “zero-phase component analysis” or short ZCA transform in machine learning) is a specific example of a **whitening transformation**. These constitute an important and widely used class of invertible location-scale transformations.

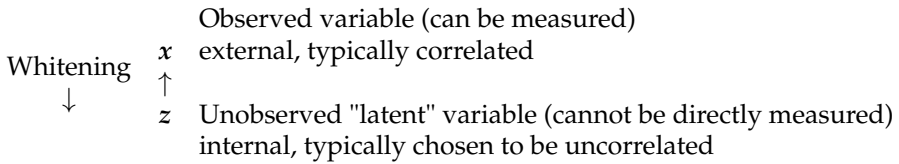
Terminology: whitening refers to the fact that after the transformation the covariance matrix is spherical, isotrop, white (\mathbf{I}_d)

Whitening is **useful in preprocessing**, to **turn multivariate problems into simple univariate models** and some **reduce the dimension in an optimal way**.

²Kobyzev et al. 2021. *Normalizing Flows: Introduction and Ideas*. [IEEE Trans. Pattern Anal. Mach. Intell.](#) **43**:3964-3979

³Papamakarios et al. 2021. *Normalizing Flows for Probabilistic Modeling and Inference*. [JMLR](#) **22**:1-64

In so-called latent variable models whitening procedures link observed and latent variables (which usually are uncorrelated and standardised random variables):



2.3.2 General whitening transformation

Starting point:

Random vector $x \sim F_x$ (not necessarily from multivariate normal).

The random variance x has some mean $E(z) = \mu$ and a positive definite (invertible) covariance matrix $\text{Var}(x) = \Sigma$. The covariance can be split into positive variances V and a positive definite invertible correlation matrix P so that $\Sigma = V^{1/2} P V^{1/2}$.

Whitening transformation:

$$\underbrace{z}_{d \times 1 \text{ vector}} = \underbrace{W}_{d \times d \text{ whitening matrix}} \underbrace{x}_{d \times 1 \text{ vector}}$$

Objective: choose W so that $\text{Var}(z) = I_d$

For Mahalanobis/ZCA whitening we already know that $W^{ZCA} = \Sigma^{-1/2}$.

In general, the whitening matrix W needs to satisfy a constraint:

$$\begin{aligned} \text{Var}(z) &= I_d \\ \implies \text{Var}(Wx) &= W \Sigma W^T = I_d \\ \implies W \Sigma W^T W &= W \end{aligned}$$

$$\implies \text{constraint on whitening matrix: } W^T W = \Sigma^{-1}$$

Clearly, the ZCA whitening matrix satisfies this constraint: $(W^{ZCA})^T W^{ZCA} = \Sigma^{-1/2} \Sigma^{-1/2} = \Sigma^{-1}$

2.3.3 Solution of whitening constraint (covariance-based)

A general way to specify a valid whitening matrix is

$$W = Q_1 \Sigma^{-1/2}$$

where Q_1 is an orthogonal matrix.⁴

Recall that an orthogonal matrix Q has the property that $Q^{-1} = Q^T$ and as a consequence $Q^T Q = Q Q^T = I$.

As a result, the above W satisfies the whitening constraint:

$$W^T W = \Sigma^{-1/2} \underbrace{Q_1^T Q_1}_I \Sigma^{-1/2} = \Sigma^{-1}$$

Note the converse is also true: any whitening whitening matrix, i.e. any W satisfying the whitening constraint, can be written in the above form as $Q_1 = W \Sigma^{1/2}$ is orthogonal by construction.

⇒ instead of choosing W , we choose the orthogonal matrix Q_1 !

- recall that orthogonal matrices geometrically represent rotations (plus reflections).
- it is now clear that there are infinitely many whitening procedures, because there are infinitely many rotations! This also means we need to find ways to choose/select among whitening procedures.
- for the Mahalanobis/ZCA transformation $Q_1^{ZCA} = I$
- **whitening** can be interpreted as **Mahalanobis transform** followed by **rotation**

2.3.4 Another solution (correlation-based)

Instead of working with the covariance matrix Σ , we can express W also in terms of the corresponding correlation matrix $P = (\rho_{ij}) = V^{-1/2} \Sigma V^{-1/2}$ where $V^{1/2}$ is the diagonal matrix containing the variances.

Specifically we can specify the whitening matrix as

$$W = Q_2 P^{-1/2} V^{-1/2}$$

It is easy to verify that this W also satisfies the whitening constraint:

$$\begin{aligned} W^T W &= V^{-1/2} P^{-1/2} \underbrace{Q_2^T Q_2}_I P^{-1/2} V^{-1/2} \\ &= V^{-1/2} P^{-1} V^{-1/2} = \Sigma^{-1} \end{aligned}$$

Conversely, any whitening matrix W can also be written in this form as $Q_2 = W V^{1/2} P^{1/2}$ is orthogonal by construction.

⁴In this Chapter we will make frequent use of orthogonal matrices. Please check Appendix A on matrices to review their most important properties!

- **Another interpretation of whitening:** first **standardising** ($V^{-1/2}$), then **decorrelation** ($P^{-1/2}$), followed by **rotation** (Q_2)
- for Mahalanobis/ZCA transformation $Q_2^{ZCA} = \Sigma^{-1/2} V^{1/2} P^{1/2}$

Both forms to write W using Q_1 and Q_2 are equally valid (and interchangeable).

Note that for the same W

$Q_1 \neq Q_2$ Two different orthogonal matrices!

and also

$$\underbrace{\Sigma^{-1/2}}_{\text{Symmetric}} \neq \underbrace{P^{-1/2} V^{-1/2}}_{\text{Not Symmetric}}$$

even though

$$\Sigma^{-1/2} \Sigma^{-1/2} = \Sigma^{-1} = V^{-1/2} P^{-1/2} P^{-1/2} V^{-1/2}$$

2.3.5 Cross-covariance and cross-correlation

A useful criterion to characterise and to distinguish among whitening transformations is the cross-covariance and cross-correlation matrix between the whitened variable z and the original variable x :

a) **Cross-covariance** $\Phi = \Sigma_{zx}$ between z and x :

$$\begin{aligned} \Phi &= \text{Cov}(z, x) = \text{Cov}(Wx, x) \\ &= W\Sigma \\ &= Q_1 \Sigma^{-1/2} \Sigma \\ &= Q_1 \Sigma^{1/2} \end{aligned}$$

Cross-covariance is linked with Q_1 ! Thus, choosing cross-covariance determines Q_1 (and vice versa).

Note that the cross-covariance matrix Φ satisfies the condition $\Phi^T \Phi = \Sigma$, which is reminiscent of the condition for W , only that now the covariance is used rather than the inverse covariance.

The whitening matrix expressed in terms of cross-covariance is $W = \Phi \Sigma^{-1}$, so as required $W^T W = \Sigma^{-1} \Phi^T \Phi \Sigma^{-1} = \Sigma^{-1}$. Furthermore, the transpose of Φ is the *inverse* of the whitening matrix, as $W^{-1} = (Q_1 \Sigma^{-1/2})^{-1} = \Sigma^{1/2} Q_1^{-1} = \Sigma^{1/2} Q_1^T = (Q_1 \Sigma^{1/2})^T = \Phi^T$.

b) **Cross-correlation** $\Psi = P_{zx}$ between z and x :

$$\begin{aligned}\Psi &= \text{Cor}(z, x) = \Phi V^{-1/2} \\ &= W \Sigma V^{-1/2} \\ &= Q_2 P^{-1/2} V^{-1/2} \Sigma V^{-1/2} \\ &= Q_2 P^{1/2}\end{aligned}$$

Cross-correlation is linked with Q_2 ! Hence, choosing cross-correlation determines Q_2 (and vice versa). The whitening matrix expressed in terms of cross-correlation is $W = \Psi P^{-1} V^{-1/2}$.

Note that the factorisation of the cross-covariance $\Phi = Q_1 \Sigma^{1/2}$ and the cross-correlation $\Psi = Q_2 P^{1/2}$ into the product of an orthogonal matrix and a positive semi-definite symmetric matrix is a so-called *polar decomposition*.

2.3.6 Inverse whitening transformation and loadings

Reverse transformation:

Recall that $z = Wx$. Therefore, the reverse transformation going from the whitened to the original variable is $x = W^{-1}z$. This can be expressed also in terms of cross-covariance and cross-correlation. With $W^{-1} = \Phi^T$ we get

$$x = \Phi^T z.$$

Furthermore, since $\Psi = \Phi V^{-1/2}$ we have $W^{-1} = V^{1/2} \Psi^T$ and

$$V^{-1/2} x = \Psi^T z.$$

The reverse whitening transformation is also known as colouring transformation (the previously discussed inverse Mahalanobis transform is one example).

Definition of loadings:

Loadings are the coefficients of the linear transformation from the latent variable back to the observed variable. If the variables are standardised to unit variance then the loadings are also called *correlation loadings*.

Hence, the cross-covariance matrix plays the role of *loadings* linking the latent variable z with the original x . Similarly, the cross-correlation matrix are the *correlation loadings* linking the (already standardised) latent variable z with the standardised x .

2.3.7 Multiple correlation coefficients

Definition of squared multiple correlation

Squared multiple correlation is a measure summarising the linear association between a response variable y (a scalar) and predictors $\mathbf{x} = (x_1, \dots, x_d)^T$. It is given by

$$\text{Cor}(y, \mathbf{x})^2 = \mathbf{P}_{yx} \mathbf{P}_{xx}^{-1} \mathbf{P}_{xy}$$

The empirical estimate of is R^2 .

See the corresponding section in [MATH20802 Statistical Methods](#).

Multiple correlation coefficients from \mathbf{z} to \mathbf{x} :

Consider the backtransformation from \mathbf{z} to \mathbf{x} . The components of \mathbf{z} are all uncorrelated. Therefore, we can compute the squared multiple correlation coefficient between each x_j and \mathbf{z} as the sum of the squared correlations $\text{Cor}(z_i, x_j)^2$:

$$\text{Cor}(\mathbf{z}, x_j)^2 = \sum_{i=1}^d \text{Cor}(z_i, x_j)^2 = \sum_{i=1}^d \psi_{ij}^2$$

In vector notation with $\mathbf{\Psi} = (\psi_{ij})$ we get

$$\begin{aligned} \left(\text{Cor}(\mathbf{z}, x_1)^2, \dots, \text{Cor}(\mathbf{z}, x_d)^2 \right)^T &= \text{Diag} \left(\mathbf{\Psi}^T \mathbf{\Psi} \right) \\ &= \text{Diag} \left(\mathbf{P}^{1/2} \mathbf{Q}_2^T \mathbf{Q}_2 \mathbf{P}^{1/2} \right) \\ &= \text{Diag}(\mathbf{P}) \\ &= (1, \dots, 1)^T \end{aligned}$$

Therefore the column sums of the matrix (ψ_{ij}^2) are all 1 regardless of the choice of \mathbf{Q}_2 :

$$\sum_{i=1}^d \psi_{ij}^2 = 1 \text{ for all } j$$

It is easy to understand why we get multiple squared correlations of value 1 — because x_j is a linear function of the z_1, \dots, z_d with no noise term, which means x_j can be predicted perfectly from \mathbf{z} with no error.

Multiple correlation coefficients from \mathbf{x} to \mathbf{z} :

For the original direction going from x_1, \dots, x_d to the z_i the corresponding squared multiple correlations $\text{Cor}(z_i, \mathbf{x})^2$ are also 1, but because the x_j are correlated we cannot simply sum the squared correlations to get $\text{Cor}(z_i, \mathbf{x})^2$ but we also need take account of the correlations among the x_j (i.e. \mathbf{P}). In vector notation:

$$\begin{aligned} \left(\text{Cor}(z_1, \mathbf{x})^2, \dots, \text{Cor}(z_d, \mathbf{x})^2 \right)^T &= \text{Diag} \left(\mathbf{\Psi} \mathbf{P}^{-1} \mathbf{\Psi}^T \right) \\ &= \text{Diag} \left(\mathbf{Q}_2 \mathbf{P}^{1/2} \mathbf{P}^{-1} \mathbf{P}^{1/2} \mathbf{Q}_2^T \right) \\ &= \text{Diag}(\mathbf{I}) \\ &= (1, \dots, 1)^T \end{aligned}$$

2.4 Natural whitening procedures

Now we discuss several strategies (maximise correlation between individual components, maximise compression, etc.) to arrive at optimal whitening transformation.

This leads to the following “natural” whitening transformations:

- **Mahalanobis** whitening, also known as **ZCA** (zero-phase component analysis) whitening in machine learning (based on covariance)
- **ZCA-cor** whitening (based on correlation)
- **PCA** whitening (based on covariance)
- **PCA-cor** whitening (based on correlation)
- **Cholesky** whitening (based on precision or covariance matrix)

Note that there are three main types (ZCA, PCA, Cholesky) with two variations each.

In the following $\mathbf{x}_c = \mathbf{x} - \boldsymbol{\mu}_x$ and $\mathbf{z}_c = \mathbf{z} - \boldsymbol{\mu}_z$ denote the mean-centered variables.

2.4.1 ZCA whitening

Aim: remove correlations but otherwise make sure that after whitening \mathbf{z} does not differ too much from \mathbf{x} . Specifically, each element z_i should be as close as possible to the corresponding element x_i :

$$\begin{aligned} z_1 &\leftrightarrow x_1 \\ z_2 &\leftrightarrow x_2 \\ z_3 &\leftrightarrow x_3 \\ &\vdots \end{aligned}$$

One possible way to implement this is to compute the expected squared difference between the two centered random vectors \mathbf{z}_c and \mathbf{x}_c .

ZCA objective function: minimise $E((\mathbf{z}_c - \mathbf{x}_c)^T(\mathbf{z}_c - \mathbf{x}_c))$ to find an optimal whitening procedure.

The ZCA objective function can be simplified as follows:

$$\begin{aligned} &= E(\mathbf{z}_c^T \mathbf{z}_c) - 2E(\mathbf{x}_c^T \mathbf{z}_c) + E(\mathbf{x}_c^T \mathbf{x}_c) \\ &= E(\text{Tr}(\mathbf{z}_c \mathbf{z}_c^T)) - 2E(\text{Tr}(\mathbf{z}_c \mathbf{x}_c^T)) + E(\text{Tr}(\mathbf{x}_c \mathbf{x}_c^T)) \\ &= \text{Tr}(E(\mathbf{z}_c \mathbf{z}_c^T)) - 2\text{Tr}(E(\mathbf{z}_c \mathbf{x}_c^T)) + \text{Tr}(E(\mathbf{x}_c \mathbf{x}_c^T)) \\ &= \text{Tr}(\text{Var}(\mathbf{z})) - 2\text{Tr}(\text{Cov}(\mathbf{z}, \mathbf{x})) + \text{Tr}(\text{Var}(\mathbf{x})) \\ &= d - 2\text{Tr}(\boldsymbol{\Phi}) + \text{Tr}(\mathbf{V}) \end{aligned}$$

The only term that depends on the whitening transformation is $-2\text{Tr}(\boldsymbol{\Phi})$ as it is a function of \mathbf{Q}_1 . Therefore we can use the following alternative objective:

ZCA equivalent objective: maximise $\text{Tr}(\Phi) = \text{Tr}(Q_1 \Sigma^{1/2})$ to find the optimal Q_1

This is the sum $\sum_{i=1}^d \text{Cov}(z_i, x_i)$ of all covariances between corresponding elements in z and x .

Solution:

- i) Apply eigendecomposition to $\Sigma = U \Lambda U^T$. Note that Λ is diagonal with positive entries $\lambda_i > 0$ as Σ is positive definite.
- ii) The objective function becomes

$$\begin{aligned} \text{Tr}(Q_1 \Sigma^{1/2}) &= \text{Tr}(Q_1 U \Lambda^{1/2} U^T) \\ &= \text{Tr}(\Lambda^{1/2} U^T Q_1 U) \\ &= \text{Tr}(\Lambda^{1/2} B) \\ &= \sum_{i=1}^d \lambda_i b_{ii}. \end{aligned}$$

Note that the product of two orthogonal matrices is itself an orthogonal matrix. Therefore, $B = U^T Q_1 U$ is an orthogonal matrix and $Q_1 = U B U^T$.

- iii) As $\lambda_i > 0$ and all $b_{ii} \in [-1, 1]$ the objective function is maximised for $b_{ii} = 1$, i.e. for $B = I$.
- iv) Thus, for ZCA the optimal Q_1 matrix is

$$Q_1^{\text{ZCA}} = I$$

The corresponding whitening matrix for ZCA is

$$W^{\text{ZCA}} = \Sigma^{-1/2}$$

and the cross-covariance matrix is

$$\Phi^{\text{ZCA}} = \Sigma^{1/2}$$

and the cross-correlation matrix

$$\Psi^{\text{ZCA}} = \Sigma^{1/2} V^{-1/2}$$

Note that $\Sigma^{1/2}$ is a symmetric positive definite matrix, hence its diagonal elements are all positive. As a result, the diagonals of Φ^{ZCA} and Ψ^{ZCA} are positive, i.e. $\text{Cov}(z_i, x_i) > 0$ and $\text{Cor}(z_i, x_i) > 0$. Hence, for ZCA two corresponding components z_i and x_i are always positively correlated!

Summary:

- ZCA/Mahalanobis transform is the unique transformation that minimises the expected total squared component-wise difference between x_c and z_c .
- As corresponding components in the whitened and original variables are always positively correlated this facilitates interpretation of the whitening variables.
- Use ZCA aka Mahalanobis whitening if you want to “just” remove correlations.

2.4.2 ZCA-Cor whitening

Aim: same as above but remove scale in x first before comparing to z

ZCA-cor objective function: minimise $E \left((z_c - V^{-1/2}x_c)^T (z_c - V^{-1/2}x_c) \right)$ to find an optimal whitening procedure.

This can be simplified as follows:

$$\begin{aligned}
 &= E(z_c^T z_c) - 2E(x_c^T V^{-1/2} z_c) + E(x_c^T V^{-1} x_c) \\
 &= \text{Tr}(\text{Var}(z)) - 2\text{Tr}(\text{Cor}(z, x)) + \text{Tr}(\text{Cor}(x, x)) \\
 &= d - 2\text{Tr}(\Psi) + d \\
 &= 2d - 2\text{Tr}(\Psi)
 \end{aligned}$$

The only term that depends on the whitening transformation via Q_2 is $-2\text{Tr}(\Psi)$ so we can use the following alternative objective instead:

ZCA-cor equivalent objective: maximise $\text{Tr}(\Psi) = \text{Tr}(Q_2 P^{1/2})$ to find optimal Q_2

This is the sum $\sum_{i=1}^d \text{Cor}(z_i, x_i)$ of all correlations between corresponding elements in z and x .

Solution: same as above for ZCA but using correlation instead of covariance:

- i) Apply eigendecomposition to $P = G\Theta G^T$ with positive diagonal Θ .
- ii) The objective function becomes $\text{Tr}(Q_2 P^{1/2}) = \text{Tr}(\Theta^{1/2} G^T Q_2 G)$
- iii) This is maximised for

$$Q_2^{\text{ZCA-Cor}} = I$$

The corresponding whitening matrix for ZCA-cor is

$$W^{\text{ZCA-Cor}} = P^{-1/2} V^{-1/2}$$

and the cross-covariance matrix is

$$\Phi^{\text{ZCA-Cor}} = P^{1/2} V^{1/2}$$

and the cross-correlation matrix is

$$\Psi^{\text{ZCA-Cor}} = P^{1/2}$$

For the ZCA-cor transformation we also have $\text{Cov}(z_i, x_i) > 0$ and $\text{Cor}(z_i, x_i) > 0$ so that two corresponding components z_i and x_i are always positively correlated!

Summary:

- ZCA-cor whitening is the unique whitening transformation maximising the total correlation between corresponding elements in x and z .
- ZCA-cor leads to interpretable z because each individual element in z is (typically strongly) positively correlated with the corresponding element in the original x .

- As ZCA-cor is explicitly constructed to maximise the total pairwise correlations it achieves the higher correlation than ZCA.
- If x is standardised to $\text{Var}(x_i) = 1$ then ZCA and ZCA-cor are identical.

2.4.3 PCA whitening

Aim: remove correlations and at the same compress information into a few latent variables. Specifically, we would like that the first latent component z_1 is maximally linked with all variables in x , followed by the second component z_2 and so on:

$$\begin{array}{llll} z_1 & \leftarrow & x_1 & \quad z_2 \leftarrow x_1 \quad \dots \\ z_1 & \leftarrow & x_2 & \quad z_2 \leftarrow x_2 \\ & & \vdots & \\ z_1 & \leftarrow & x_d & \quad z_2 \leftarrow x_d \end{array}$$

One way to measure the total link of each z_i with all x_j is the sum of the corresponding squared covariances

$$h_i = \sum_{j=1}^d \text{Cov}(z_i, x_j)^2 = \sum_{j=1}^d \phi_{ij}^2$$

In vector notation we write

$$\mathbf{h} = (h_1, \dots, h_d)^T = \text{Diag}(\Phi \Phi^T)$$

With $\Phi = Q_1 \Sigma^{1/2}$ this can be written $\mathbf{h} = \text{Diag}(Q_1 \Sigma Q_1^T)$ as a function of Q_1 .

PCA objective functions: maximise h_1, \dots, h_{d-1} in $\mathbf{h} = \text{Diag}(Q_1 \Sigma Q_1^T)$ such that $h_1 \geq h_2 \geq \dots \geq h_d$ to find an optimal optimal Q_1 and the corresponding whitening transformation.

Note that $\sum_{i=1}^d h_i = \text{Tr}(Q_1 \Sigma Q_1^T) = \text{Tr}(\Sigma)$ is constant regardless of the choice of Q_1 so there are only $d - 1$ independent h_i .

Solution:

- Apply eigendecomposition to $\Sigma = \mathbf{U} \Lambda \mathbf{U}^T$. Note that Λ is diagonal with positive entries $\lambda_1 \geq \lambda_2 \dots \geq \lambda_d > 0$ as Σ is positive definite and that the eigenvalues are already arranged non-increasing order. Also recall that \mathbf{U} is not uniquely defined — you are free to change the columns signs.
- The objective functions become $\mathbf{h} = \text{Diag}((Q_1 \mathbf{U}) \Lambda (Q_1 \mathbf{U})^T) = \text{Diag}(\mathbf{B} \Lambda \mathbf{B}^T)$ where \mathbf{B} is an orthogonal matrix, and $Q_1 = \mathbf{B} \mathbf{U}^T$.
- The optimal (maximum) values are achieved for $\mathbf{B} = \mathbf{I}$, with

$$h_i^{\text{PCA}} = \lambda_i$$

However, this is not the only solution — you can arbitrarily change the column signs of the matrix \mathbf{B} to arrive at the same maximum!

iv) The corresponding optimal value for the \mathbf{Q}_1 matrix is

$$\mathbf{Q}_1^{\text{PCA}} = \mathbf{u}^T$$

The corresponding whitening matrix is

$$\mathbf{W}^{\text{PCA}} = \mathbf{u}^T \mathbf{\Sigma}^{-1/2} = \mathbf{\Lambda}^{-1/2} \mathbf{u}^T$$

and the cross-covariance matrix is

$$\mathbf{\Phi}^{\text{PCA}} = \mathbf{\Lambda}^{1/2} \mathbf{u}^T$$

and the cross-correlation matrix is

$$\mathbf{\Psi}^{\text{PCA}} = \mathbf{\Lambda}^{1/2} \mathbf{u}^T \mathbf{V}^{-1/2}$$

Note that all of the above (i.e. $\mathbf{Q}_1^{\text{PCA}}, \mathbf{W}^{\text{PCA}}, \mathbf{\Phi}^{\text{PCA}}, \mathbf{\Psi}^{\text{PCA}}$) is not unique as we still have the sign ambiguity in the columns of \mathbf{U} (which also has absorbed the sign ambiguity of \mathbf{B})!

Identifiability:

Therefore, for identifiability reasons we need to impose a further constraint on $\mathbf{Q}_1^{\text{PCA}}$. A useful condition is to require a positive diagonal, i.e. $\text{Diag}(\mathbf{Q}_1^{\text{PCA}}) > 0$ and also $\text{Diag}(\mathbf{U}) > 0$. As a result, $\text{Diag}(\mathbf{\Phi}^{\text{PCA}}) > 0$ and $\text{Diag}(\mathbf{\Psi}^{\text{PCA}}) > 0$. With this constraint in place all pairs of x_i and z_i are positively correlated.

It is particularly important to pay attention to the sign ambiguity if different computer implementations of PCA whitening (and the related PCA approach) are used.

Proportion of total variation:

The sum of the maximised squared covariances for each latent component z_i is $\sum_{i=1}^d h_i^{\text{PCA}} = \sum_{i=1}^d \lambda_i$ and equals the total variation $\text{Tr}(\mathbf{\Sigma})$.

The fraction $\frac{\lambda_i}{\sum_{j=1}^d \lambda_j}$ is the proportional contribution of each element in \mathbf{z} to explain the total variation. Thus, low ranking components z_i with small $h_i^{\text{PCA}} = \lambda_i$ may be discarded. In fact, the aim of PCA whitening is to achieve this kind of compression and the resulting reduction in dimension in the latent space.

Summary:

- PCA whitening is a whitening transformation that maximises compression with the sum of squared cross-covariances as underlying optimality criterion.

- There are sign ambiguities in the PCA whitened variables which are inherited from the sign ambiguities in eigenvectors.
- If a positive-diagonal condition on the orthogonal matrices is imposed then these sign ambiguities are fully resolved and corresponding components z_i and x_i are always positively correlated.

2.4.4 PCA-cor whitening

Aim: same as for PCA whitening but remove scale in \mathbf{x} first. This means we use squared correlations rather than squared covariances to measure compression, i.e.

$$h_i = \sum_{j=1}^d \text{Cor}(z_i, x_j)^2 = \sum_{j=1}^d \psi_{ij}^2$$

In vector notation this becomes

$$\mathbf{h} = (h_1, \dots, h_d)^T = \text{Diag}(\Psi\Psi^T) = \text{Diag}(\mathbf{Q}_2\mathbf{P}\mathbf{Q}_2^T)$$

PCA-cor objective functions: maximise h_1, \dots, h_{d-1} in $\mathbf{h} = \text{Diag}(\mathbf{Q}_2\mathbf{P}\mathbf{Q}_2^T)$ such that $h_1 \geq h_2 \geq \dots \geq h_d$ to find an optimal optimal \mathbf{Q}_2 and the corresponding whitening transformation.

Note that $\sum_{i=1}^d h_i = \text{Tr}(\mathbf{Q}_2\mathbf{P}\mathbf{Q}_2^T) = \text{Tr}(\mathbf{P}) = d$ is constant regardless of the choice of \mathbf{Q}_2 so there are only $d - 1$ independent h_i .

Solution:

- Apply eigendecomposition to $\mathbf{P} = \mathbf{G}\mathbf{\Theta}\mathbf{G}^T$. Note that $\mathbf{\Theta}$ is diagonal with positive entries $\theta_1 \geq \theta_2 \dots \geq \theta_d > 0$ as \mathbf{P} is positive definite and that the eigenvalues are already arranged non-increasing order. Also recall that \mathbf{G} is not uniquely defined — you are free to change the columns signs.
- The objective functions become $\mathbf{h} = \text{Diag}((\mathbf{Q}_2\mathbf{G})\mathbf{\Theta}(\mathbf{Q}_2\mathbf{G})^T) = \text{Diag}(\mathbf{B}\mathbf{\Theta}\mathbf{B}^T)$ where \mathbf{B} is an orthogonal matrix, and $\mathbf{Q}_2 = \mathbf{B}\mathbf{G}^T$.
- The optimal (maximum) values are achieved for $\mathbf{B} = \mathbf{I}$, with

$$h_i^{\text{PCA-Cor}} = \theta_i$$

However, this is not the only solution — you can arbitrarily change the column signs of the matrix \mathbf{B} to arrive at the same maximum!

- The corresponding optimal value for the \mathbf{Q}_2 matrix is

$$\mathbf{Q}_2^{\text{PCA-Cor}} = \mathbf{G}^T$$

The corresponding whitening matrix is

$$\mathbf{W}^{\text{PCA-Cor}} = \mathbf{\Theta}^{-1/2} \mathbf{G}^T \mathbf{V}^{-1/2}$$

and the cross-covariance matrix is

$$\mathbf{\Phi}^{\text{PCA-Cor}} = \mathbf{\Theta}^{1/2} \mathbf{G}^T \mathbf{V}^{1/2}$$

and the cross-correlation matrix is

$$\mathbf{\Psi}^{\text{PCA-Cor}} = \mathbf{\Theta}^{1/2} \mathbf{G}^T$$

As with PCA whitening, there are sign ambiguities in the above because the column signs of \mathbf{G} can be freely chosen.

