

# **Statistics 2: Statistical Learning with Likelihood and Bayes**

MATH27720 Semester 2

Korbinian Strimmer

10 January 2026



# Table of contents

<b>Welcome</b>	<b>1</b>
Updates . . . . .	1
License . . . . .	1
<b>Preface</b>	<b>2</b>
About the author . . . . .	2
About the module . . . . .	2
Acknowledgements . . . . .	3
<b>Prerequisites</b>	<b>4</b>
Matrices and calculus . . . . .	4
Combinatorics, probability and distributions . . . . .	5
Statistics . . . . .	5
<b>I Models</b>	<b>6</b>
<b>1 Statistical learning</b>	<b>7</b>
1.1 What is statistics? . . . . .	7
1.2 Probabilistic underpinnings . . . . .	8
1.3 Model-based learning . . . . .	9
1.4 Further reading . . . . .	13
<b>2 Distributions</b>	<b>15</b>
2.1 Characteristic features . . . . .	15
2.2 Commonly used distribution families . . . . .	16
2.3 Model building . . . . .	18
2.4 ► Exponential families . . . . .	20
2.5 Further reading . . . . .	25
<b>II Information</b>	<b>26</b>
<b>3 Entropy</b>	<b>27</b>
3.1 Code length and log-loss . . . . .	27
3.2 Information entropy . . . . .	30

3.3 ► Entropy in physics . . . . .	35
3.4 Further reading . . . . .	38
<b>4 Risk and divergence</b>	<b>39</b>
4.1 Risk . . . . .	39
4.2 Divergence . . . . .	43
4.3 Further reading . . . . .	47
<b>5 Local divergence</b>	<b>49</b>
5.1 Fisher information . . . . .	49
5.2 ► Reparametrisation . . . . .	54
5.3 Further reading . . . . .	57
<b>6 Maximum entropy</b>	<b>58</b>
6.1 Information entropy with prior measure . . . . .	58
6.2 Maximum entropy principle to characterise distributions . . . . .	62
6.3 Further reading . . . . .	64
<b>III Likelihood</b>	<b>66</b>
<b>7 Principle of maximum likelihood</b>	<b>67</b>
7.1 Overview . . . . .	67
7.2 From minimum KL divergence to maximum likelihood . . . . .	69
7.3 Properties of maximum likelihood estimation . . . . .	71
7.4 Maximum likelihood estimation for regular models . . . . .	74
7.5 Further reading . . . . .	76
<b>8 Maximum likelihood estimation in practice</b>	<b>77</b>
8.1 Likelihood estimation for a single parameter . . . . .	77
8.2 Likelihood estimation for multiple parameters . . . . .	82
8.3 Further properties of ML . . . . .	86
<b>9 Observed Fisher information</b>	<b>89</b>
9.1 Definition of the observed Fisher information . . . . .	89
9.2 Observed Fisher information examples . . . . .	92
<b>10 Quadratic approximation and normal asymptotics</b>	<b>96</b>
10.1 Approximate distribution of maximum likelihood estimates . . . . .	96
10.2 Quantifying the uncertainty of maximum likelihood estimates . . . . .	102
10.3 Normal confidence intervals and tests . . . . .	105



<b>11 Likelihood-based confidence interval and likelihood ratio</b>	<b>109</b>
11.1 Likelihood-based confidence intervals and Wilks statistic	109
11.2 Generalised likelihood ratio test (GLRT)	118
<b>12 Optimality properties and conclusion</b>	<b>123</b>
12.1 Properties of maximum likelihood encountered so far	123
12.2 Summarising data and the concept of (minimal) sufficiency	124
12.3 Concluding remarks on maximum likelihood	127
<b>IV Bayesian statistics</b>	<b>131</b>
<b>13 Conditioning and Bayes rule</b>	<b>132</b>
13.1 Conditional probability	132
13.2 Bayes' theorem	133
13.3 Conditional mean and variance	133
13.4 Conditional entropy and entropy chain rules	135
13.5 Entropy bounds for the marginal variables	136
<b>14 Models with latent variables and missing data</b>	<b>138</b>
14.1 Complete data log-likelihood versus observed data log-likelihood	138
14.2 Estimation of the unobservable latent states using Bayes theorem	140
14.3 EM Algorithm	141
<b>15 Essentials of Bayesian statistics</b>	<b>143</b>
15.1 Principle of Bayesian learning	143
15.2 Some background on Bayesian statistics	149
<b>16 Bayesian learning in practice</b>	<b>154</b>
16.1 Estimating a proportion using the beta-binomial model	154
16.2 Properties of Bayesian learning	159
16.3 Estimating the mean using the normal-normal model	163
16.4 Estimating the variance using the Inverse-Wishart-normal model	165
16.5 Estimating the precision using the Wishart-normal model	168
<b>17 Bayesian model comparison</b>	<b>171</b>
17.1 Marginal likelihood as model likelihood	171
17.2 The Bayes factor for comparing two models	173
17.3 Approximate computations	175
17.4 Bayesian testing using false discovery rates	177

<b>18 Choosing priors in Bayesian analysis</b>	<b>182</b>
18.1 Choosing a prior	182
18.2 Default priors or uninformative priors	183
18.3 Empirical Bayes	185
<b>19 Optimality properties and summary</b>	<b>188</b>
19.1 Bayesian statistics in a nutshell	188
19.2 Optimality of Bayesian inference	190
19.3 Connection with entropy learning	191
19.4 Conclusion	192
<b>Bibliography</b>	<b>193</b>
<b>Appendices</b>	<b>195</b>
<b>A Statistics refresher</b>	<b>195</b>
A.1 Data and statistics as functions of data	195
A.2 Statistical learning	195
A.3 Sampling properties of a point estimator	196
A.4 Efficiency and consistency of an estimator	197
A.5 Empirical distribution	198
A.6 Law of large numbers	199
A.7 Empirical estimators	200
A.8 Sampling distribution of mean and variance estimators for normal data	201
A.9 $t$ -statistics	202
A.10 Confidence intervals	205
<b>B Further study</b>	<b>209</b>
B.1 Recommended reading	209
B.2 Additional references	209



# Welcome

These are the lecture notes for MATH27720 Statistics 2, a course for second year mathematics students at the [Department of Mathematics of the University of Manchester](#).

The course text was written by [Korbinian Strimmer](#) from 2023–2026. This version is from 10 January 2026.

If you have any questions, comments, or corrections please get in touch!<sup>1</sup>

## Updates

The lecture notes will be updated from time to time.

The most current version is found at the web page for the

- [online MATH27720 Statistics 2 lecture notes](#).

There you can also download the MATH27720 Statistics 2 lecture notes as

- [PDF in A4 format for printing](#) (double page layout), or as
- [6x9 inch PDF for use on tablets](#) (single page layout).

## License

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

---

<sup>1</sup>Email address: [korbinian.stimmer@manchester.ac.uk](mailto:korbinian.stimmer@manchester.ac.uk)



# Preface

## About the author

Hello! My name is Korbinian Strimmer and I am a Professor in Statistics. I am a member 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](#).

## About the module

The notes are intended for the MATH27720 Statistics 2 course being taught in spring 2026 at the University of Manchester.

The MATH27720 Statistics 2 module is designed to run over the course of 10 weeks. It has the following two part structure:

1. Models, information and likelihood (W1–W5)
2. Bayesian statistics (W6–W10)

This module emphasises conceptual understanding and practical methods rather than theoretical aspects. You will specifically explore the foundations of statistical learning through likelihood and Bayesian approaches, as well as the role of entropy in supporting these concepts.

The presentation in this course is non-technical and most sections and examples will be easily accessible for a year 2 mathematics student. Sections and examples marked ► are conceptually or technically more involved and can be skipped during a first reading.

If you are a student at University of Manchester enrolled in this module, you will find additional support material on [Canvas](#):

- a weekly learning plan,
- worksheets with examples and solutions (and R code), and
- exam papers of previous years.

Furthermore, a [MATH27720 online reading list](#) is hosted by the University of Manchester library.

### ! Aims of this module

In a nutshell, the two key aims of the MATH27720 Statistics 2 module are

- i) to provide a principled introduction to maximum likelihood and Bayesian statistical analysis and
- ii) to demonstrate that statistics offers a well founded and coherent theory of information, rather than just seemingly unrelated collections of “recipes” for data analysis.

The first part of the module (Weeks 1–5) we will explore the **method of maximum likelihood** both practically and more theoretically in terms of its foundations.

The second part of this module (Weeks 6–10) focuses on the **Bayesian approach to statistical estimation and inference** that can be viewed as a natural extension of likelihood-based statistical analysis that overcomes some of the limitations of maximum likelihood.

## Acknowledgements

These notes are based in part on my earlier notes for MATH20802 Statistical Methods which was last run in Spring 2023. Many thanks to [Beatriz Costa Gomes](#) for her help in creating the 2019 version of the lecture notes when I was teaching the MATH20802 module for the first time and to [Kristijonas Raudys](#) for his extensive feedback on the 2020 version.



## Prerequisites

Statistics is a mathematical science that requires practical working knowledge of a number of subfields in mathematics, most importantly in probability theory but also in vector and matrix algebra as well as in the calculus of functions of several variables.

The MATH27720 Statistics 2 module builds on the earlier mandatory probability and statistics modules offered by the University of Manchester such as MATH11712 Statistics 1 and MATH11711 Probability 1 as well as on MATH27720 Probability 2. Many students will also have attended the modules MATH20811 Practical Statistics and the MATH27711 Linear Regression Models.

Below you find a number of resources to help you to refresh your knowledge in these areas.

### Matrices and calculus

For a refresher of the essentials in matrix algebra and calculus please refer to the supplementary [Matrix and Calculus Refresher notes](#).

Below you find topics that are particularly relevant.

#### Vectors and matrices

- Vector and matrix notation
- Vector algebra
- Eigenvectors and eigenvalues, eigendecomposition
- Properties of special matrices (e.g. a real symmetric matrix has real eigenvalues)
- Positive and negative definiteness of a matrix (i.e. matrices with only positive or only negative eigenvalues)
- Matrix inverse

#### Functions

- Gradient vector
- Hessian matrix
- Conditions for a local extremum of a function
- Convex and concave functions
- Linear and quadratic approximation

#### Logarithms

In these notes the logarithmic function  $\log(x)$  will always refer to the *natural logarithm* when the base is not explicitly stated. Logarithms with base 2 and based 10 are denoted as  $\log_2(x)$  and  $\log_{10}(x)$ , respectively.

## Combinatorics, probability and distributions

For a detailed introduction of concepts in combinatorics and probability see the lecture notes for MATH11711 Probability 1 and MATH27720 Probability 2.

A review of the essentials can also be found in the supplementary [Probability and Distribution Refresher notes](#). These supplementary notes also provide details of probability distributions frequently employed in statistical analysis.

Chapter 2 briefly revisits the distributions relevant for this module.

#### Statistics

For a refresher of statistical concepts discussed in earlier statistics courses and required in this module see Appendix A.

For example Sahu (2024) offers a concise introduction to foundations in statistics relevant for Statistics 2.



# Part I

## Models

## 1 Statistical learning

Learning from data using probabilistic models lies at the heart of statistical learning. This chapter provides an overview over key foundational concepts.

### 1.1 What is statistics?

The following fundamental questions typically arise in any scientific data analysis:

- **Optimality:** How do we extract information from data as efficiently and accurately as possible?
- **Model fit:** How can we build models that accurately reflect the observed data?
- **Interpretability:** How can we construct models that reveal underlying mechanisms and remain understandable?
- **Prediction:** How do we use these models and information to make the best possible predictions?

**Statistics** is a mathematical science for reasoning about data and uncertainty. It employs **probabilistic models** to address the questions above, offering a principled framework for learning from data and extracting and for processing information under uncertainty in an optimal way. This includes model selection, approximation and inference.

**Machine learning** overlaps substantially with statistics. Rooted in computer science it frequently adopts an engineering-centric perspective and often emphasises algorithmic approaches. Some methods are non-probabilistic while many modern approaches adopt a statistical perspective.

**Data science** today comprises elements of both statistics and machine learning and brings together mathematics, computer science and domain-specific expertise (e.g. **biomedical data science**).



**Artificial intelligence** (AI) is a branch of computer science that makes substantial use of statistical and machine learning techniques, along with methods such as natural language processing and symbolic reasoning, to create systems that can generate and interpret highly complex data, including language and images, and make decisions to achieve goals.

## 1.2 Probabilistic underpinnings

### Randomness, probability and uncertainty

**Random** and **randomness** refer to unpredictable, non-deterministic outcomes or events. Equivalent, more technical terms are **stochastic** and **stochasticity**.

The degree of randomness (or uncertainty, see below) is quantified by the **probability**, or equivalently, by the **chance** of particular outcomes.

An interesting question is the source of the randomness. On a fundamental level, some phenomena are *intrinsically* random (e.g. radioactive decay, measurement outcomes in quantum theory). However, much *apparent* randomness arises from our ignorance of the underlying mechanisms. The process may be deterministic in principle but we treat it as random for convenience. For example, a coin flip is often considered random. However, in reality the outcome of a coin flip is fully determined by classical physics.

Randomness that is not intrinsic but stems from a lack of knowledge or understanding, is called **uncertainty**, and corresponding events are **uncertain**. Uncertainty generally decreases if more data and information or a better model is available.

### Probability theory versus statistics

It is important to recognise the distinct domains of probability theory and statistics.

On the one hand, **probability theory** provides the **mathematical underpinnings** of probability and chance (e.g. probability axioms, measure theory) and corresponding models for randomness and uncertainty (e.g. probability distributions, stochastic processes). Crucially, it is neutral about the sources of randomness and interpretations of probability, and may simply be viewed as **pure mathematics**.

On the other hand, **statistics** uses probabilistic approaches to **learn from observations**, thus linking real-world phenomena with mathematical models. Thus, it is a branch of **applied mathematics**. Importantly, statistics is concerned with **uncertainty** (e.g. about events, predictions, outcomes, model parameters) without assuming that the underlying process is actually random, and to make decisions and predictions under that uncertainty.

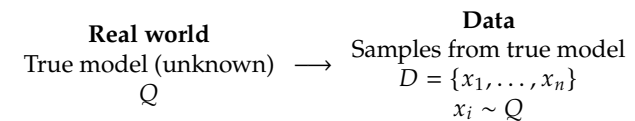
The link of statistics with the real world also leads to different interpretations of probability, with the two most common being the **frequentist interpretation** (ontological, “every probability is a long running frequency and exists independently from an observer”) and the **Bayesian interpretation** (epistemological, “probability is a degree of belief and represent the state of knowledge”), both to be discussed later.

## 1.3 Model-based learning

### Sketch of statistical learning

The aim of statistical learning is to use observed data in an optimal way to learn about the underlying mechanism of the data-generating process. Since data is typically finite but models can be in principle arbitrarily complex there may be issues of overfitting (insufficient data for the complexity of the model) but also underfitting (model is too simplistic).

We observe data  $D = \{x_1, \dots, x_n\}$  assumed to result from an underlying probabilistic model  $Q$ , the distribution for  $x$ :



The true model underlying the data-generating process is unknown and cannot be observed. However, what we can observe is data  $D$  arising from the true model  $Q$  by measuring properties of interest (our observations from experiments). Sometimes we can also perturb the model and see what the effect is (interventional study).

To explain the observed data  $D$ , and also to predict future data, we consider various competing hypotheses in the form of candidate models  $P_1, P_2, \dots$ . Often these candidates belong to a parametric family  $P(\theta)$

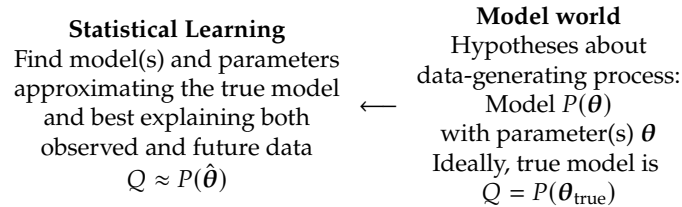


indexed by a parameter vector  $\theta$ . Individual models such as  $P(\theta_1)$  or  $P(\theta_2)$  correspond to specific parameter values  $\theta_1$  and  $\theta_2$ .

Frequently parameters are chosen such that they allow some interpretation, such as moments or other properties of the distribution. However, intrinsically parameters are just labels and may be changed by any one-to-one transformation.

For statistical learning it is important that models are **identifiable** within a family so that distinct parameters correspond to distinct distributions, as this allows models to be uniquely estimated from data. Specifically, if parameters are identifiable then  $P(\theta_1) = P(\theta_2)$  implies  $\theta_1 = \theta_2$ , and conversely if  $P(\theta_1) \neq P(\theta_2)$  then  $\theta_1 \neq \theta_2$ . If parameters and distributions are unique within a neighbourhood  $\theta_0 + \epsilon$  relative to a reference value  $\theta_0$ , rather than globally, they are **locally identifiable** at  $\theta_0$ .

The various candidate models  $P(\theta)$  in the **model world** will at best be good approximations to the true underlying data-generating model  $Q$ . In some cases the true model will be part of the model family, i.e. there exists a parameter  $\theta_{\text{true}}$  so that  $Q = P(\theta_{\text{true}})$ . However, more typically we cannot assume that the true underlying model is contained in the family. Nonetheless, even an imperfect candidate model will often provide a useful mathematical approximation and capture some important characteristics of the true model and thus will help to interpret the observed data.



**The aim of statistical learning is to identify the model(s) that explain the current data and also predict future data (i.e. outcomes of experiments that have not yet been conducted).**

Thus a good model provides a good fit to the current data (i.e. it explains current observations well) and also to the future data (i.e. it generalises well).

A large proportion of statistical theory is devoted to finding these “good” models that avoid both *overfitting* (models being too complex and not

generalising well) or *underfitting* (models being too simplistic and hence also not predicting well).

Typically, the aim is to find an approximating model whose **complexity** is well matched to that of the unknown true model and to the complexity of the observed data.

### Finding the best models

A core task in statistical learning is to identify those distributions that explain the existing data well and that also generalise well to future yet unseen observations.

In a **nonparametric setting** we may simply rely on the law of large numbers that implies that the empirical distribution  $\hat{Q}_n$  constructed from the observed data  $D$  converges to the true distribution  $Q$  if the sample size is large. We can therefore obtain an **empirical estimator**  $\hat{\theta}$  of the functional  $\theta = g(Q)$  by  $\hat{\theta} = g(\hat{Q}_n)$ , i.e. by substituting the true distribution with the empirical distribution. This allows us, e.g., to get the empirical estimate of the mean  $E_Q(x) = \mu$  by

$$\hat{E}(x) = \hat{\mu} = E_{\hat{Q}_n}(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

and of the variance  $\text{Var}(x) = \sigma^2 = E_Q((x - \mu)^2)$  by

$$\widehat{\text{Var}}(x) = \hat{\sigma}^2 = E_{\hat{Q}_n}((x - \hat{\mu})^2) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

simply by replacing the expectation with the sample average.

For parametric models we need to find estimates of the parameters that correspond to the distributions that best approximate the unknown true data-generating model. One such approach is provided by the **method of maximum likelihood**. More precisely, given a probability distribution  $P(\theta)$  with density or mass function  $p(x|\theta)$  where  $\theta$  is a parameter vector, and  $D = \{x_1, \dots, x_n\}$  are the observed iid data (i.e. independent and identically distributed), the **likelihood function** is defined as

$$L_n(\theta) = L(\theta|D) = \prod_{i=1}^n p(x_i|\theta)$$



The parameter value  $\hat{\theta}_{ML}$  that maximises the likelihood function for fixed data  $D$  is the **maximum likelihood estimate**:

$$\hat{\theta}_{ML} = \arg \max_{\theta} L_n(\theta)$$

Historically, the likelihood was introduced as “the probability of observing the data given a model with specified parameters  $\theta$ ”.

However, this view breaks down for continuous random variables, where the likelihood function is formed from densities rather than probabilities. Furthermore, for discrete random variables the likelihood function typically does not include the additional factor accounting for the permutations of samples that is needed to obtain the actual probability of the data.

Instead, we will see that estimation and inference in statistics are closely linked to information-theoretic concepts such as entropy and divergences between distributions. Specifically, maximising the likelihood function is, for large samples, equivalent to minimising the Kullback-Leibler divergence  $D_{KL}(Q, P(\theta))$  between the unknown true distribution  $Q$  and the model  $P(\theta)$ . As a consequence, the method of **maximum likelihood extends empirical estimation to parametric models**<sup>1</sup>. This connection explains both the optimality properties as well as the limitations of the maximum likelihood approach in statistical learning.

## Models and decomposition of uncertainty

When constructing statistical models we will often choose to explain some aspects of the uncertainty while intentionally ignoring others, to create simple yet effective models.

As a result, for any given model, uncertainty decomposes into the sum of

- i) **reducible uncertainty** (epistemic uncertainty): **explained by model**, and
- ii) **irreducible uncertainty** (aleatoric / residual / intrinsic uncertainty): **unexplained by model**.

<sup>1</sup>Conversely, empirical estimators are, in fact, also likelihood estimators based on an [empirical likelihood](#) function constructed from the empirical distribution.