Identifiability:

For identifiability we choose to set $\text{Diag}(\mathbf{Q}_2^{\text{PCA-Cor}}) > 0$ and also $\text{Diag}(\mathbf{G}) > 0$ so that $\text{Diag}(\mathbf{\Phi}^{\text{PCA-Cor}}) > 0$ and $\text{Diag}(\mathbf{\Psi}^{\text{PCA-Cor}}) > 0$.

Proportion of total variation:

The sum of the maximised squared correlations for each latent component z_i is $\sum_{i=1}^d h_i^{\text{PCA-Cor}} = \sum_{i=1}^d \theta_i = d$ and equals the total variation $\text{Tr}(\mathbf{P})$. Therefore the fraction $\frac{\theta_i}{\sum_{j=1}^d \theta_j} = \frac{\theta_i}{d}$ is the proportional contribution of each element in \mathbf{z} to explain the total variation.

Summary:

- PCA-cor whitening is a whitening transformation that maximises compression with the sum of squared cross-correlations as underlying optimality criterion.
- There are sign ambiguities in the PCA-cor whitened variables which are inherited from the sign ambiguities in eigenvectors.
- If a positive-diagonal condition on the orthogonal matrices is imposed then these sign ambiguities are fully resolved and corresponding components z_i and x_i are always positively correlated.
- If \mathbf{x} is standardised to $\text{Var}(x_i) = 1$, then PCA and PCA-cor whitening are identical.

2.4.5 Cholesky whitening

Cholesky matrix decomposition:

The Cholesky decomposition of a square matrix $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ requires a positive definite \mathbf{A} and is unique. \mathbf{L} is a lower triangular matrix with positive diagonal elements. Its inverse \mathbf{L}^{-1} is also lower triangular with positive diagonal elements. If \mathbf{D} is a diagonal matrix with positive elements then $\mathbf{D}\mathbf{L}$ is also a lower

triangular matrix with a positive diagonal and the Cholesky factor for the matrix DAD .

Aim in Cholesky whitening:

Find a whitening transformation such that the cross-covariance Φ and cross-correlation Ψ have triangular structure. This is a structural constraint that occurs for example in some models for time course data.

To obtain a whitening transformation we can apply the Cholesky decomposition either to the precision matrix or to the covariance matrix. Furthermore, we can apply it to the inverse correlation or the correlation matrix.

Solution 1: Apply a Cholesky decomposition to $\Sigma^{-1} = LL^T$

The resulting whitening matrix is

$$W^{\text{Chol.1}} = L^T$$

By construction, $W^{\text{Chol.1}}$ is an upper triangular matrix with positive diagonal, and it clearly satisfies the whitening constraint since $(W^{\text{Chol.1}})^T W^{\text{Chol.1}} = \Sigma^{-1}$.

The cross-covariance matrix is (with $\Sigma = (L^{-1})^T L^{-1}$)

$$\Phi^{\text{Chol.1}} = L^T \Sigma = L^{-1}$$

and the cross-correlation matrix is

$$\Psi^{\text{Chol.1}} = L^T \Sigma V^{-1/2} = L^{-1} V^{-1/2}$$

Note that both $\Phi^{\text{Chol.1}}$ and $\Psi^{\text{Chol.1}}$ are lower triangular matrices with positive diagonal elements. Hence two corresponding components z_i and x_i are always positively correlated!

Finally, the corresponding orthogonal matrices are

$$Q_1^{\text{Chol.1}} = L^T \Sigma^{1/2} = L^{-1} \Sigma^{-1/2}$$

and

$$Q_2^{\text{Chol.1}} = L^T V^{1/2} P^{1/2} = L^{-1} V^{-1/2} P^{-1/2}$$

Solution 2: Apply a Cholesky decomposition to $\Sigma = FF^T$

The resulting whitening matrix is

$$W^{\text{Chol.2}} = F^{-1}$$

By construction, $W^{\text{Chol.2}}$ is a lower triangular matrix with positive diagonal. The whitening constraint is satisfied as $(W^{\text{Chol.2}})^T W^{\text{Chol.2}} = (F^{-1})^T F^{-1} = (F^T)^{-1} F^{-1} = (FF^T)^{-1} = \Sigma^{-1}$.

The cross-covariance matrix is (recall that in general $\mathbf{W}^{-1} = \Phi^T$)

$$\Phi^{\text{Chol.2}} = \mathbf{F}^T$$

and the cross-correlation matrix is

$$\Psi^{\text{Chol.2}} = \mathbf{F}^T \mathbf{V}^{-1/2}$$

Both $\Phi^{\text{Chol.2}}$ and $\Psi^{\text{Chol.2}}$ are upper triangular matrices with positive diagonal elements. Hence two corresponding components z_i and x_i are always positively correlated!

Finally, the corresponding orthogonal matrices are

$$\mathbf{Q}_1^{\text{Chol.2}} = \mathbf{F}^T \Sigma^{-1/2} = \mathbf{F}^{-1} \Sigma^{1/2}$$

and

$$\mathbf{Q}_2^{\text{Chol.2}} = \mathbf{F}^T \mathbf{V}^{-1/2} \mathbf{P}^{-1/2} = \mathbf{F}^{-1} \mathbf{V}^{1/2} \mathbf{P}^{1/2}$$

Note that the two Cholesky whitening procedures are different with in general $\mathbf{W}^{\text{Chol.1}} \neq \mathbf{W}^{\text{Chol.2}}$.

Application to correlation instead of covariance:

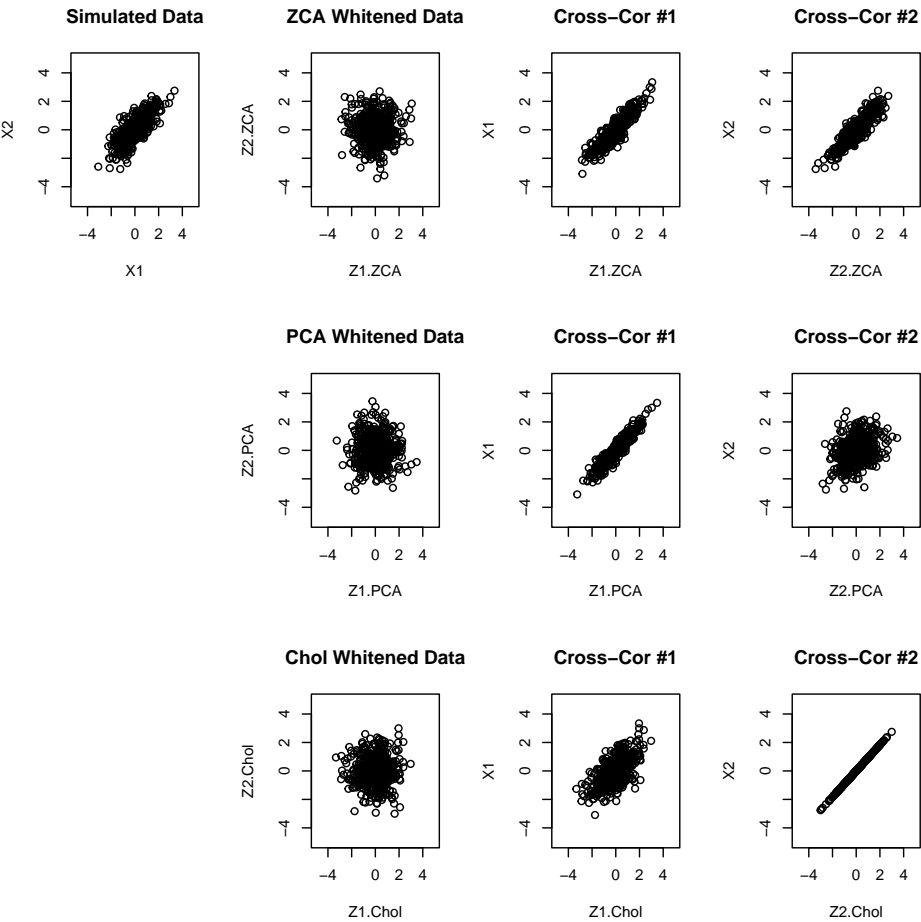
We may also wish to apply Cholesky decomposition to the (inverse) correlation rather than the (inverse) covariance matrix. Intriguingly, this does *not* lead to different whitening transforms (unlike for ZCA and PCA).

For example, the Cholesky factor for $\mathbf{P}^{-1} = \mathbf{V}^{1/2} \Sigma^{-1} \mathbf{V}^{1/2}$ is $\mathbf{V}^{1/2} \mathbf{L}$. The corresponding whitening matrix is $(\mathbf{V}^{1/2} \mathbf{L})^T \mathbf{V}^{-1/2} = \mathbf{L}^T = \mathbf{W}^{\text{Chol.1}}$.

Similarly, the Cholesky factor for $\mathbf{P} = \mathbf{V}^{-1/2} \Sigma \mathbf{V}^{-1/2}$ is $\mathbf{V}^{-1/2} \mathbf{F}$. The corresponding whitening matrix is $(\mathbf{V}^{-1/2} \mathbf{F})^{-1} \mathbf{V}^{-1/2} = \mathbf{F}^{-1} = \mathbf{W}^{\text{Chol.2}}$.

2.4.6 Comparison of ZCA, PCA and Cholesky whitening

For comparison, here are the results of ZCA, PCA and Cholesky whitening (precision matrix) applied to a simulated bivariate normal data set with correlation $\rho = 0.8$.



In column 1 you can see the simulated data as scatter plot.

Column 2 shows the scatter plots of the whitened data — as expect all three methods removed correlation produce isotropic covariance.

The three approached differ differ in the cross-correlations. Columns 3 and 4 show the cross-correlations between the first two corresponding components (x_1 and z_1 , and x_2 and z_2) for ZCA, PCA and Cholesky whitening. As expected, in ZCA both pairs show strong correlation, but this is not the case for PCA and Cholesky whitening.

2.4.7 Recap

Method	Type of usage
ZCA, ZCA-cor:	pure decorrelate, maintain similarity to original data set, interpretability

Method	Type of usage
PCA, PCA-cor:	compression, find effective dimension, reduce dimensionality, feature identification
Cholesky	time course data, triangular W , Φ and Ψ

Related models not discussed in this course:

- Factor models: essentially whitening plus an additional error term, factors have rotational freedom just like in whitening
- PLS: similar to PCA but in regression setting (with the choice of latent variables depending on the response)

2.5 Principal Component Analysis (PCA)

2.5.1 PCA transformation

Principal component analysis was proposed in 1933 by Harald Hotelling⁵ and is very closely related to **PCA whitening**. The underlying mathematics was developed earlier in 1901 by Karl Pearson⁶ for the problem of orthogonal regression.

Assume random vector x with $\text{Var}(x) = \Sigma = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. PCA is a particular orthogonal transformation of the original x such that the resulting components are orthogonal:

$$\underbrace{\mathbf{t}^{\text{PCA}}}_{\text{Principal components}} = \underbrace{\mathbf{U}^T}_{\text{Orthogonal matrix}} x$$
$$\text{Var}(\mathbf{t}^{\text{PCA}}) = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}$$

Note that while principal components are *orthogonal* they do *not* have unit variance but the variance of principal components t_i equals the eigenvalues λ_i .

Thus PCA itself is *not* a whitening procedure. However, you arrive at PCA whitening by simply by standardising the PCA components: $\mathbf{z}^{\text{PCA}} = \mathbf{\Lambda}^{-1/2}\mathbf{t}^{\text{PCA}}$

Compression properties:

The total variation is $\text{Tr}(\text{Var}(\mathbf{t}^{\text{PCA}})) = \text{Tr}(\mathbf{\Lambda}) = \sum_{j=1}^d \lambda_j$. With principle components the fraction $\frac{\lambda_i}{\sum_{j=1}^d \lambda_j}$ can be interpreted as the proportion of variation

⁵Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:417–441 (Part 1) and 24:498–520 (Part 2). <https://doi.org/10.1037/h0071325> and <https://doi.org/10.1037/h0070888>

⁶Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2:559–572.

contributed by each component in \mathbf{t}^{PCA} to the total variation. Thus, low ranking components in \mathbf{t}^{PCA} with low variation may be discarded, thus leading to a reduction in dimension.

2.5.2 Application to data

Written in terms of a data matrix \mathbf{X} instead of a random vector \mathbf{x} PCA becomes:

$$\underbrace{\mathbf{T}}_{\text{Sample version of principal components}} = \underbrace{\mathbf{X}}_{\text{Data matrix}} \mathbf{U}$$

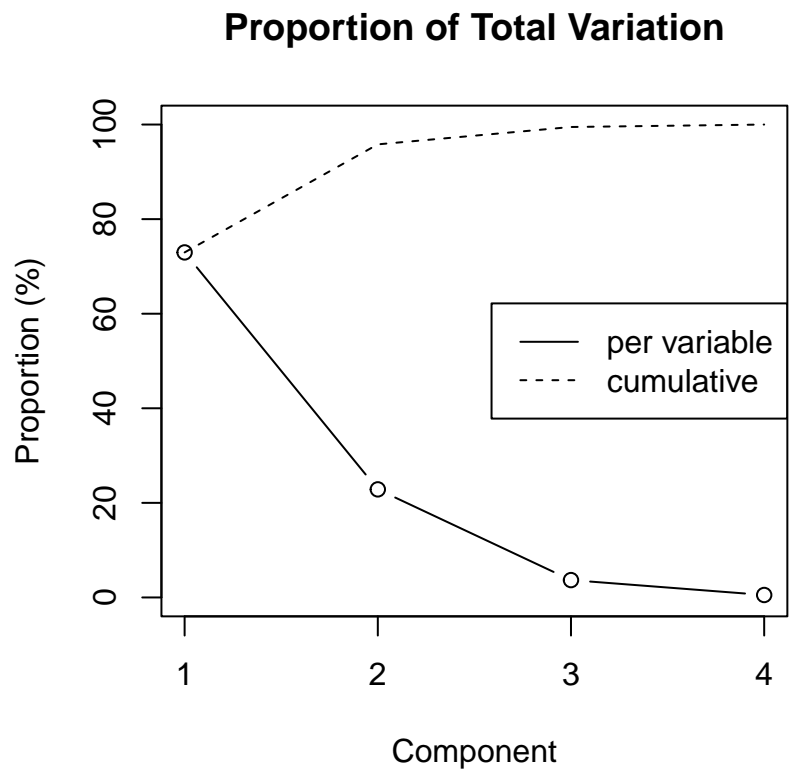
There are now two ways to obtain \mathbf{U} :

1. Estimate the covariance matrix, e.g. by $\hat{\Sigma} = \frac{1}{n} \mathbf{X}_c^T \mathbf{X}_c$ where \mathbf{X}_c is the column-centred data matrix; then apply the eigenvalue decomposition on $\hat{\Sigma}$ to get \mathbf{U} .
2. Compute the singular value decomposition of $\mathbf{X}_c = \mathbf{V} \mathbf{D} \mathbf{U}^T$. As $\hat{\Sigma} = \frac{1}{n} \mathbf{X}_c^T \mathbf{X}_c = \mathbf{U} (\frac{1}{n} \mathbf{D}^2) \mathbf{U}^T$ you can just use \mathbf{U} from the SVD of \mathbf{X}_c and there is no need to compute the covariance.

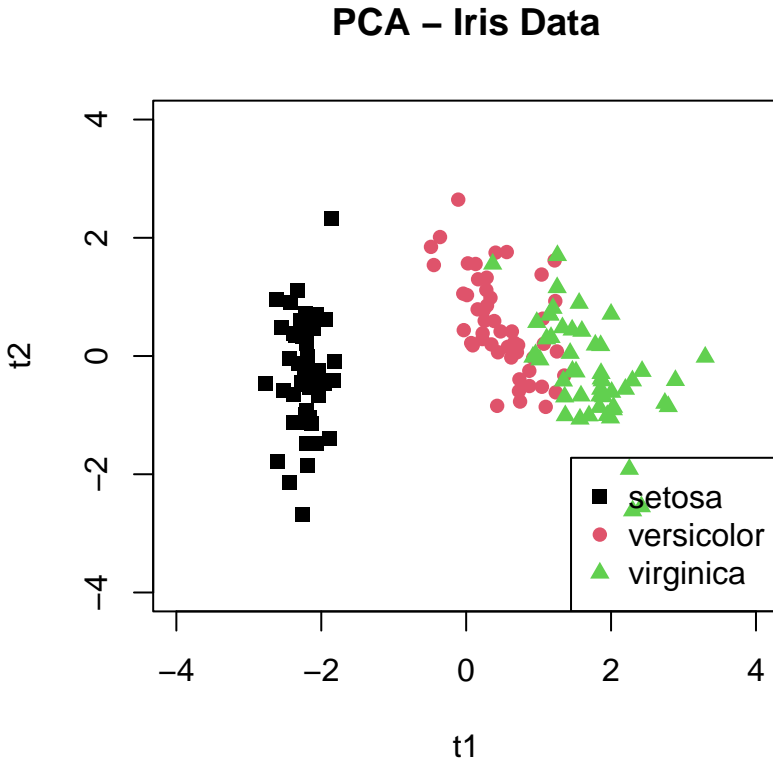
2.5.3 Iris data example

As an example we consider the famous [iris flower data set](#). It consists of data for for botanical variables (sepal length, sepal width, petal length and petal width) measured on 150 flowers from three iris species (setosa, versicolr, virginica). Thus for this data set $d = 4$ and $n = 150$.

We first standardise the data, then compute PCA components and plot the proportion of total variance contributed by each component. This shows that only two PCA components are needed to achieve 95% of the total variation:



A scatter plot plot of the the first two principal components is also informative:



This shows that there groupings among the 150 flowers, corresponding to the species, and that these groups can be characterised by the the principal components.

2.6 Correlation loadings plot to interpret PCA components

2.6.1 PCA correlation loadings

In an earlier section we have learned that for a general whitening transformation the cross-correlations $\Psi = \text{Cor}(z, x)$ play the role of correlation loadings in the inverse transformation:

$$V^{-1/2}x = \Psi^T z,$$

i.e. they are the coefficients linking the whitened variable z with the standardised original variable x . This relationship holds therefore also for PCA-whitening with $z^{\text{PCA}} = \Lambda^{-1/2}U^T x$ and $\Psi^{\text{PCA}} = \Lambda^{1/2}U^T V^{-1/2}$.

The classical PCA is not a whitening approach because $\text{Var}(t^{\text{PCA}}) \neq I$. However, we can still compute cross-correlations between the principal components t^{PCA}

and \mathbf{x} , resulting in

$$\text{Cor}(\mathbf{t}^{\text{PCA}}, \mathbf{x}) = \mathbf{\Lambda}^{1/2} \mathbf{U}^T \mathbf{V}^{-1/2} = \mathbf{\Psi}^{\text{PCA}}$$

Note these are the same as the cross-correlations for PCA-whitening since \mathbf{t}^{PCA} and \mathbf{z}^{PCA} only differ in scale.

The inverse PCA transformation is

$$\mathbf{x} = \mathbf{U} \mathbf{t}^{\text{PCA}}$$

In terms of standardised PCA components and standardised original components it becomes

$$\mathbf{V}^{-1/2} \mathbf{x} = \mathbf{\Psi}^T \mathbf{\Lambda}^{-1/2} \mathbf{t}^{\text{PCA}}$$

Thus the cross-correlation matrix $\mathbf{\Psi}$ plays the role of *correlation loadings* also in classical PCA, i.e. they are the coefficients linking the standardised PCA components with the standardised original components.

2.6.2 PCA correlation loadings plot

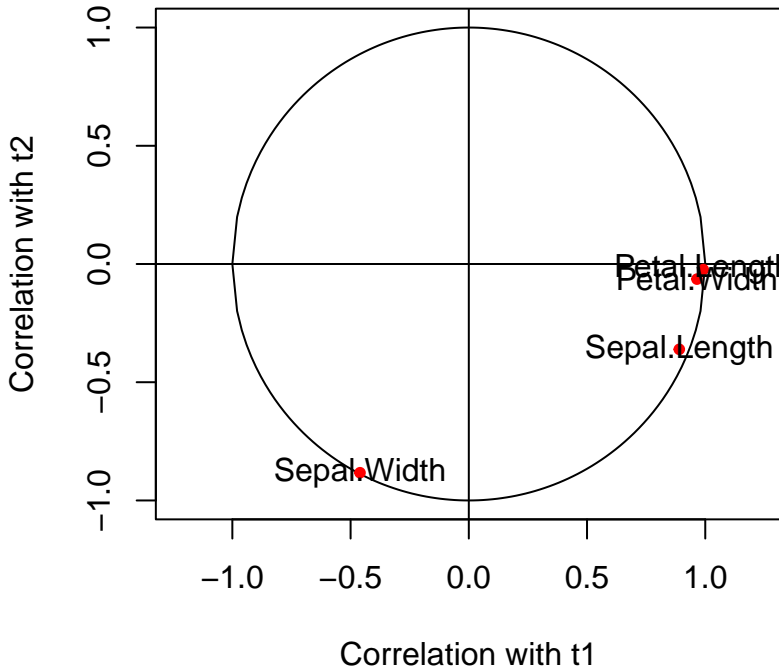
In PCA and PCA-cor whitening as well as in classical PCA the aim is compression, i.e. to find latent variables such that most of the total variation is contributed by a small number of components.

In order to be able to better interpret the top ranking PCA component we can use a visual device called *correlation loadings plot*. For this we compute the correlation between the PCA components 1 and 2 (t_1^{PCA} and t_2^{PCA}) with all original variables x_1, \dots, x_d .

For each original variable x_j we therefore have two numbers between -1 and 1, the correlation $\text{Cor}(t_1^{\text{PCA}}, x_j)$ and $\text{Cor}(t_2^{\text{PCA}}, x_j)$ that we use as coordinates to draw a point in a plane. By construction, all points have to lie within a unit circle around the origin. As the sum of the squared correlation loadings from all latent components to one specific original variable sum up to one, the sum of the squared loadings from just the first two components is also at most 1. The original variables most strongly influenced by the two latent variables will have strong correlation and thus lie near the outer circle, whereas variables that are not influenced by the two latent variables will lie near the origin.

As an example, here is the correlation loadings plot showing the cross-correlation between the first two PCA components and all four variables of the iris flower data set discussed earlier.

Correlation Loadings Plot Iris Data



The interpretation of this plot is discussed in Worksheet 4.

2.7 CCA whitening (Canonical Correlation Analysis)

Canonical correlation analysis was invented by Harald Hotelling in 1936.⁷

So far, we have looked only into whitening as a **single** vector x . In CCA whitening we consider **two** vectors x and y simultaneously:

$$\begin{array}{ll}
 x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} & y = \begin{pmatrix} y_1 \\ \vdots \\ y_q \end{pmatrix} \\
 \text{Dimension } p & \text{Dimension } q
 \end{array}
 \quad
 \begin{array}{l}
 \text{Var}(x) = \Sigma_x = V_x^{1/2} P_x V_x^{1/2} \\
 \text{Var}(y) = \Sigma_y = V_y^{1/2} P_y V_y^{1/2}
 \end{array}$$

$$\text{Whitening of } x: z_x = W_x x = Q_x P_x^{-1/2} V_x^{-1/2} x$$

$$\text{Whitening of } y: z_y = W_y y = Q_y P_y^{-1/2} V_y^{-1/2} y$$

⁷Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* **28**:321–377.

(note we use the correlation-based form of W)

Cross-correlation between z_x and z_y :

$$\text{Cor}(z_x, z_y) = Q_x K Q_y^T$$

with $K = P_x^{-1/2} P_{xy} P_y^{-1/2}$.

Idea: we can choose suitable orthogonal matrices Q_x and Q_y by putting constraints on the cross-correlation.

CCA: we aim for a *diagonal* $\text{Cor}(z_x, z_y)$ so that each component in z_x only influences one (the corresponding) component in z_y .

Motivation: pairs of “modules” represented by components of z_x and z_y influencing each other (and not anyone other module).

$$z_x = \begin{pmatrix} z_1^x \\ z_2^x \\ \vdots \\ z_p^x \end{pmatrix} \quad z_y = \begin{pmatrix} z_1^y \\ z_2^y \\ \vdots \\ z_q^y \end{pmatrix}$$

$$\text{Cor}(z_x, z_y) = \begin{pmatrix} d_1 & \dots & 0 \\ \vdots & \vdots & \\ 0 & \dots & d_m \end{pmatrix}$$

where d_i are the *canonical correlations* and $m = \min(p, q)$.

2.7.1 How to make cross-correlation matrix $\text{Cor}(z_x, z_y)$ diagonal?

- Use Singular Value Decomposition (SVD) of matrix K :

$$K = (Q_x^{\text{CCA}})^T D Q_y^{\text{CCA}}$$

where D is the diagonal matrix containing the singular values of K

- This yields orthogonal matrices Q_x^{CCA} and Q_y^{CCA} and thus the desired whitened matrices W_x^{CCA} and W_y^{CCA}
- As a result $\text{Cor}(z_x, z_y) = D$ i.e. singular values of K are identical to canonical correlations d_i !

→ Q_x^{CCA} and Q_y^{CCA} are determined by the diagonality constraint (and are different to the other previously discussed whitening methods).

Note that the signs of corresponding in columns in Q_x^{CCA} and Q_y^{CCA} are not identified. Traditionally, in an SVD the signs are chosen such that the singular

values are positive. However, if we impose positive-diagonality on Q_x^{CCA} and Q_y^{CCA} , and thus positive-diagonality on the cross-correlations Ψ_x and Ψ_y , then the canonical correlations may take on both positive and negative values.

2.7.2 Related methods

- O2PLS: similar to CCA but using orthogonal projections (thus in O2PLS the latent variables underlying x and y are not orthogonal)
- Vector correlation: aggregates the squared canonical correlations into a single overall measure of association between two random vectors x and y (see Chapter 5 on multivariate dependencies).

Chapter 3

Unsupervised learning and clustering

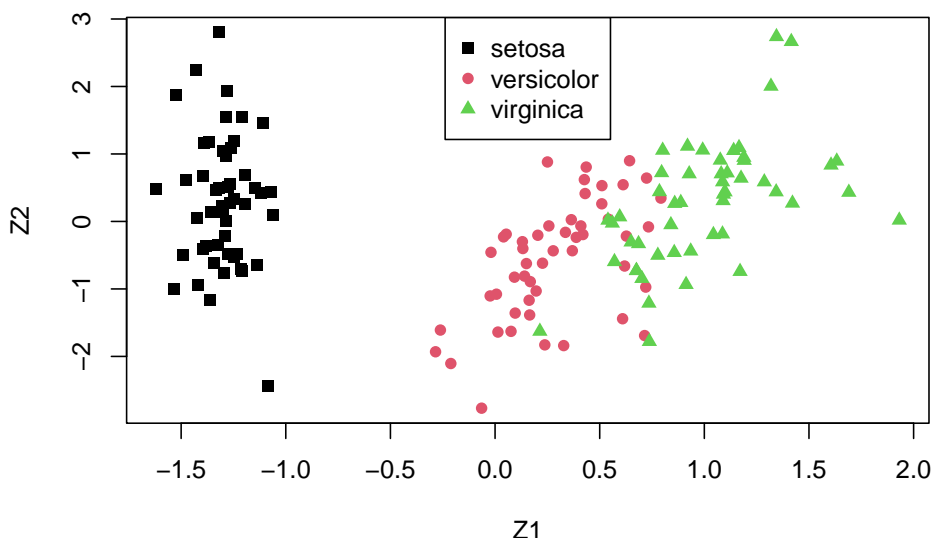
3.1 Challenges in supervised learning

3.1.1 Objective

We observe data x_1, \dots, x_n for n objects (or subjects). Each sample x_i is a vector of dimension d . Thus, for each of the n objects / subjects we have measurements on d variables. The aim of unsupervised learning is to identify patterns relating the objects/subjects based on the information available in x_i . Note that unsupervised learning *only* uses this information and nothing else.

For illustration consider the first two principal components of the Iris flower data (see e.g. Worksheet 4):

PCA Whitening – Iris Data



Clearly there is a group structure among the samples that is linked to particular patterns in the first two principal components.

Note that in this plot we have used additional information, the class labels (setosa, versicolor, virginica), to highlighting the true underlying structure (the three flower species).

In **unsupervised learning** the class labels are (assumed to be) unknown, and the aim is to infer the clustering and thus the classes labels.¹

There are many methods for clustering and unsupervised learning, both purely algorithmic as well as probabilistic. In this chapter we will study a few of the most commonly used approaches.

3.1.2 Questions and problems

In order to implement unsupervised learning we need to address a number of questions:

- how do we define clusters?
- how do we learn / infer clusters?
- how many clusters are there? (this is surprisingly difficult!)
- how can we assess the uncertainty of clusters?

Once we know the clusters we are also interested in:

- which features define / separate each cluster?

¹In contrast, in **supervised learning** (to be discussed in a subsequent chapter) the class labels are known for a subset of the data (the training data set) and are required to learn a prediction function.

(note this is a feature / variable selection problem, discussed in supervised learning).

Many of these problems and questions are highly specific to the data at hand. Correspondingly, there are many different types of methods and models for clustering and unsupervised learning.

In terms of representing the data, unsupervised learning tries to balance between the following two extremes:

- 1) all objects are grouped into a single cluster (low complexity model)
- 2) all objects are put into their own cluster (high complexity model)

In practise, the aim is to find a compromise, i.e. a model that captures the structure in the data with appropriate complexity — not too low and not too complex.

3.1.3 Why is clustering difficult?

Partitioning problem (combinatorics): How many partitions of n objects (say flowers) into K groups (say species) exists?

Answer:

$$S(n, K) = \left\{ \begin{matrix} n \\ K \end{matrix} \right\}$$

this is the “Sterling number of the second type”.

For large n :

$$S(n, K) \approx \frac{K^n}{K!}$$

Example:

n	K	Number of possible partitions
15	3	≈ 2.4 million (10^6)
20	4	≈ 2.4 billion (10^9)
\vdots		
100	5	$\approx 6.6 \times 10^{76}$

These are enormously big numbers even for relatively small problems!

⇒ Clustering / partitioning / structure discovery is not easy!

⇒ We cannot expect perfect answers or a single “true” clustering

In fact, as a model of the data many different clusterings may fit the data equally well.

⇒ We need to assess the uncertainty of the clustering

This can be done as part of probabilistic modelling or by resampling (e.g., bootstrap).

3.1.4 Common types of clustering methods

There are very many different clustering algorithms!

We consider the following two broad types of methods:

1) **Algorithmic clustering methods** (these are not explicitly based on a probabilistic model)

- K-means
- PAM
- hierarchical clustering (distance or similarity-based, divide and agglomerative)

pros: fast, effective algorithms to find at least some grouping **cons:** no probabilistic interpretation, blackbox methods

2) **Model-based clustering** (based on a probabilistic model)

- mixture models (e.g. Gaussian mixture models, GMMs, non-hierarchical)
- graphical models (e.g. Bayesian networks, Gaussian graphical models GGM, trees and networks)

pros: full probabilistic model with all corresponding advantages
cons: computationally very expensive, sometimes impossible to compute exactly.

3.2 Hierarchical clustering

3.2.1 Tree-like structures

Often, categorisations of objects are nested, i.e. there sub-categories of categories etc. These can be naturally represented by **tree-like hierarchical structures**.

In many branches of science hierarchical clusterings are widely employed, for example in evolutionary biology: see e.g.

- [Tree of Life](#) explaining the biodiversity of life
- phylogenetic trees among species (e.g. vertebrata)
- population genetic trees to describe human evolution
- taxonomic trees for plant species
- etc.

Note that when visualising hierarchical structures typically the corresponding tree is depicted facing downwards, i.e. the root of the tree is shown on the top, and the tips/leaves of the tree are shown at the bottom!

In order to obtain such a hierarchical clustering from data two opposing strategies are commonly used:

- 1) **divisive or recursive partitioning algorithms**
 - grow the tree from the root downwards
 - first determine the main two clusters, then recursively refine the clusters further.
- 2) **agglomerative algorithms**
 - grow the tree from the leaves upwards
 - successively form partitions by first joining individual object together, then recursively join groups of items together, until all is merged.

In the following we discuss a number of popular hierarchical agglomerative clustering algorithms that are based on the pairwise distances / similarities (a $n \times n$ matrix) among all data points.

3.2.2 Agglomerative hierarchical clustering algorithms

A general algorithm for agglomerative construction of a hierarchical clustering works as follows:

Initialisation:

Compute a dissimilarity / distance matrix between all pairs of objects where “objects” are single data points at this stage but later are also be sets of data points.

Iterative procedure:

- 1) identify the pair of objects with the smallest distance. These two objects are then merged together into a common set. Create an internal node in the tree to describe this coalescent event.
- 2) update the distance matrix by computing the distances between the new set and all other objects. If the new set contains all data points the procedure terminates.

For actual implementation of this algorithm two key ingredients are needed:

- 1) a distance measure $d(\mathbf{a}, \mathbf{b})$ between two individual elementary data points \mathbf{a} and \mathbf{b} .

This is typically one of the following:

- Euclidean distance $d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^d (a_i - b_i)^2} = \sqrt{(\mathbf{a} - \mathbf{b})^T (\mathbf{a} - \mathbf{b})}$
- Manhattan distance $d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^d |a_i - b_i|$
- Maximum norm $d(\mathbf{a}, \mathbf{b}) = \max_{i \in \{1, \dots, d\}} |a_i - b_i|$

In the end, making the correct choice of distance will require subject knowledge about the data!

- 2) a distance measure $d(A, B)$ between two sets of objects $A = \{a_1, a_2, \dots, a_{n_A}\}$ and $B = \{b_1, b_2, \dots, b_{n_B}\}$ of size n_A and n_B , respectively.

To determine the distance $d(A, B)$ between these two sets the following measures are often employed:

- **complete linkage** (max. distance): $d(A, B) = \max_{a_i \in A, b_i \in B} d(a_i, b_i)$
- **single linkage** (min. distance): $d(A, B) = \min_{a_i \in A, b_i \in B} d(a_i, b_i)$
- **average linkage** (avg. distance): $d(A, B) = \frac{1}{n_A n_B} \sum_{a_i \in A} \sum_{b_i \in B} d(a_i, b_i)$

3.2.3 Ward's clustering method

Another agglomerative hierarchical procedure is **Ward's minimum variance approach**. In this approach in each iteration the two sets A and B are merged that lead to the *smallest increase in within-group variation, or equivalently, 5h3 total within-group sum of squares* (cf. K-means). The centroids of the two sets is given by $\mu_A = \frac{1}{n_A} \sum_{a_i \in A} a_i$ and $\mu_B = \frac{1}{n_B} \sum_{b_i \in B} b_i$.

The within-group sum of squares for group A is

$$w_A = \sum_{a_i \in A} (a_i - \mu_A)^T (a_i - \mu_A)$$

and is computed here on the basis of the difference of the observations a_i relative to their mean μ_A . However, it is also possible to compute it from the pairwise differences between the observations using

$$w_A = \frac{1}{n_A} \sum_{a_i, a_j \in A, i < j} (a_i - a_j)^T (a_i - a_j)$$

This trick is used in Ward's clustering method by constructing a distance measure between to sets A and B as

$$d(A, B) = w_{A \cup B} - w_A - w_B.$$

Correspondingly, the distance between two elementary data points a and b is the squared Euclidean distance

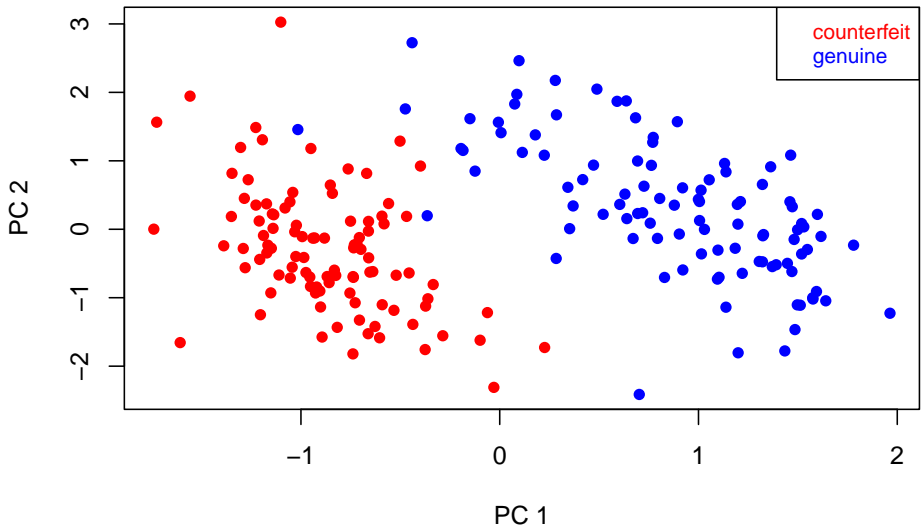
$$d(a, b) = (a - b)^T (a - b).$$

3.2.4 Application to Swiss banknote data set

This data set is reports 6 physical measurements on 200 Swiss bank notes. Of the 200 notes 100 are genuine and 100 are counterfeit. The measurements are: length, left width, right width, bottom margin, top margin, diagonal length of the bank notes.

Plotting the first two PCAs of this data shows that there are indeed two well defined groups, and that these groups correspond precisely to the genuine and counterfeit banknotes:

PCA Swiss Banknote Data

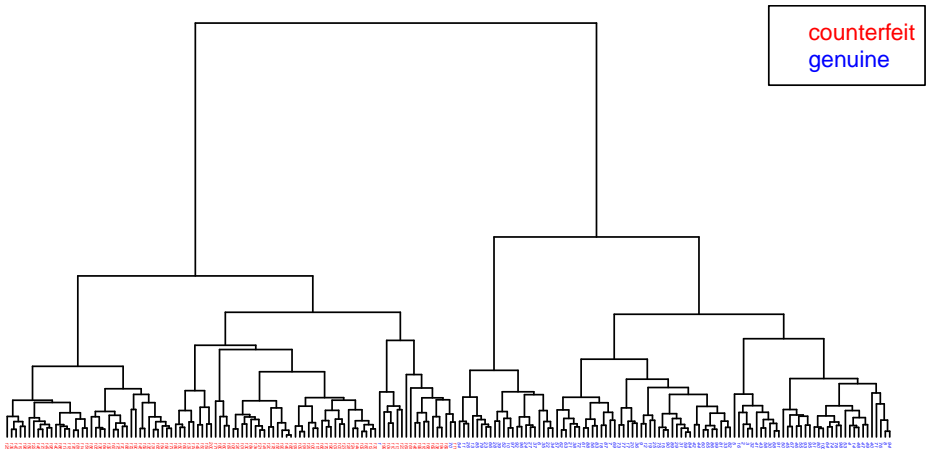


We now compare the hierarchical clusterings of the Swiss bank note data using four different methods using Euclidean distance.

An interactive [R Shiny web app](https://minerva.it.manchester.ac.uk/shiny/strimmer/hclust/) of this analysis (which also allows to explore further distance measures) is available online at <https://minerva.it.manchester.ac.uk/shiny/strimmer/hclust/>.

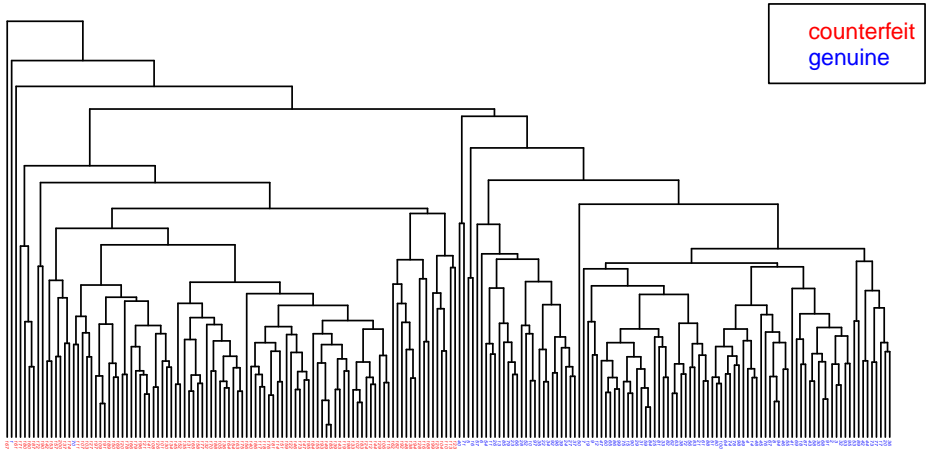
Ward.D2 (=Ward’s method):

ward.D2 + euclidean



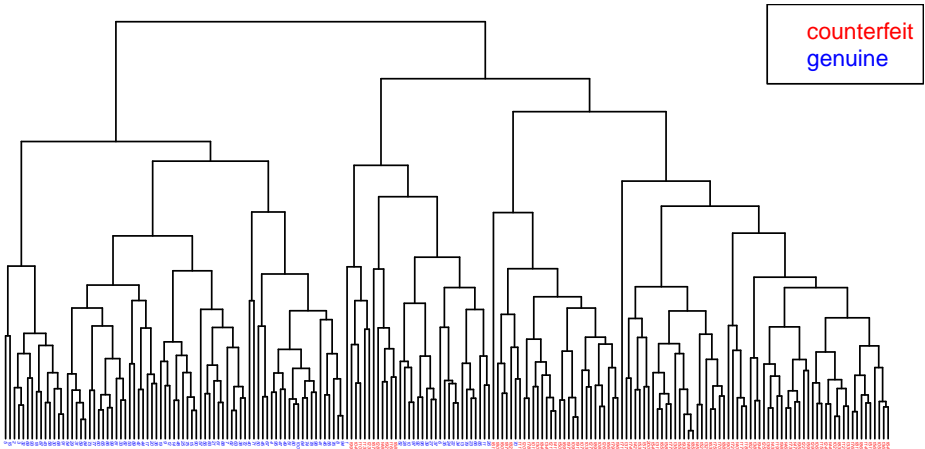
Average linkage:

average + euclidean



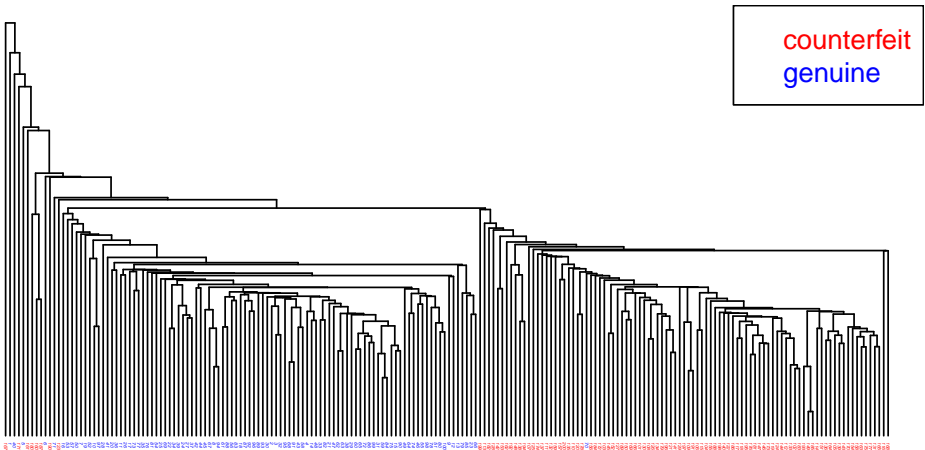
Complete linkage:

complete + euclidean



Single linkage:

single + euclidean



Result:

- All four trees / hierarchical clusterings are quite different!
- The Ward.D2 method is the only one that finds the correct grouping (except for a single error).

3.2.5 Assessment of the uncertainty of hierarchical clusterings

In practical application of hierarchical clustering methods it is essential to evaluate the stability and uncertainty of the obtained groupings. This is often done as follows using the “bootstrap”:

- Sampling with replacement is used to generate a number of so-called bootstrap data sets (say $B = 200$) similar to the original one. Specifically, we create new data matrices by repeatedly randomly selecting columns (variables) from the original data matrix for inclusion in the bootstrap data matrix. Note that we sample columns as our aim is to cluster the samples.
- Subsequently, a hierarchical clustering is computed for each of the bootstrap data sets. As a result, we now have an “ensemble” of B bootstrap trees.
- Finally, analysis of the clusters (bipartitions) shown in all the bootstrap trees allows to count the clusters that appear frequently, and also those that appear less frequently. These counts provide a measure of the stability of the clusterings appearing in the original tree.
- Additionally, from the bootstrap tree we can also compute a consensus tree containing the most stable clusters. This can be viewed as an “ensemble average” of all the bootstrap trees.

A disadvantage of this procedure is that bootstrapping trees is computationally very very expensive, as the original procedure is already time consuming but now needs to be repeated a large number of times.

3.3 K -means clustering

3.3.1 General aims

- Partition the data into K groups, with K given in advance
- The groups are non-overlapping, so each of the n data points / objects x_i is assigned to exactly one of the K groups
- maximise the homogeneity within each group (i.e. each group should contain similar objects)
- maximise the heterogeneity among the different groups (i.e. each group should differ from the other groups)

3.3.2 Algorithm

For each group $k \in \{1, \dots, K\}$ we assume a group mean μ_k .

After running K -means we will get estimates of $\hat{\mu}_k$ of the group means, as well as an allocation of each data point to one of the classes.

Initialisation:

At the start of the algorithm the n observations x_1, \dots, x_n are randomly allocated

with equal probability to one of the K groups. The resulting assignment is given by the function $C(x_i) \in \{1, \dots, K\}$. With $G_k = \{i | C(x_i) = k\}$ we denote the set of indices of the data points in cluster k , and with $n_k = |G_k|$ the number of samples in cluster k .

Iterative refinement:

- 1) estimate the group means by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in G_k} x_i$$

- 2) update group assignment: each data point x_i is (re)assigned to the group k with the nearest $\hat{\mu}_k$ (in terms of the Euclidean norm). Specifically, the assignment $C(x_i)$ is updated to

$$\begin{aligned} C(x_i) &= \arg \min_k |x_i - \hat{\mu}_k|_2 \\ &= \arg \min_k (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k) \end{aligned}$$

Steps 1 and 2 are repeated until the algorithm converges (or an upper limit of repeats is reached).

3.3.3 Properties

Despite its simplicity K -means is a surprisingly effective clustering algorithms.

The final clustering depends on the initialisation so it is often useful to run K -means several times with different starting allocations of the data points. Furthermore, non-random or non-uniform initialisations can lead to improved and faster convergence, see e.g. the [K-means++](#) algorithm.

As a result of the way the clusters are assigned in K -means the corresponding cluster boundaries form a Voronoi tessellation (cf. https://en.wikipedia.org/wiki/Voronoi_diagram) around the cluster means.

Later we will also discuss the connection of K -means with probabilistic clustering using Gaussian mixture models.

3.3.4 Choosing the number of clusters

Once the K -means clustering has been obtained it is insightful to compute:

- a) the total within-group sum of squares SSW (tot.withinss), or total unexplained sum of squares:

$$SSW = \sum_{k=1}^K \sum_{i \in G_k} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)$$

This quantity decreases with K and is zero for $K = n$. The K -means algorithm tries to minimise this quantity but it will typically only find a local minimum rather than the global one.

- b) the between-group sum of squares SSB (betweeness), or explained sum of squares:

$$SSB = \sum_{k=1}^K n_k (\hat{\mu}_k - \hat{\mu}_0)^T (\hat{\mu}_k - \hat{\mu}_0)$$

where $\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{k=1}^K n_k \hat{\mu}_k$ is the global mean of the samples. SSB increases with the number of clusters K until for $K = n$ it becomes equal to

- c) the total sum of squares

$$SST = \sum_{i=1}^n (x_i - \hat{\mu}_0)^T (x_i - \hat{\mu}_0).$$

By construction $SST = SSB + SSW$ for any K (i.e. SST is a constant independent of K).

If divide the sum of squares by the sample size n we get $T = \frac{SST}{n}$ as the *total variation*, $B = \frac{SSW}{n}$ as the *explained variation* and $W = \frac{SSW}{n}$ as the *total unexplained variation*, with $T = B + W$.

For deciding on the optimal number of clusters we can run K -means for various settings of K and then choose the smallest K for which the explained variation B is not significantly worse compared to a model with substantially larger number of clusters (see example below).

3.3.5 K -medoids aka PAM

A closely related clustering method is K -medoids or PAM (“Partitioning Around Medoids”).

This works exactly like K -means, only that

- instead of the estimated group means $\hat{\mu}_k$ one member of each group is selected as its representative (the so-called “medoid”)
- instead of squared Euclidean distance other dissimilarity measures are also allowed.

3.3.6 Application of K -means to Iris data

Scatter plots of Iris data:


```
## [1] 47.35062 44.08754 47.45019
## (between_SS / total_SS = 76.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

The corresponding total within-group sum of squares (SSW , `tot.withinss`) is

```
kmeans.out$tot.withinss
```

```
## [1] 138.8884
```

and the between-group sum of squares (SSB , `betweenss`) is

```
kmeans.out$betweenss
```

```
## [1] 457.1116
```

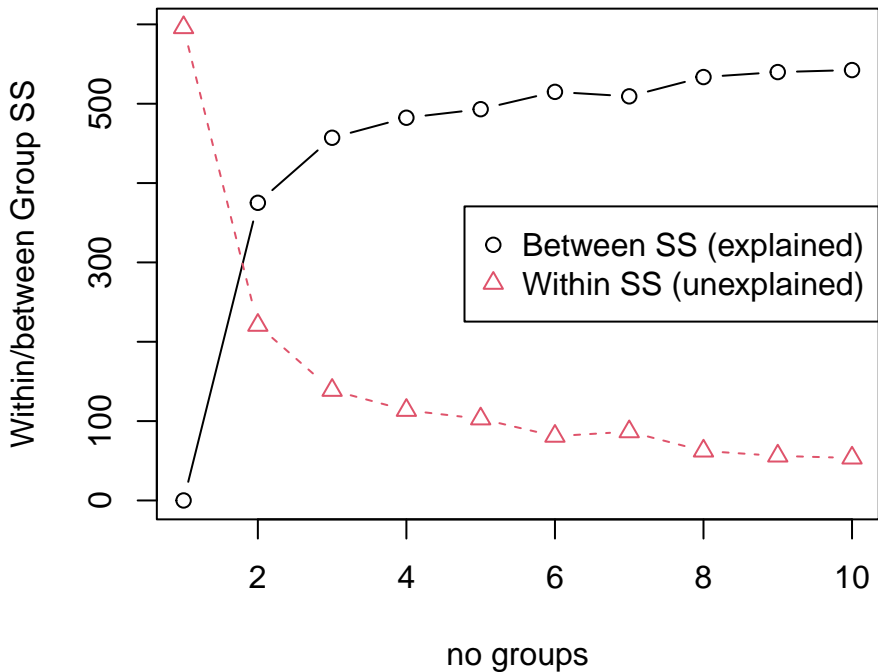
By comparing with the known class assignments we can assess the accuracy of K -means clustering:

```
table(L.iris, kmeans.out$cluster)
```

```
##
## L.iris      1  2  3
## setosa      50  0  0
## versicolor  0 39 11
## virginica   0 14 36
```

For choosing K we run K -means several times and compute within and between cluster variation in dependence of K :

K-Means Iris Data



Thus, $K = 3$ clusters seem appropriate since the the explained variation does not significantly improve (and the unexplained variation does not significantly decrease) with a further increase of the number of clusters.

3.3.7 Arbitrariness of cluster labels and label switching

It is important to realise that in unsupervised learning and clustering the labels of each group are assigned in an arbitrary fashion. Specifically, for K cluster there are $K!$ possibilities to attach the labels, corresponding to the number of permutations of K groups.

Thus, a rerun of a clustering algorithm such as K -means may return the same clustering (groupings of samples) but with different labels. This phenomenon is called “label switching”.

Therefore when comparing clusterings obtained with an algorithm you cannot just rely on the group label, you need to compare the actual members of the clusters. Likewise, if you are interested in the properties of a particular group you cannot rely on the label to identify that group.

In order to resolve the problem of label switching one may wish to relabel the clusters using additional information, such as requiring that some samples are

in specific groups (e.g.: sample 1 is always in group labeled “1”), and/or linking labels to orderings or constraints on the group characteristics (e.g.: the group with label “1” has always a smaller mean than group with label “2”).

3.4 Mixture models

3.4.1 Finite mixture model

- K groups / classes / categories, with the number K specified and finite
- each class $k \in \{1, \dots, K\}$ is modeled by its own distribution F_k with own parameters θ_k .
- density in each class: $f_k(\mathbf{x}) = f(\mathbf{x}|k)$ with $k \in 1, \dots, K$
- mixing weight of each class: $\Pr(k) = \pi_k$ with $\sum_{k=1}^K \pi_k = 1$
- joint density $f(\mathbf{x}, k) = f(\mathbf{x}|k)\Pr(k) = f_k(\mathbf{x})\pi_k$

This results in the mixture density

$$f_{\text{mix}}(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x})$$

This is also called marginal density as it arises from the joint density $f(\mathbf{x}, k)$ by marginalising k .

Very often one uses **multivariate normal components** $f_k(\mathbf{x}) = N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
 \Rightarrow **Gaussian mixture model** (GMM)

Mixture models are fundamental not just in clustering but for many other applications (e.g. classification).

Note: don't confuse *mixture model* with *mixed model* (=random effects regression model)

3.4.2 Total variance and variation of mixture model

The conditional means and variances for each class are $E(\mathbf{x}|k) = \boldsymbol{\mu}_k$ and $\text{Var}(\mathbf{x}|k) = \boldsymbol{\Sigma}_k$, and the probability of class k is given by $\Pr(k) = \pi_k$. Using the law of total expectation we can therefore obtain the mean of the mixture density as follows:

$$\begin{aligned} E(\mathbf{x}) &= E(E(\mathbf{x}|k)) \\ &= \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k \\ &= \boldsymbol{\mu}_0 \end{aligned}$$

Similarly, using the law of total variance we compute the marginal variance:

$$\underbrace{\text{Var}(\mathbf{x})}_{\text{total}} = \underbrace{\text{Var}(\mathbb{E}(\mathbf{x}|k))}_{\text{explained / between-group}} + \underbrace{\mathbb{E}(\text{Var}(\mathbf{x}|k))}_{\text{unexplained / within-group}}$$

$$\Sigma_0 = \sum_{k=1}^K \pi_k (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T + \sum_{k=1}^K \pi_k \Sigma_k$$

Thus, just like in linear regression (see [MATH20802 Statistical Methods](#)) you can decompose the total variance into an explained (between group) part and an unexplained (within group) part.

The total *variation* is given by the trace of the covariance matrix, so the above decomposition turns into

$$\begin{aligned} \text{Tr}(\Sigma_0) &= \sum_{k=1}^K \pi_k \text{Tr}((\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T) + \sum_{k=1}^K \pi_k \text{Tr}(\Sigma_k) \\ &= \sum_{k=1}^K \pi_k (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0) + \sum_{k=1}^K \pi_k \text{Tr}(\Sigma_k) \end{aligned}$$

If the covariances are replaced by their empirical estimates we obtain the $T = B + W$ decomposition of total variation familiar from K -means:

$$T = \text{Tr}(\hat{\Sigma}_0) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0)^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0)$$

$$B = \frac{1}{n} \sum_{k=1}^K n_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_0)^T (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_0)$$

$$W = \frac{1}{n} \sum_{k=1}^K \sum_{i \in G_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)$$

For a univariate mixture ($d = 1$) with $K = 2$ components we get

$$\mu_0 = \pi_1 \mu_1 + \pi_2 \mu_2,$$

$$\sigma_{\text{within}}^2 = \pi_1 \sigma_1^2 + \pi_2 \sigma_2^2 = \sigma_{\text{pooled}}^2,$$

also know as pooled variance, and

$$\begin{aligned} \sigma_{\text{between}}^2 &= \pi_1 (\mu_1 - \mu_0)^2 + \pi_2 (\mu_2 - \mu_0)^2 \\ &= \pi_1 \pi_2^2 (\mu_1 - \mu_2)^2 + \pi_2 \pi_1^2 (\mu_1 - \mu_2)^2 \\ &= \pi_1 \pi_2 (\mu_1 - \mu_2)^2 \\ &= \left(\frac{1}{\pi_1} + \frac{1}{\pi_2} \right)^{-1} (\mu_1 - \mu_2)^2 \end{aligned}$$

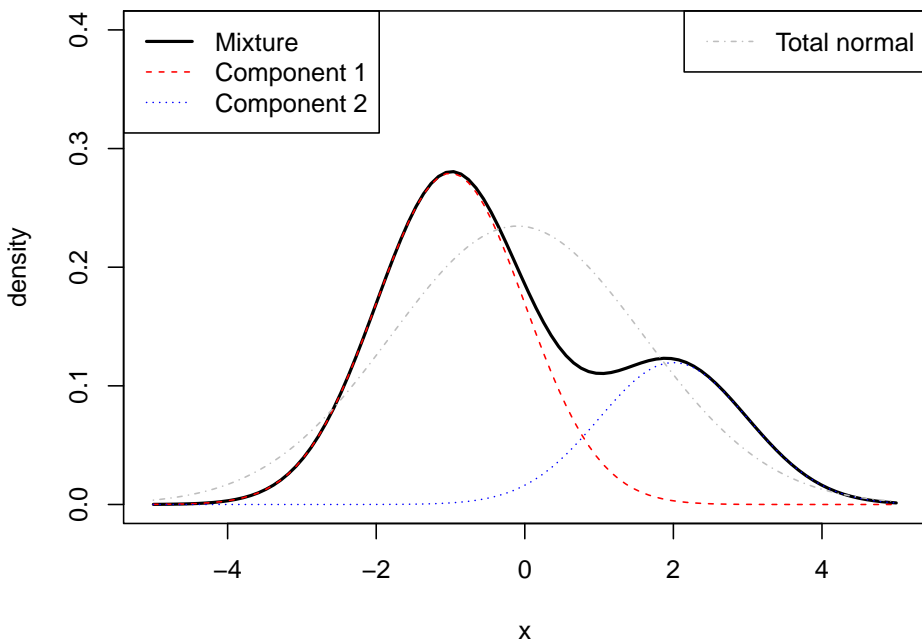
The ratio of the between-group variance and the within-group variance is proportional (by factor of n) to the squared pooled-variance t -score:

$$\frac{\sigma_{\text{between}}^2}{\sigma_{\text{within}}^2} = \frac{(\mu_1 - \mu_2)^2}{\left(\frac{1}{\pi_1} + \frac{1}{\pi_2}\right) \sigma_{\text{pooled}}^2} = \frac{t_{\text{pooled}}^2}{n}$$

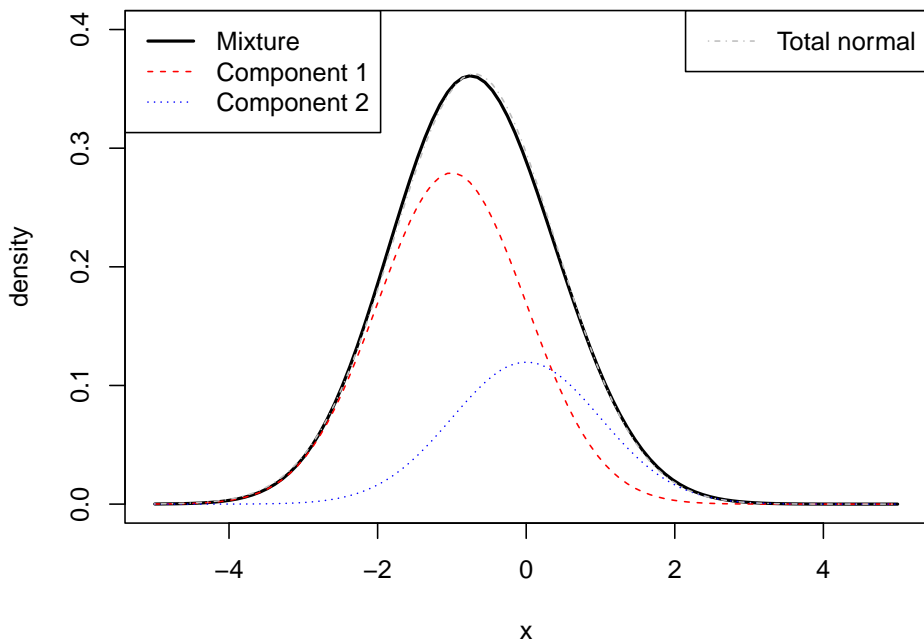
If you are familiar with ANOVA (e.g. linear models course) you will recognise this ratio as the F -score.

3.4.3 Example of mixtures

Mixtures can take on many different shapes and forms, so it is instructive to study a few examples.



In first plot we show the density of a mixture distribution consisting of two normals with $\pi_1 = 0.7$, $\mu_1 = -1$, $\mu_2 = 2$ and the two variances equal to 1 ($\sigma_1^2 = 1$ and $\sigma_2^2 = 1$). Because the two components are well-separated there are two clear modes. The plot also shows the density of a normal distribution with the same total mean ($\mu_0 = -0.1$) and variance ($\sigma_0^2 = 2.89$) as the mixture distribution. Clearly the total normal and the mixture density are very different.



However, mixtures can also look very different. For example, if the mean of the second component is adjusted to $\mu_2 = 0$ then there is only a single mode and the total normal density with $\mu_0 = -0.7$ and $\sigma_0^2 = 1.21$ is now almost indistinguishable in form from the mixture density. Thus, in this case it will be very hard (or even impossible) to identify the two peaks from data.

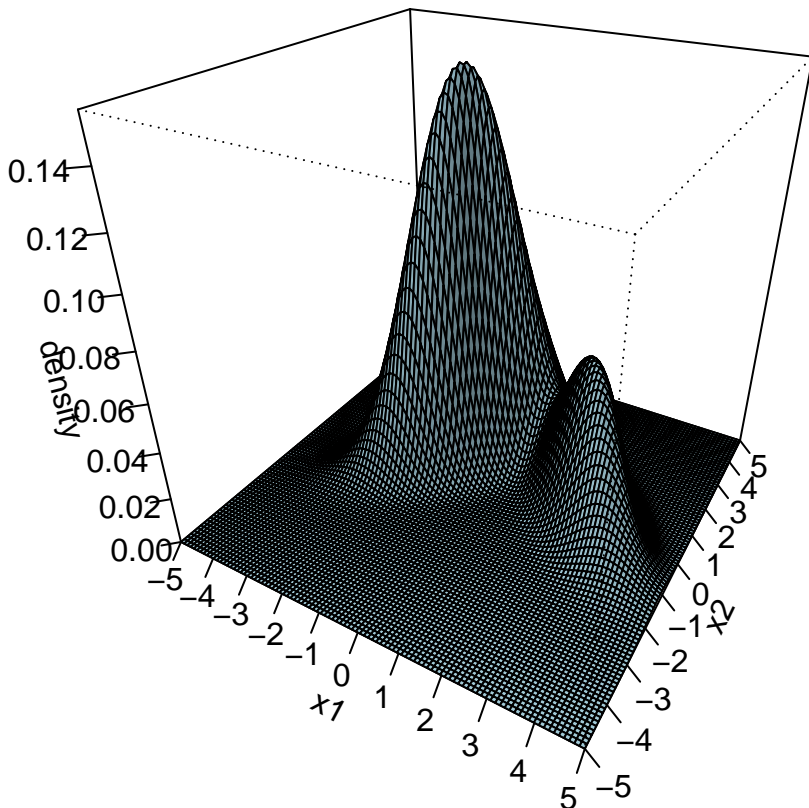
An interactive version of the above two normal component mixture is available online as [R Shiny web app](https://minerva.it.manchester.ac.uk/shiny/strimmer/mixture/) at <https://minerva.it.manchester.ac.uk/shiny/strimmer/mixture/>.

In general, learning mixture models from data can be challenging due to various issues. First, because of permutation symmetries due to the arbitrariness of group labels the group specific parameters are not identifiable without additional restrictions. Second, further identifiability issues can arise if — as in the above example — two neighboring components of the mixture model are largely overlapping and thus are too close to each other to be discriminated as two different modes. Furthermore, likelihood estimation is challenging if there are singularities in the likelihood function, for example due to singular estimated covariance matrices. However, this can be easily fixed by regularising and/or requiring sufficient sample size per group.

Mixture models need not to be univariate, in fact most mixtures we consider in this course are multivariate. For illustration, here is a plot of a mixture of two bivariate normals, with $\pi_1 = 0.7$, $\mu_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 2.5 \\ 0.5 \end{pmatrix}$

and $\Sigma_2 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$:

Mixture of 2 Bivariate Normals



3.4.4 Generative view: sampling from a mixture model

Assuming we know how to sample from the component densities $f_k(\mathbf{x})$ of the mixture model it is straightforward to set up a procedure for sampling from the mixture $f_{\text{mix}}(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x})$.

This is done in a two-step generative process:

1. draw from categorical distribution with parameters $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$:

$$\mathbf{z} \sim \text{Cat}(\boldsymbol{\pi})$$

the vector $\mathbf{z} = (z_1, \dots, z_K)^T$ indicating the group allocation. The group index k is given by $\{k : z_k = 1\}$.

2. Subsequently, sample from the component k selected in step 1:

$$\mathbf{x} \sim F_k$$

This two-stage approach is also called *latent allocation variable formulation* of a mixture model, with \mathbf{z} (or equivalently k) being the latent variable.

The two-step process needs to be repeated for each sample drawn from the mixture (i.e. every time a new latent variable \mathbf{z} is generated).

In probabilistic clustering the aim is to infer the state of \mathbf{z} for all observed samples.

3.4.5 Predicting the group allocation of a given sample

If we know the mixture model and its components we can predict the probability that an observation \mathbf{x} falls in group k via application of Bayes theorem:

$$z_k = \Pr(k|\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{f(\mathbf{x})}$$

In the above mentioned interactive Shiny app for the normal component mixture (available online at <https://minerva.it.manchester.ac.uk/shiny/strimmer/mixture/>) you can also explore the probabilities of each class.

Thus, by calculating the class probabilities using a fitted mixture model we can **perform probabilistic clustering** by assigning each sample to the class with the largest probability. Unlike in algorithmic clustering, we also get an assessment of the uncertainty of the class assignment, since for each sample \mathbf{x} we obtain the vector

$$\mathbf{z} = (z_1, \dots, z_K)^T$$

and thus can see if there are several classes with similar assignment probability. This will be the case, e.g., if \mathbf{x} lies near the boundary between two classes. Note that $\sum_{k=1}^K z_k = 1$.

3.4.6 Variation 1: Infinite mixture model

It is possible to construct mixture models with infinitely many components!

Most commonly known example is Dirichlet process mixture model (DPM):

$$f_{\text{mix}}(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k f_k(\mathbf{x})$$

with $\sum_{k=1}^{\infty} \pi_k = 1$ and where the weight π_k are taken from a infinitely dimensional Dirichlet distribution (=Dirichlet process).

DPMs are useful for clustering since with them it is not necessary to determine the number of clusters a priori (since it by definition has infinitely many!). Instead, the number of clusters is a by-product of the fit of the model to observed data.

Related: “Chinese restaurant process” - https://en.wikipedia.org/wiki/Chinese_restaurant_process

This describes an algorithm for the allocation process of samples (“persons”) to the groups (“restaurant tables”) in a DPM.

See also “stick-breaking process”: https://en.wikipedia.org/wiki/Dirichlet_process#The_stick-breaking_process

3.4.7 Variation 2: Semiparametric mixture model with two classes

A very common model is the following two-component univariate mixture model

$$f_{\text{mix}}(x) = \pi_0 f_0(x) + (1 - \pi_0) f_A(x)$$

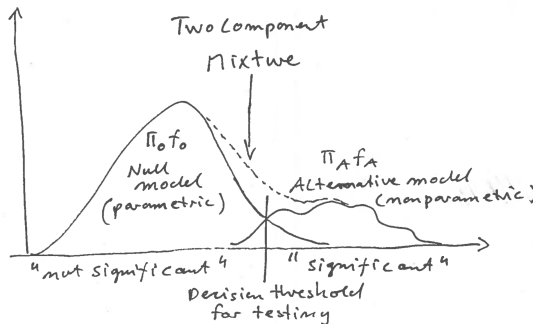
- f_0 : null model, typically parametric such as normal distribution
- f_A : alternative model, typically nonparametric
- π_0 : prior probability of null model

Using Bayes theorem this allows to compute probability that an observation x belongs to the null model:

$$\Pr(\text{Null}|x) = \frac{\pi_0 f_0(x)}{f(x)}$$

This is called the *local false discovery rate*.

The semi-parametric mixture model is the foundation for statistical testing which is based on defining decision thresholds to separate null model (“not significant”) from alternative model (“significant”):



See [MATH20802 Statistical Methods](#) for more details.

3.5 Fitting mixture models to data

3.5.1 Direct estimation of mixture model parameters

Given data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ with observations from n samples we would like to fit the mixture model $f_{\text{mix}}(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x})$ and learn its parameters θ , for example by maximising the corresponding marginal log-likelihood function with regard to θ :

$$\log L(\theta|\mathbf{X}) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k f_k(\mathbf{x}_i) \right)$$

For a Gaussian mixture model the parameters are $\theta = \{\pi, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K\}$.

However, in practise evaluation and optimisation of this likelihood function may be difficult due a number of reasons:

- The form of the log-likelihood function prevents analytic simplifications (note the sum inside the logarithm).
- Because of the symmetries due to exchangeability of cluster labels the likelihood function is multimodal (note this is also linked to the general problem of label switching and non-identifiability of cluster labels in mixtures).
- Furthermore, the likelihood in Gaussian mixture models can become singular if one of the fitted covariance matrices becomes singular. Note this can be adressed by using regularisation (Bayes, penalised ML, etc.).

The above log-likelihood function is also called the *observed data* log-likelihood, or the *incomplete data* log-likelihood, in contrast to the *complete data* log-likelihood described further below.

3.5.2 Estimating mixture model parameters using the EM algorithm

The mixture model may be viewed as an *incomplete* or *missing* data problem: here the missing data are the group allocations $\mathbf{k} = (k_1, \dots, k_n)^T$ belonging to each sample $\mathbf{x}_1, \dots, \mathbf{x}_n$.

If we would know which sample comes from which group the estimation of the parameters θ would indeed be straightforward using the so-called *complete data log-likelihood* based on the joint distribution $f(\mathbf{x}, \mathbf{k}) = \pi_k f_k(\mathbf{x})$

$$\log L(\theta|\mathbf{X}, \mathbf{k}) = \sum_{i=1}^n \log (\pi_{k_i} f_{k_i}(\mathbf{x}_i))$$

The idea of the EM algorithm (Dempster et al. 1977) is to exploit the simplicity of the complete data likelihood and to obtain estimates of θ by first finding the probability distribution z_{ik} of the latent variable k_i , and then using this distribution to compute and optimise the corresponding expected complete-data log-likelihood. Specifically, the z_{ik} contain the *probabilities* of each class for each sample i and thus provide a *soft assignment* of classes rather than a 0/1 *hard assignment* (as in the K -means algorithm or in the generative latent variable view of mixture models).

In the EM algorithm we iterate between the

- 1) estimation the probabilistic distribution z_{ik} for the group allocation latent parameters using the current estimate of the parameters θ (obtained in step 2)
- 2) maximisation of the expected complete data log-likelihood to estimate the parameters θ . The expectation is taken with regard to the distribution z_{ik} (obtained in step 1).

Specifically, the EM algorithm applied to model-based clustering proceeds as follows:

- 1) **Initialisation:** Start with a guess of the parameters $\theta^{(1)}$, then continue with “E” Step, Part A. Alternatively, start with a guess of $z_{ik}^{(1)}$, then continue with “E” Step, Part B. The initialisation may be derived from some prior information, e.g., from running K -means, or simply be at random.
- 2) **E “expectation” step** — Part A: Use Bayes’ theorem to compute new probabilities of allocation for all the samples x_i :

$$z_{ik}^{(b+1)} \leftarrow \frac{\pi_k f_k(x_i)}{f(x_i)}$$

Note that to obtain $z_{ik}^{(b+1)}$ the current value $\theta^{(b)}$ of the parameters is required.

— Part B: Construct the expected complete data log-likelihood function using the weights $z_{ik}^{(b+1)}$:

$$Q^{(b+1)}(\theta|X) = \sum_{i=1}^n \sum_{k=1}^K z_{ik}^{(b+1)} \log(\pi_k f_k(x_i))$$

- 3) **M “maximisation” step** — Maximise the expected complete data log-likelihood to update the mixture model parameters θ :

$$\theta^{(b+1)} \leftarrow \arg \max_{\theta} Q^{(b+1)}(\theta|X)$$

- 4) Repeat with “E” Step until convergence of parameters $\theta^{(b)}$ of the mixture model.

It can be shown that the output $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \dots$ of the EM algorithm converges to the estimate $\hat{\theta}$ found when maximising the marginal log-likelihood. Since maximisation of the expected complete data log-likelihood is often much easier (and analytically tractable) than maximisation of the observed data log-likelihood function the EM algorithm is the preferred approach in this case.

To avoid singularities in the expected log-likelihood function we may wish to adopt a Bayesian approach (or use regularised/penalised ML) for estimating the parameters in the M-step.

3.5.3 EM algorithm for multivariate normal mixture model

For a GMM the EM algorithm can be written down analytically:

E-step:

$$z_{ik} = \frac{\hat{\pi}_k N(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{f(\mathbf{x}_i)}$$

M-step:

$$\hat{n}_k = \sum_{i=1}^n z_{ik}$$

$$\hat{\pi}_k = \frac{\hat{n}_k}{n}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\hat{n}_k} \sum_{i=1}^n z_{ik} \mathbf{x}_i$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{\hat{n}_k} \sum_{i=1}^n z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

Note that the estimators $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ are weighted versions of the usual empirical estimators (with weights z_{ik} being the soft assignment of classes resulting from the Bayesian updating).

In Worksheet 7 you can find a simple R implementation of the EM algorithm for univariate normal mixtures.

3.5.4 Connection with K -means clustering method

The K -means algorithm is very closely related to probabilistic clustering with a Gaussian mixture models.

Specifically, the class assignment in K -means is

$$C(\mathbf{x}_i) = \arg \min_k (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)$$

If in a Gaussian mixture model the probabilities π_k of all classes are assumed to be identical (i.e. $\pi_k = \frac{1}{K}$) and the covariances $\boldsymbol{\Sigma}_k$ are all of the same spherical form $\sigma^2 \mathbf{I}$, i.e. no dependence on groups, no correlation and identical variance for all variables, then the soft assignment for the class allocation becomes

$$\log(z_{ik}) = -\frac{1}{2\sigma^2} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k) + C$$

where C is a constant depending on \mathbf{x}_i but not on k . In order to select a hard class allocation based on z_{ik} we may then use the rule

$$\begin{aligned} C(\mathbf{x}_i) &= \arg \max_k \log(z_{ik}) \\ &= \arg \min_k (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k) \end{aligned}$$

Thus, K -means can be viewed as an algorithm to provide hard classifications for a simple restricted Gaussian mixture model.

3.5.5 Choosing the number of classes

Since GMMs operate in a likelihood framework we can use penalised likelihood model selection criteria to choose among different models (i.e. GMMs with different numbers of classes).

The most popular choices are AIC (Akaike Information Criterion) and BIC (Bayesian Information criterion) defined as follows:

$$\text{AIC} = -2 \log L + 2K$$

$$\text{BIC} = -2 \log L + K \log(n)$$

Instead of maximising the log-likelihood we minimise AIC and BIC.

Note that in both criteria more complex models with more parameters (in this case groups) are penalised over simpler models in order to prevent overfitting.

\implies find optimal number of groups K .

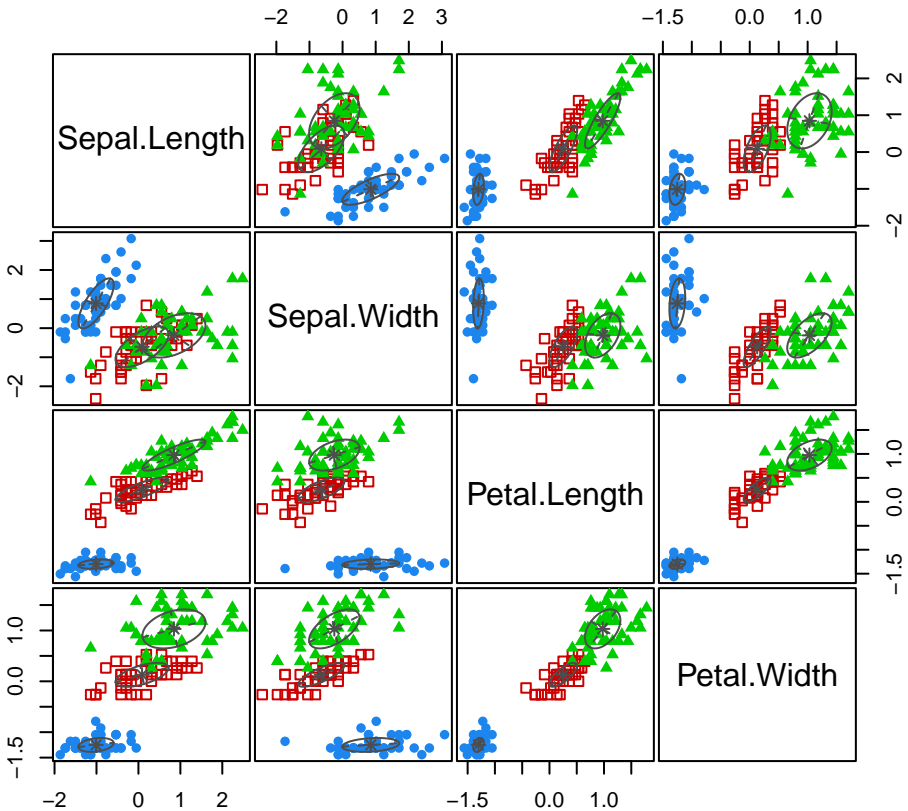
Another way of choosing optimal numbers of clusters is by cross-validation (see later chapter on supervised learning).

3.5.6 Application of GMMs to Iris flower data

We now explore the application of Gaussian mixture models to the Iris flower data set we also investigated with PCA and K-means.

First, we fit a GMM with 3 clusters, using the R software “mclust.”²

```
library("mclust")
gmm3 = Mclust(X.iris, G=3, verbose=FALSE)
plot(gmm3, what="classification")
```



The “mclust” software has used the following model when fitting the mixture:

```
gmm3$modelName
```

```
## [1] "VVV"
```

Here “VVV” is the name used by the “mclust” software for a model allowing for an individual unrestricted covariance matrix Σ_k for each class k .

²L. Scrucca L. et. al. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 8:205–233. See <https://journal.r-project.org/archive/2016/RJ-2016-021/> and <https://mclust-org.github.io/mclust/>

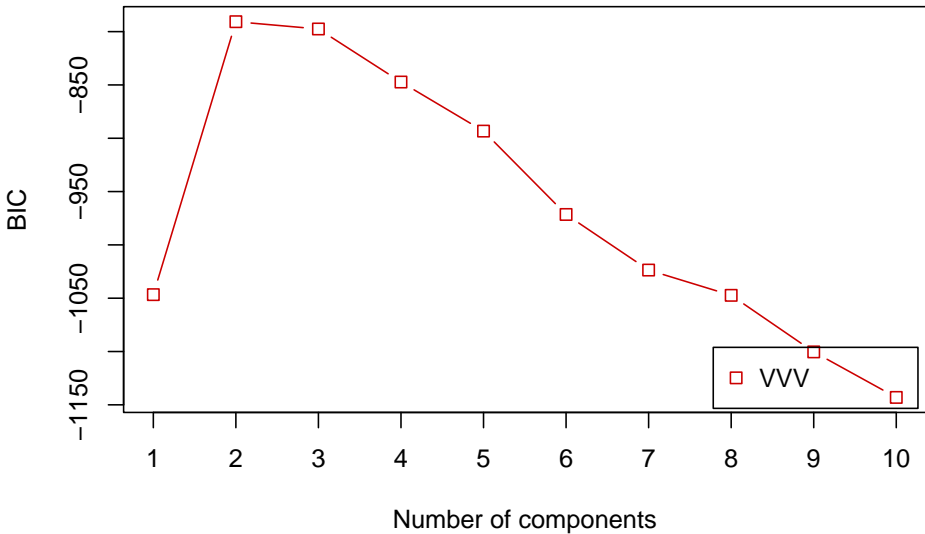
This GMM has a substantially lower misclassification error compared to K -means with the same number of clusters:

```
table(gmm3$classification, L.iris)
```

```
##      L.iris
##      setosa versicolor virginica
##  1      50          0          0
##  2       0         45          0
##  3       0          5         50
```

Note that in “mclust” the BIC criterion is defined with the opposite sign ($\text{BIC}_{\text{mclust}} = 2 \log L - K \log(n)$), thus we need to find the *maximum* value rather than the smallest value.

If we compute BIC for various numbers of groups we find that the model with the best $\text{BIC}_{\text{mclust}}$ is a model with 2 clusters but the model with 3 cluster has nearly as good a BIC:



Chapter 4

Supervised learning and classification

4.1 Introduction

4.1.1 Supervised learning vs. unsupervised learning

Unsupervised learning:

Starting point:

- unlabeled data x_1, \dots, x_n .

Aim: find labels y_1, \dots, y_n to attach to each sample x_i .

For discrete labels y unsupervised learning is called *clustering*.

Supervised learning:

Starting point:

- labeled *training data*: $\{x_1^{\text{train}}, y_1^{\text{train}}\}, \dots, \{x_n^{\text{train}}, y_n^{\text{train}}\}$
- In addition, we have unlabeled *test data*: x^{test}

Aim: use training data to learn a function, say $h(x)$, to predict the label corresponding to the test data. The predictor function may provide a soft (probabilistic) assignment or a hard assignment of a class label to a test sample.

For y discrete supervised learning is called *classification*. For continuous y the label is called response and supervised learning becomes *regression*.

Thus, supervised learning is a two-step procedure:

- 1) Learn predictor function $h(x)$ using the training data x_i^{train} plus labels y_i^{train} .

- 2) Predict the label y^{test} for the test data x^{test} using the estimated classifier function: $\hat{y}^{\text{test}} = \hat{h}(x^{\text{test}})$.

4.1.2 Terminology

The function $h(x)$ that predicts the class y is called a *classifier*.

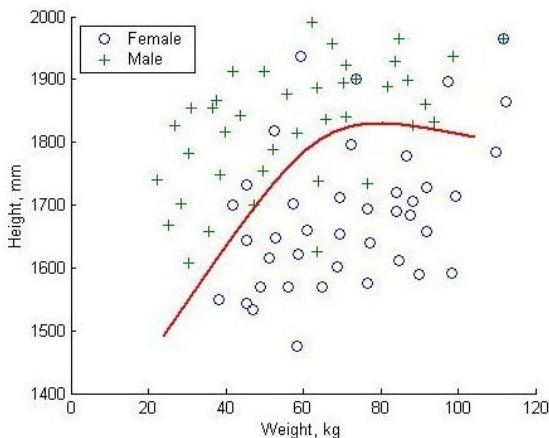
There are many types of classifiers, we focus here primarily on probabilistic classifiers (i.e. those that output probabilities for each possible class/label).

The challenge is to find a classifier that

- explains the current training data well *and*
- that also generalises well to future unseen data.

Note that it is relatively easy to find a predictor that explains the training data but especially in high dimensions (i.e. with many predictors) there is often overfitting and then the predictor does not generalise well!

The *decision boundary* between the classes is defined as the set of all x for which the class assignment by the predictor $h(x)$ switches from one class to another.



In general, simple decision boundaries are preferred over complex decision boundaries to avoid overfitting.

Some commonly used probabilistic methods for classifications:

- QDA (quadratic discriminant analysis)
- LDA (linear discriminant analysis)
- DDA (diagonal discriminant analysis),
- Naive Bayes classification
- logistic regression

Common non-probabilistic methods include:

- SVM (support vector machine),
- random forest
- neural networks

Depending on how the classifiers are trained there are many variations of the above methods, e.g. Fisher discriminant analysis, regularised LDA, shrinkage discriminant analysis etc.

4.2 Bayesian discriminant rule or Bayes classifier

Same setup as with mixture models:

- K groups with K prespecified
- each group has its own distribution F_k with own parameters θ_k
- the density of each class is $f_k(\mathbf{x}) = f(\mathbf{x}|k)$.
- prior probability of group k is $\Pr(k) = \pi_k$ with $\sum_{k=1}^K \pi_k = 1$
- marginal density is the mixture $f(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x})$

The posterior probability of group k is then

$$\Pr(k|\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{f(\mathbf{x})}$$

This already provides a “soft” classification

$$\mathbf{h}(\mathbf{x}^{\text{test}}) = (\Pr(k = 1|\mathbf{x}^{\text{test}}), \dots, \Pr(k = K|\mathbf{x}^{\text{test}}))^T$$

where each possible class $k \in \{1, \dots, K\}$ is assigned a probability to be the label for the test sample \mathbf{x} .

The *discriminant function* is the logarithm of the posterior probability:

$$d_k(\mathbf{x}) = \log \Pr(k|\mathbf{x}) = \log \pi_k + \log f_k(\mathbf{x}) - \log f(\mathbf{x})$$

Since we use d_k to compare the different classes k we can simplify the discriminant function by dropping all constant terms that do not depend on k — in the above this is the term $\log f(\mathbf{x})$. Hence we get for the Bayes discriminant function

$$d_k(\mathbf{x}) = \log \pi_k + \log f_k(\mathbf{x}).$$

For subsequent “hard” classification $h(\mathbf{x}^{\text{test}})$ we then select the group/label for which the value of the discriminant function is maximised:

$$\hat{y}^{\text{test}} = h(\mathbf{x}^{\text{test}}) = \arg \max_k d_k(\mathbf{x}^{\text{test}}).$$

The discriminant functions $d_k(\mathbf{x})$ can be mapped back to the probabilistic class assignment by using the softargmax function (also known as softmax function):

$$\Pr(k|\mathbf{x}) = \frac{\exp(d_k(\mathbf{x}))}{\sum_{c=1}^K \exp(d_c(\mathbf{x}))} = \frac{\exp(d_k(\mathbf{x}) - d_{\max})}{\sum_{c=1}^K \exp(d_c(\mathbf{x}) - d_{\max})}.$$

Note that subtracting $d_{\max} = \max\{d_1(x), \dots, d_K(x)\}$ avoids numerical overflow problems when computing the exponential by standardising the maximum of the discriminant functions to zero.

We have already encountered the Bayes classifier in the EM algorithm to predict the state of the latent variables (soft assignment) and in the K -means algorithm (hard assignment), see previous Chapter and also Worksheet 7.

4.3 Normal Bayes classifier

4.3.1 Quadratic discriminant analysis (QDA) and Gaussian assumption

Quadratic discriminant analysis (QDA) is a special case of the Bayes classifier when all densities are multivariate normal with $f_k(x) = N(x|\mu_k, \Sigma_k)$.

This leads to the discriminant function for QDA:

$$d_k^{QDA}(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \log \det(\Sigma_k) + \log(\pi_k)$$

There are a number of noteworthy things here:

- Again terms are dropped that do not depend on k , such as $-\frac{d}{2} \log(2\pi)$.
- Note the appearance of the Mahalanobis distance between x and μ_k in the last term — recall $d^{Mahalanobis}(x, \mu|\Sigma) = (x - \mu)^T \Sigma^{-1}(x - \mu)$.
- The **QDA discriminant function is quadratic in x** - hence its name!
This implies that the **decision boundaries for QDA classification are quadratic** (i.e. parabolas in two dimensional settings).

For Gaussian models specifically it can be useful to multiply the discriminant function by -2 to get rid of the factor $-\frac{1}{2}$, but note that in that case we then need to find the minimum of the discriminant function rather than the maximum:

$$d_k^{QDA(v2)}(x) = (x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \det(\Sigma_k) - 2 \log(\pi_k)$$

In the literature you will find both versions of Gaussian discriminant functions so you need to check carefully which convention is used. In the following we will use the first version only.

Decision boundaries for the QDA classifier can be either linear or nonlinear (quadratic curve). The decision boundary between any two classes i and j require that $d_i^{QDA}(x) = d_j^{QDA}(x)$, or equivalently $d_i^{QDA}(x) - d_j^{QDA}(x) = 0$, which is a quadratic equation.

4.3.2 Linear discriminant analysis (LDA)

LDA is a special case of QDA, with the assumption of common overall covariance across all groups: $\Sigma_k = \Sigma$.

This leads to a simplified discriminant function:

$$d_k^{LDA}(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) + \log(\pi_k)$$

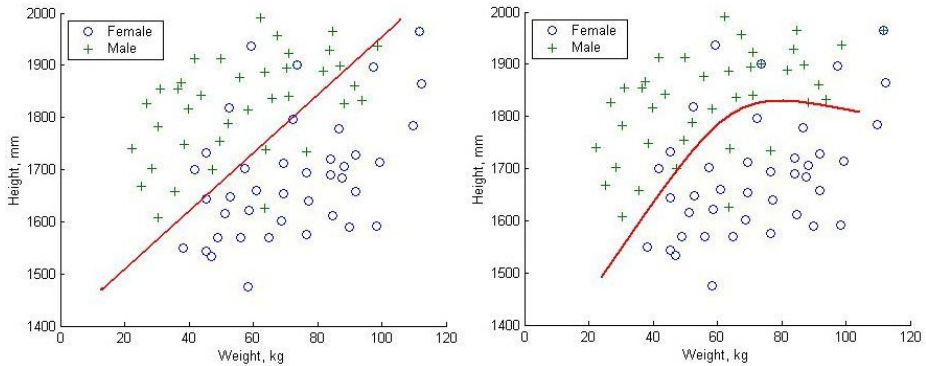
Note that term containing the log-determinant is now gone, and that LDA is essentially now a method that tries to minimize the Mahalanobis distance (while taking also into account the prior class probabilities).

The above function can be further simplified, by noting that the quadratic term $x^T \Sigma^{-1} x$ does not depend on k and hence can be dropped:

$$\begin{aligned} d_k^{LDA}(x) &= \mu_k^T \Sigma^{-1} x - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \\ &= b^T x + a \end{aligned}$$

Thus, the **LDA discriminant function is linear in x , and hence the resulting decision boundaries are linear** as well (i.e. straight lines in two-dimensional settings).

Comparison of linear decision boundaries of LDA (left) compared with QDA (right):



Note that logistic regression (cf. GLM module) takes on exactly the above linear form and is indeed closely linked with the LDA classifier.

4.3.3 Diagonal discriminant analysis (DDA)

For DDA we start with the same setting as for LDA, but now we simplify the model even further by additionally requiring a **diagonal covariance** containing only the variances (thus we assume that all correlations among the predictors x_1, \dots, x_d are zero):

$$\Sigma = V = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{pmatrix}$$

This simplifies the inversion of Σ as

$$\Sigma^{-1} = V^{-1} = \begin{pmatrix} \sigma_1^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d^{-2} \end{pmatrix}$$

and leads to the discriminant function

$$\begin{aligned} d_k^{DDA}(\mathbf{x}) &= \boldsymbol{\mu}_k^T V^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^T V^{-1} \boldsymbol{\mu}_k + \log(\pi_k) \\ &= \sum_{j=1}^d \frac{\mu_{k,j} x_j - \mu_{k,j}^2 / 2}{\sigma_d^2} + \log(\pi_k) \end{aligned}$$

As special case of LDA, the **DDA classifier is a linear classifier** and thus has linear decision boundaries.

The **Bayes classifier** (using any distribution) **assuming uncorrelated predictors** is also known as the **naive Bayes classifier**.

Hence, **DDA is a naive Bayes classifier** assuming underlying Gaussian distributions.

However, don't let you misguide because of the name "naive": in fact DDA and other "naive" Bayes classifier are often very effective classifiers, especially in high-dimensional settings!

4.4 The training step — learning QDA, LDA and DDA classifiers from data

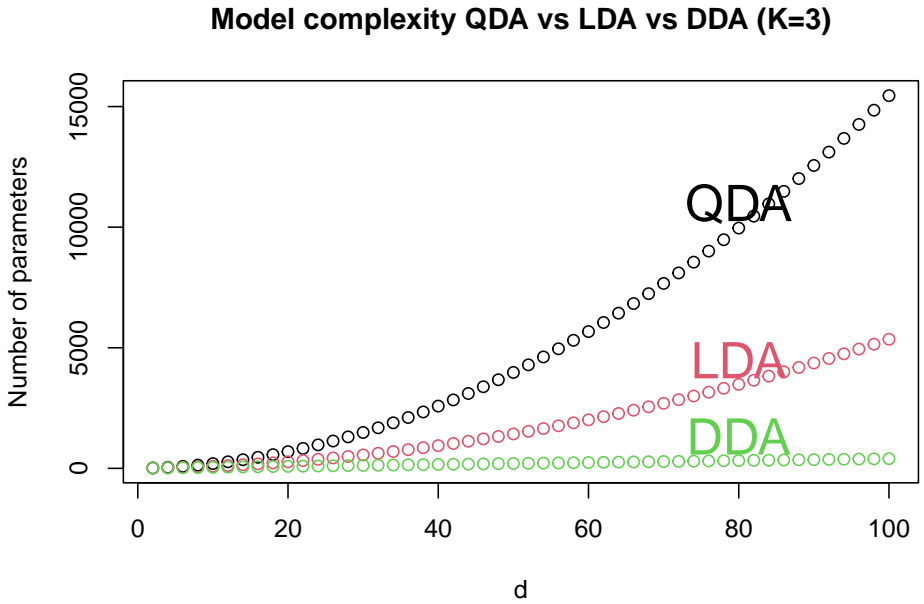
4.4.1 Number of model parameters

In order to predict the class for new data using any of the above discriminant functions we need to first learn the underlying parameters from the training data $\mathbf{x}_i^{\text{train}}$ and y_i^{train} :

- For QDA, LDA and DDA we need to learn π_1, \dots, π_K with $\sum_{k=1}^K \pi_k = 1$ and the mean vectors $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$
- For QDA we additionally require $\Sigma_1, \dots, \Sigma_K$
- For LDA we need Σ
- For DDA we estimate $\sigma_1^2, \dots, \sigma_d^2$.

Overall, the total number of parameters to be estimated when learning the discriminant functions from training data is as follows:

- QDA: $K - 1 + Kd + K \frac{d(d+1)}{2}$
- LDA: $K - 1 + Kd + \frac{d(d+1)}{2}$
- DDA: $K - 1 + Kd + d$



4.4.2 Estimating the discriminant / predictor function

For QDA, LDA and DDA we learn the predictor by estimating the parameters of the discriminant function from the training data.

4.4.2.1 Large sample size

If the sample size of the training data set is sufficiently large compared to the model dimensions we can use maximum likelihood (ML) to estimate the model parameters. To be able use ML we need a larger sample size for QDA and LDA (because full covariances need to be estimated) but for DDA relatively small sample size can be sufficient (which explains why “naive” Bayes methods are very popular in practise).

To obtain the parameters estimates we use the known labels y_i^{train} to sort the samples x_i^{train} into the corresponding classes, and then apply the standard ML estimators. Let $g_k = \{i : y_i^{\text{train}} = k\}$ be the set of all indices of training sample belonging to group k , n_k the sample size in group k

The ML estimates of the class probabilities are the frequencies

$$\hat{\pi}_k = \frac{n_k}{n}$$

and the ML estimate of the group means $k = 1, \dots, K$ are

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in g_k} x_i^{\text{train}}.$$

The ML estimate of the global mean μ_0 (i.e. if we assume there is only a single class and ignore the group labels) is

$$\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^n x_i^{\text{train}} = \sum_{k=1}^K \hat{\pi}_k \hat{\mu}_k$$

Note the global mean is identical to the pooled mean (i.e. weighted average of the individual group means).

The ML estimates for the covariances Σ_k for QDA are

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i \in g_k} (x_i^{\text{train}} - \hat{\mu}_k)(x_i^{\text{train}} - \hat{\mu}_k)^T$$

In order to get the ML estimate of the pooled variance Σ for use with LDA we compute

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in g_k} (x_i^{\text{train}} - \hat{\mu}_k)(x_i^{\text{train}} - \hat{\mu}_k)^T = \sum_{k=1}^K \hat{\pi}_k \hat{\Sigma}_k$$

Note that the pooled variance Σ differs (substantially!) from the global variance Σ_0 that results from simply ignoring class labels and that is computed as

$$\hat{\Sigma}_0^{ML} = \frac{1}{n} \sum_{i=1}^n (x_i^{\text{train}} - \hat{\mu}_0)(x_i^{\text{train}} - \hat{\mu}_0)^T$$

You will recognise the above from the variance decomposition in mixture models, with Σ_0 being the total variance and the pooled Σ the unexplained/with-in group variance.

4.4.2.2 Small sample size

If the dimension d is large compared to the sample size then the number of parameters in the predictor function grows fast. Especially QDA but also LDA is data hungry and ML estimation becomes an ill-posed problem.

As discussed in Section 1.5 in this instance we need to use a regularised estimator for the covariance(s) such as estimators derived in the framework of penalised ML, Bayesian learning, shrinkage estimation etc. This also ensures that the estimated covariance matrices are positive definite (which is automatically guaranteed only for DDA if all variances are positive).

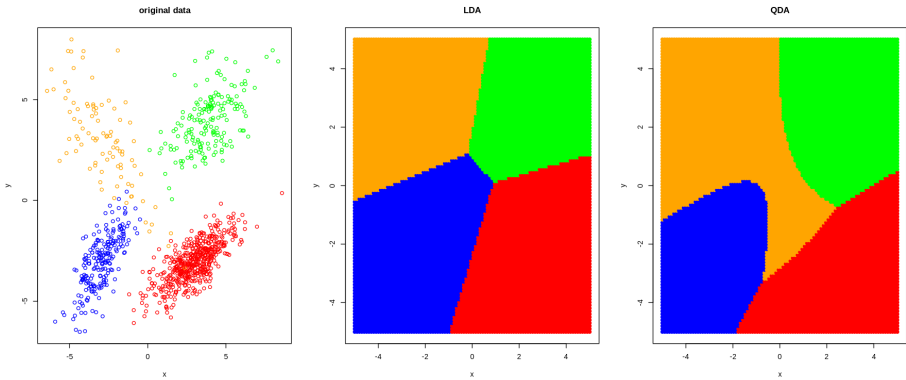
Furthermore, in small sample setting it is advised to reduce the number of parameters of the model. Thus using LDA or DDA is preferred over QDA. This can also prevent overfitting and lead to a predictor that generalises better.

To analyse high-dimensional data in the worksheets we will employ a regularised version of LDA and DDA using Stein-type shrinkage estimation as discussed in Section 1.5 and implemented in the R package “sda”.

4.4.3 Comparison of estimated decision boundaries: LDA vs. QDA

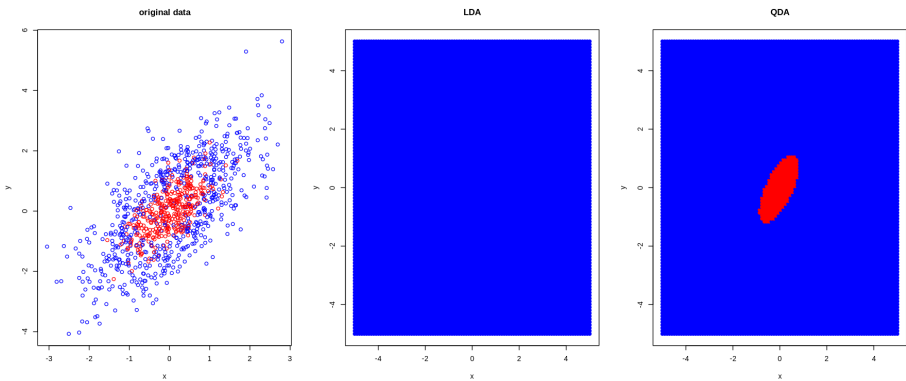
We compare two simple scenarios using simulated data.

Non-nested case ($K = 4$):



Both LDA and QDA clearly separate the 4 classes. Note the curved decision boundaries for QDA and the linear decision boundaries for LDA.

Nested case ($K = 2$):



In the nested case LDA fails to separate the two classes because there is no way to separate two nested classes with a simple linear boundary.

4.5 Goodness of fit and variable ranking

As in linear regression (cf. “Statistical Methods” module) we are interested in finding out whether the fitted mixture model is an appropriate model, and which

particular predictor(s) x_j from $\mathbf{x} = (x_1, \dots, x_d)^T$ are responsible prediction the outcome, i.e. for categorizing a sample into group k .

In order to study these problem it is helpful to rewrite the discriminant function to highlight the influence (or importance) of each predictor.

We focus on linear methods (LDA and DDA) and first look at the simple case $K = 2$ and then generalise to more than two groups.

4.5.1 LDA with $K = 2$ classes

For two classes using the LDA discriminant rule will choose group $k = 1$ if $d_1^{LDA}(\mathbf{x}) > d_2^{LDA}(\mathbf{x})$, or equivalently, if

$$\Delta_{12}^{LDA} = d_1^{LDA}(\mathbf{x}) - d_2^{LDA}(\mathbf{x}) > 0$$

Since $d_k(\mathbf{x})$ is the log-posterior (plus/minus identical constants) Δ_{12}^{LDA} is in fact the **log-posterior odds of class 1 versus class 2** (see Statistical Methods, Bayesian inference).

The difference Δ_{12}^{LDA} is

$$\underbrace{\Delta_{12}^{LDA}}_{\text{log posterior odds}} = \underbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right)}_{\text{log Bayes factor } \log B_{12}} + \underbrace{\log \left(\frac{\pi_1}{\pi_2} \right)}_{\text{log prior odds}}$$

Note that since we only consider simple non-composite models here the log-Bayes factor is identical with the log-likelihood ratio!

The log Bayes factor $\log B_{12}$ is known as the *weight of evidence* in favour of F_1 given \mathbf{x} . The *expected weight of evidence* assuming \mathbf{x} is indeed from F_1 is the Kullback-Leibler discrimination information in favour of group 1, i.e. the KL divergence of from distribution F_2 to F_1 :

$$E_{F_1}(\log B_{12}) = D_{KL}(F_1, F_2) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}\Omega^2$$

This yields, apart of a scale factor, a population version of the Hotelling T^2 statistic defined as

$$T^2 = c^2(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$$

where $c = (\frac{1}{n_1} + \frac{1}{n_2})^{-1/2} = \sqrt{n\pi_1\pi_2}$ is a sample size dependent factor (for $SD(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$). T^2 is a measure of fit of the underlying two-component mixture.

Using the whitening transformation with $\mathbf{z} = \mathbf{W}\mathbf{x}$ and $\mathbf{W}^T\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ we can rewrite the log Bayes factor as

$$\begin{aligned} \log B_{12} &= \left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{W}^T \right) \left(\mathbf{W} \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right) \right) \\ &= \boldsymbol{\omega}^T \boldsymbol{\delta}(\mathbf{x}) \end{aligned}$$

i.e. as the product of two vectors:

- $\delta(x)$ is the whitened x (centered around average means) and
- $\omega = (\omega_1, \dots, \omega_d)^T = W(\mu_1 - \mu_2)$ gives the weight of each whitened component $\delta(x)$ in the log Bayes factor.

A large positive or negative value of ω_j indicates that the corresponding whitened predictor is relevant for choosing a class, whereas small values of ω_j close to zero indicate that the corresponding ZCA whitened predictor is unimportant. Furthermore, $\omega^T \omega = \sum_{j=1}^d \omega_j^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) = \Omega^2$, i.e. the squared ω_j^2 provide a component-wise decomposition of the overall fit Ω^2 .

Choosing ZCA-cor as whitening transformation with $W = P^{-1/2}V^{-1/2}$ we get

$$\omega^{ZCA-cor} = P^{-1/2}V^{-1/2}(\mu_1 - \mu_2)$$

A better understanding of $\omega^{ZCA-cor}$ is provided by comparing with the two-sample t -statistic

$$\hat{\tau} = c\hat{V}^{-1/2}(\hat{\mu}_1 - \hat{\mu}_2)$$

With τ the population version of $\hat{\tau}$ we can define

$$\tau^{adj} = P^{-1/2}\tau = c\omega^{ZCA-cor}$$

as correlation-adjusted t -scores (cat scores). With $(\hat{\tau}^{adj})^T \hat{\tau}^{adj} = T^2$ we can see that the cat scores offer a component-wise decomposition of Hotelling's T^2 .

Note the choice of ZCA-cor whitening is to ensure that the whitened components are interpretable and stay maximally correlated to the original variables. However, you may also choose for example PCA whitening in which case the $\omega^T \omega$ provide the variable importance for the PCA whitened variables.

For DDA, which assumes that correlations among predictors vanish, i.e. $P = I_d$, we get

$$\Delta_{12}^{DDA} = \underbrace{\left((\mu_1 - \mu_2)^T V^{-1/2} \right)}_{c^{-1}\tau^T} \underbrace{\left(V^{-1/2} \left(x - \frac{\mu_1 + \mu_2}{2} \right) \right)}_{\text{centered standardised predictor}} + \log \left(\frac{\pi_1}{\pi_2} \right)$$

Similarly as above, the t -score τ determines the impact of the standardised predictor in Δ^{DDA} .

Consequently, in DDA we can rank predictors by the squared t -score. Recall that in standard linear regression with uncorrelated predictors we can find the most important predictors by ranking the squared marginal correlations – ranking by (squared) t -scores in DDA is the exact analogy but for discrete response.

4.5.2 Multiple classes

For more than two classes we need to refer to the so-called **pooled centroids formulation** of DDA and LDA (introduced by Tibshirani 2002).

The pooled centroid is given by $\mu_0 = \sum_{k=1}^K \pi_k \mu_k$ — this is the centroid if there would be only a single class. The corresponding probability (for a single class) is $\pi_0 = 1$ and the distribution is called F_0 .

The LDA discriminant function for this “group 0” is

$$d_0^{LDA}(x) = \mu_0^T \Sigma^{-1} x - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0$$

and the log posterior odds for comparison of group k with the pooled group 0 is

$$\begin{aligned} \Delta_k^{LDA} &= d_k^{LDA}(x) - d_0^{LDA}(x) \\ &= \log B_{k0} + \log(\pi_k) \\ &= \omega_k^T \delta_k(x) + \log(\pi_k) \end{aligned}$$

with

$$\omega_k = W(\mu_k - \mu_0)$$

and

$$\delta_k(x) = W(x - \frac{\mu_k + \mu_0}{2})$$

The expected log Bayes factor is

$$E_{F_k}(\log B_{k0}) = KL(F_k || F_0) = \frac{1}{2}(\mu_k - \mu_0)^T \Sigma^{-1}(\mu_k - \mu_0) = \frac{1}{2} \Omega_k^2$$

With scale factor $c_k = (\frac{1}{n_k} - \frac{1}{n})^{-1/2} = \sqrt{n \frac{\pi_k}{1 - \pi_k}}$ (for $SD(\hat{\mu}_k - \hat{\mu}_0)$, with the minus sign before $\frac{1}{n}$ due to correlation between $\hat{\mu}_k$ and pooled mean $\hat{\mu}_0$) we get as correlation-adjusted t -score for comparing mean of group k with the pooled mean

$$\tau_k^{adj} = c_k \omega_k^{ZCA-cor}.$$

For the two class case ($K = 2$) we get with $\mu_0 = \pi_1 \mu_1 + \pi_2 \mu_2$ for the mean difference $(\mu_1 - \mu_0) = \pi_2(\mu_1 - \mu_2)$ and with $c_1 = \sqrt{n \frac{\pi_1}{\pi_2}}$ this yields

$$\tau_1^{adj} = \sqrt{n \pi_1 \pi_2} P^{-1/2} V^{-1/2} (\mu_1 - \mu_2),$$

i.e. the exact same score as in the two-class setting.

4.6 Variable selection and cross-validation

In the previous we saw that in DDA the natural score for **ranking features** with regard to their relevance in separating the classes is the (squared) t -score, and for LDA a whitened version such as the squared correlation-adjusted t -score (based on ZCA-cor whitening) may be used. Once such a ranking has been established the question of a **suitable cutoff** arises, i.e. how many features need (or should) be retained in a model.

For large and high-dimensional models **feature selection can also be viewed as a form of regularisation and also dimension reduction**. Specifically, there may be many variables/ features that do not contribute to the class prediction. Despite having in principle no effect on the outcome the presence of these “null variables” can nonetheless deteriorate (sometimes dramatically!) the overall predictive accuracy of a trained predictor, because they add noise and increase the model dimension. Therefore, variables that do not contribute to prediction should be filtered out in order to be able to construct good prediction models and classifiers.

4.6.1 Choosing a threshold by multiple testing using false discovery rates

The most simple way to determine a cutoff threshold is to use a standard technique for multiple testing.

For each predictor variable x_1, \dots, x_d we have a corresponding test statistic measuring the influence of this variable on the response, for example the t -scores and related statistics discussed in the previous section. In addition to providing an overall ranking the set of all these statistics can be used to determine a suitable cutoff by trying to separate two populations of predictor variables:

- “Null” variables that do not contribute to prediction
- “Alternative” variables that are linked to prediction

As discussed in the “Statistical Methods” module last term (Part 2 - Section 8) this can be done as follows:

- The distribution of the observed test statistics z_i is assumed to follow a two-component mixture where $F_0(z)$ and $F_A(z)$ are the distributions corresponding to the null and the alternative, $f_0(z)$ and $f_a(z)$ the densities, and π_0 and $\pi_A = 1 - \pi_0$ are the weights:

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_a(z)$$

- The null model is typically from a parametric family (e.g. normal around zero and with a free variance parameter) whereas the alternative is often modelled nonparametrically.

- After fitting the mixture model, often assuming some additional constraints to make the mixture identifiable, one can compute false discovery rates (FDR) as follows:

Local FDR:

$$\widehat{fdr}(z_i) = \hat{\Pr}(\text{null}|z_i) = \frac{\hat{\pi}_0 \hat{f}_0(z_i)}{\hat{f}(z_i)}$$

Tail-area-based FDR (=q-value):

$$\widehat{Fdr}(z_i) = \hat{\Pr}(\text{null}|Z > z_i) = \frac{\hat{\pi}_0 \hat{F}_0(z_i)}{\hat{F}(z_i)}$$

Note these are essentially p -values adjusted for multiple testing (by a variant of the Benjamini-Hochberg method).

By thresholding false discovery rates it is possible to identify those variables that clearly belong to each of the two groups but also those features that cannot easily be discriminated to fall into either group:

- “alternative” variables have low local FDR, e.g. $\widehat{fdr}(z_i) \leq 0.2$
- “null” variables have high local FDR, e.g. $\widehat{fdr}(z_i) \geq 0.8$
- features that cannot easily classified as null or alternative, e.g. $0.2 < \widehat{fdr}(z_i) < 0.8$

For feature selection in prediction settings we generally aim to remove only those variable that clearly belong to the null group, leaving all others in the model.

4.6.2 Quantifying prediction error

Another and more direct way to compare models is to compare their predictive performance by quantification of prediction error. Specifically, we are interested in the relative performance of models with diverse sets of predictor. variables.

A measure of predictor error compares the predicted label \hat{y} with the true label y for a validation data set. A validation data set contains both the x_i and the associated label y_i but unlike the training data it has not been used for learning the predictor function.

For continuous response often the squared loss is used:

$$\text{err}(\hat{y}, y) = (\hat{y} - y)^2$$

For binary outcomes one often employs the 0/1 loss:

$$\text{err}(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y} \neq y \\ 0, & \text{otherwise} \end{cases}$$

Alternatively, any other quantity derived from the confusion matrix (containing TP, TN, FP, FN) can be used.

The mean prediction error is the expectation

$$PE = E(\text{err}(\hat{y}, y))$$

and thus the empirical mean prediction error is

$$\widehat{PE} = \frac{1}{m} \sum_{i=1}^m \text{err}(\hat{y}_i, y_i)$$

where m is the sample size of the validation data set.

More generally, we can also quantify prediction error in the framework of so-called **proper scoring rules**, where the whole probabilistic forecast is taken into account (e.g. the individual probabilities for each class, rather than just the selected most probable class). A commonly used scoring rule is the negative log-probability (“surprise”), and the expected surprise is the cross-entropy (cf. Statistical Methods module). So this leads back to entropy and likelihood (see [MATH20802 Statistical Methods](#)).

Once we have an estimate of the prediction error of a model we can use it to compare and choose among a set of candidate models, selecting those with a sufficiently low prediction error.

4.6.3 Estimation of prediction error without validation data using cross-validation

Unfortunately, quite often we do not have separate validation data available to evaluate a classifier.

In this case we need to rely on a simple algorithmic procedure called **cross-validation**.

Outline of cross-validation:

- 1) split the samples in the training data into a number (say K) parts (“folds”).
- 2) use each of the K folds as validation data and the other $K - 1$ folds as training data.
- 3) average over the resulting K individual estimates of prediction error, to get an overall aggregated predictor error, along with an error.

Note that in each case one part of the data is reserved for validation and *not* used for training the predictor.

We choose K such that the folds are not too small (to allow estimation of prediction error) but also not too large (to make sure that we actually are able to train a reliable classifier from the remaining data). A typical value for K is 5 or 10, so

that 80% respectively 90% of the samples are used for training and the other 20% or 10% for validation.

If $K = n$ there are as many folds as there are samples and the validation data set consists only of a single data point. This is called “leave one out” cross-validation (LOOCV). There are analytic approximations for the prediction error obtained by LOOCV so that this approach is computationally inexpensive for some standard models (including regression).

In a number of worksheets cross-validation is employed to evaluate classification models to demonstrate in practise that feature selection is useful to construct compact models with only a small number of variables that nonetheless generalise and predict well.

Further reading:

To study the technical details of cross-validation: read **Section 5.1 Cross-Validation** in James et al. (2021) *An introduction to statistical learning with applications in R (2nd edition)*. Springer.

Chapter 5

Multivariate dependencies

5.1 Measuring the linear association between two sets of random variables

5.1.1 Outline

In this section we discuss how to measure the total linear association between two sets of random variables $\mathbf{x} = (x_1, \dots, x_p)^T$ and $\mathbf{y} = (y_1, \dots, y_q)^T$. We assume a joint correlation matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_x & \mathbf{P}_{xy} \\ \mathbf{P}_{yx} & \mathbf{P}_y \end{pmatrix}$$

with cross-correlation matrix $\mathbf{P}_{xy} = \mathbf{P}_{yx}^T$ and the within-group group correlations \mathbf{P}_x and \mathbf{P}_y . If the cross-correlations vanish, $\mathbf{P}_{xy} = 0$, then the two random vectors are uncorrelated, and the joint correlation matrix becomes a diagonal block matrix

$$\mathbf{P}_{\text{indep}} = \begin{pmatrix} \mathbf{P}_x & 0 \\ 0 & \mathbf{P}_y \end{pmatrix}.$$

5.1.2 Special cases

In linear regression model the *squared multiple correlation* or *coefficient of determination*

$$\text{Cor}(\mathbf{x}, \mathbf{y})^2 = \mathbf{P}_{yx} \mathbf{P}_x^{-1} \mathbf{P}_{xy}$$

is a standard measure to describe the strength of total linear association between the predictors \mathbf{x} and the response \mathbf{y} . If $\mathbf{P}_{xy} = 0$ then $\text{Cor}(\mathbf{x}, \mathbf{y})^2 = 0$.

If there is only a single predictor x then $\mathbf{P}_{xy} = \rho$ and $\mathbf{P}_x = 1$ and therefore the squared multiple correlation reduces to the squared Pearson correlation

$$\text{Cor}(x, y)^2 = \rho^2.$$

5.1.3 Rozeboom vector correlation

For the general case with two random vectors we are looking for a scalar quantity that quantifies the divergence of the general joint correlation matrix \mathbf{P} from the joint correlation matrix $\mathbf{P}_{\text{indep}}$ assuming uncorrelated \mathbf{x} and \mathbf{y} .

One such quantity is Hotelling's *vector alienation coefficient* (given in his 1936 CCA paper¹)

$$\begin{aligned} a(\mathbf{x}, \mathbf{y}) &= \frac{\det(\mathbf{P})}{\det(\mathbf{P}_{\text{indep}})} \\ &= \frac{\det(\mathbf{P})}{\det(\mathbf{P}_x) \det(\mathbf{P}_y)} \end{aligned}$$

With $\mathbf{K} = \mathbf{P}_x^{-1/2} \mathbf{P}_{xy} \mathbf{P}_y^{-1/2}$ (cf. Chapter 2, Canonical correlation analysis) the vector alienation coefficient can be written (using the Weinstein-Aronszajn determinant identity and the formula for the determinant for block-structured matrices, see Appendix) as

$$\begin{aligned} a(\mathbf{x}, \mathbf{y}) &= \det(\mathbf{I}_p - \mathbf{K}\mathbf{K}^T) \\ &= \det(\mathbf{I}_q - \mathbf{K}^T\mathbf{K}) \\ &= \prod_{i=1}^m (1 - \lambda_i^2) \end{aligned}$$

where the λ_i are the singular values of \mathbf{K} , i.e. the canonical correlations for the pair \mathbf{x} and \mathbf{y} .

If $\mathbf{P}_{xy} = 0$ and thus \mathbf{x} and \mathbf{y} are uncorrelated then $\mathbf{P} = \mathbf{P}_{\text{indep}}$ and thus by construction the vector alienation coefficient $a(\mathbf{x}, \mathbf{y}) = 1$. Hence, it is itself not a generalisation of the squared multiple correlation.

Instead, Rozeboom (1965) proposed to use as squared *vector correlation* the complement

$$\begin{aligned} \text{Cor}(\mathbf{x}, \mathbf{y})^2 &= \rho_{xy}^2 = 1 - a(\mathbf{x}, \mathbf{y}) \\ &= 1 - \det(\mathbf{I}_p - \mathbf{K}\mathbf{K}^T) \\ &= 1 - \det(\mathbf{I}_q - \mathbf{K}^T\mathbf{K}) \\ &= 1 - \prod_{i=1}^m (1 - \lambda_i^2) \end{aligned}$$

If $\mathbf{P}_{xy} = 0$ then $\text{Cor}(\mathbf{x}, \mathbf{y})^2 = 0$. Moreover, if either $p = 1$ or $q = 1$ the squared vector correlation reduces to the corresponding squared multiple correlation, and for both $p = 1$ and $q = 1$ it becomes squared Pearson correlation.

¹Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28:321–377.

A more general way to measure multivariate association is mutual information (MI) which not only covers linear but also non-linear association. As we will discuss in a subsequent section the Rozeboom vector correlation arises naturally in MI for the multivariate normal distribution.

5.1.4 RV coefficient

Another common approach to measure association between two random vectors is the RV coefficient introduced by Robert and Escoufier (1976) as

$$RV(\mathbf{x}, \mathbf{y}) = \frac{\text{Tr}(\Sigma_{xy}\Sigma_{yx})}{\sqrt{\text{Tr}(\Sigma_x^2)\text{Tr}(\Sigma_y^2)}}$$

It is easier to compute than the Rozeboom vector correlation since it is based on the matrix trace rather than matrix determinant.

For $q = p = 1$ the RV coefficient reduces to the squared correlation. However, the RV coefficient does *not* reduce to the multiple correlation coefficient for $q = 1$ and $p > 1$, and therefore the RV coefficient cannot be considered a coherent generalisation of Pearson and multiple correlation to the case of two random vectors.

5.2 Mutual information as generalisation of correlation

5.2.1 Definition of mutual information

Recall the definition of Kullback-Leibler divergence, or relative entropy, between two distributions:

$$D_{\text{KL}}(F, G) := E_F \log \left(\frac{f(\mathbf{x})}{g(\mathbf{x})} \right)$$

Here F plays the role of the reference distribution and G is an approximating distribution, with f and g being the corresponding density functions (see [MATH20802 Statistical Methods](#)).

The *Mutual Information* (MI) between two random variables \mathbf{x} and \mathbf{y} is defined as the KL divergence between the corresponding joint distribution and the product distribution:

$$MI(\mathbf{x}, \mathbf{y}) = D_{\text{KL}}(F_{\mathbf{x}, \mathbf{y}}, F_{\mathbf{x}}F_{\mathbf{y}}) = E_{F_{\mathbf{x}, \mathbf{y}}} \log \left(\frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{x})f(\mathbf{y})} \right).$$

Thus, MI measures how well the joint distribution can be approximated by the product distribution (which would be the appropriate joint distribution if \mathbf{x} and \mathbf{y} are independent). Since MI is an application of KL divergence it shares all its properties. In particular, $MI(\mathbf{x}, \mathbf{y}) = 0$ implies that the joint distribution and

product distributions are the same. Hence the two random variables x and y are independent if the mutual information vanishes.

5.2.2 Mutual information between two normal variables

The KL divergence between two multivariate normal distributions F_{ref} and F is

$$D_{\text{KL}}(F_{\text{ref}}, F) = \frac{1}{2} \left\{ (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{ref}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{ref}}) + \text{Tr} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\text{ref}} \right) - \log \det \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\text{ref}} \right) - d \right\}$$

This allows compute the mutual information $\text{MI}_{\text{norm}}(x, y)$ between two univariate random variables x and y that are correlated and assumed to be jointly bivariate normal. Let $\mathbf{z} = (x, y)^T$. The joint bivariate normal distribution is characterised by the mean $\mathbb{E}(\mathbf{z}) = \boldsymbol{\mu} = (\mu_x, \mu_y)^T$ and the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix}$$

where $\text{Cor}(x, y) = \rho$. If x and y are independent then $\rho = 0$ and

$$\boldsymbol{\Sigma}_{\text{indep}} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}.$$

The product

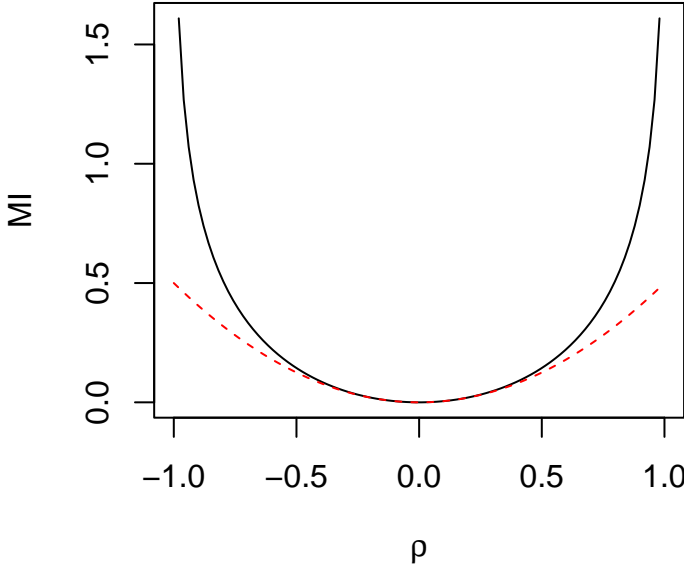
$$\mathbf{A} = \boldsymbol{\Sigma}_{\text{indep}}^{-1} \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \frac{\sigma_y}{\sigma_x} \\ \rho \frac{\sigma_x}{\sigma_y} & 1 \end{pmatrix}$$

has trace $\text{Tr}(\mathbf{A}) = 2$ and determinant $\det(\mathbf{A}) = 1 - \rho^2$.

With this the mutual information between x and y can be computed as

$$\begin{aligned} \text{MI}_{\text{norm}}(x, y) &= D_{\text{KL}}(N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\text{indep}})) \\ &= \frac{1}{2} \left\{ \text{Tr} \left(\boldsymbol{\Sigma}_{\text{indep}}^{-1} \boldsymbol{\Sigma} \right) - \log \det \left(\boldsymbol{\Sigma}_{\text{indep}}^{-1} \boldsymbol{\Sigma} \right) - 2 \right\} \\ &= \frac{1}{2} \left\{ \text{Tr}(\mathbf{A}) - \log \det(\mathbf{A}) - 2 \right\} \\ &= -\frac{1}{2} \log(1 - \rho^2) \\ &\approx \frac{\rho^2}{2} \end{aligned}$$

Thus $\text{MI}_{\text{norm}}(x, y)$ is a one-to-one function of the squared correlation ρ^2 between x and y :



For small values of correlation $2 \text{MI}_{\text{norm}}(x, y) \approx \rho^2$.

5.2.3 Mutual information between two normally distributed random vectors

The mutual information $\text{MI}_{\text{norm}}(x, y)$ between two multivariate normal random vector x and y can be computed in a similar fashion as in the bivariate case.

Let $z = (x, y)^T$ with dimension $d = p + q$. The joint multivariate normal distribution is characterised by the mean $E(z) = \mu = (\mu_x^T, \mu_y^T)^T$ and the covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{pmatrix}.$$

If x and y are independent then $\Sigma_{xy} = 0$ and

$$\Sigma_{\text{indep}} = \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma_y \end{pmatrix}.$$

The product

$$\begin{aligned} A &= \Sigma_{\text{indep}}^{-1} \Sigma = \begin{pmatrix} I_p & \Sigma_x^{-1} \Sigma_{xy} \\ \Sigma_y^{-1} \Sigma_{yx} & I_q \end{pmatrix} \\ &= \begin{pmatrix} I_p & V_x^{-1/2} P_x^{-1} P_{xy} V_y^{1/2} \\ V_y^{-1/2} P_y^{-1} P_{yx} V_x^{1/2} & I_q \end{pmatrix} \end{aligned}$$

has trace $\text{Tr}(A) = d$ and determinant

$$\begin{aligned} \det(A) &= \det(I_p - K K^T) \\ &= \det(I_q - K^T K) \end{aligned}$$

with $K = P_x^{-1/2} P_{xy} P_y^{-1/2}$. With $\lambda_1, \dots, \lambda_m$ the singular values of K (i.e. the canonical correlations between x and y) we get

$$\det(A) = \prod_{i=1}^m (1 - \lambda_i^2)$$

The mutual information between x and y is then

$$\begin{aligned} \text{MI}_{\text{norm}}(x, y) &= D_{\text{KL}}(N(\mu, \Sigma), N(\mu, \Sigma_{\text{indep}})) \\ &= \frac{1}{2} \left\{ \text{Tr}(\Sigma_{\text{indep}}^{-1} \Sigma) - \log \det(\Sigma_{\text{indep}}^{-1} \Sigma) - d \right\} \\ &= \frac{1}{2} \left\{ \text{Tr}(A) - \log \det(A) - d \right\} \\ &= -\frac{1}{2} \sum_{i=1}^m \log(1 - \lambda_i^2) \end{aligned}$$

Note that $\text{MI}_{\text{norm}}(x, y)$ is the sum of the MIs resulting from the individual canonical correlations λ_i with the same functional form as in the bivariate normal case.

By comparison with the squared Rozeboom vector correlation coefficient ρ_{xy}^2 we recognize that

$$\text{MI}_{\text{norm}}(x, y) = -\frac{1}{2} \log(1 - \rho_{xy}^2) \approx \frac{1}{2} \rho_{xy}^2$$

Thus, in the multivariate case $\text{MI}_{\text{norm}}(xy)$ has again exactly the same functional relationship with the vector correlation $\rho_{x,y}^2$ as the $\text{MI}_{\text{norm}}(x, y)$ for two univariate variables with squared Pearson correlation ρ^2 .

Thus, Rozebooms vector correlation is directly linked to mutual information for jointly multivariate normally distributed variables.

5.2.4 Using MI for variable selection

In principle, MI can be computed for any distribution and model and thus applies to both normal and non-normal models, and to both linear and nonlinear relationships.

In very general way we may denote by $F_{y|x}$ we denote a predictive model for y conditioned on x and F_y is the marginal distribution of y without predictors. Note that the predictive model can assume any form (incl. nonlinear). Typically $F_{y|x}$ is a complex model and F_y a simple model (no predictors).

Then mutual information between x and y can be also understood as expected KL divergence between the conditional and marginal distributions:

$$E_{F_x} D_{\text{KL}}(F_{y|x}, F_y) = \text{MI}(x, y)$$

This can be shown as follows. The KL divergence between $F_{y|x}$ and F_y is given by

$$D_{\text{KL}}(F_{y|x}, F_y) = E_{F_{y|x}} \log \left(\frac{f(y|x)}{f(y)} \right),$$

which is a random variable since it depends on x . Taking the expectation with regard to F_x (the distribution of x) we get

$$E_{F_x} D_{\text{KL}}(F_{y|x}, F_y) = E_{F_x} E_{F_{y|x}} \log \left(\frac{f(y|x)f(x)}{f(y)f(x)} \right) = E_{F_{x,y}} \log \left(\frac{f(x,y)}{f(y)f(x)} \right) = \text{MI}(x, y).$$

Because of this link of MI with conditioning the MI between response and predictor variables is often used for variable and feature selection in general models.

5.2.5 Other measures of general dependence

Besides mutual information there are others measures of general dependence between multivariate random variables.

The two most important ones that have been proposed in the recent literature are i) [distance correlation](#) and ii) the [maximal information coefficient](#) (MIC and MIC_e).

5.3 Graphical models

5.3.1 Purpose

Graphical models combine features from

- graph theory
- probability
- statistical inference

The literature on graphical models is huge, we focus here only on two commonly used models:

- DAGs (directed acyclic graphs), all edges are directed, no directed loops (i.e. no cycles, hence “acyclic”)
- GGM (Gaussian graphical models), all edges are undirected

Graphical models provide probabilistic models for trees and for networks, with random variables represented by nodes in the graphs, and branches representing conditional dependencies. In this regard they generalise both the tree-based clustering approaches as well as the probabilistic non-hierarchical methods (GMMs).

However, the class of graphical models goes much beyond simple unsupervised learning models. It also includes regression, classification, time series models etc. See e.g. the reference book by Murphy (2012).

5.3.2 Basic notions from graph theory

- Mathematically, a graph $G = (V, E)$ consists of a set of vertices or nodes $V = \{v_1, v_2, \dots\}$ and a set of branches or edges $E = \{e_1, e_2, \dots\}$.
- Edges can be undirected or directed.
- Graphs containing only directed edges are directed graphs, and likewise graphs containing only undirected edges are called undirected graphs. Graphs containing both directed and undirected edges are called partially directed graphs.
- A path is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence.
- A graph is connected when there is a path between every pair of vertices.
- A cycle is a path in a graph that connects a node with itself.
- A connected graph with no cycles is called a tree.
- The degree of a node is the number of edges it connects with. If edges are all directed the degree of a node is the sum of the in-degree and out-degree, which counts the incoming and outgoing edges, respectively.
- External nodes are nodes with degree 1. In a tree-structured graph these are also called leaves.

Some notions are only relevant for graphs with directed edges:

- In a directed graph the parent node(s) of vertex v is the set of nodes $\text{pa}(v)$ directly connected to v via edges directed from the parent node(s) towards v .
- Conversely, v is called a child node of $\text{pa}(v)$. Note that a parent node can have several child nodes, so v may not be the only child of $\text{pa}(v)$.
- In a directed tree graph, each node has only a single parent, except for one particular node that has no parent at all (this node is called the root node).
- A DAG, or directed acyclic graph, is a directed graph with no directed cycles. A (directed) tree is a special version of a DAG.

5.3.3 Probabilistic graphical models

A graphical model uses a graph to describe the relationship between random variables x_1, \dots, x_d . The variables are assumed to have a joint distribution with density/mass function $\Pr(x_1, x_2, \dots, x_d)$. Each random variable is placed in a node of the graph.

The structure of the graph and the type of the edges connecting (or not connecting) any pair of nodes/variables is used to describe the conditional dependencies, and to simplify the joint distribution.

Thus, a graphical model is in essence a visualisation of the joint distribution using structural information from the graph helping to understand the mutual relationship among the variables.

5.3.4 Directed graphical models

In a **directed graphical model** the graph structure is assumed to be a DAG (or a directed tree, which is also a DAG).

Then the joint probability distribution can be factorised into a *product of conditional probabilities* as follows:

$$\Pr(x_1, x_2, \dots, x_d) = \prod_i \Pr(x_i | \text{pa}(x_i))$$

Thus, the overall joint probability distribution is specified by local conditional distributions and the graph structure, with the directions of the edges providing the information about parent-child node relationships.

Probabilistic DAGs are also known as “Bayesian networks”.

Idea: by trying out all possible trees/graphs and fitting them to the data using maximum likelihood (or Bayesian inference) we hope to be able identify the graph structure of the data-generating process.

Challenges

- 1) in the tree/network the internal nodes are usually not known, and thus have to be treated as *latent* variables.

Answer: To impute the states at these nodes we may use the EM algorithm as in GMMs (which in fact can be viewed as graphical models, too!).

- 2) If we treat the internal nodes as unknowns we need to marginalise over the internal nodes, i.e. we need to sum / integrate over all possible set of states of the internal nodes!

Answer: This can be handled very effectively using the **Viterbi algorithm** which is essentially an application of the generalised distributive law. In particular for tree graphs this means that the summations occurs locally at each nodes and propagates recursively accross the tree.

- 3) In order to infer the tree or network structure the space of all trees or networks need to be explored. This is not possible in an exhaustive fashion unless the number of variables in the tree is very small.

Answer: Solution: use heuristic approaches for tree and network search!

- 4) Furthermore, there exist so-called “equivalence classes” of graphical models, i.e. sets of graphical models that share the same joint probability distribution. Thus, all graphical models within the same equivalence class cannot be distinguished from observational data, even with infinite sample size!

Answer: this is a fundamental mathematical problem of identifiability so there is now way around this issue. However, on the positive side, this also implies that the search through all graphical models can be restricted to finding the so-called “essential graph” (e.g. <https://projecteuclid.org/euclid.aos/1031833662>)

Conclusion: using directed graphical models for structure discovery is very time consuming and computationally demanding for anything but small toy data sets.

This also explains why heuristic and non-model based approaches (such as hierarchical clustering) are so popular even though full statistical modelling is in principle possible.

5.3.5 Undirected graphical models

Another class of graphical models are models that contain only undirected edges. These **undirected graphical models** are used to represent the pairwise conditional (in)dependencies among the variables in the graph, and the resulting model is therefore also called **conditional independence graph**.

If x_i and x_j two selected random variables/nodes, and the set $\{x_k\}$ represents all other variables/nodes with $k \neq i$ and $k \neq j$. We say that variables x_i and x_j are conditionally independent given all the other variables $\{x_k\}$

$$x_i \perp\!\!\!\perp x_j | \{x_k\}$$

if the joint probability density of x_i, x_j and x_k factorises as

$$\Pr(x_1, x_2, \dots, x_d) = \Pr(x_i | \{x_k\}) \Pr(x_j | \{x_k\}) \Pr(\{x_k\}).$$

or equivalently

$$\Pr(x_i, x_j | \{x_k\}) = \Pr(x_i | \{x_k\}) \Pr(x_j | \{x_k\}).$$

In a corresponding conditional independence graph, there is no edge between x_i and x_j , as in such a graph *missing edges correspond to conditional independence* between the respective non-connected nodes.

5.3.5.1 Gaussian graphical model

Assuming that x_1, \dots, x_d are jointly normally distributed, i.e. $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it turns out that it is straightforward to identify the pairwise conditional independencies. From $\boldsymbol{\Sigma}$ we first obtain the precision matrix

$$\boldsymbol{\Omega} = (\omega_{ij}) = \boldsymbol{\Sigma}^{-1}.$$

Crucially, it can be shown that $\omega_{ij} = 0$ implies $x_i \perp\!\!\!\perp x_j | \{x_k\}$! Hence, from the precision matrix $\boldsymbol{\Omega}$ we can directly read off all the pairwise conditional independencies among the variables x_1, x_2, \dots, x_d !

Often, the covariance matrix $\boldsymbol{\Sigma}$ is dense (few zeros) but the corresponding precision matrix $\boldsymbol{\Omega}$ is sparse (many zeros).

The conditional independence graph computed for normally distributed variables is called a **Gaussian graphical model**, or **GGM**. A further alternative name is **covariance selection model**.

5.3.5.2 Related quantity: partial correlation

From the precision matrix $\boldsymbol{\Omega}$ we can also compute the matrix of pairwise full conditional *partial correlations*:

$$\rho_{ij|\text{rest}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

which is essentially the standardised precision matrix (similar to correlation but with an extra minus sign!)

The partial correlations lie in the range between -1 and +1, $\rho_{ij|\text{rest}} \in [-1, 1]$, just like standard correlations.

If \mathbf{x} is multivariate normal then $\rho_{ij|\text{rest}} = 0$ indicates conditional independence between x_i and x_j .

Regression interpretation: partial correlation is the correlation that remains between the two variables if the effect of the other variables is “regressed away”.

In other words, the partial correlation is exactly equivalent to the correlation between the residuals that remain after regressing x_i on the variables $\{x_k\}$ and x_j on $\{x_k\}$.

5.3.6 Algorithm for learning GGMs

From the above we can devise a simple algorithm to learn Gaussian graphical model (GGM) from data:

1. Estimate covariance $\hat{\Sigma}$ (in such a way that it is invertible!)
2. Compute corresponding partial correlations
3. If $\hat{\rho}_{ij|\text{rest}} \approx 0$ then there is (approx). conditional independence between x_i and x_j .
In practise this is done by statistical testing for vanishing partial correlations. If there are many edges we also need adjustment for simultaneous multiple testing since all edges are tested in parallel.

5.3.7 Example: exam score data (Mardia et al 1979:)

Correlations (rounded to 2 digits):

##	mechanics	vectors	algebra	analysis	statistics
## mechanics	1.00	0.55	0.55	0.41	0.39
## vectors	0.55	1.00	0.61	0.49	0.44
## algebra	0.55	0.61	1.00	0.71	0.66
## analysis	0.41	0.49	0.71	1.00	0.61
## statistics	0.39	0.44	0.66	0.61	1.00

Partial correlations (rounded to 2 digits):

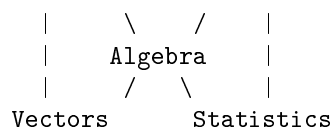
##	mechanics	vectors	algebra	analysis	statistics
## mechanics	1.00	0.33	0.23	0.00	0.02
## vectors	0.33	1.00	0.28	0.08	0.02
## algebra	0.23	0.28	1.00	0.43	0.36
## analysis	0.00	0.08	0.43	1.00	0.25
## statistics	0.02	0.02	0.36	0.25	1.00

Note that that there are no zero correlations but there are **four partial correlations close to 0**, indicating **conditional independence** between:

- analysis and mechanics,
- statistics and mechanics,
- analysis and vectors, and
- statistics and vectors.

Thus, of 10 possible edges four are missing, and thus the conditional independence graph looks as follows:

Mechanics Analysis



Chapter 6

Nonlinear and nonparametric models

In the last part of the module we discuss methods that go beyond the linear methods prevalent in classical multivariate statistics.

Relevant textbooks:

The lectures for much of this part of the module follow selected chapters from the following three text books:

- James et al. (2021) *An introduction to statistical learning with applications in R (2nd edition)*. Springer.
- Rogers and Girolami (2017) *A first course in machine learning (2nd edition)*. CRC Press.

Please study the relevant section and chapters as indicated below in each subsection!

6.1 Limits of linear models and correlation

Linear models are very effective tools. However, it is important to recognise their limits especially when modelling complex nonlinear relationships.

6.1.1 Correlation only measures linear dependence

A very simple demonstration of this is given by the following example. Assume x is a normally distributed random variable with $x \sim N(0, 1)$. From x we construct a second random variable $y = x^2$ — thus y fully depends on x with no added extra noise. What is the correlation between x and y ?

Let's answer this question by running a small computer simulation:

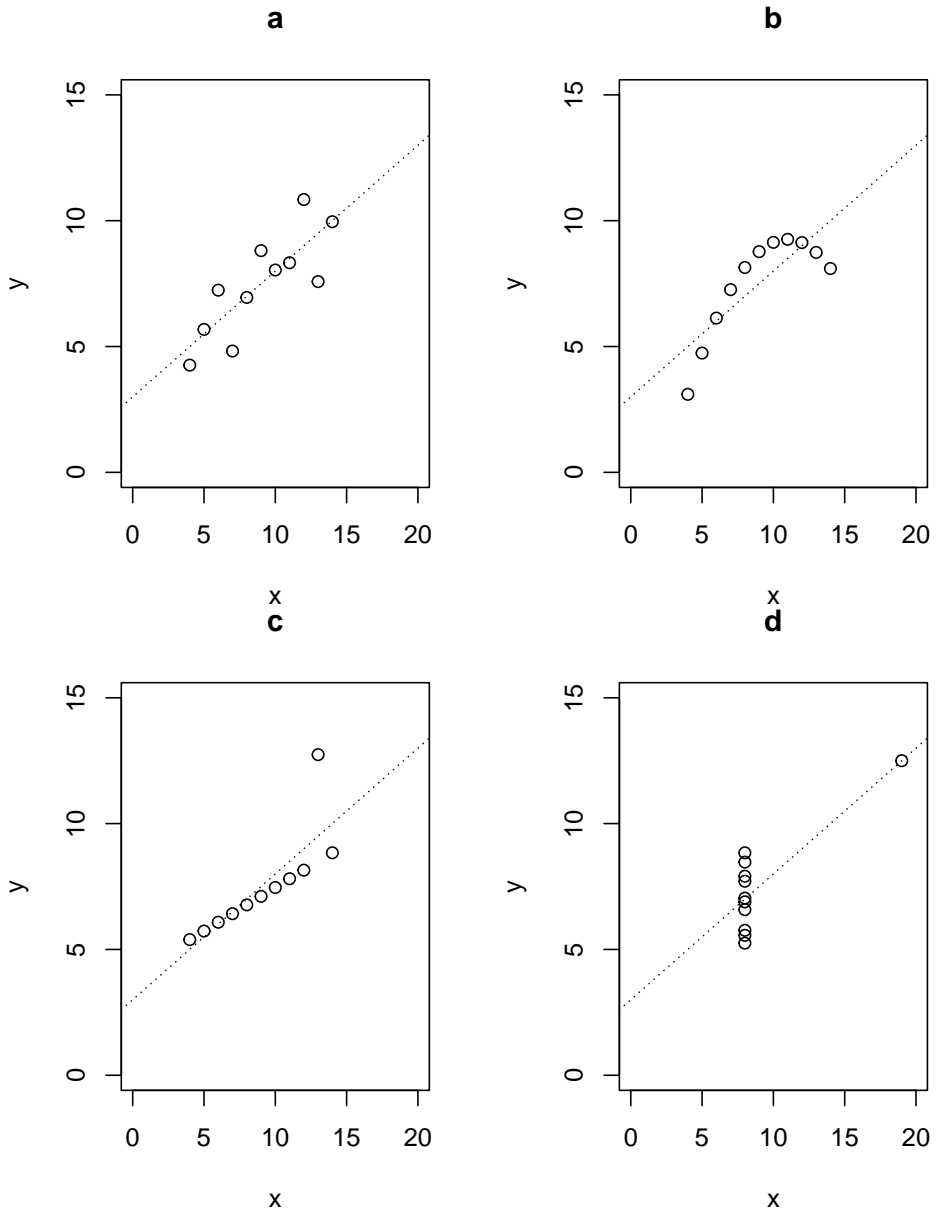
```
x=rnorm(10000)
y = x^2
cor(x,y)
```