Crucially, even with arbitrarily large amounts of data, the total uncertainty cannot always be eliminated fully, since the residual uncertainty depends on the employed model. Consequently, to reduce the unexplained uncertainty the model itself must be changed, but whether this is at all desirable is a different matter (taking into account model complexity, interpretability, etc.).

For example, in linear regression or classification, the decomposition of uncertainty is expressed by the law of total variance with decomposes **total variance** into **explained variance** (“signal”, between-group variance) and **unexplained variance** (“noise”, within-group variance). Additional data improves the accuracy of estimates of the residual and the explained variance but does not eliminate the model’s unexplained variance. To reduce the residual error you need to change the model, e.g. by adding further covariates.

Importantly, intrinsic uncertainty should not be confused with intrinsic randomness: the latter is a claim about nature (fundamental non-determinism), whereas the first is a property of the model (residual uncertainty). Hence, unexplained uncertainty does not imply intrinsic randomness.

## 1.4 Further reading

The book “Ten Great Ideas About Chance” by Diaconis and Skyrms (2018) provides an overview of the history and philosophical foundations of probability as well as discussing various interpretations of probability.

The book “Probability Theory: The Logic of Science” by Jaynes (2003) advocates that probability theory, as an extension of logic, is the natural framework for scientific reasoning.

The popular science book “The Theory That Would Not Die” (McGrayne 2011) focuses on the history of Bayes’ theorem and its importance in statistics.

For a quick recap of essential statistical concepts introduced in earlier statistical modules in year 1 and 2 see [Appendix A](#).



## 💡 A bit of history

Some important milestones in the development of learning from data are highlighted below:

- **Bayesian statistics** dates back to [Thomas Bayes \(1701–1761\)](#) and his famous posthumous [1763 essay](#) and was further developed by [Pierre-Simon Laplace \(1749–1827\)](#) in the early 19th century.
- **Maximum likelihood** was developed by [Ronald A. Fisher \(1890–1962\)](#) in the early 20th century, with a seminal paper published in 1922 ([Fisher 1922](#)).
- Links of statistical learning with **entropy** were established in the 1940s with roots going back to discoveries in statistical physics in the 1870s. The close link of physics and statistical learning has recently been underlined by the [2024 Nobel Prize in Physics](#) awarded to [John J. Hopfield \(1933–\)](#) and [Geoffrey E. Hinton \(1947–\)](#) for advances in artificial neural networks.
- The first artificial **neural network** model appeared in the 1950s as a nonlinear input-output mapping without probabilistic foundations. Progress continued through the 1980s and accelerated after 2010 with **deep learning**. Today neural networks are widely used for complex data analysis and underpin many generative AI systems. Modern views treat them as high-dimensional nonlinear statistical models.
- From the 1960s onwards the foundations of **computational learning theory** were developed, including later the well-known “support vector machine” (a non-probabilistic model). Other important advances include “ensemble learning” and corresponding algorithmic approaches to classification such as “random forests”.
- Classical statistics focused on settings with few variables and large sample sizes. Since about 2000, large-scale genomics and other high-dimensional datasets have driven rapid advances in statistics and machine learning to develop new methods to handle **high-dimensional data** (many variables with moderate sample sizes) and **big data** (many variables and large sample sizes).

## 2 Distributions

Choosing appropriate distributions for statistical modelling is a crucial aspect of probabilistic data analysis. This chapter explores various factors to consider when selecting suitable distributions and also lists frequently used families.

### 2.1 Characteristic features

Distributions can be differentiated by a number of characteristics.

Firstly, by the **type of random variable**:

- discrete versus continuous
- univariate versus multivariate

Secondly, by the **support** of the random variable, with typical ranges such as:

- finite discrete support, e.g.  $\{1, 2, \dots, n\}$
- infinite discrete support, e.g.  $\{1, 2, \dots\}$
- $[0, 1]$
- $[-\infty, \infty]$
- $[0, \infty]$

The choice of support will depend on the intended use of the random variable in the model. Common applications include:

- proportion
- location
- scale
- mean
- variance
- spread
- concentration
- shape
- rate



- (squared) correlation

These interpretations apply both to the random variable itself but also to the parameter of a distribution family. For instance, we might select a distribution whose outcomes can be interpreted as proportions (e.g. the beta distribution). Alternatively, we might choose a distribution family in which a parameter itself represents a proportion (e.g. the Bernoulli distribution).

A third consideration may be the general **shape** of the distribution:

- symmetric or asymmetric
- left or right skewed
- short tails or long tails
- unimodal or multimodal

A further characteristic of a distribution family is the **number of parameters**, with choices such as

- single parameter
- multiple parameters
- multiple types of parameters (e.g. location+scale)
- nonparametric model (i.e. highly parametric model)

A distribution **family** consists of a finite or infinite set of distributions that correspond to specific instances of parameter values.

Models with many parameters or nonparametric models are employed when there is an abundance of data and they typically make weak assumptions about the data-generating process. Conversely, simpler parametric models are generally preferred for smaller sample sizes, and typically make stronger assumptions about how the data were generated.

In data analysis, we aim to build models that are complex enough to capture the essential features of the data, but simple enough to avoid overfitting. Consequently, when two models have similar explanatory and predictive power, the one with fewer parameters is generally preferred.

## 2.2 Commonly used distribution families

In this module we will often make use of the following common univariate distributions:

- 1) Binomial distribution  $\text{Bin}(n, \theta)$ , with support  $\{0, 1, \dots, n\}$ .

Special case ( $n = 1$ ):

- Bernoulli distribution  $\text{Ber}(\theta)$ , with support  $\{0, 1\}$ .

- 2) Beta distribution  $\text{Beta}(\alpha, \beta)$ , with support  $[0, 1]$ .

- 3) Normal distribution  $N(\mu, \sigma^2)$ , with support  $[-\infty, \infty]$ .

Special cases:

- standard normal distribution  $N(0, 1)$ .
- delta distribution  $\delta(\mu)$  representing a point mass at  $\mu$  ( $N(\mu, \varepsilon \sigma^2)$  with  $\varepsilon \rightarrow 0$ ).

- 4) Gamma distribution  $\text{Gam}(\alpha, \theta)$ , with support  $[0, \infty]$ . It is also known as univariate Wishart distribution  $\text{Wis}(s^2, k)$ .

Special cases of the gamma/Wishart distribution:

- scaled chi-squared distribution  $s^2 \chi_k^2$  (discrete  $k$ )
- chi-squared distribution  $\chi_k^2$  (discrete  $k$ ,  $s^2 = 1$ )
- exponential distribution  $\text{Exp}(\theta)$  ( $\alpha = 1$ )

- 5) Inverse gamma distribution  $\text{IGam}(\alpha, \beta)$ , with support  $[0, \infty]$ . Also known as univariate inverse Wishart distribution  $\text{IWis}(\psi, k)$ .

All of the distributions above belong to the class of **exponential families** (see Section 2.4). These families have many properties that are highly convenient for statistical analysis.

- 6) Location-scale  $t$ -distribution  $t_v(\mu, \tau^2)$ , with support  $[-\infty, \infty]$ .

Special cases of the location-scale  $t$ -distribution:

- Student's  $t$ -distribution  $t_v$
- Cauchy distribution  $\text{Cau}(\mu, \tau)$

The location-scale  $t$ -distribution is generalisation of the normal distribution but with more probability mass in the tails. Depending on the choice of the degrees of freedom  $v$ , not all moments of the distribution may exist. Furthermore, it's not an exponential family.

For all of the above univariate distribution there exist corresponding multivariate variants. In this module we will make use of the following multivariate distributions:



- 1) Multinomial distribution  $\text{Mult}(n, \theta)$ , generalising the binomial distribution.

Special case ( $n = 1$ ):

- Categorical distribution  $\text{Cat}(\theta)$ , generalising the Bernoulli distribution.

- 2) Multivariate normal distribution  $N(\mu, \Sigma)$ , generalising the univariate normal distribution.

Special cases:

- standard multivariate normal distribution  $N(0, I)$ .
- multivariate delta distribution  $\delta(\mu)$  representing a point mass at  $\mu$  ( $N(\mu, \varepsilon \Sigma)$  with  $\varepsilon \rightarrow 0$ ).

A distribution family can be parametrised in multiple equivalent ways. Typically, there is a standard parametrisation, and also a mean parametrisation, where one of the parameters can be interpreted as the mean of the data. Sometimes, the same distribution is referred to by different names and there are various default parametrisations.

Importantly, any parametrisation is a matter of choice and simply provides as an alternative means to index the elementary distributions within the family. However, certain parametrisations may be more interpretable or offer computational advantages.

## 2.3 Model building

### Choosing the right distribution

When choosing a distribution we typically aim to align the characteristics of the distribution with those of the observations. For instance, if the data exhibit long tails, we will need to use a long-tailed model. Additionally, there may be a mechanistic rationale, such as a physical law, suggesting that the underlying process follows a particular model.

In many cases, the central limit theorem justifies using a normal distribution.

Another approach to selecting a distribution family is to fix certain properties of the distribution, such as its mean and variance, and then selecting the family that maximises the spread of the probability mass. This method is closely linked to the principle of maximum entropy,

which will be discussed in more detail in Chapter 6. It also helps to explain why exponential families are often preferred in statistical modelling.

### Complex models

Statistical analysis often uses models that are composed of many random variables. These models can be complex, with hierarchical or network-like structures that connect observed and latent variables, and potentially nonlinear functional relationships. Even so, the most sophisticated statistical models are constructed from simpler, more fundamental components.

Specifically, the large class of *graphical models* provide a principled means to form complex joint distributions for observed and unobserved random variables built from more elementary components. This includes *regression models*, *mixture models* and *compound models* (continuous version of mixture models) as well as more general network-like and hierarchically structured models.

In these complex models some of the underlying elementary distributions will serve to model the observed output while others represent internal variables or account for the uncertainty regarding a parameter (in a Bayesian context).

The advanced statistical course units in year 3 and year 4 will introduce many types of these models, for instance related to

- multivariate statistics and machine learning
- temporal and spatial modelling, and
- generalised linear and nonparametric regression.

### Iterative refinement

Models and distributions are best considered as approximations of the true unknown data-generating process. The aim of data analysis is therefore to find models that capture the essential properties at an appropriate level of detail<sup>1</sup>.

This is typically done in an iterative fashion, either starting from a simple model and increasing complexity as needed, or alternatively, starting

<sup>1</sup>In fact, processes at one length or time scale can often be modelled independently of those at other scales. This general phenomenon is known as *decoupling of scales*.



from a highly parametrised model and simplifying it. In either approach systematic methods are required to compare different models, to assess their fit to data and how well they predict future observations. Statistics provides principled tools for all these tasks.

## 2.4 ► Exponential families

### Overview

Many commonly used distributions in statistics are exponential families, including core examples such as the Bernoulli distribution and the normal distribution.

Exponential families are central in probability and statistics. They support effective statistical learning using likelihood and Bayesian approaches, enable data reduction via minimal sufficiency and provide the basis for generalised linear models. Furthermore, exponential families often allow to generalise results established for specific cases, such as the normal distribution, to a broader domain.

### Definition

An **exponential family**  $P(\eta)$  arises by **exponential tilting** of a **base distribution**  $B$  with (typically unnormalised) **base function**  $h(x)$  toward the inner product  $\langle \eta^T t(x) \rangle$  of the **canonical statistics**  $t(x)$  and the **canonical parameters**  $\eta$ . This yields a pdmf of the form

$$p(x|\eta) = \underbrace{h(x)}_{\text{base}} \underbrace{e^{\langle \eta, t(x) \rangle}}_{\text{exponential tilt}} / \underbrace{z(\eta)}_{\text{normaliser}}$$

The base pdmf is obtained at  $\eta = 0$  yielding  $b(x) = p(x|\eta = 0) = h(x)/z(0)$ . If  $h(x)$  is already a normalised pdmf then  $z(0) = 1$  and  $b(x) = h(x)$ .

The above presentation of exponential families assumes a univariate random variable (scalar  $x$ ) but also applies to multivariate random variables (vector  $x$  or matrix  $X$ ).

Likewise, canonical statistics and parameters are written as vectors but these may also be scalars or matrices (or a combination of both). The use of inner product notation  $\langle, \rangle$  includes all these cases, vectorising matrices as required, recalling that  $\langle A, B \rangle = \text{Tr}(A^T B) = \text{Vec}(A)^T \text{Vec}(B)$ .

### Canonical statistics and canonical parameters

The canonical statistics  $t(x)$  are transformations of  $x$ , usually simple functions such the identity ( $x$ ), the square ( $x^2$ ), the inverse ( $1/x$ ) or the logarithm ( $\log x$ ). Typically, the dimension of  $t(x)$  is small.

The canonical statistics  $t(x)$  may be affinely dependent. If this is the case there is a vector  $\eta_0$  for which

$$\langle \eta_0, t(x) \rangle = \text{const.}$$

If the elements in  $t(x)$  are affinely independent the representation of the exponential family is **minimal** or **complete**, otherwise the representation is **non-minimal** or **overcomplete**.

For each canonical statistic there is a corresponding canonical parameter so the dimensions and shape of  $t(x)$  and  $\eta$  match.

In a minimal representation the canonical parameters of the exponential family are **identifiable** and hence distinct parameter settings for  $\eta$  yield distinct distributions. Conversely, in a non-minimal or overcomplete representation there are redundant elements in the canonical parameters  $\eta$  and the distributions within the exponential family are **not identifiable**. Specifically, there will be multiple  $\eta$  yielding the same underlying distribution.

The canonical parameters  $\eta$  are typically some transformation of the conventional parameters  $\theta$ .

### Partition function

The **normaliser** or **partition function**  $z(\eta)$  ensures that  $p(x|\eta)$  integrates to one, with

$$z(\eta) = \int_x e^{\langle \eta, t(x) \rangle} h(x) dx$$

For discrete  $x$  replace the integral by a sum.

The set of values of  $\eta$  for which  $z(\eta) < \infty$ , and hence for which  $p(x|\eta)$  is well defined, comprises the parameter space of the exponential family. Some choices of  $h(x)$  and  $t(x)$  do not yield a finite normalising factor for any  $\eta$  and hence these cannot be used to form an exponential family.

The **log-normaliser** or **log-partition function**

$$a(\eta) = \log z(\eta)$$



allows to compute the cumulants of the canonical statistics. In particular, its gradient yields the mean

$$\begin{aligned} E(\mathbf{t}(x)) &= \boldsymbol{\mu}_t = \nabla a(\boldsymbol{\eta}) \\ &= \frac{\nabla z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})} \end{aligned}$$

and the Hessian matrix the variance

$$\begin{aligned} \text{Var}(\mathbf{t}(x)) &= \boldsymbol{\Sigma}_t = \nabla \nabla^T a(\boldsymbol{\eta}) \\ &= \frac{\nabla \nabla^T z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})} - \left( \frac{\nabla z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})} \right) \left( \frac{\nabla z(\boldsymbol{\eta})}{z(\boldsymbol{\eta})} \right)^T \end{aligned}$$

For a minimal exponential family  $\boldsymbol{\Sigma}_t$  is a positive definite matrix and invertible, whereas for non-minimal representations the covariance matrix is positive semi-definite and not invertible.

The means  $\boldsymbol{\mu}_t$  of the canonical statistics  $\mathbf{t}(x)$  provide a further parametrisation of exponential families, in addition to the canonical parameters  $\boldsymbol{\eta}$  and the conventional parameters  $\boldsymbol{\theta}$ , and are called the **expectation parameters**.

## Examples

**Example 2.1.** Logit and logistic function:

The logit and logistic functions are useful for describing the Bernoulli distribution as exponential family (Example 2.2).

The **logit** function is

$$\text{logit } p = \log \left( \frac{p}{1-p} \right)$$

also known as the the log-odds function. It maps the interval  $[0, 1]$  to the interval  $[-\infty, +\infty]$ .

Its inverse is the **logistic** function

$$\text{logit}^{-1}(\alpha) = (\exp \alpha) / (1 + \exp \alpha)$$

and maps the interval  $[-\infty, +\infty]$  to the interval  $[0, 1]$ .

**Example 2.2.** ► Bernoulli distribution  $\text{Ber}(\theta)$  as exponential family:

The Bernoulli distribution  $\text{Ber}(\theta)$  can be specified in exponential family form as follows:

- $x \in \{0, 1\}$
- canonical statistic  $t(x) = x$
- base function  $h(x) = 1$
- canonical parameter  $\eta$

This results in the partition function

$$z(\eta) = \sum_{x \in \{0,1\}} e^{\eta x} = 1 + e^{\eta}$$

and the log-partition function  $a(\eta) = \log z(\eta)$ , both of which are defined for  $\eta \in \mathbb{R}$ .

The first and second derivatives of the partition function  $z(\eta)$  are  $z'(\eta) = e^{\eta}$  and  $z''(\eta) = e^{\eta}$ .

The mean  $\mu_t$  of the canonical statistic  $t(x) = x$  is given by

$$\begin{aligned} \mu_t &= a'(\eta) = \frac{z'(\eta)}{z(\eta)} = \frac{e^{\eta}}{1 + e^{\eta}} \\ &= \text{logit}^{-1}(\eta) \\ &= \theta \end{aligned}$$

Since  $z(x) = x$  for the Bernoulli distribution the expectation parameter  $\mu_t$  corresponds to the conventional mean parameter  $E(x) = \theta \in [0, 1]$ . In the above  $\mu_t$  and  $\theta$  are obtained from  $\eta$  by the **inverse logit** function (also known as the **logistic** function).

Conversely, the canonical parameter  $\eta$  can be computed from  $\theta$  (or  $\mu_t$ ) by the **logit** function

$$\eta = \text{logit } \theta = \log \left( \frac{\theta}{1-\theta} \right)$$

The variance  $\sigma_t^2$  of the canonical statistic  $t(x) = x$  is

$$\begin{aligned} \sigma_t^2 &= a''(\eta) = \frac{z''(\eta)}{z(\eta)} - \left( \frac{z'(\eta)}{z(\eta)} \right)^2 \\ &= \frac{e^{\eta}}{(1 + e^{\eta})^2} \\ &= \theta(1 - \theta) \end{aligned}$$

The pmf of the Bernoulli distribution is

$$p(x|\eta) = \frac{e^{\eta x}}{z(\eta)} = \frac{e^{\eta x}}{1 + e^{\eta}}$$



which in terms of the conventional parameter  $\theta$  takes on the familiar form

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x} \\ = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

**Example 2.3.** ► Normal distribution  $N(\mu, \sigma^2)$  as exponential family:

The two-parameter normal distribution  $N(\mu, \sigma^2)$  can be written in exponential family form as follows:

- $x \in \mathbb{R}$
- canonical statistics  $\mathbf{t}(x) = (x, x^2)^T$
- base function  $h(x) = 1$
- canonical parameters  $\boldsymbol{\eta} = (\eta_1, \eta_2)^T$

This results in the partition function

$$z(\boldsymbol{\eta}) = \left(-\frac{\pi}{\eta_2}\right)^{1/2} \exp\left(-\frac{\eta_1^2}{4\eta_2}\right)$$

and the log-partition function

$$a(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \log\left(-\frac{\pi}{\eta_2}\right)$$

which are defined for  $\eta_1 \in \mathbb{R}$  and  $\eta_2 \in \mathbb{R}^+$ .

The mean  $\boldsymbol{\mu}_t$  of the canonical statistics  $\mathbf{t}(x) = (x, x^2)^T$  is given by

$$\boldsymbol{\mu}_t = \nabla a(\boldsymbol{\eta}) \\ = \begin{pmatrix} -\frac{\eta_1}{2\eta_2} \\ \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \end{pmatrix} \\ = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix}$$

The conventional parameters  $\mu = E(x)$  and  $\sigma^2 = \text{Var}(x)$  are thus directly linked to the expectation parameters  $\boldsymbol{\mu}_t$  and can be obtained from the canonical parameters by  $\mu = -\frac{\eta_1}{2\eta_2}$  and  $\sigma^2 = -\frac{1}{2\eta_2}$ . Conversely, we have

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$$

The covariance matrix of the canonical statistics  $\mathbf{t}(x)$  is

$$\boldsymbol{\Sigma}_t = \begin{pmatrix} \text{Var}(x) & \text{Cov}(x, x^2) \\ \text{Cov}(x^2, x) & \text{Var}(x^2) \end{pmatrix} \\ = \nabla \nabla^T a(\boldsymbol{\eta}) \\ = \begin{pmatrix} -\frac{1}{2\eta_2} & \frac{\eta_1}{2\eta_2^2} \\ \frac{\eta_1}{2\eta_2^2} & \frac{\eta_2 - \eta_1^2}{2\eta_2^3} \end{pmatrix} \\ = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 2\sigma^4 + 4\mu^2\sigma^2 \end{pmatrix}$$

The pdf in terms of  $\boldsymbol{\eta}$  is

$$p(x|\boldsymbol{\eta}) = \left(-\frac{\pi}{\eta_2}\right)^{-1/2} \exp\left(\eta_1 x + \eta_2 x^2 + \frac{\eta_1^2}{4\eta_2}\right)$$

which in terms of the standard parameters  $\mu$  and  $\sigma^2$  takes on the familiar form

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

## 2.5 Further reading

For details of the distributions listed above and additional background on exponential families see the supplementary [Probability and Distribution Refresher notes](#).