```
## [1] 0.02760381
```

Thus, correlation is (almost) zero even though x and y are fully dependent! This is because correlation only measures linear association!

6.1.2 Anscombe data sets

Using correlation, and more generally linear models, blindly can thus hide the underlying complexity of the analysed data. This is demonstrated by the classic “Anscombe quartet” of data sets (F. J. Anscombe. 1973. Graphs in statistical analysis. The American Statistician 27:17-21, <http://dx.doi.org/10.1080/00031305.1973.10478966>):



As evident from the scatter plots the relationship between the two variables x and y is very different in the four cases! However, intriguingly all four data sets share exactly the same linear characteristics and summary statistics:

- Means $m_x = 9$ and $m_y = 7.5$
- Variances $s_x^2 = 11$ and $s_y^2 = 4.13$
- Correlation $r = 0.8162$

- Linear model fit with intercept $a = 3.0$ and slope $b = 0.5$

Thus, in actual data analysis it is always a **good idea to inspect the data visually** to get a first impression whether using a linear model makes sense.

In the above only data “a” follows a linear model. Data “b” represents a quadratic relationship. Data “c” is linear but with an outlier that disturbs the linear relationship. Finally data “d” also contains an outlier but also represent a case where y is (apart from the outlier) is not dependent on x .

In the Worksheet 10 a more recent version of the Anscombe quartet will be analysed in the form of the “datasauRus” dozen - 13 highly nonlinear datasets that all share the same linear characteristics.

6.1.3 Alternatives

- Mutual information (based on relative entropy) as discussed in previous chapter
- Other measures designed to capture nonlinear association, such as [distance correlation](#). and ii) the [maximal information coefficient](#) (MIC and MIC_e).

6.2 Random forests

Another widely used approach for prediction in nonlinear settings is the method of random forests.

Relevant reading:

Please read: James et al. (2013) **Chapter 8 “Tree-Based Methods”**

Specifically:

- Section 8.1 The Basics of Decision Trees
- Section 8.2.1 Bagging
- Section 8.2.2 Random Forests

6.2.1 Stochastic vs. algorithmic models

Two cultures in statistical modelling: stochastic vs. algorithmic models

Classic discussion paper by Leo Breiman (2001): Statistical modeling: the two cultures. *Statistical Science* 16:199–231. <https://projecteuclid.org/euclid.ss/1009213726>

This paper has recently be revisited in the following discussion paper by Efron (2020) and discussants: Prediction, estimation, and attribution. *JASA* 115:636–677. <https://doi.org/10.1080/01621459.2020.1762613>

6.2.2 Random forests

Invented by Breimann in 1996.

Basic idea:

- A single decision tree is unreliable and unstable (weak predictor/classifier).
- Use bootstrap to generate multiple decision trees (=“forest”)
- Average over predictions from all tree (=“bagging”, bootstrap aggregation)

The averaging procedure has the effect of variance stabilisation. Intriguingly, averaging across all decision trees dramatically improves the overall prediction accuracy!

The Random Forests approach is an example of an **ensemble method** (since it is based on using an “ensemble” of trees).

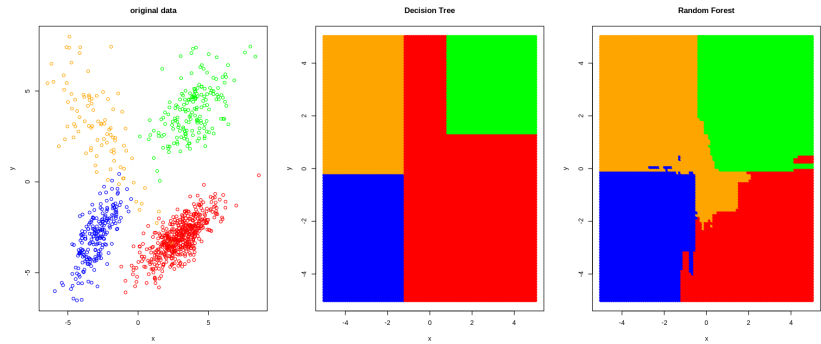
Variations: boosting, XGBoost (<https://xgboost.ai/>)

Random forests will be applied in Worksheet 10.

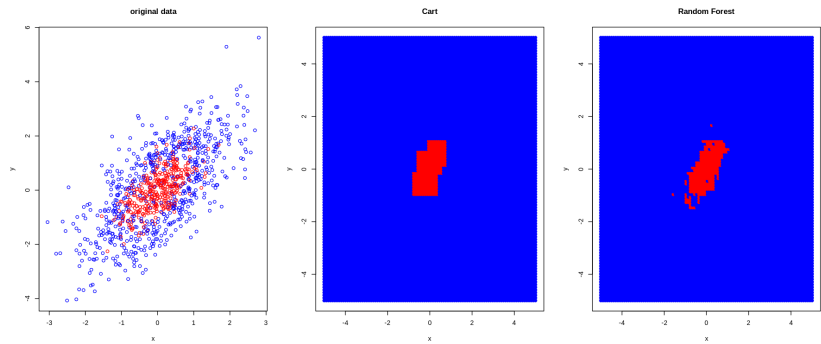
They are computationally expensive but typically perform very well!

6.2.3 Comparison of decision boundaries: decision tree vs. random forest

Non-nested case:



Nested case:



Compare also with the decision boundaries for LDA and QDA (previous chapter).

6.3 Gaussian processes

Gaussian processes offer another nonparametric approach to model nonlinear dependencies. They provide a probabilistic model for the unknown nonlinear function.

Relevant reading:

Please read: Rogers and Girolami (2017) **Chapter 8: Gaussian processes**.

6.3.1 Main concepts

- Gaussian processes (GPs) belong to the family of **Bayesian nonparametric models**
- Idea:
 - start with prior over a function (!),
 - then condition on observed data to get posterior distribution (again over a function)
- GPs use an infinitely dimensional multivariate normal distribution as prior

6.3.2 Conditional multivariate normal distribution

GPs make use of the fact that marginal and conditional distributions of a multivariate normal distribution are also multivariate normal.

Multivariate normal distribution:

$$z \sim N_d(\mu, \Sigma)$$

Assume:

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_2 \end{pmatrix}$$

with corresponding dimensions d_1 and d_2 and $d_1 + d_2 = d$.

Marginal distributions:

Any subset of z is also multivariate normally distributed. Specifically,

$$z_1 \sim N_{d_1}(\mu_1, \Sigma_1)$$

and

$$z_2 \sim N_{d_2}(\mu_2, \Sigma_2)$$

Conditional multivariate normal:

The conditional distribution is also multivariate normal:

$$z_1|z_2 = z_{1|2} \sim N_{d_1}(\mu_{1|2}, \Sigma_{1|2})$$

with

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_2^{-1}(z_2 - \mu_2)$$

and

$$\Sigma_{1|2} = \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{12}^T$$

$z_{1|2}$ and $\mu_{1|2}$ have dimension $d_1 \times 1$ and $\Sigma_{1|2}$ has dimension $d_1 \times d_1$, i.e. the same dimension as the unconditioned variables.

6.3.3 Covariance functions and kernels

The GP prior is a infinitely dimensional multivariate normal with mean zero and the **covariance specified by a function** $k(x, x')$:

A widely used covariance function is

$$k(x, x') = \text{Cov}(x, x') = \sigma^2 e^{-\frac{(x-x')^2}{2l^2}}$$

This is known as the **squared-exponential kernel** or **Radial-basis function (RBF) kernel**.

Note that this kernel implies

- $k(x, x) = \text{Var}(x) = \sigma^2$ and
- $\text{Cor}(x, x') = e^{-\frac{(x-x')^2}{2l^2}}$.

The parameter l in the RBF kernel is the length scale parameter and describes the “wigglyness” or smoothness of the resulting function. Small values of l mean more complex, more wiggly functions, and low autocorrelation.

There are many other kernel functions, including linear, polynomial or periodic kernels.

6.3.4 GP model

Nonlinear regression in the GP approach is conceptually very simple:

- start with multivariate prior
- then condition on the observed data
- the resulting conditional multivariate normal can used to predict the function values at any unobserved values
- the conditional variance can be used to compute credible intervals for predictions.

GP regression also provides a direct link with classical Bayesian linear regression (using a linear kernel).

Drawbacks: computationally expensive ($O(n^3)$) because of the matrix inversion)

6.3.5 Gaussian process example

We now show how to apply Gaussian processes in R justing using standard matrix calculations.

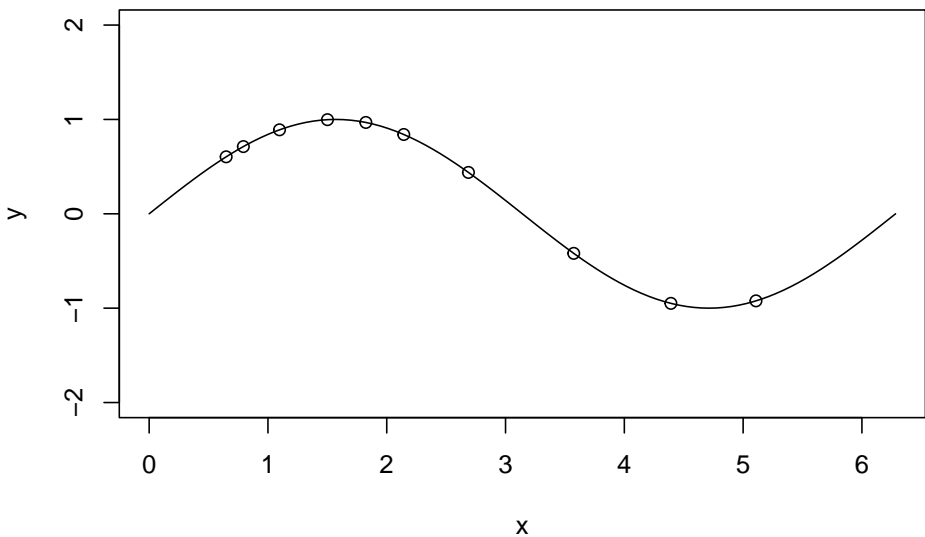
Our aim is to estimate the following nonlinear function from a number of observations. Note that initially we assume that there is no additional noise (so the observations lie directly on the curve):