A recent textbook on exponential families is Efron (2022).

### 💡 A bit of history

The concept of exponential families in statistics was independently developed in the 1930s by [Georges Darmois \(1888–1960\)](#), [Edwin J. G. Pitman \(1897–1993\)](#), and [Bernard Koopman \(1900–1981\)](#). However, exponential families were introduced earlier in statistical mechanics in the 1870s by [Josiah W. Gibbs \(1839–1903\)](#) and [Ludwig Boltzmann \(1844–1906\)](#).

In statistics, exponential families were motivated by identifying distributions allowing minimal sufficient statistics whereas in physics the objective was to find distributions maximising entropy.



## Part II

# Information

## 3 Entropy

Entropy, a fundamental concept that originated in physics, is central also in information theory and statistical learning. This chapter introduces information entropy through log-loss and combinatorics.

### 3.1 Code length and log-loss

#### Information storage units

When we record information on paper or on a computer, we need an alphabet, whose symbols act as units of stored information:

- For an alphabet of size  $b = 2$  (e.g. symbols 0 and 1) the storage unit is called **bit** (binary digit). To represent  $K = 256$  possible states requires  $\log_2 256 = 8$  bits (or 1 byte) of storage.
- For an alphabet of size  $b = 10$  (e.g. Arabic numerals) the unit is called **dit** (decimal digit). To represent  $K = 100$  possible states requires  $\log_{10} 100 = 2$  dits.

Bit and dit are **units of information** to describe information content. One dit is equal to  $\log_2 10 \approx 3.32$  bits.

If the natural logarithm is used ( $b = e$ ) the unit of information is called **nat**. This is the standard unit of information used in statistics. It is equal to  $\log_2 e \approx 1.44$  bits.

#### Constant code lengths

We now assume a discrete variable  $x$  with  $K$  possible outcomes  $\Omega = \{\omega_1, \dots, \omega_K\}$ . In order to record a realised state of  $x$  using an alphabet of size  $b$  we require

$$S = \log_b K$$



units of information.  $S$  is called the **code length** or the **cost** to describe the state of  $x$ . The base  $b$  can be larger or smaller than  $K$ .

In the above we have tacitly assumed that storage size, code length, and cost requirements are fixed and identical for all possible  $K$  states. This can be made more explicit by writing

$$S = -\log_b \left( \frac{1}{K} \right)$$

where  $1/K$  is the equal probability of each of the  $K$  states.

### Variable code lengths

In practice the  $K$  states are usually not equally likely. For example,  $x$  might represent letters in a text message, each occurring with different frequency. To account for that we use a distribution  $P$  with pmf  $p(x)$  and allow variable code lengths per state.

Generalising the case of constant code lengths, we use the negative logarithm to map the probability of a state  $x$  to a corresponding cost and code length:

$$S(x, P) = -\log p(x)$$

This is called the **log-loss** or the **logarithmic scoring rule**. Under this rule, more probable states under model  $P$  get shorter codes (smaller loss/cost) and less probable states get longer codes (larger loss/cost).

As we will see in the examples (Example 3.3, Example 3.4 and Example 3.5) using variable code lengths lets the expected code length be much smaller than for a fixed-length encoding.

The log-loss underpins much of statistical estimation and inference through likelihood and Bayesian methods. It is the leading example of a **proper scoring rule**, a broader class of loss functions suitable for principled statistical analysis (Note 1).

Scoring rules are only defined up to an equivalence class  $S^{\text{equiv}}(x, P) = kS(x, P) + c(x)$ , where  $k$  is a positive scaling factor. All **equivalent scoring rules** have the same minimiser for fixed  $x$ .

In the equivalence class for the log-loss, the scaling factor  $k$  determines the base of the logarithm. In the following, we set  $b = e$  and use the natural logarithm and nats as units of information throughout. We will also apply the log-loss to both discrete and continuous variables  $x$  and corresponding models  $P$ .

#### i Note 1: Proper scoring rules

The logarithmic scoring rule  $S(x, P) = -\log p(x)$  is a particularly important loss function to evaluate probabilistic models. It is unique as the only **local** strictly proper scoring rule, solely depending on the value of the pmf at the observed outcome, and not on any other features of the distribution.

However, instead of the negative logarithm one may also employ other suitable loss functions as scoring rule, while still preserving the useful properties of the log-loss. This leads to the wider class of **proper scoring rules**. In machine learning these are related to **Bregman divergences**.

Using proper scoring rules other than the log-loss may be preferable, e.g., for computational convenience and simplicity, for numerical stability for robustness to outliers or interpretability.

#### Example 3.1. Surprise and log-odds:

The negative logarithm of a probability or density is called the **surprise** or **surprisal**. The log-loss thus returns the surprise for the outcomes of a random variable.

For a discrete random variable, the surprise  $-\log p$  to observe an outcome with probability  $p = 1$  is zero, and conversely the surprise to observe an outcome with probability  $p = 0$  is infinite.

#### The log-odds

$$\log \left( \frac{p}{1-p} \right) = -\log(1-p) - (-\log p)$$

is the difference between the surprise of the complementary event (probability  $1-p$ ) and the surprise of the event (probability  $p$ ):

#### Example 3.2. Log-loss for a normal model:

If we quote in the log-loss  $S(x, P) = -\log p(x)$  the normal distribution  $P = N(\mu, \sigma^2)$  with density  $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)$  we get

$$S(x, P) = \frac{1}{2} \left( \log(2\pi\sigma^2) + \frac{(x-\mu)^2}{\sigma^2} \right)$$

For a normal model with fixed variance the log-loss is thus equivalent to the squared distance between  $x$  and  $\mu$ .



## 3.2 Information entropy

### Entropy of a distribution

The expectation of the log-loss

$$\begin{aligned} H(P) &= \mathbb{E}_P(S(x, P)) \\ &= -\mathbb{E}_P(\log p(x)) \end{aligned}$$

is the **information entropy**  $H(P)$  of the distribution  $P$ .

The notation  $H(P)$  emphasises that entropy is a functional of the distribution  $P$ . Note that in this expression  $P$  serves two distinct roles, first as the data-generating model (the expectation is taken with respect to  $P$ ) and second as the model  $P$  quoted in the log-loss.

Information entropy is **only defined up to an affine transformation with a positive scaling factor** as the underlying loss function is itself also only defined up to an equivalence class

$$H(P)^{\text{equiv}} = -k \mathbb{E}_P(\log p(x)) + c$$

The scaling factor  $k$  determines the units (or the base of the logarithm) and the additive constant  $c = \mathbb{E}_P(c(x))$  can be chosen to fix a reference value (e.g. typically either set the maximum or minimum entropy to zero).

#### **i** Note 6: Other names for information entropy

Information entropy is also called **Shannon entropy** (for discrete distributions) and **differential entropy** (for continuous distribution).

In physics, the formula for information entropy is primarily known as **Gibbs entropy**, but also referred to as **Boltzmann-Gibbs entropy** (BG entropy) or **Boltzmann-Gibbs-Shannon entropy** (BGS entropy) entropy. Physical entropy also carries the **Boltzmann constant**  $k_B = 1.380649 \times 10^{-23} \text{ J/K}$  as scaling factor, giving physical entropy units of energy/temperature.<sup>1</sup>

We will mostly use the term **information entropy**.

<sup>1</sup>The Boltzmann constant  $k_B$  is one of **seven defining constants** in the **International System of Units**.

### Concavity

A key property of the **entropy**  $H(P)$  is that it is **strictly concave** in  $P$ . This means that

$$H(P_\lambda) > (1 - \lambda)H(P_0) + \lambda H(P_1)$$

for the mixture  $P_\lambda = (1 - \lambda)P_0 + \lambda P_1$  with  $0 < \lambda < 1$  and  $P_0 \neq P_1$ , and follows from the fact that  $-x \log(x)$  is a strictly concave function.

Strict concavity implies that mixing states (represented by  $P_0$  and  $P_1$ ) raises the entropy. Further, it ensures that the maximum is unique (if a maximum exists).

Concavity is a result of the properness of the log-loss (Section 4.1). Specifically, the entropy is (strictly) concave because the underlying scoring rule is (strictly) proper.

### Information entropy for discrete distributions

The information entropy  $H(P)$  of a discrete probability distribution  $P$  with pmf  $p(x)$  with  $x \in \Omega$  is the **Shannon entropy**:

$$H(P) = - \sum_x p(x) \log p(x)$$

By construction, since all  $p(x) \in [0, 1]$  and hence  $-\log p(x) \geq 0$ , the Shannon entropy is **non-negative**.

**Example 3.3.** Entropy of the categorical distribution:

For  $P = \text{Cat}(\mathbf{p})$  with class probabilities  $\mathbf{p} = (p_1, \dots, p_K)^T$  the information entropy of  $P$  is

$$H(P) = - \sum_{k=1}^K p_k \log p_k$$

The maximum entropy ( $\log K$ ) is achieved for the discrete uniform distribution (Example 3.4) and the minimum entropy (0) for a concentrated categorical distribution (Example 3.5). Hence for a categorical distribution  $P$  with  $K$  categories we have

$$0 \leq H(P) \leq \log K$$



**Example 3.4.** Entropy of the discrete uniform distribution:

The discrete uniform distribution is  $U_K = \text{Cat}(\mathbf{p} = \mathbf{1}_K/K)$  so that  $p_1 = p_2 = \dots = p_K = \frac{1}{K}$ . Then

$$H(U_K) = - \sum_{k=1}^K \frac{1}{K} \log \left( \frac{1}{K} \right) = \log K$$

Note that  $\log K$  is the largest value the information entropy can assume with  $K$  classes (cf. Example 6.3) and indicates maximum spread of probability mass.

**Example 3.5.** Entropy of a categorical distribution with concentrated probability mass:

Let  $p_1 = 1$  and  $p_2 = p_3 = \dots = p_K = 0$ . Using  $0 \times \log 0 = 0$  we obtain for the information entropy of  $P = \text{Cat}(\mathbf{p})$

$$H(P) = 1 \times \log 1 + 0 \times \log 0 + \dots = 0$$

Note that zero is the smallest value that the entropy can assume for a categorical distribution and that it corresponds to maximum concentration of probability mass.

**Example 3.6.** Entropy of the geometric distribution:

With  $P = \text{Geom}(\theta)$  and  $\theta \in [0, 1]$ , support  $x \in \{1, 2, \dots\}$  and pmf  $p(x|\theta) = \theta(1 - \theta)^{x-1}$ , with  $E(x) = 1/\theta$  and  $\text{Var}(x) = (1 - \theta)/(\theta^2)$ , the entropy is

$$\begin{aligned} H(P) &= -E(\log \theta + (x - 1) \log(1 - \theta)) \\ &= -\log \theta + \left( \frac{1}{\theta} - 1 \right) \log(1 - \theta) \\ &= -\frac{\theta \log \theta + (1 - \theta) \log(1 - \theta)}{\theta} \end{aligned}$$

Using the identity  $0 \times \log 0 = 0$  we see that the entropy of the geometric distribution for  $\theta = 1$  equals 0, i.e. it achieves the minimum possible information entropy for a discrete distribution. Conversely, as  $\theta \rightarrow 0$  it diverges to infinity.

## Information entropy for continuous distributions

The entropy  $H(P)$  of a continuous distribution  $P$  with pdf  $p(x)$  is the **differential entropy**:

$$H(P) = - \int_x p(x) \log p(x) dx$$

Differential entropy **can be negative** because the logarithm is taken of a density value, and unlike probabilities, densities are not constrained to be less than or equal to one.

Under a **change of variables**, such as from  $x$  to  $y$ , the shape of the pdf typically changes, so that differential entropy changes as well and is **not invariant**:  $H(P_y) \neq H(P_x)$ .

**Example 3.7.** Entropy of the uniform distribution:

Consider the uniform distribution  $P_x = \text{Unif}(0, a)$  with  $a > 0$ , support  $x \in [0, a]$  and density  $p(x) = 1/a$ . The corresponding entropy is

$$\begin{aligned} H(P_x) &= - \int_0^a \frac{1}{a} \log \left( \frac{1}{a} \right) dx \\ &= \frac{1}{a} \log a \int_0^a dx \\ &= \log a. \end{aligned}$$

The range of the entropy is unbounded for a positive  $a$ , and for  $0 < a < 1$  the entropy assumes a negative value.

**Example 3.8.** Differential entropy under a change of variables:

Starting with the uniform distribution  $P_x = \text{Unif}(0, a)$  from Example 3.7 the variable  $x$  is changed to  $y = x^2$  yielding the distribution  $P_y$  with support from 0 to  $a^2$  and density  $p(y) = 1/(2a\sqrt{y})$ .

The corresponding entropy is

$$\begin{aligned} H(P_y) &= \int_0^{a^2} \frac{1}{2a\sqrt{y}} \log(2a\sqrt{y}) dy \\ &= \left[ \sqrt{y}/a (\log(2a\sqrt{y}) - 1) \right]_{y=0}^{y=a^2} \\ &= \log(2a^2) - 1. \end{aligned}$$



The range of entropy is unbounded, and is negative for  $0 < a < \sqrt{e/2} \approx 1.1658$ . As expected  $H(P_y) \neq H(P_x)$  as differential entropy is not invariant under a change of variables.

**Example 3.9.** Entropy of the normal distribution:

The log-density of the univariate normal distribution  $P = N(\mu, \sigma^2)$  is  $\log p(x|\mu, \sigma^2) = -\frac{1}{2} \left( \log(2\pi\sigma^2) + \frac{(x-\mu)^2}{\sigma^2} \right)$  with  $\sigma^2 > 0$ . The corresponding entropy is with  $E((x - \mu)^2) = \sigma^2$

$$\begin{aligned} H(P) &= -E(\log p(x|\mu, \sigma^2)) \\ &= \frac{1}{2} (\log(2\pi\sigma^2) + 1) . \end{aligned}$$

Importantly, the entropy of a normal distribution depends only on its variance  $\sigma^2$ , not on its mean  $\mu$ . This is intuitively clear as the variance controls the concentration of probability mass. A large variance means that the probability mass is more spread out and thus less concentrated around the mean.

Since the variance  $\sigma^2$  ranges from 0 to infinity, the entropy of the normal distribution is unbounded ranging from negative to positive infinity, taking on the value of zero at  $\sigma^2 < 1/(2\pi e) \approx 0.0585$ .

**Example 3.10.** ▶ Entropy of an exponential family:

Assume  $P(\eta)$  is an exponential family with canonical parameters  $\eta$ , canonical statistics  $t(x)$  and log-partition function  $a(\eta)$  with log-pdmf  $\log p(x|\eta) = \langle \eta, t(x) \rangle + \log h(x) - a(\eta)$ .

Then with  $E_{P(\eta)}(t(x)) = \mu_t(\eta)$  the entropy is

$$\begin{aligned} H(P(\eta)) &= -E_{P(\eta)}(\log p(x|\eta)) \\ &= a(\eta) - \langle \eta, \mu_t(\eta) \rangle - E_{P(\eta)}(\log h(x)) \end{aligned}$$

### Interpretation of entropy

Information entropy  $H(P)$  can be interpreted as the **expected cost** or **expected code length** when the data are generated according to model  $P$  (“sender”, “encoder”) and the same model  $P$  is used to describe the data (“receiver”, “decoder”). As will be discussed in Chapter 4 in more detail, using different models for “sending” and “receiving” always increases the cost, hence the entropy is a lower bound on that cost.

Furthermore, information entropy  $H(P)$  quantifies the **dispersal of probability mass** within the distribution  $P$ . When probability mass is concentrated, entropy is low (Example 3.5). Conversely, when the probability mass is more broadly distributed, the entropy is high (Example 3.4).

The notion of entropy as measure of spread is also supported in continuous settings, e.g. the entropy of the normal distribution (Example 3.9) is a monotonically increasing function of the variance.

A distribution  $P$  with low entropy and concentrated probability mass is highly informative. Conversely, a distribution with high entropy and dispersed probability mass is uninformative. This is the basis of the principle of maximum entropy (Chapter 6).

## 3.3 ▶ Entropy in physics

### Entropy as macroscopic phenomenon

The concept of entropy was introduced in the mid-19th century within **thermodynamics**, a phenomenological theory that quantifies heat, work, energy, and the laws governing their transformations. Entropy determines how much of the internal energy is available to do work. Low entropy implies energy is concentrated and available for work, and high entropy means energy is spread out and not available for work. Systems at equilibrium are characterised by maximum entropy (subject to constraints).

In **statistical mechanics**, entropy and the laws of thermodynamics are explained as emergent, probabilistic consequences of the behaviour of many-particle systems. A **microstate** provides a complete specification of the current state of the system which evolves by transitioning from one such state to the other.

In the **Gibbs approach**, hypothetical ensembles of systems are considered, represented by a probability distribution over all microstates. If that distribution is concentrated on relatively few microstates the entropy is low. Conversely, if it is spread over many microstates the entropy is large. Once in equilibrium, the distribution and its entropy is stationary. Its information entropy corresponds the conventional thermodynamic entropy.



In the **Boltzmann approach**, entropy arises from coarse graining and the introduction of **macrostates** (e.g., particle counts or energy levels). The entropy of a particular system (not of an ensemble!) is computed from the size of its current macrostate as determined by the number of underlying microstates compatible with it. This gives rise to both the standard formula for information entropy discussed in this chapter (see Example 3.11) and a further generalisation introduced in Section 6.1 (see Example 6.2).

Equilibrium corresponds to the typical macrostate with overwhelmingly many compatible microstates (maximum entropy) and is the state in which the system spends most of its time. Entropy fluctuates as the system evolves and continues to do so even in equilibrium. This explains macroscopic irreversibility as a statistical phenomenon.

Both the Gibbs and Boltzmann approaches support the interpretation of entropy as a measure of dispersal and use the same functional form of entropy.

A further common physical interpretation of entropy is as a measure of disorder, with low entropy corresponding to “order” and high entropy to “disorder”. However, this interpretation is now discouraged because those everyday terms are ambiguous. Many apparently ordered systems can have large entropy, while some disordered-looking systems can have low entropy.

Finally, in physics the concavity of entropy is a crucial property as it underpins the [second law of thermodynamics](#).

### ► Combinatorial derivation

Information entropy can also be obtained by combinatorial considerations. This was the route taken by Boltzmann in 1877 in the early days of statistical mechanics.

**Example 3.11.** ► Combinatorial derivation of information entropy:

Let  $D = \{n_1, \dots, n_K\}$  be counts for  $K$  classes with  $n = \sum_{k=1}^K n_k$  and  $\hat{Q} = \text{Cat}(\hat{q})$  be the corresponding empirical categorical distribution with class frequencies  $\hat{q}_k = n_k/n$ .

The entropy of  $\hat{Q}$  is  $H(\hat{Q}) = -\sum_{k=1}^K \hat{q}_k \log \hat{q}_k$  (Example 3.3).

The macrostate corresponds to the counts  $D$  or equivalently the distribution  $\hat{Q}$ . Each particular allocation of the  $n$  elements to the  $K$  groups constitutes a microstate.

The number of allocations of  $n$  distinct items to  $K$  groups compatible with the given counts  $D$  is given by the **multinomial coefficient**

$$W_K = \binom{n}{n_1, \dots, n_K} = \frac{n!}{n_1! n_2! \dots n_K!}$$

This represents the **multiplicity** of the macrostate.

Using the de Moivre-Stirling formula we can, for large  $n$ , approximate the factorial according to

$$\log n! \approx n \log n - n$$

and get

$$\begin{aligned} \log W_K &= \log n! - \sum_{k=1}^K \log n_k! \\ &\approx n \log n - n - \sum_{k=1}^K (n_k \log n_k - n_k) \\ &= \sum_{k=1}^K n_k \log n - \sum_{k=1}^K n_k \log n_k \\ &= -n \sum_{k=1}^K \frac{n_k}{n} \log \left( \frac{n_k}{n} \right) \\ &= -n \sum_{k=1}^K \hat{q}_k \log \hat{q}_k \\ &= nH(\hat{Q}) \end{aligned}$$

Therefore, information entropy and the multinomial coefficient are directly related

$$H(\hat{Q}) \approx \frac{1}{n} \log W_K$$

with equality for  $n \rightarrow \infty$ .



The number of permutations  $W_K$  is far greater when items are uniformly “spread out” over the  $K$  categories than when they are “concentrated”. Hence, the combinatorial perspective supports the interpretation of entropy as a measure of dispersal.

See also Example 6.2 for a related derivation of information entropy with regard to a prior measure (Section 6.1).

### 3.4 Further reading

For a justification of the logarithmic loss function see, e.g., Good (1952) and Bernardo (1979).

A brief overview of proper scoring rules is found in the supplementary [Probability and Distribution Refresher notes](#).

Akaike (1985) provides an historical account from a statistical perspective of how Boltzmann arrived in 1877 at the formula for information entropy.

Lemons (2013) offers a friendly introduction to entropy in physics and in information theory. This book also discusses the combinatorial derivation of entropy (Example 3.11) as well as various interpretations and applications of entropy.