```
truefunc = function(x) sin(x)
XLIM = c(0, 2*pi)
YLIM = c(-2, 2)

n2 = 10
x2 = runif(n2, min=XLIM[1], max=XLIM[2])
y2 = truefunc(x2) # no noise

curve( truefunc(x), xlim=XLIM, ylim=YLIM, xlab="x", ylab="y",
       main="True Function")
points(x2, y2)
```

True Function



Next we use the RFB kernel as the prior covariance and also assume prior mean 0:

```

# RBF kernel
rbfkernel = function(xa, xb, s2=1, l=1/2) s2*exp(-1/2*(xa-xb)^2/l^2)
kfun.mat = function(xavec, xbvec, FUN=rbfkernel)
  outer(X=as.vector(xavec), Y=as.vector(xbvec), FUN=FUN)

# prior mean
mu.vec = function(x) rep(0, length(x))

```

We can now visualise the functions samples from the multivariate normal prior:

```

# grid of x-values
n1 = 100
x1 = seq(XLIM[1], XLIM[2], length.out=n1)

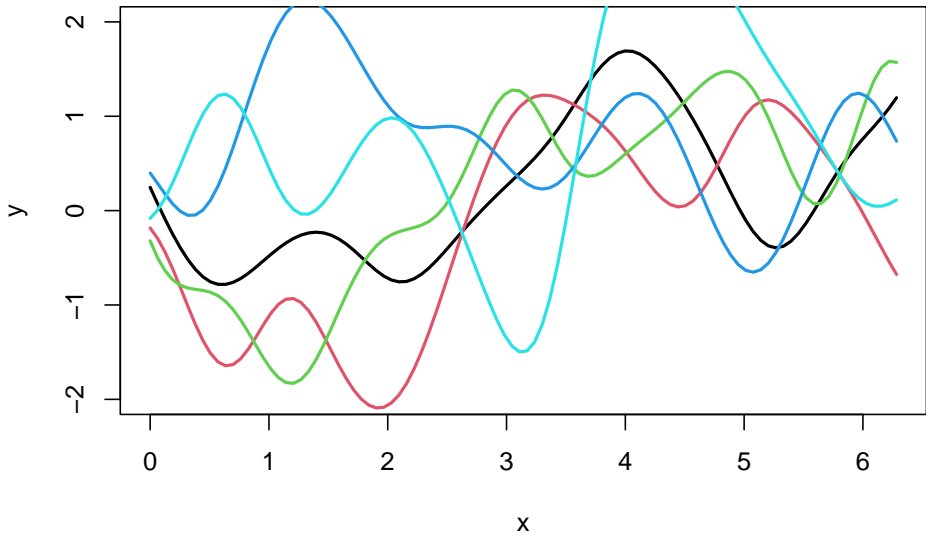
# unconditioned covariance and mean (unobserved samples x1)
K1 = kfun.mat(x1, x1)
m1 = mu.vec(x1)

## sample functions from GP prior
B = 5
library("MASS") # for mvrnorm
y1r = t(mvrnorm(B, mu = m1, Sigma=K1))

plot(x1, y1r[,1], type="l", lwd=2, ylab="y", xlab="x", ylim=YLIM,
     main="Prior Functions (RBF Kernel with l=1/2)")
for(i in 2:B)
  lines(x1, y1r[,i], col=i, lwd=2)

```

Prior Functions (RBF Kernel with $l=1/2$)



Now we compute the posterior mean and variance by conditioning on the observations:

```
# unconditioned covariance and mean (observed samples x2)
K2 = kfun.mat(x2, x2)
m2 = mu.vec(x2)
iK2 = solve(K2) # inverse

# cross-covariance
K12 = kfun.mat(x1, x2)

# Conditioning: x1 conditioned on x2

# conditional mean
m1.2 = m1 + K12 %*% iK2 %*% (y2 - m2)

# conditional variance
K1.2 = K1 - K12 %*% iK2 %*% t(K12)
```

Now we can plot the posterior mean and upper and lower bounds of a 95% credible interval:

```
# upper and lower CI
upper.bound = m1.2 + 1.96*sqrt(diag(K1.2))
lower.bound = m1.2 - 1.96*sqrt(diag(K1.2))

plot(x1, m1.2, type="l", xlim=XLIM, ylim=YLIM, col="red", lwd=3,
```

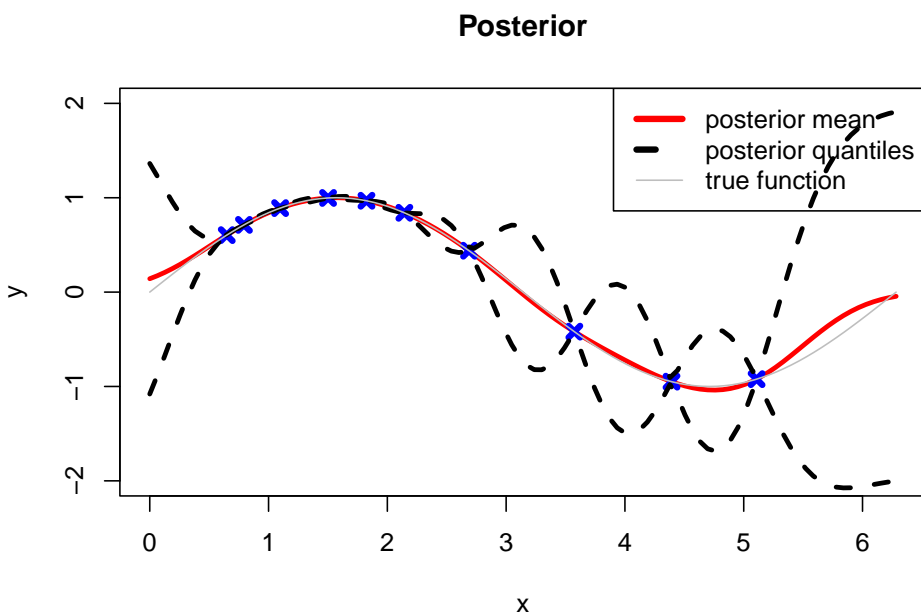
```

ylab="y", xlab = "x", main = "Posterior")

points(x2,y2,pch=4,lwd=4,col="blue")
lines(x1,upper.bound,lty=2,lwd=3)
lines(x1,lower.bound,lty=2,lwd=3)
curve(truefunc(x), xlim=XLIM, add=TRUE, col="gray")

legend(x="topright",
      legend=c("posterior mean", "posterior quantiles", "true function"),
      lty=c(1, 2, 1),lwd=c(4, 4, 1), col=c("red","black", "gray"), cex=1.0)

```



Finally, we can take into account noise at the measured data points by adding an error term:

```

# add some noise
sdeps = 0.1
K2 = K2 + sdeps^2*diag(1,length(x2))

# update
iK2 = solve(K2) # inverse
m1.2 = m1 + K12 %*% iK2 %*% (y2 - m2)
K1.2 = K1 - K12 %*% iK2 %*% t(K12)
upper.bound = m1.2 + 1.96*sqrt(diag(K1.2))
lower.bound = m1.2 - 1.96*sqrt(diag(K1.2))

plot(x1, m1.2, type="l", xlim=XLIM, ylim=YLIM, col="red", lwd=3,

```

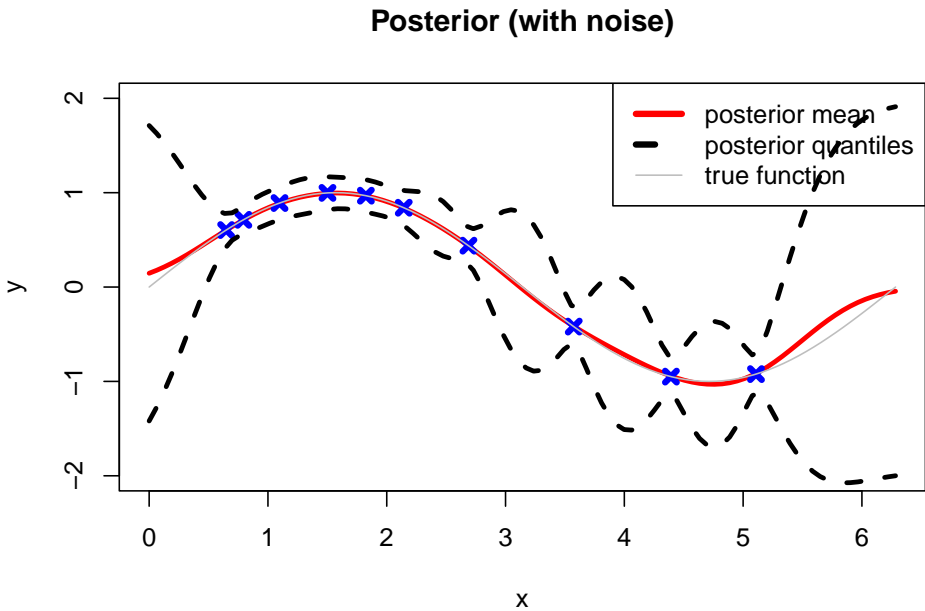
```

ylab="y", xlab = "x", main = "Posterior (with noise)")

points(x2,y2,pch=4,lwd=4,col="blue")
lines(x1,upper.bound,lty=2,lwd=3)
lines(x1,lower.bound,lty=2,lwd=3)
curve(truefunc(x), xlim=XLIM, add=TRUE, col="gray")

legend(x="topright",
  legend=c("posterior mean", "posterior quantiles", "true function"),
  lty=c(1, 2, 1),lwd=c(4, 4, 1), col=c("red","black", "gray"), cex=1.0)

```



Note that in the vicinity of data points the CIs are small and the further away from data the more uncertain the estimate of the underlying function becomes.

6.4 Neural networks

Another highly important class of models for nonlinear prediction (and nonlinear function approximation) are neural networks.

Relevant reading:

Please read: Hastie, Tibshirani, and Friedman (2009) **Chapter 11 “Neural networks”** and James et al. (2021) **Chapter 10 “Deep Learning”**

6.4.1 History

Neural networks are actually relatively old models, going back to the 1950s!

Three phases of neural networks (NN)

- 1950/60: replicating functions of neurons in the brain (perceptron)
- 1980/90: neural networks as universal function approximators
- 2010—today: deep learning

The first phase was biologically inspired, the second phase focused on mathematical properties, and the current phase is pushed forward by advances in computer science and numerical optimisation:

- backpropagation algorithm
- auto-differentiation,
- stochastic gradient descent
- use of GPUs and TPUs (e.g. for linear algebra)
- availability and development of deep learning packages:
 - Theano (University of Montreal), now Theano-PyMC/Aesara (PyMC3)
 - TensorFlow (Google),
 - Flax / JAX (Google),
 - MXNet (Amazon),
 - PyTorch (Facebook),
 - PaddlePaddle (Baidu) etc.

and high-level wrappers:

- Keras (for Tensorflow, MXNet, Theano)
- PyTorch-Lightning (for PyTorch)

6.4.2 Neural networks

Neural networks are essentially stacked systems of linear regressions, mapping input nodes (random variables) to outputs (response nodes). Each internal layer

corresponds to internal latent variables. Each layer is connected with the next layer by **non-linear activation functions**.

- feedforward single layer NN
- stacked nonlinear multiple regression with hidden variables
- optimise by empirical risk minimisation

It can be shown that NN can approximate any arbitrary non-linear function mapping input and output.

“Deep” neural networks have many layers, and their optimisation requires advanced techniques (see above).

Neural networks are very highly parameterised models and require typically a lot of data for training.

Some of the statistical aspects of NN are not well understood: in particular it is known that NN overfit the data but can still generalise well. On the other hand, it is also known that NN can also be “fooled”, i.e. prediction can be unstable (adversarial examples).

Current statistical research on NN focuses on interpretability and on links with Bayesian inference and models (e.g. GPs). For example:

- <https://link.springer.com/book/10.1007/978-3-030-28954-6>
- <https://arxiv.org/abs/1910.12478>

6.4.3 Learning more about deep learning

A good place to learn more about deep learning and about the actual implementations in computer code on various platforms is the book “Dive into deep learning” by Zhang et al. (2020) available online at <https://d2l.ai/>

Appendix A

Brief refresher on matrices

This is intended a very short recap of some essentials you need to know about matrices. We will frequently make use of matrix calculations. Matrix notation helps to make multivariate equations simpler and to better understand the underlying concepts.

For more details please consult the lecture notes of earlier modules (e.g. linear algebra).

In this course we mostly work with **real matrices**, i.e. we assume all matrix elements are real numbers. However, one important matrix decomposition — the eigenvalues decomposition — can yield complex-valued matrices when applied to real matrices. Below we will point out when this is the case.

A.1 Matrix basics

A.1.1 Matrix notation

In matrix notation we distinguish between scalars, vectors, and matrices:

Scalar: x , X , lower or upper case, plain type.

Vector: x , lower case, bold type. In handwriting an arrow \vec{x} indicates a vector.

In component notation we write $\mathbf{x} = (x_1, \dots, x_d)^T = (x_i)^T$. By convention, a vector is a column vector, i.e. the elements are arranged in a column and the index (here i) refers to the row of the column. If you transpose a column vector it becomes a row vector $\mathbf{x}^T = (x_1, \dots, x_d) = (x_i)$ and the index now refers to the column.

Matrix: X , upper case, bold type. In handwriting an underscore \underline{X} indicates a matrix.

In component notation we write $\mathbf{X} = (x_{ij})$. By convention, the first index (here i) of the scalar elements x_{ij} denotes the row and the second index (here j) the column of the matrix. Assuming that n is the number of rows and d is the number of columns you can also view the matrix $\mathbf{X} = (x_j) = (\mathbf{z}_i)^T$ as being composed of column vectors $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$ or of row vectors $\mathbf{z}_i^T = (x_{i1}, \dots, x_{id})$.

A (column) vector is a matrix of size $d \times 1$. A row vector is a matrix of size $1 \times d$. A scalar is a matrix of size 1×1 .

A.1.2 Random matrix

A **random matrix** (vector) is a matrix (vector) whose elements are random variables.

Note that the standard notation used in univariate statistics to distinguish random variables and their realisations (i.e. upper versus lower case) does not work in multivariate statistics. Therefore, you need to determine from the context whether a quantity represents a random variable, or whether it is a constant.

A.1.3 Special matrices

\mathbf{I}_d is the identity matrix. It is a square matrix of size $d \times d$ with the diagonal filled with 1 and off-diagonals filled with 0.

$$\mathbf{I}_d = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & & 1 \end{pmatrix}$$

$\mathbf{1}$ is a matrix that contains only 1s. Most often it is used in the form of a column vector with d rows:

$$\mathbf{1}_d = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

A diagonal matrix is a matrix where all off-diagonal elements are zero.

A triangular matrix is a square matrix whose elements either below or above the diagonal are all zero (upper vs. lower triangular matrix).

A.2 Simple matrix operations

A.2.1 Matrix addition and multiplication

Matrices behave much like common numbers. For example, we can add matrices $C = A + B$ and multiply matrices $C = AB$.

For **matrix addition** $C = A + B$ we add the corresponding elements $c_{ij} = a_{ij} + b_{ij}$. For matrix addition A and B must have the same dimensions, i.e. the same number of rows and columns.

The **dot product**, or **scalar product**, between two vectors a and b is given by $a \cdot b = a^T b = ab^T = \sum_{i=1}^d a_i b_i$.

Matrix multiplication $C = AB$ is obtained by setting $c_{ij} = \sum_{k=1}^m a_{ik} b_{kj}$ where m is the number of columns of A and the number of rows in B . Thus, C contains all possible dot products of the row vectors in A with the column vectors in B . For matrix multiplication the number of columns in A must match the number of rows in B . Note that matrix multiplication is in general (for $d > 1$) not commutative, i.e. $AB \neq BA$.

A.2.2 Matrix transpose

The matrix transpose $t(A) = A^T$ interchanges rows and columns. The transpose is a linear operator $(A + B)^T = A^T + B^T$ and applied to a matrix product it reverses the ordering, i.e. $(AB)^T = B^T A^T$.

If $A = A^T$ then A is symmetric (and square).

By construction given a rectangular A the matrices $A^T A$ and AA^T are symmetric with non-negative diagonal.

A.3 Matrix summaries

A.3.1 Matrix trace

The trace of the matrix is the sum of the diagonal elements $\text{Tr}(A) = \sum a_{ii}$.

A useful identity for the matrix trace is

$$\text{Tr}(AB) = \text{Tr}(BA)$$

with for two vectors becomes

$$a^T b = \text{Tr}(ba^T).$$

A.3.2 Euclidean norm

The dot product $a \cdot a = a^T a = aa^T = \sum_{i=1}^d a_i^2 = \|a\|_2^2$ yields the the **squared length** or the **squared Euclidean norm** of the vector a .

The **squared Frobenius norm** is a generalisation of the Euclidean vector norm to a rectangular matrix $A = (a_{ij})$ and is the sum of the squares of all its elements a_{ij} . Using the trace it can be written as

$$\begin{aligned} \|A\|_F^2 &= \sum_{i,j} a_{ij}^2 \\ &= \text{Tr}(A^T A) = \text{Tr}(A A^T). \end{aligned}$$

A.3.3 Determinant of a matrix

If A is a square matrix the determinant $\det(A)$ is a scalar measuring the volume spanned by the column vectors in A with the sign determined by the orientation of the vectors.

If $\det(A) \neq 0$ the matrix A is non-singular or non-degenerate. Conversely, if $\det(A) = 0$ the matrix A is singular or degenerate.

One way to compute the determinant of a matrix A is the Laplace cofactor expansion approach that proceeds recursively based on the determinants of the submatrices $A_{-i,-j}$ obtained by deleting row i and column j from A . Specifically, at each level we compute the

- 1) cofactor expansion either
 - a) along the i -th row — pick any row i :

$$\det(A) = \sum_{j=1}^d a_{ij}(-1)^{i+j} \det(A_{-i,-j}), \text{ or}$$

- b) along the j -th column — pick any j :

$$\det(A) = \sum_{i=1}^d a_{ij}(-1)^{i+j} \det(A_{-i,-j})$$

- 2) Then repeat until the submatrix is a scalar a and $\det(a) = a$.

The recursive nature of this algorithm leads to a complexity of order $O(d!)$ so it is not practical except for very small d . Therefore, in practice other more efficient algorithms for computing determinants are used but these still have algorithmic complexity in the order of $O(d^3)$ so for large dimensions obtaining determinants is very expensive.

However, some specially structured matrices do allow for very fast calculation. In particular, it turns out that the determinant of a triangular matrix (which includes diagonal matrices) is simply the product of the diagonal elements.

For a two-dimensional matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ the determinant is $\det(A) = a_{11}a_{22} - a_{12}a_{21}$.

For a block-structured square matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where the matrices on the diagonal A_{11} and A_{22} are themselves square but A_{21} and A_{12} can have any shape, the determinant is

$$\det(A) = \det(A_{22}) \det(C_1) = \det(A_{11}) \det(C_2)$$

with the (Schur complement of A_{22})

$$C_1 = A_{11} - A_{12}A_{22}^{-1}A_{21}$$

and (Schur complement of A_{11})

$$C_2 = A_{22} - A_{21}A_{11}^{-1}A_{12}$$

Note that C_1 and C_2 are square matrices.

For a block-diagonal matrix A with $A_{12} = 0$ and $A_{21} = 0$ the determinant is $\det(A) = \det(A_{11}) \det(A_{22})$.

Determinants have a multiplicative property,

$$\det(AB) = \det(BA) = \det(A) \det(B).$$

For scalar a this becomes $\det(aB) = a^d \det(B)$ where d is the dimension of B .

Another important identity is

$$\det(I_n + AB) = \det(I_m + BA)$$

where A is a $n \times m$ and B is a $m \times n$ matrix. This is called the Weinstein-Aronszajn determinant identity (also credited to Sylvester).

A.4 Matrix inverse

A.4.1 Inversion of square matrix

If A is a square matrix then the inverse matrix A^{-1} is a matrix such that

$$A^{-1}A = AA^{-1} = I.$$

Only non-singular matrices with $\det(A) \neq 0$ are invertible.

As $\det(A^{-1}A) = \det(I) = 1$ the determinant of the inverse matrix equals the inverse determinant,

$$\det(A^{-1}) = \det(A)^{-1}.$$

The transpose of the inverse is the inverse of the transpose as

$$\begin{aligned}(A^{-1})^T &= (A^{-1})^T A^T (A^T)^{-1} \\ &= (AA^{-1})^T (A^T)^{-1} = (A^T)^{-1}.\end{aligned}$$

The inverse of a matrix product $(AB)^{-1} = B^{-1}A^{-1}$ is the product of the individual matrix inverses in reverse order.

There are many different algorithms to compute the inverse of a matrix (which is essentially a problem of solving a system of equations). The computational complexity of matrix inversion is of the order $O(d^3)$ where d is the dimension of A . Therefore matrix inversion is very costly in higher dimensions.

Example A.1. Inversion of a 2×2 matrix:

The inverse of the matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

A.4.2 Inversion of structured matrices

However, for specially structured matrices inversion can be done effectively:

- The inverse of a **diagonal matrix** is another diagonal matrix obtained by inverting the diagonal elements.
- More generally, the inverse of a **block-diagonal matrix** is obtained by individually inverting the blocks along the diagonal.

The **Woodbury matrix identity** simplifies the inversion of matrices that can be written as $A + UBV$ where A and B are both square and U and V are suitable rectangular matrices:

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Typically, the inverse A^{-1} is either already known or can be easily obtained and the dimension of B is much lower than that of A .

The class of matrices that can be most easily inverted are **orthogonal matrices** whose inverse is obtained simply by transposing the matrix.

A.5 Orthogonal matrices

A.5.1 Properties

An orthogonal matrix Q is a square matrix with the property that $Q^T = Q^{-1}$, i.e. the transpose is also the inverse. This implies that $QQ^T = Q^TQ = I$.

Both the column and the row vectors in Q all have length 1. This implies that each element q_{ij} of Q can only take a value in the interval $[-1, 1]$.

The identity matrix I is the simplest example of an orthogonal matrix.

The squared Euclidean and Frobenius norm is preserved when a vector \mathbf{a} or matrix A is multiplied with an orthogonal matrix Q :

$$\|Q\mathbf{a}\|_2^2 = (Q\mathbf{a})^T Q\mathbf{a} = \mathbf{a}^T \mathbf{a} = \|\mathbf{a}\|_2^2$$

and

$$\|QA\|_F^2 = \text{Tr}((QA)^T QA) = \text{Tr}(A^T A) = \|A\|_F^2$$

Multiplication of Q with a vector results in a new vector of the same length but with a change in direction (unless $Q = I$). An orthogonal matrix Q can thus be interpreted geometrically as an operator performing rotation, reflection and/or permutation.

The product $Q_3 = Q_1 Q_2$ of two orthogonal matrices Q_1 and Q_2 yields another orthogonal matrix as $Q_3 Q_3^T = Q_1 Q_2 (Q_1 Q_2)^T = Q_1 Q_2 Q_2^T Q_1^T = I$.

The determinant $\det(Q)$ of an orthogonal matrix is either +1 or -1, because $QQ^T = I$ and thus $\det(Q) \det(Q^T) = \det(Q)^2 = \det(I) = 1$.

The set of all orthogonal matrices of dimension d together with multiplication form a group called the orthogonal group $O(d)$. The subset of orthogonal matrices with $\det(Q) = 1$ are called rotation matrices and form with multiplication the special orthogonal group $SO(d)$. Orthogonal matrices with $\det(Q) = -1$ are rotation-reflection matrices.

A.5.2 Generating orthogonal matrices

In two dimensions ($d = 2$) all orthogonal matrices R representing rotations with $\det(R) = 1$ are given by

$$R(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

and those representing rotation-reflections G with $\det(G) = -1$ by

$$G(\theta) = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix}.$$

Every orthogonal matrix of dimension $d = 2$ can be represented as the product of at most two rotation-reflection matrices because

$$R(\theta) = G(\theta) G(0) = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Thus, the matrix G is a generator of two-dimensional orthogonal matrices. Note that $G(\theta)$ is symmetric, orthogonal and has determinant -1.

More generally, and applicable in arbitrary dimension, the role of generator is taken by the Householder reflection matrix

$$\mathbf{Q}_{HH}(\mathbf{v}) = \mathbf{I} - 2\mathbf{v}\mathbf{v}^T$$

where \mathbf{v} is a vector of unit length (with $\mathbf{v}^T\mathbf{v} = 1$) orthogonal to the reflection hyperplane. Note that $\mathbf{Q}_{HH}(\mathbf{v}) = \mathbf{Q}_{HH}(-\mathbf{v})$. By construction the matrix $\mathbf{Q}_{HH}(\mathbf{v})$ is symmetric, orthogonal and has determinant -1.

It can be shown that any d -dimensional orthogonal matrix \mathbf{Q} can be represented as the product of at most d Householder reflection matrices. The two-dimensional generator $G(\theta)$ is recovered as the Householder matrix $\mathbf{Q}_{HH}(\mathbf{v})$ with $\mathbf{v} = \begin{pmatrix} -\sin \frac{\theta}{2} \\ \cos \frac{\theta}{2} \end{pmatrix}$ or $\mathbf{v} = \begin{pmatrix} \sin \frac{\theta}{2} \\ -\cos \frac{\theta}{2} \end{pmatrix}$.

A.5.3 Permutation matrix

A special type of an orthogonal matrix is a permutation matrix \mathbf{P} created by permuting rows and/or columns of the identity matrix \mathbf{I} . Thus, each row and column of \mathbf{P} contains exactly one entry of 1, but not necessarily on the diagonal.

If a permutation matrix \mathbf{P} is multiplied with a matrix \mathbf{A} it acts as an operator permuting the columns (\mathbf{AP}) or the rows (\mathbf{PA}). For a set of d elements there exist $d!$ permutations. Thus, for dimension d there are $d!$ possible permutation matrices (including the identity matrix).

The determinant of a permutation matrix is either +1 or -1. The product of two permutation matrices yields another permutation matrix.

Symmetric permutation matrices correspond to self-inverse permutations (i.e. the permutation matrix is its own inverse), and are also called permutation involutions. They can have determinant +1 and -1.

A transposition is a permutation where only two elements are exchanged. Thus, in a transposition matrix \mathbf{T} exactly two rows and/or columns are exchanged compared to identity matrix \mathbf{I} . Transpositions are self-inverse, and transposition matrices are symmetric. There are $\frac{d(d-1)}{2}$ different transposition matrices. The determinant of a transposition matrix is $\det(\mathbf{T}) = -1$.

Note that the transposition matrix is an instance of a Householder matrix $\mathbf{Q}_{HH}(\mathbf{v})$ with vector \mathbf{v} filled with zeros except for two elements that have value $\frac{\sqrt{2}}{2}$ and $-\frac{\sqrt{2}}{2}$.

Any permutation of d elements can be generated by a series of at most $d - 1$ transpositions. Correspondingly, any permutation matrix \mathbf{P} can be constructed by multiplication of the identity matrix with at most $d - 1$ transposition matrices. If the number of transpositions is even then $\det(\mathbf{P}) = 1$ otherwise for an uneven number $\det(\mathbf{P}) = -1$. This is called the sign or signature of the permutation.

The set of all permutations form the symmetric group S_d , the subset of even permutations (with positive sign and $\det(P) = 1$) the alternating group A_d .

A.6 Eigenvalues and eigenvectors

A.6.1 Definition

Assume a square symmetric matrix A of size $d \times d$. A vector $\mathbf{u} \neq 0$ is called an eigenvector of the matrix A and λ the corresponding eigenvalue if

$$A\mathbf{u} = \mathbf{u}\lambda.$$

A.6.2 Finding eigenvalues and vectors

To find the eigenvalues and eigenvector the *eigenequation* is rewritten as

$$(A - I\lambda)\mathbf{u} = 0.$$

For any solution $\mathbf{u} \neq 0$ the corresponding eigenvalue λ must make the matrix $A - I\lambda$ singular, i.e. the determinant must vanish

$$\det(A - I\lambda) = 0.$$

This is the *characteristic equation* of the matrix A , and its solution yields d not necessarily distinct and also potentially complex eigenvalues $\lambda_1, \dots, \lambda_d$.

If there are complex eigenvalues, for a real matrix those eigenvalues come in conjugate pairs. Specifically, for a complex $\lambda_1 = re^{i\phi}$ there will also be a corresponding complex eigenvalue $\lambda_2 = re^{-i\phi}$.

Given the eigenvalues we then solve the eigenequation for the corresponding non-zero eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_d$. Note that eigenvectors of real matrices can have complex components. Also the eigenvector is only defined by the eigenequation up to a scalar. By convention eigenvectors are therefore typically standardised to unit length but this still leaves a sign ambiguity for real eigenvectors and implies that complex eigenvectors are defined only up to a factor with modulus 1.

A.6.3 Eigenequation in matrix notation

With the matrix

$$\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$$

containing the standardised eigenvectors in the columns and the diagonal matrix

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}$$

containing the eigenvalues (typically sorted in order of magnitude) the eigenvalue equation can be written as

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}.$$

Note that if eigenvalues are not in order, we can always apply a permutation matrix \mathbf{P} to sort them in order, such that $\mathbf{U}' = \mathbf{U}\mathbf{P}$ reorders the eigenvectors and $\mathbf{\Lambda}' = \mathbf{P}^T \mathbf{\Lambda} \mathbf{P}$ the eigenvalues, with

$$\mathbf{A}\mathbf{U}' = \mathbf{A}\mathbf{U}\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{P} = \mathbf{U}\mathbf{P}\mathbf{P}^T \mathbf{\Lambda} \mathbf{P} = \mathbf{U}' \mathbf{\Lambda}'.$$

A.6.4 Defective matrix

In most cases the eigenvectors \mathbf{u}_i will be linearly independent so that they form a basis to span a d dimensional space.

However, if this is not the case and the matrix \mathbf{A} does not have a complete basis of eigenvectors, then the matrix is called defective. In this case the matrix \mathbf{U} containing the eigenvectors is singular and $\det(\mathbf{U}) = 0$.

An example of a defective matrix is $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ which has determinant 1 so that it can be inverted and its column vectors do form a complete basis but has only one distinct eigenvector $(1, 0)^T$ so that the eigenvector basis is incomplete.

A.6.5 Eigenvalues of a diagonal or triangular matrix

In the special case that \mathbf{A} is diagonal or a triangular matrix the eigenvalues are easily determined. This follows from the simple form of their determinants as the product of the diagonal elements. Hence for these matrices the characteristic equation becomes $\prod_i^d (a_{ii} - \lambda) = 0$ and has solution $\lambda_i = a_{ii}$, i.e. the eigenvalues are equal to the diagonal elements.

A.6.6 Eigenvalues and vectors of a symmetric matrix

If \mathbf{A} is symmetric, i.e. $\mathbf{A} = \mathbf{A}^T$, then its eigenvalues and eigenvectors have special properties:

- i) all eigenvalues of \mathbf{A} are real,
- ii) the eigenvectors are orthogonal, i.e. $\mathbf{u}_i^T \mathbf{u}_j = 0$ for $i \neq j$, and real. Thus, the matrix \mathbf{U} containing the standardised orthonormal eigenvectors is orthogonal.
- iii) \mathbf{A} is never defective as \mathbf{U} forms a complete basis.

A.6.7 Eigenvalues of orthogonal matrices

The eigenvalues of an orthogonal matrix Q are not necessarily real but they all have modulus 1 and lie on the unit circle. Thus, the eigenvalues of Q all have the form $\lambda = e^{i\phi} = \cos \phi + i \sin \phi$.

In any real matrix complex eigenvalues come in conjugate pairs. Hence if an orthogonal matrix Q has the complex eigenvalue $e^{i\phi}$ it also has a complex eigenvalue $e^{-i\phi} = \cos \phi - i \sin \phi$. The product of these two conjugate eigenvalues is 1. Thus, an orthogonal matrix of uneven dimension has at least one real eigenvalue (+1 or -1).

The eigenvalues of a Householder matrix $Q_{HH}(v)$ are all real (recall that it is symmetric!). In fact, in dimension d its eigenvalues are -1 (one time) and 1 ($d - 1$ times). Since a transposition matrix T is a special Householder matrix they have the same eigenvalues.

A.6.8 Positive definite matrices

If all eigenvalues of a square matrix A are real and $\lambda_i \geq 0$ then A is called *positive semi-definite*. If all eigenvalues are strictly positive $\lambda_i > 0$ then A is called *positive definite*.

Note that a matrix does not need to be symmetric to be positive definite, e.g. $\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}$ has positive eigenvalues 5 and 1. It also has a complete set of eigenvectors and is diagonalisable.

A symmetric matrix A is positive definite if the quadratic form $x^T A x > 0$ for any non-zero x , and it is positive semi-definite if $x^T A x \geq 0$. This holds also the other way around: a symmetric positive definite matrix (with positive eigenvalues) has a positive quadratic form, and a symmetric positive semi-definite matrix (with non-negative eigenvalues) a non-negative quadratic form.

A symmetric positive definite matrix always has a positive diagonal (this can be seen by setting x above to a unit vector with 1 at a single position, and 0 at all other elements). However, just requiring a positive diagonal is too weak to ensure positive definiteness of a symmetric matrix, for example $\begin{pmatrix} 1 & 10 \\ 10 & 1 \end{pmatrix}$ has a negative eigenvalue of -9. On the other hand, a symmetric matrix is indeed positive definite if it is strictly diagonally dominant, i.e. if all its diagonal elements are positive and are larger than the absolute value of any of the corresponding row or column elements. However, diagonal dominance is too restrictive as criterion to characterise all symmetric positive definite matrices, since there are many symmetric matrices that are positive definite but not diagonally dominant, such as $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$.

Finally, the sum of a symmetric positive semi-definite matrix A and a symmetric

positive definite matrix B is itself symmetric positive definite because the corresponding quadratic form $x^T(A+B)x = x^T Ax + x^T Bx > 0$ is positive. Similarly, the sum of two symmetric positive (semi)-definite matrices is itself symmetric positive (semi)-definite.

A.7 Matrix decompositions

A.7.1 Diagonalisation and eigenvalue decomposition

If A is a square non-defective matrix then U is invertible and we can rewrite the eigenvalue equation to

$$A = U\Lambda U^{-1}.$$

This is called the eigendecomposition, or spectral decomposition, of A and equivalently

$$\Lambda = U^{-1}AU$$

is the diagonalisation of A . Thus any matrix A that is not defective is diagonalisable using eigenvalue decomposition.

A.7.2 Orthogonal eigenvalue decomposition

For symmetric A this becomes

$$A = U\Lambda U^T$$

with real eigenvalues and orthogonal matrix U and

$$\Lambda = U^T A U.$$

This special case is known as the orthogonal diagonalisation of A .

The orthogonal decomposition for symmetric A is unique apart from the signs of the eigenvectors. In order to make it fully unique one needs to impose further restrictions (e.g. require a positive diagonal of U). Note that this can be particularly important in computer application where the sign can vary depending on the specific implementation of the underlying numerical algorithms.

A.7.3 Singular value decomposition

The **singular value decomposition** (SVD) is a generalisation of the orthogonal eigenvalue decomposition for symmetric matrices.

Any (!) rectangular matrix A of size $n \times d$ can be factored into the product

$$A = UDV^T$$

where U is a $n \times n$ orthogonal matrix, V is a second $d \times d$ orthogonal matrix and D is a diagonal but rectangular matrix of size $n \times d$ with $m = \min(n, d)$ real

diagonal elements d_1, \dots, d_m . The d_i are called singular values, and appear along the diagonal in D by order of magnitude.

The SVD is unique apart from the signs of the columns vectors in U , V and D (you can freely specify the column signs of any two of the three matrices). By convention the signs are chosen such that the singular values in D are all non-negative, which leaves ambiguity in columns signs of U and V . Alternatively, one may fix the columns signs of U and V , e.g. by requiring a positive diagonal, which then determines the sign of the singular values (thus allowing for negative singular values as well).

If A is symmetric then the SVD and the orthogonal eigenvalue decomposition coincide (apart from different sign conventions for singular values, eigenvalues and eigenvectors).

Since $A^T A = V D^T D V^T$ and $A A^T = U D D^T U^T$ the squared singular values correspond to the eigenvalues of $A^T A$ and $A A^T$. It also follows that $A^T A$ and $A A^T$ are both positive semi-definite symmetric matrices, and that V and U contain the respective sets of eigenvectors.

A.7.4 Polar decomposition

Any square matrix A can be factored into the product

$$A = QB$$

of an orthogonal matrix Q and a symmetric positive semi-definite matrix B .

This follows from the SVD of A given as

$$\begin{aligned} A &= U D V^T \\ &= (U V^T) (V D V^T) \\ &= Q B \end{aligned}$$

with non-negative D . Note that this decomposition is unique as the sign ambiguities in the columns of U and V cancel out in Q and B .

A.7.5 Cholesky decomposition

A symmetric positive definite matrix A can be decomposed into a product of a triangular matrix L with its transpose

$$A = L L^T.$$

Here, L is a lower triangular matrix with positive diagonal elements.

This decomposition is unique and is called **Cholesky factorisation**. It is often used to check whether a symmetric matrix is positive definite as it is algorithmically less demanding than eigenvalue decomposition.

Note that some implementations of the Cholesky decomposition (e.g. R) use upper triangular matrices K with positive diagonal so that $A = K^T K$ and $L = K^T$.

A.8 Matrix summaries based on eigenvalues and singular values

A.8.1 Trace and determinant computed from eigenvalues

The eigendecomposition $A = U\Lambda U^{-1}$ allows to establish a link between trace and determinant and eigenvalues.

Specifically,

$$\begin{aligned}\text{Tr}(A) &= \text{Tr}(U\Lambda U^{-1}) = \text{Tr}(\Lambda U^{-1}U) \\ &= \text{Tr}(\Lambda) = \sum_{i=1}^d \lambda_i\end{aligned}$$

thus the trace of a square matrix A is equal to the *sum* of its eigenvalues. Likewise,

$$\begin{aligned}\det(A) &= \det(U) \det(\Lambda) \det(U^{-1}) \\ &= \det(\Lambda) = \prod_{i=1}^d \lambda_i\end{aligned}$$

therefore the determinant of A is the *product* of the eigenvalues.

The relationship between eigenvalues and the trace and the determinant is demonstrated here for diagonalisable non-defective matrices. However, it does hold also in general for any matrix. This can be shown by using certain non-diagonal matrix decompositions (e.g. Jordan decomposition).

As a result, if any of the eigenvalues is equal to zero then $\det(A) = 0$ and as hence A is singular and not invertible.

The trace and determinant of a real matrix are always real even though the individual eigenvalues may be complex.

A.8.2 Rank and condition number

The rank is the dimension of the space spanned by both the column and row vectors. A rectangular matrix of dimension $n \times d$ will have rank of at most $m = \min(n, d)$, and if the maximum is indeed achieved then it has full rank.

The condition number describes how well- or ill-conditioned a full rank matrix is. For example, for a square matrix a large condition number implies that

the matrix is close to being singular and thus ill-conditioned. If the condition number is infinite then the matrix is not full rank.

The rank and condition of a matrix can both be determined from the m singular values d_1, \dots, d_m of a matrix obtained by SVD:

- i) The rank is number of non-zero singular values.
- ii) The condition number is the ratio of the largest singular value divided by the smallest singular value (absolute values if signs are allowed).

If a square matrix A is singular then the condition number is infinite, and it will not have full rank. On the other hand, a non-singular square matrix, such as a positive definite matrix, has full rank.

A.9 Functions of symmetric matrices

We focus on *symmetric* square matrices $A = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ which are always diagonalisable with real eigenvalues $\mathbf{\Lambda}$ and orthogonal eigenvectors \mathbf{U} .

A.9.1 Definition of a matrix function

Assume a real-valued function $f(a)$ of a real number a . Then the corresponding matrix function $f(A)$ is defined as

$$f(A) = \mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^T = \mathbf{U} \begin{pmatrix} f(\lambda_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & f(\lambda_d) \end{pmatrix} \mathbf{U}^T$$

where the function $f(a)$ is applied to the eigenvalues of A . By construction $f(A)$ is real, symmetric and has real eigenvalues $f(\lambda_i)$.

Examples:

Example A.2. Matrix power: $f(a) = a^p$ (with p a real number)

Special cases of matrix power include :

- Matrix inversion: $f(a) = a^{-1}$
Note that if the matrix A is singular, i.e. contains one or more eigenvalues $\lambda_i = 0$, then A^{-1} is not defined and therefore A is not invertible.

However, a so-called *pseudoinverse* can still be computed, by inverting the non-zero eigenvalues, and keeping the zero eigenvalues as zero.

- Matrix square root: $f(a) = a^{1/2}$
Since there are multiple solutions to the square root there are also multiple matrix square roots. The principal matrix square root is obtained by using the positive square roots of all the eigenvalues. Thus the **principal matrix square root** of a positive semi-definite matrix is also positive semi-definite and it is unique.

Example A.3. Matrix exponential: $f(a) = \exp(a)$

Note that because $\exp(a) \geq 0$ for all real a the matrix $\exp(A)$ is positive semi-definite. Thus, the matrix exponential can be used to generate positive semi-definite matrices.

If A and B commute, i.e. if $AB = BA$, then $\exp(A + B) = \exp(A)\exp(B)$. However, this is not the case otherwise!

Example A.4. Matrix logarithm: $f(a) = \log(a)$

As the logarithm requires $a > 0$ the matrix A needs to be positive definite for $\log(A)$ to be defined.

A.9.2 Identities for the matrix exponential and logarithm

The above give rise to useful identities:

- 1) For any symmetric matrix A we have

$$\det(\exp(A)) = \exp(\text{Tr}(A))$$

because $\prod_i \exp(\lambda_i) = \exp(\sum_i \lambda_i)$ where λ_i are the eigenvalues of A .

- 2) If we take the logarithm on both sides and replace $\exp(A) = B$ we get another identity for a symmetric positive definite matrix B :

$$\log \det(B) = \text{Tr}(\log(B))$$

because $\log(\prod_i \lambda_i) = \sum_i \log(\lambda_i)$ where λ_i are the eigenvalues of B .

A.10 Matrix calculus

A.10.1 First order vector derivatives

A.10.1.1 Gradient

The **nabla operator** (also known as **del operator**) is the *row* vector

$$\nabla = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_d} \right) = \frac{\partial}{\partial \mathbf{x}}$$

containing the first order partial derivative operators.

The **gradient** of a scalar-valued function $f(\mathbf{x})$ with vector argument $\mathbf{x} = (x_1, \dots, x_d)^T$ is also a *row* vector (with d columns) and can be expressed using the nabla operator

$$\begin{aligned} \nabla f(\mathbf{x}) &= \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_d} \right) \\ &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \text{grad} f(\mathbf{x}). \end{aligned}$$

Note the various notations for the gradient.

Example A.5. $f(x) = \mathbf{a}^T x + b$. Then $\nabla f(x) = \frac{\partial f(x)}{\partial x} = \mathbf{a}^T$.

Example A.6. $f(x) = x^T x$. Then $\nabla f(x) = \frac{\partial f(x)}{\partial x} = 2x^T$.

Example A.7. $f(x) = x^T A x$. Then $\nabla f(x) = \frac{\partial f(x)}{\partial x} = x^T (A + A^T)$.

A.10.1.2 Jacobian matrix

For a vector-valued function

$$f(x) = (f_1(x), \dots, f_m(x))^T.$$

the computation of the gradient of each component yields the **Jacobian matrix** (with m rows and d columns)

$$\begin{aligned} Df(x) &= \begin{pmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_m(x) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial f_i(x)}{\partial x_j} \end{pmatrix} \\ &= \frac{\partial f(x)}{\partial x} \end{aligned}$$

Again, note the various notations for the Jacobian matrix!

Example A.8. $f(x) = Ax + b$. Then $Df(x) = \frac{\partial f(x)}{\partial x} = A$.

If $m = d$ then the Jacobian matrix is a square matrix and this allows to compute the **Jacobian determinant**

$$\det Df(x) = \det \left(\frac{\partial f(x)}{\partial x} \right)$$

If $y = f(x)$ is an invertible function with $x = f^{-1}(y)$ then the Jacobian matrix is invertible and the inverted matrix is in fact the Jacobian of the inverse function!

This allows to compute the Jacobian determinant of the backtransformation as the inverse of the Jacobian determinant the original function:

$$\det Df^{-1}(y) = (\det Df(x))^{-1}$$

or in alternative notation

$$\det Dx(y) = \frac{1}{\det Dy(x)}$$

A.10.2 Second order vector derivatives

The matrix of all second order partial derivatives of scalar-valued function with vector-valued argument is called the **Hessian matrix** and is computed by double application of the nabla operator:

$$\begin{aligned}\nabla^T \nabla f(\mathbf{x}) &= \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \dots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} & \dots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2} \end{pmatrix} \\ &= \left(\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j} \right) = \left(\frac{\partial}{\partial \mathbf{x}} \right)^T \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}.\end{aligned}$$

By construction it is square and symmetric.

A.10.3 First order matrix derivatives

The derivative of a scalar-valued function $f(\mathbf{X})$ with regard to a matrix argument \mathbf{X} can also be defined and results in a matrix with transposed dimensions compared to \mathbf{X} .

Two important specific examples are:

Example A.9. $\frac{\partial \text{Tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}$

Example A.10. $\frac{\partial \log \det(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial \text{Tr}(\log \mathbf{X})}{\partial \mathbf{X}} = \mathbf{X}^{-1}$

Appendix B

Further study

In this module we can only touch the surface of the field of multivariate statistics and machine learning. If you would like to study further I recommend the following books below as a starting point.

B.1 Recommended reading

For multivariate statistics and machine learning:

- Härdle and Simar (2015) *Applied multivariate statistical analysis. 4th edition.* Springer.
- Marden (2015) *Multivariate Statistics: Old School*
- Rogers and Girolami (2017) *A first course in machine learning (2nd Edition).* Chapman and Hall / CRC.
- James et al. (2021) *An introduction to statistical learning with applications in R (2nd edition).* Springer.

B.2 Advanced reading

Additional (advanced) reference books for probabilistic machine learning are:

- Bishop (2006) *Pattern recognition and machine learning.* Springer.
- Hastie, Tibshirani, and Friedman (2009) *The elements of statistical learning: data mining, inference, and prediction.* Springer.
- Murphy (2012) *Machine learning: a probabilistic perspective.* MIT Press.

You can find further suggestions on my list of [online textbooks in statistics and machine learning](#).

Bibliography

- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer. <https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/>.
- Härdle, W. K., and L. Simar. 2015. *Applied Multivariate Statistical Analysis*. Berlin: Springer.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer. <https://web.stanford.edu/~hastie/ElemStatLearn/>.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer. <https://www.statlearning.com>.
- . 2021. *An Introduction to Statistical Learning with Applications in R*. 2nd ed. Springer. <https://www.statlearning.com>.
- Marden, J. I. 2015. *Multivariate Statistics: Old School*. CreateSpace. <http://stat.istics.net/Multivariate>.
- Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Rogers, S., and M. Girolami. 2017. *A First Course in Machine Learning*. 2nd ed. Chapman; Hall / CRC.
- Zhang, A., Z. C. Lipton, M. Li, and A. J. Smola. 2020. *Dive into Deep Learning*. <https://d2l.ai>.