#### 💡 A bit of history

Entropy as a thermodynamic quantity was introduced in 1865 by [Rudolph Clausius \(1822-1888\)](#), along with the first and second law of thermodynamics.

The probabilistic and combinatorial underpinnings of entropy were developed in the period from 1870 to 1900 by [Josiah W. Gibbs \(1839-1903\)](#) and [Ludwig Boltzmann \(1844-1906\)](#). The specific formula for information entropy was found by Boltzmann in 1877. By the mid-20th century the notion of information entropy was established in statistics. [Ralph V. L. Hartley \(1888-1970\)](#) and [Irving J. Good \(1916-2009\)](#) argued for logarithmic measures of information and [Claude Shannon \(1916-2001\)](#) introduced entropy in information theory ([Shannon 1948](#)).

## 4 Risk and divergence

This chapter introduces two measures of risk and divergence: the mean log-loss, also known as cross-entropy, and the Kullback-Leibler (KL) divergence, also known as relative entropy.

### 4.1 Risk

#### Mean log-loss or cross-entropy

The **risk** is defined as the **expected loss**. For the logarithmic scoring rule  $S(x, P) = -\log p(x)$  the risk of  $P$  under  $Q$  is the **mean log-loss**

$$H(Q, P) = -E_Q \log p(x)$$

which is also known as **cross-entropy** (see Note 6).

The risk is a functional of two distributions  $Q$  and  $P$ , where  $Q$  represents the data-generating process (note the expectation  $E_Q$  with regard to  $Q$ ) and the distribution  $P$  is the model that is evaluated on the observations. Hence, the two arguments are not interchangeable and  $H(Q, P)$  is *not symmetric* with regard to  $Q$  and  $P$ .

The risk, like entropy, is **only defined up to an affine transformation with a positive scaling factor**, yielding the equivalence class

$$H(Q, P)^{\text{equiv}} = -k E_Q (\log p(x)) + c$$

For fixed  $Q$  all **equivalent risks** have the same minimiser for  $P$ .

#### **i** Note 6: Terminology for cross-entropy

The term “cross-entropy” is used inconsistently in the literature, so we will generally use “mean log-loss” or “risk” instead. In contemporary usage, “cross-entropy” typically refers to the mean log-loss (a risk to be minimised). However, many authors



have also used “cross-entropy” to mean the KL divergence (relative entropy).

### Mixture preserving

The risk is **mixture preserving in  $Q$**  meaning that

$$H(Q_\lambda, P) = (1 - \lambda)H(Q_0, P) + \lambda H(Q_1, P)$$

for the mixture  $Q_\lambda = (1 - \lambda)Q_0 + \lambda Q_1$  with  $0 < \lambda < 1$  and  $Q_0 \neq Q_1$ . This follows from the linearity of expectation.

Hence, unlike entropy  $H(Q)$ , the mean log-loss  $H(Q, P)$  is *not* concave in  $Q$ .

### Properness inequality

By construction, for  $P = Q$  the mean log-loss reduces to the information entropy  $H(Q) = H(Q, Q)$ . Moreover,  $H(Q, P)$  is uniquely minimised for  $P = Q$  so that the mean log-loss also satisfies the **properness inequality**

$$H(Q, P) \geq H(Q)$$

with equality only if  $P = Q$ . This is also known as **Gibbs’ inequality**.

That information entropy is the unique lower bound of the mean log-loss can be established using **Jensen’s inequality** for convex functions (recall that  $-\log x$  is strictly convex).

The logarithmic scoring rule (log-loss) is called **strictly proper** because the associated risk (mean log-loss) satisfies the **properness inequality**.

Properness is what makes the log-scoring rule (and indeed other proper scoring rules) useful as it allows to identify the underlying true model by risk minimisation.

If a scoring rule is (strictly) proper, then the associated entropy is (strictly) concave (see Section 3.2).

### Interpretation

The mean log-loss can be interpreted as the expected cost or expected code length when the data are generated according to model  $Q$  (“sender”, “encoder”) and but we employ model  $P$  to describe the data (“receiver”, “decoder”).

As a consequence of properness, when data are generated (encoded) under model  $Q$  and described (decoded) using model  $P$  there is always an extra cost, or penalty, to employ the approximating model  $P$  rather than the correct model  $Q$ .

### Mean log-loss for discrete and continuous distributions

For two discrete distributions  $Q$  and  $P$  with pmf  $q(x)$  and  $p(x)$  with  $x \in \Omega$  the mean log-loss is computed as the weighted sum

$$H(Q, P) = - \sum_{x \in \Omega} q(x) \log p(x)$$

Similarly, for two continuous distributions  $Q$  and  $P$  with pdfs  $q(x)$  and  $p(x)$  we compute the integral

$$H(Q, P) = - \int_x q(x) \log p(x) dx$$

Like differential entropy the **mean log-loss is not invariant under a change of variables** for continuous random variables, such as from  $x$  to  $y$ , hence  $H(Q_y, P_y) \neq H(Q_x, P_x)$ .

### Examples

**Example 4.1.** Mean log-loss for two normal distributions:

Assume  $P_1 = N(\mu_1, \sigma_1^2)$  and  $P_2 = N(\mu_2, \sigma_2^2)$ . The mean log-loss is

$$\begin{aligned} H(P_1, P_2) &= -E_{P_1}(\log p(x|\mu_2, \sigma_2^2)) \\ &= \frac{1}{2} E_{P_1} \left( \log(2\pi\sigma_2^2) + \frac{(x - \mu_2)^2}{\sigma_2^2} \right) \\ &= \frac{1}{2} \left( \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} + \log(2\pi\sigma_2^2) \right) \end{aligned}$$



using  $E_{P_1}((x - \mu_2)^2) = (\mu_1 - \mu_2)^2 + \sigma_1^2$ .

**Example 4.2.** Normal mean log-loss reduces to normal differential entropy:

The mean log-loss for two normal distributions  $P_1$  and  $P_2$  is (Example 4.1)

$$H(P_1, P_2) = \frac{1}{2} \left( \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} + \log(2\pi\sigma_2^2) \right)$$

Setting  $P_1 = P_2 = P$  with  $\mu_1 = \mu_2 = \mu$  and  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  the mean log-loss degenerates to

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

which is the normal differential entropy (Example 3.9).

**Example 4.3.** Empirical mean log-loss and log-likelihood

In the **empirical mean log-loss** the expectation is taken with regard to the empirical distribution  $\hat{Q}_n$ . For a model family  $P(\theta)$  it is

$$\begin{aligned} H(\hat{Q}_n, P(\theta)) &= -E_{\hat{Q}_n}(\log p(x|\theta)) \\ &= -\frac{1}{n} \sum_{i=1}^n \log p(x_i|\theta) \\ &= -\frac{1}{n} \ell_n(\theta) \end{aligned}$$

The empirical risk is equal to the negative log-likelihood  $\ell_n(\theta)$  standardised by the sample size  $n$ . Conversely, the **log-likelihood** is the **negative empirical risk based on the log-loss multiplied by sample size  $n$** .

Therefore, maximising the log-likelihood is equal to minimising the empirical risk based on the log-loss

**Example 4.4.** ► Mean log-loss for an exponential family:

Assume  $P(\eta)$  is an exponential family with canonical parameters  $\eta$ , canonical statistics  $t(x)$  and log-partition function  $a(\eta)$  with log-pdmf  $\log p(x|\eta) = \langle \eta, t(x) \rangle + \log h(x) - a(\eta)$ .

Then with  $E_{P(\eta_1)}(t(x)) = \mu_t(\eta_1)$  the mean log-loss is

$$\begin{aligned} H(P(\eta_1), P(\eta_2)) &= -E_{P(\eta_1)}(\log p(x|\eta_2)) \\ &= a(\eta_2) - \langle \eta_2, \mu_t(\eta_1) \rangle - E_{P(\eta_1)}(\log h(x)) \end{aligned}$$

For  $\eta_1 = \eta_2 = \eta$  it reduces to

$$H(P(\eta)) = a(\eta) - \langle \eta, \mu_t(\eta) \rangle - E_{P(\eta)}(\log h(x))$$

the entropy of an exponential family (Example 3.10).

## 4.2 Divergence

### Kullback-Leibler divergence

The **divergence**<sup>1</sup> between the two distributions  $Q$  and  $P$  based on the logarithmic scoring rule is the difference between the mean log-loss (risk) and the information entropy (minimum risk). This yields the **Kullback-Leibler (KL) divergence** between  $Q$  and  $P$ :

$$\begin{aligned} D_{\text{KL}}(Q, P) &= H(Q, P) - H(Q) \\ &= E_Q \log \left( \frac{q(x)}{p(x)} \right) \end{aligned}$$

By construction, as a consequence of properness, the KL divergence is always non-negative,  $D_{\text{KL}}(Q, H) \geq 0$ . Furthermore,  $D_{\text{KL}}(Q, P) = 0$  only for  $P = Q$  (strict properness)

The use of the term “divergence” rather than “distance” is a reminder that  $Q$  and  $P$  are not interchangeable.

Using the KL divergence, we can write the mean log-loss as

$$\begin{aligned} H(Q, P) &= \underbrace{D_{\text{KL}}(Q, P)}_{\geq 0} + H(Q) \\ &\geq H(Q) \end{aligned}$$

which also includes the properness inequality.

The divergence is only **defined up to a positive scaling factor  $k$**  that effectively determines the base of the logarithm:

$$D_{\text{KL}}(Q, P)^{\text{equiv}} = H(Q, P)^{\text{equiv}} - H(Q)^{\text{equiv}} = k E_Q \log \left( \frac{q(x)}{p(x)} \right)$$

<sup>1</sup>The **divergence between distributions** is unrelated to the **divergence vector operator** from vector calculus.



**i** Note 6: Terminology for KL divergence

Various notations exist for KL divergence, we use  $D_{\text{KL}}(Q, P)$  but  $\text{KL}(Q||P)$  and  $I^{\text{KL}}(Q; P)$  are also common.

The KL divergence is also called **KL information**, **KL information number** or **discrimination information**. Other names include **information divergence** or **I-divergence**. Some authors call  $2D_{\text{KL}}(Q, P) = D(Q, P)$  the **deviance** of  $P$  from  $Q$ .

KL divergence is frequently termed **relative entropy**, but this label is ambiguous as some authors use it for the **negative of KL divergence** (i.e. for information entropy with prior measure, see Section 6.1). Shannon (1948) also used the “relative” entropy to mean entropy standardised by its maximum.

**Convexity**

By construction,  $D_{\text{KL}}(Q, P)$  is strictly convex in  $Q$  as  $H(Q)$  is strictly concave in  $Q$  and  $H(Q, P)$  is mixture-preserving.

**Interpretation**

The KL divergence  $D_{\text{KL}}(Q, P)$  can be interpreted as the excess risk or additional cost incurred if model  $P$  is used to describe data that actually follow  $Q$ . If  $Q$  and  $P$  are identical there is no extra cost and  $D_{\text{KL}}(Q, P) = 0$ . However, if they differ then there is an additional cost and  $D_{\text{KL}}(Q, P) > 0$ .

**Invariance under a change of variables**

A further crucial property of KL divergence is its **invariance property** with respect to reparametrisation of the sample space. Specifically,  $D_{\text{KL}}(Q, P)$  is **invariant under general invertible variable transformations**, so that  $D_{\text{KL}}(Q_y, P_y) = D_{\text{KL}}(Q_x, P_x)$  under a change of the random variable from  $x$  to  $y$ .

This is a remarkable property as it holds not just for discrete but also for continuous random variables. For comparison, recall that both differential entropy as well as mean log-loss for continuous random variables are not invariant under a change of variables.

In the definition of KL divergence the expectation is taken over a *ratio of two densities*. The invariance is created because the Jacobian determinant changes both densities in the same way under variable transformation and thus cancel out<sup>2</sup>.

**Data processing inequality**

More generally, the KL divergence also satisfies the **data processing inequality** (DPI). Applying a transformation (stochastic or deterministic, possibly coarsening) that produces  $y$  from  $x$ , cannot increase the divergence, so that  $D_{\text{KL}}(Q_x, P_x) \geq D_{\text{KL}}(Q_y, P_y)$ . Note that a change of variables is a special case of data processing (with identity).

Divergences between distributions satisfying the DPI (and hence being invariant under a change of variables) form the class of ***f*-divergences**. The KL divergence is the *only* divergence induced by a proper scoring rule (i.e. the only Bregman divergence) that is also an *f*-divergence.

**Further properties**

The KL divergence and mean log-loss inherit a number of further useful properties from proper scoring rules. We will not cover these in this text. For example, there are various **decompositions** for the risk and the divergence satisfies a **generalised Pythagorean theorem**.

In summary, KL divergence (and the underlying log-loss) stands out among divergences between distributions that is features many unique and desirable properties in a single quantity. It is therefore not surprising that it plays a central role in statistics and machine learning.

**Examples**

**Example 4.5.** KL divergence between two Bernoulli distributions  $B_1 = \text{Ber}(\theta_1)$  and  $B_2 = \text{Ber}(\theta_2)$ :

The “success” probabilities for the two distributions are  $\theta_1$  and  $\theta_2$ , respectively, and the complementary “failure” probabilities are  $1 - \theta_1$  and  $1 - \theta_2$ . With this we get for the KL divergence

$$D_{\text{KL}}(B_1, B_2) = \theta_1 \log \left( \frac{\theta_1}{\theta_2} \right) + (1 - \theta_1) \log \left( \frac{1 - \theta_1}{1 - \theta_2} \right)$$

<sup>2</sup>For a proof of the invariance property of KL divergence see Worksheet E1.



**Example 4.6.** KL divergence between two categorical distributions with  $K$  classes:

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

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

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

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

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

**Example 4.7.** KL divergence between two univariate normals with different means and common variance:

Assume  $P_1 = N(\mu_1, \sigma^2)$  and  $P_2 = N(\mu_2, \sigma^2)$ . Setting  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  in the more general case of the KL divergence between two normals (Example 4.9) yields

$$D_{\text{KL}}(P_1, P_2) = \frac{1}{2\sigma^2} (\mu_1 - \mu_2)^2$$

which, apart from a scale factor, is the **squared Euclidean distance** or **squared loss** between the two means  $\mu_1$  and  $\mu_2$ . Note that in this case the KL divergence is symmetric with regard to the two mean parameters.

**Example 4.8.** KL divergence between two univariate normals with common mean and different variances:

Assume  $P_1 = N(\mu, \sigma_1^2)$  and  $P_2 = N(\mu, \sigma_2^2)$ . Setting  $\mu_1 = \mu_2 = \mu$  in the more general case of the KL divergence between two normals (Example 4.9) yields

$$D_{\text{KL}}(P_1, P_2) = \frac{1}{2} \left( \frac{\sigma_1^2}{\sigma_2^2} - \log \left( \frac{\sigma_1^2}{\sigma_2^2} \right) - 1 \right)$$

This is a convex function of the ratio  $\sigma_1^2/\sigma_2^2$  of the two variances. Apart from the scale factor this is known as **Stein's loss** between the two variances  $\sigma_1^2$  and  $\sigma_2^2$ .

**Example 4.9.** KL divergence between two univariate normals with different means and variances:

Assume  $P_1 = N(\mu_1, \sigma_1^2)$  and  $P_2 = N(\mu_2, \sigma_2^2)$ . Then with Example 3.9 (entropy of normal) and Example 4.1 (mean log-loss between two normals) we get

$$\begin{aligned} D_{\text{KL}}(P_1, P_2) &= H(P_1, P_2) - H(P_1) \\ &= \frac{1}{2} \left( \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \log \left( \frac{\sigma_1^2}{\sigma_2^2} \right) - 1 \right) \end{aligned}$$

For the two special cases of equal variances ( $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ) and equal means ( $\mu_1 = \mu_2 = \mu$ ) see Example 4.7 and Example 4.8.

**Example 4.10.** ► KL divergence between two members of an exponential family:

Assume  $P(\eta)$  is an exponential family with canonical parameters  $\eta$ , canonical statistics  $t(x)$  and log-partition function  $a(\eta)$  with log-pdmf  $\log p(x|\eta) = \langle \eta, t(x) \rangle + \log h(x) - a(\eta)$ .

Then with  $E_{P(\eta_1)}(t(x)) = \mu_t(\eta_1)$  the KL divergence between two distributions  $P(\eta_1)$  and  $P(\eta_2)$  in this family is

$$\begin{aligned} D_{\text{KL}}(P(\eta_1), P(\eta_2)) &= E_{P(\eta_1)}(\log p(x|\eta_1) - \log p(x|\eta_2)) \\ &= \langle \eta_1 - \eta_2, \mu_t(\eta_1) \rangle - (a(\eta_1) - a(\eta_2)) \end{aligned}$$

This can also be obtained by computing

$$D_{\text{KL}}(P(\eta_1), P(\eta_2)) = H(P(\eta_1), P(\eta_2)) - H(P(\eta_1))$$

using the results from Example 3.10 and Example 4.4.

## 4.3 Further reading

A brief overview of proper scoring rules along with corresponding divergences is found in the supplementary [Probability and Distribution Refresher notes](#).

A broad review of scoring rules, including a discussion of properness, is found, e.g., in Winkler (1996).

See also: [Bregman divergences \(Wikipedia\)](#) and [f-divergences \(Wikipedia\)](#)



## 💡 A bit of history

Solomon Kullback (1907–1994) and Richard Leibler (1914–2003) introduced KL divergence in statistics in Kullback and Leibler (1951) under the term “discrimination information”. An earlier application of KL divergence in 1940/1941 is credited to Alan Turing (1912–1954) by Good (1979). Negative KL divergence as entropy with regard to a prior measure was discovered in 1878 by Ludwig Boltzmann (1844–1906) in statistical mechanics (see Section 6.1).

## 5 Local divergence

This chapter introduces the Fisher information matrix as the local curvature (the Hessian matrix) of the Kullback-Leibler divergence, serving as the local second-order sensitivity matrix for model parameters.

### 5.1 Fisher information

#### Local quadratic approximation of KL divergence

The Kullback-Leibler (KL) number measures the divergence between two distributions. We now study the KL divergence of two distributions within a parametric family separate only by some small  $\epsilon$ .

Specifically, we consider

$$\begin{aligned} D_{\text{KL}}(P(\theta), P(\theta + \epsilon)) &= E_{P(\theta)} (\log p(x|\theta) - \log p(x|\theta + \epsilon)) \\ &= h(\theta + \epsilon) \end{aligned}$$

where  $\theta$  is kept constant and  $\epsilon$  is varying. Assuming that the pdmf  $p(x|\theta)$  is **twice differentiable** with regard to  $\theta$  we can approximate the function  $h(\theta + \epsilon)$  quadratically by

$$h(\theta + \epsilon) \approx h(\theta) + \nabla h(\theta)^T \epsilon + \frac{1}{2} \epsilon^T \nabla \nabla^T h(\theta) \epsilon$$

From the familiar properties of the KL divergence we conclude

- $D_{\text{KL}}(P(\theta), P(\theta + \epsilon)) \geq 0$  and
- with equality only if  $\epsilon = 0$ .

Thus, by construction the function  $h(\theta + \epsilon)$  assumes a minimum at  $\epsilon = 0$  with  $h(\theta) = 0$  and a vanishing gradient  $\nabla h(\theta) = 0$ . Therefore, in the quadratic approximation of  $h(\theta + \epsilon)$  around  $\theta$  the first two terms (constant and linear) vanish and only the quadratic term remains:

$$h(\theta + \epsilon) \approx \frac{1}{2} \epsilon^T \nabla \nabla^T h(\theta) \epsilon$$



Furthermore, the Hessian matrix of  $h(\boldsymbol{\theta} + \boldsymbol{\varepsilon})$  evaluated at  $\boldsymbol{\varepsilon} = 0$  is given by

$$\begin{aligned}\nabla\nabla^T h(\boldsymbol{\theta}) &= -\mathbb{E}_{P(\boldsymbol{\theta})} \nabla\nabla^T \log p(\mathbf{x}|\boldsymbol{\theta}) \\ &= \mathcal{I}_P(\boldsymbol{\theta})\end{aligned}$$

This matrix  $\mathcal{I}_P(\boldsymbol{\theta})$  is known as the **Fisher information**. The index  $P$  serves as a reminder of the underlying model. It is also called **expected Fisher information** to emphasise that it is computed as the mean Hessian of the negative log-pdmf. The Fisher information matrix is always **symmetric** and **positive semidefinite**.

With its help the KL divergence can be locally approximated by

$$D_{\text{KL}}(P(\boldsymbol{\theta}), P(\boldsymbol{\theta} + \boldsymbol{\varepsilon})) \approx \frac{1}{2} \boldsymbol{\varepsilon}^T \mathcal{I}_P(\boldsymbol{\theta}) \boldsymbol{\varepsilon}$$

We may also vary the first argument in the KL divergence. It is straightforward to show that this leads to the same approximation to second order in  $\boldsymbol{\varepsilon}$ :

$$D_{\text{KL}}(P(\boldsymbol{\theta} + \boldsymbol{\varepsilon}), P(\boldsymbol{\theta})) \approx \frac{1}{2} \boldsymbol{\varepsilon}^T \mathcal{I}_P(\boldsymbol{\theta}) \boldsymbol{\varepsilon}$$

Hence, although the KL divergence is not symmetric in its arguments in general, it is symmetric to second order (locally symmetric).

#### **i** Note 5: ► Fisher information as metric tensor

In **information geometry** probability distributions are studied using tools from differential geometry. From this geometric perspective, smoothly parametrised distribution families  $P(\boldsymbol{\theta})$  are viewed as **manifolds**. In KL divergence geometry the Fisher information  $\mathcal{I}_P(\boldsymbol{\theta})$  serves as **metric tensor**, measuring local distances between nearby distributions.

Other types of divergences among distributions induce related geometries, with local metrics similarly obtained by quadratic approximation.

### Parameter identifiability

For a **regular model** the Fisher information is **positive definite** (with only positive eigenvalues) and hence **parameters are locally identifiable**.

Recall that a positive definite Hessian implies that  $h(\boldsymbol{\theta} + \boldsymbol{\varepsilon})$  has a true minimum at  $\boldsymbol{\theta}$ .

Conversely, for a **singular statistical model** the Fisher information matrix is **singular** (some or all of its eigenvalues vanish) at some parameter values. This indicates **local non-identifiability** arising, e.g., from overparametrisation, parameters linked by exact constraints, lower dimensional latent structure, parameters on boundaries or other regularity failures.

### Additivity of Fisher information

We may wish to compute the Fisher information based on a set of independent identically distributed (iid) random variables.

Assume that a random variable  $x \sim P(\boldsymbol{\theta})$  has log-pdmf  $\log p(x|\boldsymbol{\theta})$  and Fisher information  $\mathcal{I}_P(\boldsymbol{\theta})$ . The Fisher information  $\mathcal{I}_{P_{x_1} \dots P_{x_n}}(\boldsymbol{\theta})$  for a set of iid random variables  $x_1, \dots, x_n \sim P(\boldsymbol{\theta})$  is computed from the joint log-pdmf  $\log p(x_1, \dots, x_n|\boldsymbol{\theta}) = \sum_i \log p(x_i|\boldsymbol{\theta})$ . This yields

$$\begin{aligned}\mathcal{I}_{P_{x_1} \dots P_{x_n}}(\boldsymbol{\theta}) &= -\mathbb{E}_{P_{x_1} \dots P_{x_n}(\boldsymbol{\theta})} \nabla\nabla^T \sum_{i=1}^n \log p(x_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^n \mathcal{I}_{P_{x_i}}(\boldsymbol{\theta}) = n \mathcal{I}_{P_x}(\boldsymbol{\theta})\end{aligned}$$

Hence, the Fisher information for a set of  $n$  iid random variables equals  $n$  times the Fisher information of a single variable.

### Invariance property of Fisher information

Like KL divergence the **Fisher information is invariant under a change of variables in the sample space**, such as from  $x$  to  $y$  and from distribution  $P_x$  to  $P_y$  so that  $\mathcal{I}_{P_x}(\boldsymbol{\theta}) = \mathcal{I}_{P_y}(\boldsymbol{\theta})$ . This follows from the invariance of KL divergence under change of variables.

More specifically, when the random variable is changed from  $x$  to  $y$  the density will gain a factor in the form of a Jacobian determinant associated with this transformation. However, since this factor does not depend on the model parameters, it does not affect the first and second derivatives of the log-density with regard to the model parameters.



See also Section 7.4 for related sample space invariance of the gradient and curvature of the log-likelihood and Chapter 9 for the sample invariance of observed Fisher information.

### Scalar examples — single parameter models

**Example 5.1.** Fisher information for the Bernoulli distribution:

The log-pmf for the Bernoulli distribution  $\text{Ber}(\theta)$  is

$$\log p(x|\theta) = x \log \theta + (1-x) \log(1-\theta)$$

where  $\theta$  is the probability of “success”. The second derivative with regard to the parameter  $\theta$  is

$$\frac{d^2}{d\theta^2} \log p(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

Since  $E(x) = \theta$  we get as Fisher information

$$\begin{aligned} \mathcal{I}_P(\theta) &= -E\left(\frac{d^2}{d\theta^2} \log p(x|\theta)\right) \\ &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

**Example 5.2.** Quadratic approximations of the KL divergence between two Bernoulli distributions:

From Example 4.5 we have as KL divergence

$$D_{\text{KL}}(\text{Ber}(\theta_1), \text{Ber}(\theta_2)) = \theta_1 \log\left(\frac{\theta_1}{\theta_2}\right) + (1-\theta_1) \log\left(\frac{1-\theta_1}{1-\theta_2}\right)$$

and from Example 5.1 the corresponding Fisher information.

The quadratic approximation implies that

$$D_{\text{KL}}(\text{Ber}(\theta), \text{Ber}(\theta + \varepsilon)) \approx \frac{\varepsilon^2}{2} \mathcal{I}_{\text{Ber}}(\theta) = \frac{\varepsilon^2}{2\theta(1-\theta)}$$

and also that

$$D_{\text{KL}}(\text{Ber}(\theta + \varepsilon), \text{Ber}(\theta)) \approx \frac{\varepsilon^2}{2} \mathcal{I}_{\text{Ber}}(\theta) = \frac{\varepsilon^2}{2\theta(1-\theta)}$$

In Worksheet E1 this is verified by using a second order Taylor series applied to the KL divergence.

**Example 5.3.** Fisher information for the normal distribution with known variance.

The log-pdf for  $N(\mu, \sigma^2)$  is

$$\log p(x|\mu, \sigma^2) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (x - \mu)^2 - \frac{1}{2} \log(2\pi)$$

The second derivative with respect to  $\mu$  is

$$\frac{d^2}{d\mu^2} \log p(x|\mu, \sigma^2) = -\frac{1}{\sigma^2}$$

Therefore the Fisher information is

$$\mathcal{I}_P(\mu) = \frac{1}{\sigma^2}$$

### Matrix examples — multiple parameter models

**Example 5.4.** Fisher information for the normal distribution.

The log-pdf for  $N(\mu, \sigma^2)$  is

$$\log p(x|\mu, \sigma^2) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (x - \mu)^2 - \frac{1}{2} \log(2\pi)$$

The gradient with respect to  $\mu$  and  $\sigma^2$  (!) is the vector

$$\nabla \log p(x|\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} (x - \mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} (x - \mu)^2 \end{pmatrix}$$

Hint for calculating the gradient: replace  $\sigma^2$  by  $v$  and then take the partial derivative with regard to  $v$ , then substitute back.

The corresponding Hessian matrix is

$$\nabla \nabla^T \log p(x|\mu, \sigma^2) = \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{1}{\sigma^4} (x - \mu) \\ -\frac{1}{\sigma^4} (x - \mu) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} (x - \mu)^2 \end{pmatrix}$$

As  $E(x) = \mu$  we have  $E(x - \mu) = 0$ . Furthermore, with  $E((x - \mu)^2) = \sigma^2$  we see that  $E\left(\frac{1}{\sigma^6} (x - \mu)^2\right) = \frac{1}{\sigma^4}$ . Therefore the Fisher information matrix as the negative expected Hessian matrix is

$$\mathcal{I}_P(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$



**Example 5.5.** ► Fisher information for the canonical parameter of an exponential family:

Review Section 2.4 before studying this example.

Assume  $P(\eta)$  is an exponential family with canonical parameter vector  $\eta$ , canonical statistics  $t(x)$  and log-partition function  $a(\eta)$  with log-pdmf  $\log p(x|\eta) = \eta^T t(x) + \log h(x) - a(\eta)$ .

If we take second derivatives with regard to  $\eta$  all terms except for the last vanish:

$$\nabla \nabla^T \log p(x|\eta) = -\nabla \nabla^T a(\eta) = -\Sigma_t(\eta)$$

Then the Fisher information is

$$\begin{aligned} \mathcal{I}_P(\eta) &= -E_{P(\eta)} \nabla \nabla^T \log p(x|\eta) \\ &= E_{P(\eta)} \Sigma_t(\eta) \\ &= \Sigma_t(\eta) \end{aligned}$$

Hence, the Fisher information for the canonical parameter in an exponential family is the variance of the canonical statistics.

See Example 5.8 for a related but different result for the expectation parameters.

## 5.2 ► Reparametrisation

### ► Fisher information under reparametrisation

The Fisher information  $\mathcal{I}_P(\theta)$  depends on the parameter  $\theta$ . If we use a different parametrisation of the underlying distribution family, say  $P(\zeta)$  instead of  $P(\theta)$ , with a map  $\theta(\zeta)$  from  $\zeta$  to  $\theta$ , then the Fisher information changes according to the chain rule in calculus.

To find the resulting Fisher information in terms of the new parameter  $\zeta$  we need to use the Jacobian matrix  $D\theta(\zeta)$ . This matrix contains the gradients for each component of the map  $\theta(\zeta)$  in its rows:

$$D\theta(\zeta) = \begin{pmatrix} \nabla^T \theta_1(\zeta) \\ \nabla^T \theta_2(\zeta) \\ \vdots \end{pmatrix}$$

With the above the Fisher information for  $\theta$  is then transformed to the Fisher information for  $\zeta$  applying the chain rule for the Hessian matrix:

$$\mathcal{I}_P(\zeta) = (D\theta(\zeta))^T \mathcal{I}_P(\theta)|_{\theta=\theta(\zeta)} D\theta(\zeta)$$

This type of transformation is also known as *covariant transformation*.

### Examples

**Example 5.6.** ► Fisher information for the Bernoulli distribution from canonical to conventional parametrisation by change of variables:

From Example 2.2 and Example 5.5 the Fisher information for the Bernoulli distribution using the canonical parameter  $\eta$  is

$$\mathcal{I}_P(\eta) = a''(\eta) = \frac{e^\eta}{(e^\eta + 1)^2}$$

The map to the canonical parameter  $\eta$  from the conventional parameter  $\theta$  is the logit function  $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$  with Jacobian

$$D\eta(\theta) = \eta(\theta)' = \frac{1}{\theta(1-\theta)}$$

Using the chain rule to obtain the Fisher information for  $\theta$  with

$$\mathcal{I}_P(\eta)|_{\eta=\eta(\theta)} = \theta(1-\theta)$$

yields

$$\mathcal{I}_P(\theta) = (D\eta(\theta))^2 \mathcal{I}_P(\eta)|_{\eta=\eta(\theta)} = \frac{1}{\theta(1-\theta)}$$

which agrees with the result obtained by direct calculation in Example 5.1.

Note that the Fisher information for the mean parameter  $\theta$  is the inverse of the Fisher information for the canonical parameter  $\eta$ .

**Example 5.7.** ► Fisher information for the normal distribution from canonical to conventional parametrisation by change of variables:

From Example 2.3 and Example 5.5 the Fisher information matrix for the normal distribution using canonical parameters  $\eta$  is

$$\mathcal{I}_P(\eta) = \nabla \nabla^T a(\eta) = \begin{pmatrix} \frac{1}{1-2\eta_2} & \frac{2\eta_1}{(1-2\eta_2)^2} \\ \frac{2\eta_1}{(1-2\eta_2)^2} & \frac{4\eta_1^2 - 4\eta_2 + 2}{(1-2\eta_2)^3} \end{pmatrix}$$



The map to the canonical parameter  $\eta$  from the conventional parameters  $\theta = c(\mu, \sigma^2)^T$  is  $\eta = (\eta_1, \eta_2)^T = (\frac{\mu}{\sigma^2}, \frac{1}{2} - \frac{1}{2\sigma^2})^T$  with Jacobian matrix

$$D\eta(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & -\frac{\mu}{\sigma^4} \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

With

$$\mathcal{I}_P(\eta)|_{\eta=\eta(\theta)} = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{pmatrix}$$

this yields the Fisher information for  $\theta = c(\mu, \sigma^2)^T$  as

$$\begin{aligned} \mathcal{I}_P(\theta) &= (D\eta(\theta))^T \mathcal{I}_P(\eta)|_{\eta=\eta(\theta)} D\eta(\theta) \\ &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \end{aligned}$$

which agrees with the result obtained by direct calculation in Example 5.4.

**Example 5.8.** ► Fisher information for the expectation parameter of an exponential family:

An alternative parametrisation of an exponential family is provided by the means  $\mu_t$  of the canonical statistics  $t(x)$ . These expectation parameters are given by  $\mu_t(\eta) = \nabla a(\eta)$ .

The Jacobian for the transformation  $\mu_t(\eta)$  is  $D\mu_t(\eta) = D\nabla a(\eta) = \nabla \nabla^T a(\eta) = \Sigma_t(\eta)$ . Hence, the Jacobian for the inverse transformation  $\eta(\mu_t)$  is  $D\eta(\mu_t) = \Sigma_t(\eta(\mu_t))^{-1}$ .

The Fisher information transforms covariantly under change of parameter. With

$$\mathcal{I}_P(\eta)|_{\eta=\eta(\mu_t)} = \Sigma_t(\eta(\mu_t))$$

this yields the Fisher information for the expectation parameters  $\mu_t$

$$\begin{aligned} \mathcal{I}_P(\mu_t) &= (D\eta(\mu_t))^T \mathcal{I}_P(\eta)|_{\eta=\eta(\mu_t)} D\eta(\mu_t) \\ &= \Sigma_t(\eta(\mu_t))^{-1} \end{aligned}$$

Hence the Fisher information for the expectation parameters  $\mu_t$  is the inverse of the variance of the canonical statistics, and therefore the inverse of the Fisher information for the canonical parameters  $\eta$ .

This relationship has been encountered before in Example 5.6 for the special case of the Bernoulli distribution.

## 5.3 Further reading

Amari (2016) is a recent book and standard reference on information geometry.

For metrics associated with proper scoring rules see Dawid and Musio (2014).

### 💡 A bit of history

The Fisher information matrix was originally introduced by [Ronald A. Fisher \(1890–1962\)](#) in the 1925 to measure the precision of an estimator.

[C. Radhakrishna Rao \(1920–2023\)](#) showed in 1945 that the Fisher information matrix defines a local metric tensor on the parameter space and established its interpretation as a local sensitivity measure.

This insight later helped lead to the development of information geometry and to the study of singular or non-regular models in statistics.



## 6 Maximum entropy

This chapter introduces information entropy with prior measure, equal to the negative KL divergence and closely linked to Boltzmann entropy. Subsequently, the principle of maximum entropy is used to characterise distributions, naturally leading to exponential families.

### 6.1 Information entropy with prior measure

#### Entropy of a distribution with regard to a reference

The **information entropy with prior measure** of the distribution  $Q$  with reference to the distribution  $P$  is given by

$$\begin{aligned} G(Q, P) &= H(Q) - H(Q, P) \\ &= -E_Q \log \left( \frac{q(x)}{p(x)} \right) \\ &= -D_{\text{KL}}(Q, P) \end{aligned}$$

Note that this quantity equals the *negative* of the KL divergence between  $Q$  and  $P$  (Section 4.2).

We denote information entropy with prior measure by  $G(Q, P)$ , not to be confused with the mean log-loss  $H(Q, P)$ .

Standard information entropy  $H(Q)$  (see Section 3.2) is a special case of information entropy with prior measure  $G(Q, P)$  when the reference distribution  $P$  is uniform (see Example 6.1).

Standard information entropy  $H(Q)$  can be derived by combinatorial arguments (see Example 3.11). Information entropy with prior measure  $G(Q, P)$  is a natural generalisation of standard information entropy  $H(Q)$  and follows a related derivation (see Example 6.2).

#### **i** Note 6: Other names for information entropy with prior measure

Information entropy with regard to a reference distribution is known in physics as **generalised Boltzmann-Gibbs-Shannon entropy** (generalised BGS entropy).

Occasionally, it is called **relative BGS entropy** but that is ambiguous because relative entropy commonly refers to KL divergence. Many other proposals for “generalised” entropies exist in the literature. Hence it is recommended to use “information entropy with respect to a prior measure” for  $G(Q, P)$  rather than “relative” or “generalised” entropy.

#### Properties

Information entropy with prior measure shares its properties with KL divergence, such as invariance under change of variables, since the only difference is the sign.

As a result, it is always non-positive,  $G(Q, P) \leq 0$ , with the *maximum*  $G(Q, P) = 0$  achieved only for  $Q = P$ .

Furthermore, the negative sign compared to KL divergence means that information entropy with prior measure  $G(Q, P)$  is **strictly concave in  $Q$** , exactly like information entropy  $H(Q)$ . This ensures that optimisation with regard to  $Q$  yields a unique maximum (it one exists).

$G(Q, P)$  is equivalent to information entropy  $H(Q)$  when the reference  $P$  is the uniform distribution (Example 6.1).

**Example 6.1.** Information entropy with uniform prior measure:

Assume a distribution  $Q$  and the uniform distribution  $U$  with constant pmf  $u(x) = c$  over the domain of  $Q$ .

The mean log-loss  $H(Q, U)$  is

$$H(Q, U) = -E_Q(\log u(x)) = -\log c = \text{const.}$$

and does not depend on  $Q$ .

The information entropy of  $Q$  with reference to  $U$  is therefore

$$G(Q, U) = H(Q) - H(Q, U) = H(Q) + \text{const.}$$

Thus, the standard information entropy  $H(Q)$  is equivalent to information entropy  $G(Q, U)$  if the reference is the uniform distribution.



## Interpretation

$G(Q, P)$  measures the spread of probability mass relative to a reference distribution  $P$ .

This generalises from information entropy  $H(Q)$  which measures the spread of probability mass relative to the uniform distribution.

### ► Entropy as log-probability

As discussed in Section 3.3 Boltzmann's key insight in statistical mechanics was that entropy is a logarithmic measure of the "size" or "volume" of a macrostate.

Taking this volume to be proportional to the number of microstates compatible with the observed macrostate yields information entropy (see Example 3.11). A natural generalisation is to **measure the relative volume a macrostate by its probability** (recall that probability is a normalised positive measure). This leads to information entropy with prior measure (Example 6.2) and the general notion by Boltzmann of **entropy as log-probability**.

Specifically, for a macrostate represented by counts  $D = \{n_1, \dots, n_K\}$  for  $K$  classes with  $n = \sum_{k=1}^K n_k$ , or equivalently by  $\hat{Q} = \text{Cat}(\hat{q})$  with  $\hat{q}_k = n_k/n$ , the logarithm of the multiplicity  $W_K$  yields, for large  $n$ , the information entropy

$$\frac{1}{n} \log W_K \approx H(\hat{Q})$$

If instead we use the logarithm of the probability of the macrostate the above relation leads to information entropy with prior measure  $P$  with

$$\frac{1}{n} \log \Pr(D|p) \approx G(\hat{Q}, P)$$

where  $p = (p_1, \dots, p_K)^T$  are prior probabilities of each class with  $P = \text{Cat}(p)$ . See Example 6.2 for details.

The above link between probability and entropy written in terms of KL divergence as

$$\Pr(D|p) \approx e^{-n D_{\text{KL}}(\hat{Q}, P)}$$

forms the basis of **large deviations theory** (in probability) and the **method of types** (in information theory).

**Example 6.2.** ► Combinatorial derivation of information entropy with prior measure:

This is an extension of the combinatorial derivation of the standard information entropy (Example 3.11).

Assume  $\hat{Q} = \text{Cat}(\hat{q})$  is an empirical categorical distribution based on observed counts  $D = \{n_1, \dots, n_K\}$ , representing the macrostate, and  $P = \text{Cat}(p)$  is a second categorical distribution with class probabilities  $p = (p_1, \dots, p_K)^T$ . For large  $n$  we may use the multinomial coefficient  $W_K = \binom{n}{n_1, \dots, n_K}$  to obtain the entropy of  $\hat{Q}$  via  $\log W_K \approx nH(\hat{Q})$  (Example 3.11).

The standardised multinomial log-probability of  $D$  given model  $p$  then yields

$$\begin{aligned} \frac{1}{n} \log p(D|p) &= \frac{1}{n} \log \left( W_K \times \prod_{k=1}^K p_k^{n_k} \right) \\ &= \frac{1}{n} \left( \log W_K + \sum_{k=1}^K n_k \log p_k \right) \\ &\approx H(\hat{Q}) + \frac{1}{n} \sum_{k=1}^K n_k \log p_k \\ &= H(\hat{Q}) + \sum_{k=1}^K \hat{q}_k \log p_k \\ &= H(\hat{Q}) - H(\hat{Q}, P) \\ &= G(\hat{Q}, P) \end{aligned}$$

and hence

$$\frac{1}{n} \log p(D|p) \approx G(\hat{Q}, P)$$

Standard information entropy  $H(\hat{Q})$  implicitly assumes that every possible microstate, regardless of its associated macrostate, has the same probability (equal to  $1/K^n$ ) and that the probabilities of each class are equal ( $p_k = 1/K$ ). In contrast, information entropy  $G(\hat{Q}, P)$  with prior measure  $P$  only assumes that microstates belonging to the same macrostate have identical probability (equal to  $\prod_{k=1}^K p_k^{n_k}$ ) and also allows for unequal class probabilities  $p_k$ .



## 6.2 Maximum entropy principle to characterise distributions

### Rationale of maximum entropy

In physics, systems at equilibrium are characterised by maximum entropy, subject to constraints. For example, in statistical mechanics maximum entropy characterised the most typical macrostate (see Section 3.3).

As both entropy with prior measure  $G(Q, P)$  and standard entropy  $H(Q)$  are strictly concave in  $Q$  (with the reference  $P$  fixed or implicitly set to the uniform distribution), the maximum-entropy distribution is unique (it exists).

As discussed in Chapter 3 and Section 6.1, **large entropy** implies that the **distribution is spread out** (relative to the reference  $P$ ). Correspondingly, **maximum-entropy distributions** can be considered **minimally informative** about a random variable.

Intriguingly, many distributions (and distribution families) commonly used in statistical modelling are characterised by maximum entropy. For example:

- 1) The **discrete uniform distribution** is the unique maximum-entropy distribution among all categorical distributions. See Example 6.3.
- 2) The **exponential distribution** is the unique maximum-entropy distribution among all continuous distributions supported in  $[0, \infty]$  with a specified mean. See Example 6.4.
- 3) The **normal distribution** is the unique maximum-entropy distribution among all continuous distributions supported in  $[-\infty, \infty]$  with a specified mean and variance.
- 4) More generally, **exponential families** also follow from the principle of maximum entropy subject to specified expectation parameters. See Example 6.5.

Therefore, there is a very close link between the principle of maximum entropy and widely used choices for models in statistics and machine learning.

### Examples

**Example 6.3.** Discrete uniform distribution as maximum-entropy distribution:

Let  $Q = \text{Cat}(\mathbf{q})$  be a general categorical distribution with  $K$  classes and corresponding class probabilities  $\mathbf{q} = (q_1, \dots, q_K)^T$  and  $P = \text{Cat}(\mathbf{p})$  be the discrete uniform distribution with equal class probabilities  $\mathbf{p} = (p_1 = 1/K, \dots, p_K = 1/K)$ .

We now show that the entropy  $H(Q)$  is maximised uniquely for  $Q = P$ , i.e. if  $Q$  is the discrete uniform distribution.

First, note that  $H(Q)$  is equivalent to  $G(Q, P)$  since  $P$  is the uniform distribution (Example 6.1). Hence, maximising  $H(Q)$  is equivalent to maximising  $G(Q, P)$  with regard to  $Q$ . The entropy  $G(Q, P) = 0$  achieves its maximum uniquely for  $Q = P$ , thus  $H(Q)$  is also maximised uniquely for  $Q = P$ .

Alternatively, we can also show directly that  $H(P) \geq H(Q)$  with equality only for  $Q = P$ . i.e. that the entropy of  $Q$  is always smaller than the entropy of  $P$  unless  $Q = P$ . The entropy of the discrete uniform distribution  $P$  is  $H(P) = \log K$  (see Example 3.4). The mean log-loss between  $Q$  and  $P$  is  $H(Q, P) = -\sum_{k=1}^K q_k \log p_k = \log K$ . Gibbs' inequality states that  $H(Q, P) \geq H(Q)$ , with equality only if  $P = Q$ . Since in our case  $H(Q, P) = H(P) = \log K$  it follows directly that  $H(P) \geq H(Q)$ .

The uniqueness of the maximum follows from Gibbs' inequality and from the strict concavity of information entropy.

**Example 6.4.** Exponential distribution as maximum-entropy distribution:

Assume  $Q$  is a continuous distribution for  $x$  with support  $[0, \infty]$  and with specified mean  $E_Q(x) = \theta$ . Furthermore,  $P = \text{Exp}(\theta)$  is an exponential distribution with scale parameter  $\theta$  and log-density  $\log p(x|\theta) = x/\theta - \log \theta$  and  $E_P(x) = \theta$ .

We now show that the entropy  $H(Q)$  is maximised uniquely for  $Q = P$ , i.e. if  $Q$  is an exponential distribution with specified mean.

The information entropy of  $P$  is  $H(P) = -E_P(\log p(x|\theta)) = 1 + \log \theta$  as  $E_P(x) = \theta$ . The mean log-loss is  $H(Q, P) = -E_Q(\log p(x|\theta)) = 1 + \log \theta$  as  $E_Q(x) = \theta$ . Gibbs' inequality states that  $H(Q, P) \geq H(Q)$ . Since in our case  $H(Q, P) = H(P)$  it follows directly that  $H(P) \geq H(Q)$ .



Therefore, the exponential distribution  $P$  achieves maximum entropy, and the distribution  $Q$  will have lower entropy than  $P$ , unless  $Q = P$ . The maximum is unique because of Gibb's inequality and the strict concavity of information entropy.

**Example 6.5.** ▶ Exponential families as maximum-entropy distributions:

Exponential families (see Section 2.4) are also characterised by achieving maximum entropy with specified expectation parameters.

Specifically, assume  $B$  is a base distribution with pmf  $b(x)$  and  $P = P(\eta)$  is an exponential family constructed by exponential tilting  $B$  with log-pmf  $p(x|\eta) = \langle \eta, t(x) \rangle + \log b(x) - a(\eta)$ . The expectation parameters are specified as  $E_P(t(x)) = \mu_t$  where  $t(x)$  are the canonical statistics. Let  $Q$  be a distribution on the same domain as  $B$  and  $P$  subject to the constraint that  $E_Q(t(x)) = \mu_t$ .

We now show that the entropy  $G(Q, B)$  is maximised uniquely for  $Q = P$ , i.e. if  $Q$  is an exponential family constructed from  $B$  with the specified expectation parameters.

The entropy of  $P$  is  $H(P) = -\langle \eta, \mu_t \rangle + H(P, B) + a(\eta)$ . The mean log-loss between  $Q$  and  $P$  is  $H(Q, P) = -\langle \eta, \mu_t \rangle + H(Q, B) + a(\eta)$ . Gibbs' inequality states that  $H(Q, P) \geq H(Q)$ . Since in our case  $H(Q, P) = H(P) - H(P, B) + H(Q, P)$  it follows directly that  $H(P) - H(P, B) \geq H(Q) - H(Q, P)$  and thus  $G(P, B) \geq G(Q, B)$ .

Therefore, the exponential family  $P$  achieves maximum entropy and the distribution  $Q$  will have lower entropy than  $P$ , unless  $Q = P$ . The maximum is unique because of Gibb's inequality and the strict concavity of information entropy.

## 6.3 Further reading

See Akaike (1985) for a historical account of Boltzmann entropy as log-probability and the discovery of information entropy with prior measure by Boltzmann in 1878.

Jaynes (2003) provides a broad discussion of the principle of maximum entropy as a foundation for statistical learning.

See also: [Maximum-entropy probability distributions \(Wikipedia\)](#)

### 💡 A bit of history

The concept of entropy as log-probability is due to [Ludwig Boltzmann \(1844–1906\)](#). Boltzmann discovered the form of information entropy with prior measure in 1878, following the earlier 1877 formulation of standard information entropy.

[Max Planck \(1958–1947\)](#) was the first to explicitly state the fundamental relationship between entropy and log-probability (with credit to Boltzmann), using it to derive the formula for black-body radiation in 1900.

Maximum entropy originated in thermodynamics as an empirical characterisation of equilibrium states. [Josiah W. Gibbs \(1839–1903\)](#) used the principle of maximum entropy subject to constraints to explicitly derive the canonical distribution (the physics name for exponential families) to represent equilibrium states in statistical mechanics.

[Solomon Kullback \(1907–1994\)](#) and [Richard Leibler \(1914–2003\)](#) advocated the principle of minimum KL divergence in statistics.

[Edwin T. Jaynes \(1922–1998\)](#) unified information theory and statistical mechanics and promoted the principle of maximum entropy as a general method for statistical learning and updating uncertainty.



## Part III

# Likelihood

## 7 Principle of maximum likelihood

### 7.1 Overview

#### Outline of maximum likelihood estimation

**Maximum likelihood** is a very general method for fitting probabilistic models to data, generalising the earlier method of least-squares. Consequently, it plays a very important role in statistics.

In a nutshell, the starting points in a maximum likelihood analysis are

- i) the observed data  $D = \{x_1, \dots, x_n\}$  with  $n$  independent and identically distributed (iid) samples, with the ordering irrelevant, and a
- ii) a model  $P(\theta)$  with corresponding probability density or probability mass function  $p(x|\theta)$  and parameters  $\theta$

From model and data the likelihood function (note upper case “L”) is constructed as

$$L_n(\theta) = L(\theta|D) = \prod_{i=1}^n p(x_i|\theta)$$

Equivalently, the log-likelihood function (note lower case “ℓ”) is

$$\ell_n(\theta) = \ell(\theta|D) = \sum_{i=1}^n \log p(x_i|\theta)$$

In the above notation, we either explicitly mention the data  $D$  or use an index to remind us of the sample size (here  $n$ ).

The likelihood is multiplicative and the log-likelihood additive over the samples  $x_i$  because of the iid assumption.



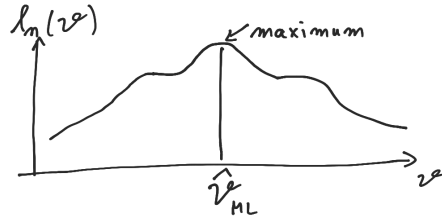


Figure 7.1: Finding the maximum likelihood estimate by maximisation of the (log)-likelihood function.

The maximum likelihood estimate (MLE)  $\hat{\theta}^{ML}$  is then found by maximising the (log)-likelihood function with regard to the parameters  $\theta$  (see Figure 7.1):

$$\hat{\theta}_{ML} = \arg \max_{\theta} \ell_n(\theta)$$

Hence, once the model is chosen and data are collected, finding the MLE and thus fitting the model to the data is an *optimisation problem*.

Depending on the complexity of the likelihood function and the number of the parameters finding the maximum likelihood can be very difficult. On the other hand, for likelihood functions constructed from common distribution families, such as exponential families, maximum likelihood estimation is very straightforward and can even be done analytically (this is the case for most examples we encounter in this course).

In practice in application to more complex models the optimisation required for maximum likelihood analysis is done on the computer, typically on the log-likelihood rather than on the likelihood function in order to avoid problems with the computer representation of small floating point numbers. Suitable optimisation algorithm may rely only on function values without requiring derivatives, or use in addition gradient and possibly curvature information. In recent years there has been a lot of progress in high-dimensional optimisation using combined numerical and analytical approaches (e.g. using automatic differentiation) and stochastic approximations (e.g. stochastic gradient descent).

### Origin of the method of maximum likelihood

Historically, the likelihood has often been interpreted and justified as the probability of the data given the model. However, while providing

an intuitive understanding this is not strictly correct. First, this interpretation only applies to discrete random variables. Second, since the samples  $x_1, \dots, x_n$  are typically **exchangeable** (i.e. permutation invariant) even in this case one would still need to add a factor accounting for the multiplicity of the possible orderings of the samples to obtain the correct probability of the data. Third, the interpretation of likelihood as probability of the data completely breaks down for continuous random variables because in that case  $p(x|\theta)$  is a density, not a probability.

Next, we will see that maximum likelihood estimation is a well-justified method that arises naturally from an entropy perspective. More specifically, the maximum likelihood estimate corresponds to the distribution  $P(\theta)$  that is closest in terms of KL divergence to the unknown true data-generating model as represented by the observed data and the empirical distribution.

## 7.2 From minimum KL divergence to maximum likelihood

### The KL divergence between true model and approximating model

Assume we have observations  $D = \{x_1, \dots, x_n\}$ . The data are sampled from  $Q$ , the true but unknown data-generating distribution. We also specify a family of distributions  $P(\theta)$  indexed by  $\theta$  to approximate  $Q$ .

The KL divergence  $D_{KL}(Q, P(\theta))$  measures the divergence of the approximation  $P(\theta)$  from the unknown true model  $Q$ . It can be written as

$$\begin{aligned} D_{KL}(Q, P(\theta)) &= H(Q, P(\theta)) - H(Q) \\ &= \underbrace{-E_Q \log p(x|\theta)}_{\text{mean log-loss}} - \underbrace{(-E_Q \log q(x))}_{\text{entropy of } Q, \text{ does not depend on } \theta} \end{aligned}$$

However, since we do not know  $Q$  we cannot actually compute this divergence. Nonetheless, we may use the empirical distribution  $\hat{Q}_n$  — a function of the observed data — as approximation for  $Q$ , and in this way we arrive at an approximation for  $D_{KL}(Q, P(\theta))$  that becomes more and more accurate with growing sample size.



Recall the “Law of Large Numbers” :

- The empirical distribution  $\hat{Q}_n$  based on observed data  $D = \{x_1, \dots, x_n\}$  converges strongly (almost surely) to the true underlying distribution  $Q$  as  $n \rightarrow \infty$ :

$$\hat{Q}_n \xrightarrow{a.s.} Q$$

- Correspondingly, for  $n \rightarrow \infty$  the average  $E_{\hat{Q}_n}(h(x)) = \frac{1}{n} \sum_{i=1}^n h(x_i)$  converges to the expectation  $E_Q(h(x))$ .

Hence, for large sample size  $n$  we can approximate the mean log-loss and as a result the KL divergence. The mean log-loss  $H(Q, P(\theta))$  is approximated by the **empirical mean log-loss** where the expectation is taken with regard to  $\hat{Q}_n$  rather than  $Q$  (see Example 4.3):

$$\begin{aligned} H(Q, P(\theta)) &\approx H(\hat{Q}_n, P(\theta)) \\ &= -\frac{1}{n} \ell_n(\theta) \end{aligned}$$

The empirical risk is equal to the negative log-likelihood standardised by the sample size  $n$ .

### Minimum KL divergence and maximum likelihood

If we knew  $Q$  we would simply minimise  $D_{KL}(Q, P(\theta))$  to find the particular model  $P(\theta)$  that is closest to the true model, or equivalently, we would minimise the mean log-loss  $H(Q, P(\theta))$ . However, since we actually don't know  $Q$  this is not possible.

However, for large sample size  $n$  when the empirical distribution  $\hat{Q}_n$  is a good approximation for  $Q$ , we can use the results from the previous section. Thus, instead of minimising the KL divergence  $D_{KL}(Q, P(\theta))$  we simply minimise  $H(\hat{Q}_n, P(\theta))$  which is the same as maximising the log-likelihood  $\ell_n(\theta)$ .

Conversely, this implies that maximising the likelihood with regard to the  $\theta$  is equivalent (asymptotically for large  $n$ !) to minimising the KL divergence of the approximating model and the unknown true model.

$$\begin{aligned} \hat{\theta}_{ML} &= \arg \max_{\theta} \ell_n(\theta) \\ &= \arg \min_{\theta} H(\hat{Q}_n, P(\theta)) \\ &\approx \arg \min_{\theta} D_{KL}(Q, P(\theta)) \end{aligned}$$

Therefore, the reasoning behind the method of **maximum likelihood** is that it minimises a **large sample approximation of the KL divergence** of the candidate model  $P(\theta)$  from the unknown true model  $Q$ . In other words, **maximum likelihood estimators are minimum empirical KL divergence estimators**.

As the KL divergence is a functional of the true distribution  $Q$  **maximum likelihood provides empirical estimators for parametric models**.

As a consequence of the close link of maximum likelihood and KL divergence maximum likelihood inherits for large  $n$  (and only then!) all the optimality properties from KL divergence.

## 7.3 Properties of maximum likelihood estimation

### Consistency of maximum likelihood estimates

One important property of the method of maximum likelihood is that in general it produces **consistent estimates**. This means that estimates are well behaved so that they become more accurate with more data and in the limit of infinite data converge to the true parameters.

Specifically, if the true underlying model  $Q$  is contained in the set of specified candidates models  $P(\theta)$

$$\underbrace{Q}_{\text{true model}} \subset \underbrace{P(\theta)}_{\text{specified models}}$$

there is a parameter  $\theta_{\text{true}}$  for which  $Q = P(\theta_{\text{true}})$ .



Maximisation of the likelihood function is, for large  $n$ , equivalent to minimising the KL divergence, and  $D_{\text{KL}}(Q, P(\theta)) \rightarrow 0$  implies  $P(\theta) \rightarrow Q$ , hence

$$\hat{\theta}_{ML} \xrightarrow{\text{large } n} \theta_{\text{true}}$$

Thus, given sufficient data, the maximum likelihood estimate of the model parameters will converge to the true parameter values. Note that the above assumes the model family, and therefore the number of parameters, is fixed.

As a consequence of consistency, **maximum likelihood estimates are asymptotically unbiased**. As we will see in the examples, they can still be biased in finite samples.

Note that even if the candidate model family  $P(\theta)$  is **misspecified** (so that it does not contain the actual true model), the maximum likelihood estimate is still optimal, for large sample size, in the sense in that it will identify the distribution in the family that is closest in terms of KL divergence.

Finally, inconsistent maximum likelihood estimates can occur, but only in non-regular situations, for example when the MLE lies on a boundary or when there are singularities in the likelihood function. Furthermore, inconsistency can arise when the number of parameters grows with the sample size and hence the information per parameter does not sufficiently increase (as in the well-known Neyman–Scott paradox).

### Invariance property of the maximum likelihood

Maximum likelihood is **invariant under reparametrisation**, i.e. under transformation of the model parameters the achieved maximum likelihood stays the same. This property is shared by KL divergence minimisation as the achieved minimum KL divergence is also invariant under a change of parameters.

A parameter is just an arbitrary label to index a specific distribution within a distribution family, and changing that label does not affect the maximum (likelihood) or the minimum (KL divergence). For example, consider a function  $h_x(x)$  with a maximum at  $x_{\text{max}} = \arg \max h_x(x)$ . Now relabel the argument using  $y = g(x)$  where  $g$  is an invertible function. Then the function in terms of  $y$  is  $h_y(y) = h_x(g^{-1}(y))$ . Clearly, this function has a maximum at  $y_{\text{max}} = g(x_{\text{max}})$  since  $h_y(y_{\text{max}}) = h_x(g^{-1}(y_{\text{max}})) = h_x(x_{\text{max}})$ . Furthermore, the achieved maximum value is the same.

In application to maximum likelihood, assume we transform a parameter  $\theta$  into another parameter  $\omega$  using an invertible function so that  $\omega = g(\theta)$ . Then the maximum likelihood estimate  $\hat{\omega}_{ML}$  of the new parameter  $\omega$  is simply the transformation of the maximum likelihood estimate  $\hat{\theta}_{ML}$  of the original parameter  $\theta$  with  $\hat{\omega}_{ML} = g(\hat{\theta}_{ML})$ . The achieved maximum likelihood is the same in both cases.

The invariance property can be very useful in practice because it is often easier (and sometimes numerically more stable) to maximise the likelihood using a different set of parameters.

See Worksheet L1 for an example application of the invariance property.

### Sufficient statistics

Another important concept are so-called sufficient statistics to summarise the information available in the data about a parameter in a model.

A statistic  $t(D)$  is called a **sufficient statistic** for the model parameters  $\theta$  if the corresponding likelihood function can be written using only  $t(D)$  in the terms that involve  $\theta$  such that

$$L(\theta|D) = h(t(D), \theta) k(D),$$

where  $h()$  and  $k()$  are positive-valued functions. This is known as the **Fisher-Pearson factorisation**. Equivalently on log-scale this becomes

$$\ell(\theta|D) = \log h(t(D), \theta) + \log k(D).$$

By construction, estimation and inference about  $\theta$  based on the factorised likelihood  $L(\theta)$  is mediated through the sufficient statistic  $t(D)$  and does not require knowledge of the original data  $D$ . Instead, the sufficient statistic  $t(D)$  contains all the information in  $D$  required to learn about the parameter  $\theta$ .

Note that a **sufficient statistic always exists** since the data  $D$  are themselves sufficient statistics, with  $t(D) = D$ . However, in practice one aims to find sufficient statistics that summarise the data  $D$  and hence provide data reduction. This will become clear in the examples below.

Furthermore, sufficient statistics are **not unique** since applying a one-to-one transformation to  $t(D)$  yields another sufficient statistic.

Therefore, if the MLE  $\hat{\theta}_{ML}$  of  $\theta$  exists and is unique then **the MLE is a unique function of the sufficient statistic  $t(D)$** . If the MLE is not unique then it can be chosen to be function of  $t(D)$ .



## 7.4 Maximum likelihood estimation for regular models

### Regular models

A regular model is one that is well-behaved and well-suited for model fitting by optimisation. In particular this requires that:

- the support does not depend on the parameters,
- the model is identifiable (in particular the model is not over-parametrised and has a minimal set of parameters),
- the density/probability mass function and hence the log-likelihood function is twice differentiable everywhere with regard to the parameters,
- the maximum (peak) of the likelihood function lies inside the parameter space and not at a boundary,
- the second derivative of the log-likelihood at the maximum is negative and not zero (for multiple parameters: the Hessian matrix at the maximum is negative definite and not singular)

Most models considered in this course are regular.

### Maximum likelihood estimation in regular models

For a regular model maximum likelihood estimation and the necessary optimisation is greatly simplified by being able to using gradient and curvature information.

In order to maximise  $\ell_n(\theta)$  one may use the **score function**  $S_n(\theta)$  which is the first derivative of the log-likelihood function with regard to the parameters<sup>1</sup>.

$$S_n(\theta) = \frac{d\ell_n(\theta)}{d\theta} \quad \text{scalar parameter } \theta: \text{ first derivative}$$

$$S_n(\theta) = \nabla \ell_n(\theta) \quad \text{multivariate parameter } \theta: \text{ gradient}$$

<sup>1</sup>The score function  $S_n(\theta)$  as the gradient of the log-likelihood function must not be confused with the scoring rule  $S(x, P)$ .

In this case a necessary (but not sufficient) condition for the MLE is that

$$S_n(\hat{\theta}_{ML}) = 0$$

To demonstrate that the log-likelihood function actually achieves a maximum at  $\hat{\theta}_{ML}$  the curvature at the MLE must be negative, i.e. that the log-likelihood must be locally concave at the MLE.

In the case of a single parameter (scalar  $\theta$ ) this requires to check that the second derivative of the log-likelihood function with regard to the parameter is negative:

$$H_n(\theta) = \frac{d^2 \ell_n(\theta)}{d\theta^2}$$

and

$$H_n(\hat{\theta}_{ML}) < 0$$

In the case of a parameter vector (multivariate  $\theta$ ) one first computes the Hessian matrix (matrix of second order derivatives) of the log-likelihood function

$$H_n(\theta) = \nabla \nabla^T \ell_n(\theta)$$

For a multivariate parameter vector  $\theta$  of dimension  $d$  the Hessian is a matrix of size  $d \times d$ . Then one needs to verify that the Hessian matrix is negative definite at the MLE:

$$H_n(\hat{\theta}_{ML}) < 0,$$

i.e. all its eigenvalues must be negative.

### Invariance of score function and second derivative of the log-likelihood

The score function  $S_n(\theta)$  is **invariant under a transformation of the sample space**. Assume  $x$  has log-density  $\log f_x(x|\theta)$  then the log-density for  $y$  is

$$\log f_y(y|\theta) = \log | \det(Dx(y)) | + \log f_x(x(y)|\theta)$$

where  $Dx(y)$  is the Jacobian matrix of the inverse transformation  $x(y)$ . When taking the derivative of the log-likelihood function with regard to the parameter  $\theta$  the first term containing the Jacobian determinant



vanishes. Hence the score function  $S_n(\theta)$  is not affected by a change of variables.

As a consequence, the second derivative of log-likelihood function with regard to  $\theta$  is also invariant under transformations of the sample space.

## 7.5 Further reading

The developments that lead to the theory of maximum likelihood are reviewed, e.g., in Aldrich (1997), Stiegler (2007) and Hand (2015).

See Kullback (1959) for a view of maximum likelihood and corresponding statistical procedures in terms of minimising KL divergence.

### A bit of history

The method of maximum likelihood as unified statistical theory was developed by [Ronald A. Fisher](#) in the early 20th century. A key seminal paper is Fisher (1922).

The equivalence of maximum likelihood estimation and minimisation of Kullback–Leibler divergence was first made explicit in the 1950s, see Kullback (1959).

# 8 Maximum likelihood estimation in practice

Next, maximum likelihood estimation is illustrated on a number of examples. Among others we discuss three basic problems, namely how to estimate a proportion, the mean and the variance in the likelihood framework. We also consider an example of non-regular model (Example 8.4).

## 8.1 Likelihood estimation for a single parameter

In the following we illustrate likelihood estimation for models with a single parameter. In this case the score function and the second derivative of the log-likelihood are all scalar-valued like the log-likelihood function itself.

**Example 8.1.** Maximum likelihood estimation for the Bernoulli model:

We aim to estimate the true proportion  $\theta$  in a Bernoulli experiment with binary outcomes, say the proportion of “successes” vs. “failures” or of “heads” vs. “tails” in a coin tossing experiment.

- Bernoulli model  $\text{Ber}(\theta)$ :  $\Pr(\text{“success”}) = \theta$  and  $\Pr(\text{“failure”}) = 1 - \theta$ .
- The “success” is indicated by outcome  $x = 1$  and the “failure” by  $x = 0$ .
- We conduct  $n$  trials and record  $n_1$  successes and  $n - n_1$  failures.
- Parameter:  $\theta$  probability of “success”.

What is the MLE of  $\theta$ ?

- the observations  $D = \{x_1, \dots, x_n\}$  take on values 0 or 1.
- the average of the data points is  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n_1}{n}$ .



- the probability mass function (PMF) of the Bernoulli distribution  $\text{Ber}(\theta)$  is:

$$p(x|\theta) = \theta^x(1-\theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

- log-PMF:

$$\log p(x|\theta) = x \log \theta + (1-x) \log(1-\theta)$$

- log-likelihood function:

$$\begin{aligned} \ell_n(\theta) &= \sum_{i=1}^n \log p(x_i|\theta) \\ &= n_1 \log \theta + (n - n_1) \log(1 - \theta) \\ &= n (\bar{x} \log \theta + (1 - \bar{x}) \log(1 - \theta)) \end{aligned}$$

Note that the log-likelihood depends on the data only via  $\bar{x}$ . Thus,  $t(D) = \bar{x}$  is a **sufficient statistic** for the parameter  $\theta$ . In fact it is also a **minimal sufficient statistic** as will be discussed in more detail later.

- Score function:

$$S_n(\theta) = \frac{d\ell_n(\theta)}{d\theta} = n \left( \frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta} \right)$$

- Maximum likelihood estimate: Setting  $S_n(\hat{\theta}_{ML}) = 0$  yields as solution

$$\hat{\theta}_{ML} = \bar{x} = \frac{n_1}{n}$$

With  $H_n(\theta) = \frac{dS_n(\theta)}{d\theta} = -n \left( \frac{\bar{x}}{\theta^2} + \frac{1-\bar{x}}{(1-\theta)^2} \right) < 0$  the optimum corresponds indeed to the maximum of the (log-)likelihood function as this is negative for  $\hat{\theta}_{ML}$  (and indeed for any  $\theta$ ).

The maximum likelihood estimator of  $\theta$  is therefore identical to the frequency of the successes among all observations.

Note that to analyse the coin tossing experiment and to estimate  $\theta$  we may equally well use the binomial distribution  $\text{Bin}(n, \theta)$  as model for the number of successes. This results in the same MLE for  $\theta$  but the likelihood function based on the binomial PMF includes the binomial coefficient. However, as it does not depend on  $\theta$  it disappears in the score function and has no influence in the derivation of the MLE.

**Example 8.2.** Maximum likelihood estimation for the normal distribution with unknown mean and known variance:

- $x \sim N(\mu, \sigma^2)$  with  $E(x) = \mu$  and  $\text{Var}(x) = \sigma^2$
- the parameter to be estimated is  $\mu$  whereas  $\sigma^2$  is known.

What's the MLE of the parameter  $\mu$ ?

- the data  $D = \{x_1, \dots, x_n\}$  are all real in the range  $x_i \in [-\infty, \infty]$ .
- the average  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is real as well.
- Density:

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

- Log-Density:

$$\log p(x|\mu) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

- Log-likelihood function:

$$\begin{aligned} \ell_n(\mu) &= \sum_{i=1}^n \log p(x_i|\mu) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad \underbrace{-\frac{n}{2} \log(2\pi\sigma^2)}_{\text{constant term, does not depend on } \mu, \text{ can be removed}} \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^2 - 2x_i\mu + \mu^2) + \text{const.} \\ &= \frac{n}{\sigma^2} (\bar{x}\mu - \frac{1}{2}\mu^2) \quad \underbrace{-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2}_{\text{another constant term}} + \text{const.} \end{aligned}$$

Note how the non-constant terms of the log-likelihood depend on the data only through  $\bar{x}$ . Hence  $t(D) = \bar{x}$  this is a **sufficient statistic** for  $\mu$ .

- Score function:

$$S_n(\mu) = \frac{n}{\sigma^2} (\bar{x} - \mu)$$



- Maximum likelihood estimate:

$$S_n(\hat{\mu}_{ML}) = 0 \Rightarrow \hat{\mu}_{ML} = \bar{x}$$

- With  $H_n(\mu) = \frac{dS_n(\mu)}{d\mu} = -\frac{n}{\sigma^2} < 0$  the optimum is indeed the maximum

The constant term in the log-likelihood function collects all terms that do not depend on the parameter of interest. After taking the first derivative with regard to the parameter the constant term disappears thus it has no influence in maximum likelihood estimation. **Therefore constant terms can be dropped from the log-likelihood function.**

**Example 8.3.** Maximum likelihood estimation for the normal distribution with known mean and unknown variance:

- $x \sim N(\mu, \sigma^2)$  with  $E(x) = \mu$  and  $\text{Var}(x) = \sigma^2$
- $\sigma^2$  needs to be estimated whereas the mean  $\mu$  is known

What's the MLE of  $\sigma^2$ ?

- the data  $D = \{x_1, \dots, x_n\}$  are all real in the range  $x_i \in [-\infty, \infty]$ .
- the average of the squared centred data  $\overline{(x - \mu)^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \geq 0$  is non-negative.
- Density:

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

- Log-Density:

$$\log p(x|\sigma^2) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x - \mu)^2}{2\sigma^2}$$

- Log-likelihood function:

$$\begin{aligned} \ell_n(\sigma^2) &= \ell_n(\mu, \sigma^2) = \sum_{i=1}^n \log p(x_i|\sigma^2) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad \underbrace{-\frac{n}{2} \log(2\pi)}_{\text{constant not depending on } \sigma^2} \\ &= -\frac{n}{2} \log \sigma^2 - \frac{n}{2\sigma^2} \overline{(x - \mu)^2} + C \end{aligned}$$

Note how the log-likelihood function depends on the data only through  $\overline{(x - \mu)^2}$ . Hence  $t(D) = \overline{(x - \mu)^2}$  is a sufficient statistic for  $\sigma^2$ .

- Score function:

$$S_n(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4} \overline{(x - \mu)^2}$$

Note that to obtain the score function the derivative needs to be taken with regard to the variance parameter  $\sigma^2$  — not with regard to  $\sigma$ ! As a trick, relabel  $\sigma^2 = v$  in the log-likelihood function, then take the derivative with regard to  $v$ , then backsubstitute  $v = \sigma^2$  in the final result.

- Maximum likelihood estimate:

$$S_n(\hat{\sigma}_{ML}^2) = 0 \Rightarrow$$

$$\hat{\sigma}_{ML}^2 = \overline{(x - \mu)^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

- To confirm that we actually have maximum we need to verify that the second derivative of log-likelihood at the optimum is negative. With  $H_n(\sigma^2) = \frac{dS_n(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^4} \left( \frac{2}{\sigma^2} \overline{(x - \mu)^2} - 1 \right)$  and hence  $H_n(\hat{\sigma}_{ML}^2) = -\frac{n}{2} \left( \hat{\sigma}_{ML}^2 \right)^{-2} < 0$  the optimum is indeed the maximum.

**Example 8.4.** Uniform distribution with upper bound  $\theta$ :

This is an example of a non-regular model, as the parameter  $\theta$  determines the support of the model.

- $x \sim \text{Unif}(0, \theta)$  with  $\theta > 0$
- the data  $D = \{x_1, \dots, x_n\}$  are all real in the range  $x_i \in [0, \theta]$ .
- by  $x_{[i]}$  we denote the *ordered* observations with  $0 \leq x_{[1]} < x_{[2]} < \dots < x_{[n]} \leq \theta$  with  $x_{[n]} = \max(x_1, \dots, x_n)$ .

We would like to obtain the maximum likelihood estimator  $\hat{\theta}_{ML}$ .

- The probability density function of  $\text{Unif}(0, \theta)$  is

$$p(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } x \in [0, \theta] \\ 0 & \text{otherwise.} \end{cases}$$



- the corresponding the log-density is

$$\log p(x|\theta) = \begin{cases} -\log \theta & \text{if } x \in [0, \theta] \\ -\infty & \text{otherwise.} \end{cases}$$

- the log-likelihood function is

$$\ell_n(\theta) = \begin{cases} -n \log \theta & \text{for } \theta \geq x_{[n]} \\ -\infty & \text{otherwise} \end{cases}$$

since all observed data  $D = \{x_1, \dots, x_n\}$  lie in the interval  $[0, \theta]$ . Note that the log-likelihood is a function of  $x_{[n]}$  only so this single data point is the sufficient statistic  $t(D) = x_{[n]}$ .

- the log-likelihood function remains at value  $-\infty$  until  $\theta = x_{[n]}$ , where it jumps to  $-n \log x_{[n]}$  and then it decreases monotonically with increasing  $\theta > x_{[n]}$ . Hence the log-likelihood function has a maximum at  $\hat{\theta}_{ML} = x_{[n]}$ .
- Due to the discontinuity at  $x_{[n]}$  the log-likelihood  $\ell_n(\theta)$  is **not differentiable** at  $\hat{\theta}_{ML}$ . and hence the maximum cannot be found by setting the score function equal to zero as in a regular model.
- In addition, **there is no quadratic approximation around  $\hat{\theta}_{ML}$**  and therefore the **observed Fisher information cannot be computed** either.

## 8.2 Likelihood estimation for multiple parameters

If there are several parameters likelihood estimation is conceptually no different from the case of a single parameter. However, the score function is now vector-valued and the second derivative of the log-likelihood is a matrix-valued function.

**Example 8.5.** Normal distribution with mean and variance both unknown:

- $x \sim N(\mu, \sigma^2)$  with  $E(x) = \mu$  and  $\text{Var}(x) = \sigma^2$
- both  $\mu$  and  $\sigma^2$  need to be estimated.

What's the MLE of the parameter vector  $\theta = (\mu, \sigma^2)^T$ ?

- the data  $D = \{x_1, \dots, x_n\}$  are all real in the range  $x_i \in [-\infty, \infty]$ .
- the average  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is real as well.
- the average of the squared data  $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2 \geq 0$  is non-negative.

- Density:

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

- Log-Density:

$$\log f(x|\mu, \sigma^2) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

- Log-likelihood function:

$$\begin{aligned} \ell_n(\theta) &= \ell_n(\mu, \sigma^2) = \sum_{i=1}^n \log f(x_i|\mu, \sigma^2) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \underbrace{\frac{n}{2} \log(2\pi)}_{\text{constant not depending on } \mu \text{ or } \sigma^2} \\ &= -\frac{n}{2} \log \sigma^2 - \frac{n}{2\sigma^2} (\overline{x^2} - 2\bar{x}\mu + \mu^2) + C \end{aligned}$$

Note how the log-likelihood function depends on the data only through  $\bar{x}$  and  $\overline{x^2}$ . Hence,  $t(D) = (\bar{x}, \overline{x^2})^T$  are sufficient statistics for  $\theta$ .

- Score function  $S_n$ , gradient of  $\ell_n(\theta)$ :

$$\begin{aligned} S_n(\theta) &= S_n(\mu, \sigma^2) \\ &= \nabla \ell_n(\mu, \sigma^2) \\ &= \begin{pmatrix} \frac{n}{\sigma^2} (\bar{x} - \mu) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4} (\overline{x^2} - 2\bar{x}\mu + \mu^2) \end{pmatrix} \end{aligned}$$

- Maximum likelihood estimate:

$$S_n(\hat{\theta}_{ML}) = 0 \Rightarrow$$

$$\hat{\theta}_{ML} = \begin{pmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}_{ML}^2 \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \overline{x^2} - \bar{x}^2 \end{pmatrix}$$



The ML estimate of the variance can also be written  $\widehat{\sigma}_{ML}^2 = \overline{x^2 - \bar{x}^2} = \overline{(x - \bar{x})^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ .

- To confirm that we actually have a maximum we need to verify that the eigenvalues of the Hessian matrix at the optimum are all negative. This is indeed the case, for details see Example 9.4.

**Example 8.6.** ► Maximum likelihood estimates of the parameters of the multivariate normal distribution:

The results from Example 8.5 can be generalised to the multivariate normal distribution:

- $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $E(\mathbf{x}) = \boldsymbol{\mu}$  and  $\text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$
- both  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  need to be estimated.

With

- the data  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  containing real vector-valued observations,

the maximum likelihood can be written as follows:

MLE for the mean:

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

MLE for the covariance:

$$\underbrace{\widehat{\boldsymbol{\Sigma}}_{ML}}_{d \times d} = \frac{1}{n} \sum_{k=1}^n \underbrace{(\mathbf{x}_k - \bar{\mathbf{x}})}_{d \times 1} \underbrace{(\mathbf{x}_k - \bar{\mathbf{x}})^T}_{1 \times d}$$

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

With  $\overline{\mathbf{x}\mathbf{x}^T} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T$  we can also write

$$\widehat{\boldsymbol{\Sigma}}_{ML} = \overline{\mathbf{x}\mathbf{x}^T} - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$$

Hence, the MLEs correspond to the well-known empirical estimates.

The derivation of the MLEs is discussed in more detail in the module [MATH38161 Multivariate Statistics and Machine Learning](#).

**Example 8.7.** ► Maximum likelihood estimation of the parameters of an exponential family model:

Review Section 2.4 before studying this example. Remarkably, below is a very general result summarising all the the previous examples in one single unified framework.

- $x \sim P(\boldsymbol{\eta})$  exponential family with canonical parameter vector  $\boldsymbol{\eta}$

What's the MLE of the parameters  $\boldsymbol{\eta}$ ?

We observe:

- the data  $D = \{x_1, \dots, x_n\}$
- the corresponding canonical statistics  $\mathbf{t}(x_1), \dots, \mathbf{t}(x_n)$
- their average  $\overline{\mathbf{t}(x)} = \frac{1}{n} \sum_{i=1}^n \mathbf{t}(x_i)$

Estimating the parameters:

- Log-Density:

$$\log p(x|\boldsymbol{\eta}) = \boldsymbol{\eta}^T \mathbf{t}(x) + \log h(x) - a(\boldsymbol{\eta})$$

where  $h(x)$  is base function and  $a(\boldsymbol{\eta})$  is the log-partition function.

- Log-likelihood function:

$$\begin{aligned} \ell_n(\boldsymbol{\eta}) &= \sum_{i=1}^n \log p(x_i|\boldsymbol{\eta}) \\ &= -na(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{i=1}^n \mathbf{t}(x_i) + \sum_{i=1}^n \log h(x_i) \\ &= -n \left( a(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \overline{\mathbf{t}(x)} \right) + \text{const.} \end{aligned}$$

Note how the non-constant terms of the log-likelihood depend on the data only through  $\overline{\mathbf{t}(x)}$ . Hence the average of the canonical statistics are **sufficient statistics** for  $\boldsymbol{\eta}$ . Furthermore, recall that the log-partition function  $a(\boldsymbol{\eta})$  is convex, and as a result the log-likelihood function is concave in the canonical parameters  $\boldsymbol{\eta}$ .

- Score function:

$$\begin{aligned} S_n(\boldsymbol{\eta}) &= \nabla \ell_n(\boldsymbol{\eta}) \\ &= -n \left( \nabla a(\boldsymbol{\eta}) + \overline{\mathbf{t}(x)} \right) \\ &= -n \left( \boldsymbol{\mu}_t(\boldsymbol{\eta}) + \overline{\mathbf{t}(x)} \right) \end{aligned}$$



where  $\mu_t(\eta) = E(t(x))$  are the means of the canonical statistics, given in dependence of the canonical parameters. The expectation parameters  $\mu_t$  are often preferred over the canonical parameters  $\eta$ .

- Maximum likelihood estimate:

$$S_n(\hat{\eta}_{ML}) = 0 \Rightarrow \mu_t(\hat{\eta}_{ML}) = \hat{\mu}_t^{ML} = \overline{t(x)}$$

Hence, the **average canonical statistics are the MLEs of the expectation parameters of the exponential family**. This is the generalisation of earlier specific examples such as that the MLE of the mean of a normal distribution and the MLE of the mean parameter of the Bernoulli distribution are both given by the average of the data.

To obtain the MLE for the canonical parameter we solve the above for  $\hat{\eta}_{ML}$ :

$$\hat{\eta}_{ML} = \mu_t^{-1}(\overline{t(x)})$$

- To confirm that we actually have a maximum we need to verify that the eigenvalues of the Hessian matrix at the optimum are all negative. This is indeed the case, for details see Example 9.5.

### 8.3 Further properties of ML

#### Relationship of maximum likelihood with least squares estimation

In Example 8.2 the form of the log-likelihood function is a function of the sum of squared differences. Maximising  $\ell_n(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$  is equivalent to *minimising*  $\sum_{i=1}^n (x_i - \mu)^2$ . Hence, finding the mean by **maximum likelihood assuming a normal model is equivalent to least-squares estimation!**

Note that least-squares estimation has been in use at least since the early 1800s<sup>1</sup> and thus predates maximum likelihood (1922). Due to its simplicity it is still very popular in particular in regression and the link with maximum likelihood and normality allows to understand why it usually works well.

<sup>1</sup>Stigler, S. M. 1981. *Gauss and the invention of least squares*. Ann. Statist. 9:465–474. <https://doi.org/10.1214/aos/1176345451>

See also Example 3.2 and Example 4.7 for further links of the normal distribution with squared error.

#### Bias of maximum likelihood estimates

Example 8.5 is interesting because it shows that maximum likelihood can result in both biased and as well as unbiased estimators.

Recall that  $x \sim N(\mu, \sigma^2)$ . As a result

$$\hat{\mu}_{ML} = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

with  $E(\hat{\mu}_{ML}) = \mu$  and

$$\widehat{\sigma}_{ML}^2 \sim W_1\left(s^2 = \frac{\sigma^2}{n}, k = n - 1\right)$$

(see Section A.8) with mean  $E(\widehat{\sigma}_{ML}^2) = \frac{n-1}{n} \sigma^2$ .

Therefore, the MLE of  $\mu$  is unbiased as

$$\text{Bias}(\hat{\mu}_{ML}) = E(\hat{\mu}_{ML}) - \mu = 0$$

In contrast, however, the MLE of  $\sigma^2$  is negatively biased because

$$\text{Bias}(\widehat{\sigma}_{ML}^2) = E(\widehat{\sigma}_{ML}^2) - \sigma^2 = -\frac{1}{n} \sigma^2$$

Thus, in the case of the variance parameter of the normal distribution the MLE is *not* recovering the well-known unbiased estimator of the variance

$$\widehat{\sigma}_{UB}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \widehat{\sigma}_{ML}^2$$

In other words, the unbiased variance estimate is not a maximum likelihood estimate!

Therefore it is worth keeping in mind that maximum likelihood can result in biased estimates for finite  $n$ . For large  $n$ , however, the bias disappears as MLEs are consistent.



### ► Minimal sufficient statistics and maximal data reduction

In all the examples discussed above the sufficient statistic was typically either  $\bar{x}$  and  $\overline{x^2}$  (or both). This is not a coincidence since all of the examples are exponential families with canonical statistics  $x$  and  $x^2$ , and in exponential families a sufficient statistic can be obtained as the average of the canonical statistics.

Specifically,  $t(D) = \overline{t(x)}$ , see Example 8.7.

Crucially, in the above examples the identified sufficient statistics are also **minimal sufficient statistics** where the dimension of sufficient statistic is equal to the dimension of the parameter vector, and as such as low as possible. Minimal sufficient statistics provide maximal data reduction as will be discussed later.

## 9 Observed Fisher information

### 9.1 Definition of the observed Fisher information

Visual inspection of the log-likelihood function (e.g. Figure 9.1) suggests that it contains more information about the parameter  $\theta$  than just the location of its maximum at  $\hat{\theta}_{ML}$ .

In particular, in a regular model the **curvature** of the log-likelihood function at the MLE is linked to the accuracy of  $\hat{\theta}_{ML}$ : a flat log-likelihood surface near the maximum (low curvature) makes the optimal parameter less well determined (and more difficult to find numerically). Conversely, if the log-likelihood surface is sharply peaked (strong curvature) then the maximum point is well defined. The curvature can be quantified by the matrix of second-order derivatives (Hessian matrix) of the log-likelihood function.

Accordingly, the **observed Fisher information** (matrix) is defined as the negative Hessian of the log-likelihood function  $\ell_n(\theta)$  at the MLE  $\hat{\theta}_{ML}$ :

$$J_n(\hat{\theta}_{ML}) = -H(\hat{\theta}_{ML}) = -\nabla\nabla^T \ell_n(\hat{\theta}_{ML})$$

Sometimes this is called the “observed information”. To avoid confusion with the **Fisher information**

$$\mathcal{I}_P(\theta) = -E_{P(\theta)} (\nabla\nabla^T \log p(x|\theta))$$

introduced earlier (Section 5.1) it is necessary to always use the qualifier “observed” when referring to  $J_n(\hat{\theta}_{ML})$ .

The observed Fisher information plays an important role in quantifying the uncertainty of a maximum likelihood estimate.



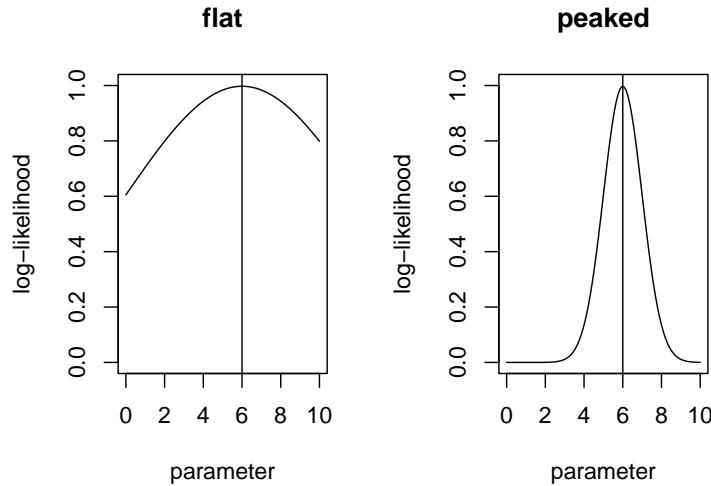


Figure 9.1: Flat and peaked log-likelihood function.

### Transformation properties

As a consequence of the invariance of the score function and curvature function the **observed Fisher information is invariant under transformations of the sample space**. This is the same invariance also shown by the Fisher information and by the KL divergence.

► Like the Fisher information the observed Fisher information (as a Hessian matrix) transforms covariantly under change of model parameters — see Section 5.2.

### Relationship between observed and expected Fisher information

The observed Fisher information  $J_n(\hat{\theta}_{ML})$  and the expected Fisher information  $\mathcal{I}_P(\theta)$  are related but also two clearly different entities.

Curvature based:

- Both types of Fisher information are based on computing second order derivatives (Hessian matrix), thus both are based on the curvature of a function.

Transformation properties:

- Both quantities are invariant under changes of the parametrisation of the sample space.
- ► Both transform covariantly when changing the parameter of the distribution.

Data-based vs. model only:

- The observed Fisher information is computed from the log-likelihood function. Therefore it takes both the model and the observed data  $D$  into account and explicitly depends on the sample size  $n$ . It contains estimates of the parameters but not the parameters themselves. While the curvature of the log-likelihood function may be computed for any point of the log-likelihood function the observed Fisher information specifically refers to the curvature at the MLE  $\hat{\theta}_{ML}$ . It is linked to the (asymptotic) variance of the MLE (see the examples and as will be discussed in more detail later).
- In contrast, the expected Fisher information is derived directly from the log-density of the model family. It does not depend on the observed data, and thus does not depend on sample size. It is a property of the model family  $P(\theta)$  alone. It makes sense and can be computed at any  $\theta$ . It describes the local geometry of the space of the model family, and is the local approximation of KL information.

Large sample equivalence:

- Assume that for large sample size  $n$  the MLE converges to  $\hat{\theta}_{ML} \rightarrow \theta_0$ . It follows from the construction of the observed Fisher information and the law of large numbers that asymptotically for large sample size  $J_n(\hat{\theta}_{ML}) \rightarrow n\mathcal{I}_P(\theta_0)$  (i.e. the Fisher information for a set of iid random variables, see Section 5.1).

► Finite sample equivalence for exponential families:

- In a very important class of models, namely for **exponential families**, we find that  $J_n(\hat{\theta}_{ML}) = n\mathcal{I}_P(\hat{\theta}_{ML})$  is valid also for finite



sample size  $n$ . The general case is shown in Example 5.5 and Example 9.5 (canonical parameter) and Example 5.8 and Example 9.6 (expectation parameter).

This can also be directly seen from special instances of exponential families such as the Bernoulli distribution (Example 5.1 and Example 9.1), the normal distribution with one parameter (Example 5.3 and Example 9.2) and the normal distribution with two parameters (Example 5.4 and Example 9.4).

- However, exponential families are an exception. In a general model  $J_n(\hat{\theta}_{ML}) \neq n\mathcal{I}_P(\hat{\theta}_{ML})$  for finite sample size  $n$ . As an example consider the location-scale  $t$ -distribution  $t_\nu(\mu, \tau^2)$  with unknown median parameter  $\mu$  and known scale parameter  $\tau^2$  and given degree of freedom  $\nu$ . This is not an exponential family model (unless  $\nu \rightarrow \infty$  when it becomes the normal distribution). It can be shown that the Fisher information is  $\mathcal{I}_P(\mu) = \frac{\nu+1}{\nu+3} \frac{1}{\tau^2}$  but the observed Fisher information  $J_n(\hat{\mu}_{ML}) \neq n \frac{\nu+1}{\nu+3} \frac{1}{\tau^2}$  with a maximum likelihood estimate  $\hat{\mu}_{ML}$  that can only be computed numerically with no closed form available.

## 9.2 Observed Fisher information examples

### Scalar examples — single parameter models

**Example 9.1.** Observed Fisher information for the Bernoulli model  $\text{Ber}(\theta)$ :

We continue Example 8.1. The second derivative of the log-likelihood function is

$$H_n(\theta) = \frac{dS_n(\theta)}{d\theta} = -n \left( \frac{\bar{x}}{\theta^2} + \frac{1-\bar{x}}{(1-\theta)^2} \right)$$

The observed Fisher information is therefore

$$\begin{aligned} J_n(\hat{\theta}_{ML}) &= -H(\hat{\theta}_{ML}) \\ &= n \left( \frac{\bar{x}}{\hat{\theta}_{ML}^2} + \frac{1-\bar{x}}{(1-\hat{\theta}_{ML})^2} \right) \\ &= n \left( \frac{1}{\hat{\theta}_{ML}} + \frac{1}{1-\hat{\theta}_{ML}} \right) \\ &= \frac{n}{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})} \end{aligned}$$

The inverse of the observed Fisher information is:

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

Compare this with  $\text{Var}\left(\frac{x}{n}\right) = \frac{\theta(1-\theta)}{n}$  for  $x \sim \text{Bin}(n, \theta)$ .

**Example 9.2.** Observed Fisher information for the normal distribution with unknown mean and known variance:

This is the continuation of Example 8.2. The second derivative of the log-likelihood function is

$$H_n(\mu) = \frac{dS_n(\mu)}{d\mu} = -\frac{n}{\sigma^2}$$

The observed Fisher information at the MLE is therefore

$$J_n(\hat{\mu}_{ML}) = -H_n(\hat{\mu}_{ML}) = \frac{n}{\sigma^2}$$

and the inverse of the observed Fisher information is

$$J_n(\hat{\mu}_{ML})^{-1} = \frac{\sigma^2}{n}$$

For  $x_i \sim N(\mu, \sigma^2)$  we have  $\text{Var}(x_i) = \sigma^2$  and hence  $\text{Var}(\bar{x}) = \sigma^2/n$ , which is equal to the inverse observed Fisher information.

**Example 9.3.** Observed Fisher information for the normal distribution with known mean and unknown variance:

This is the continuation of Example 8.3. The second derivative of the log-likelihood function is

$$H_n(\sigma^2) = \frac{dS_n(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^4} \left( \frac{2}{\sigma^2} (\bar{x} - \mu)^2 - 1 \right)$$

Correspondingly, the observed Fisher information is

$$J_n(\hat{\sigma}_{ML}^2) = -H_n(\hat{\sigma}_{ML}^2) = \frac{n}{2} \left( \hat{\sigma}_{ML}^2 \right)^{-2}$$

and its inverse is

$$J_n(\hat{\sigma}_{ML}^2)^{-1} = \frac{2}{n} \left( \hat{\sigma}_{ML}^2 \right)^2$$



With  $x_i \sim N(\mu, \sigma^2)$  the empirical variance  $\widehat{\sigma}_{ML}^2$  follows a one-dimensional Wishart distribution

$$\widehat{\sigma}_{ML}^2 \sim \text{Wis}\left(s^2 = \frac{\sigma^2}{n}, k = n - 1\right)$$

(see Section A.8) and hence has variance  $\text{Var}(\widehat{\sigma}_{ML}^2) = \frac{n-1}{n} \frac{2\sigma^4}{n}$ . For large  $n$  this becomes  $\text{Var}(\widehat{\sigma}_{ML}^2) \approx \frac{2}{n} (\sigma^2)^2$  which is (apart from the “hat”) the inverse of the observed Fisher information.

### Matrix examples — multiple parameter models

**Example 9.4.** Observed Fisher information for the normal distribution with mean and variance parameter:

This is the continuation of Example 8.5.

The Hessian matrix of the log-likelihood function is

$$\begin{aligned} H_n(\mu, \sigma^2) &= \nabla \nabla^T \ell_n(\mu, \sigma^2) \\ &= \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^4}(\bar{x} - \mu) \\ -\frac{n}{\sigma^4}(\bar{x} - \mu) & \frac{n}{2\sigma^4} - \frac{n}{\sigma^6}(\bar{x}^2 - 2\mu\bar{x} + \mu^2) \end{pmatrix} \end{aligned}$$

The negative Hessian at the MLE, i.e. at  $\hat{\mu}_{ML} = \bar{x}$  and  $\widehat{\sigma}_{ML}^2 = \bar{x}^2 - \bar{x}^2$ , yields the **observed Fisher information matrix**:

$$\begin{aligned} J_n(\hat{\mu}_{ML}, \widehat{\sigma}_{ML}^2) &= -H(\hat{\mu}_{ML}, \widehat{\sigma}_{ML}^2) \\ &= \begin{pmatrix} \frac{n}{\widehat{\sigma}_{ML}^2} & 0 \\ 0 & \frac{n}{2(\widehat{\sigma}_{ML}^2)^2} \end{pmatrix} \end{aligned}$$

The observed Fisher information matrix is diagonal with positive entries. Therefore its eigenvalues are all positive as required for a maximum, because for a diagonal matrix the eigenvalues are the entries on the diagonal.

The inverse of the observed Fisher information matrix is

$$J_n(\hat{\mu}_{ML}, \widehat{\sigma}_{ML}^2)^{-1} = \begin{pmatrix} \frac{\widehat{\sigma}_{ML}^2}{n} & 0 \\ 0 & \frac{2(\widehat{\sigma}_{ML}^2)^2}{n} \end{pmatrix}$$

**Example 9.5.** ▶ Observed Fisher information for the canonical parameter of an exponential family:

This is the continuation of Example 8.7.

The Hessian matrix of the log-likelihood function is

$$\begin{aligned} \nabla \nabla^T l_n(\eta) &= -n \nabla \nabla^T a(\eta) \\ &= -n \Sigma_t(\eta) \end{aligned}$$

with  $\text{Var}(t(x)) = \Sigma_t$ .

Assuming a minimal exponential family with non-redundant canonical parameters and strictly convex log-partition function  $a(\eta)$  the covariance matrix  $\Sigma_t$  will be positive definite and therefore the eigenvalues of the Hessian will all be positive as required for a maximum. Conversely, if the model is non-minimal the Hessian will be singular, i.e. there will be some eigenvalues of value zero indicating that there is dependence among the parameters.

The negative Hessian at the MLE yields the **observed Fisher information matrix** for the canonical parameter:

$$J_n(\hat{\eta}_{ML}) = n \Sigma_t(\hat{\eta}_{ML})$$

Compare with the corresponding expected Fisher information (Example 5.5).

**Example 9.6.** ▶ Observed Fisher information for the expectation parameter of an exponential family:

Instead of the observed Fisher information for the canonical parameter  $\eta$  we are often rather more interested in the Fisher information for the expectation parameter  $\mu_t$ .

Similarly as for the Fisher information (see Example 5.8) we may transform the observed Fisher information to the expectation parameter via covariant transformation and get

$$J_n(\hat{\mu}_t^{ML}) = n \Sigma_t(\eta(\hat{\mu}_t^{ML}))^{-1}$$

The inverse of the observed Fisher information matrix is

$$J_n(\hat{\mu}_t^{ML})^{-1} = \frac{1}{n} \Sigma_t(\eta(\hat{\mu}_t^{ML}))$$



## 10 Quadratic approximation and normal asymptotics

### 10.1 Approximate distribution of maximum likelihood estimates

#### Quadratic log-likelihood of the multivariate normal model

Assume we observe a single sample  $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with known covariance. Noting that the multivariate normal density is

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

the corresponding log-likelihood for  $\boldsymbol{\mu}$  is

$$\ell_1(\boldsymbol{\mu}) = C - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

where  $C$  is a constant that does not depend on  $\boldsymbol{\mu}$ . Note that the log-likelihood is a quadratic function (both for  $\mathbf{x}$  and  $\boldsymbol{\mu}$ ) and the maximum of the function lies at  $\boldsymbol{\mu} = \mathbf{x}$  with value  $C$ .

#### Quadratic approximation of a log-likelihood function

Now consider the quadratic approximation of a general log-likelihood function  $\ell_n(\boldsymbol{\theta})$  for  $\boldsymbol{\theta}$  around the MLE  $\hat{\boldsymbol{\theta}}_{ML}$  (Figure 10.1).

We assume the underlying model is regular and that  $\nabla \ell_n(\hat{\boldsymbol{\theta}}_{ML}) = 0$ , i.e. the gradient at the maximum vanishes. The Taylor series approximation of scalar-valued function  $f(\mathbf{x})$  around  $\mathbf{x}_0$  is

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla^T f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla \nabla^T f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \dots$$

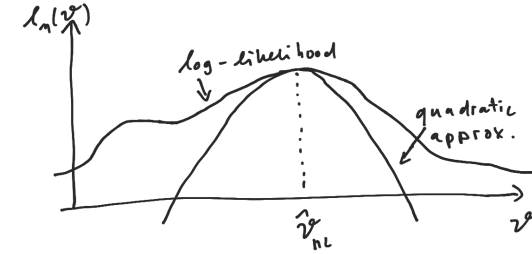


Figure 10.1: Quadratic approximation of the log-likelihood function.

Applied to the log-likelihood function this yields

$$\ell_n(\boldsymbol{\theta}) \approx \ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta})^T J_n(\hat{\boldsymbol{\theta}}_{ML})(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta})$$

This is a quadratic function with maximum at  $(\hat{\boldsymbol{\theta}}_{ML}, \ell_n(\hat{\boldsymbol{\theta}}_{ML}))$ . Note the appearance of the observed Fisher information  $J_n(\hat{\boldsymbol{\theta}}_{ML})$  in the quadratic term. There is no linear term because of the vanishing gradient at the MLE.

Crucially, this approximated log-likelihood takes the same form as if  $\hat{\boldsymbol{\theta}}_{ML}$  was sampled from a multivariate normal distribution with mean  $\boldsymbol{\theta}$  and the covariance given by the *inverse* of the observed Fisher information.

Note that this requires a regular model with positive definite observed Fisher information matrix so that  $J_n(\hat{\boldsymbol{\theta}}_{ML})$  is actually invertible!

**Example 10.1.** Quadratic approximation of the log-likelihood for a proportion:

From Example 8.1 we have the log-likelihood

$$\ell_n(\theta) = n(\bar{x} \log \theta + (1 - \bar{x}) \log(1 - \theta))$$

and the MLE

$$\hat{\theta}_{ML} = \bar{x}$$

and from Example 9.1 the observed Fisher information

$$J_n(\hat{\theta}_{ML}) = \frac{n}{\bar{x}(1 - \bar{x})}$$



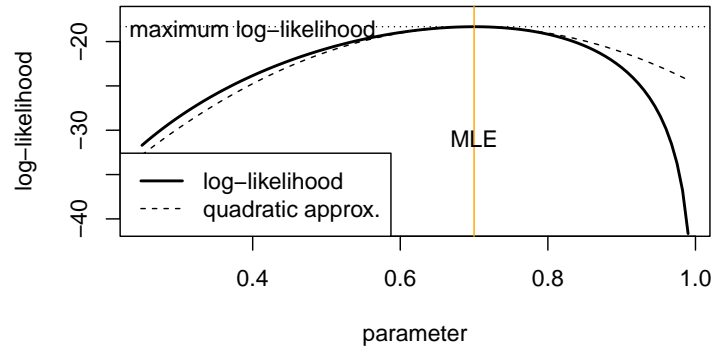


Figure 10.2: Quadratic approximation of the log-likelihood for a Bernoulli model.

The log-likelihood at the MLE is

$$\ell_n(\hat{\theta}_{ML}) = n(\bar{x} \log \bar{x} + (1 - \bar{x}) \log(1 - \bar{x}))$$

This allows us to construct the quadratic approximation of the log-likelihood around the MLE as

$$\begin{aligned} \ell_n(\theta) &\approx \ell_n(\hat{\theta}_{ML}) - \frac{1}{2} J_n(\hat{\theta}_{ML})(\theta - \hat{\theta}_{ML})^2 \\ &= n \left( \bar{x} \log \bar{x} + (1 - \bar{x}) \log(1 - \bar{x}) - \frac{(\theta - \bar{x})^2}{2\bar{x}(1 - \bar{x})} \right) \\ &= C + \frac{\bar{x}\theta - \frac{1}{2}\theta^2}{\bar{x}(1 - \bar{x})/n} \end{aligned}$$

The constant  $C$  does not depend on  $\theta$ , its function is to match the approximate log-likelihood at the MLE with that of the corresponding original log-likelihood. The approximate log-likelihood takes on the form of a normal log-likelihood (Example 8.2) for one observation of  $\hat{\theta}_{ML} = \bar{x}$  from  $N\left(\theta, \frac{\bar{x}(1-\bar{x})}{n}\right)$ .

Figure 10.2 shows the Bernoulli log-likelihood function and its quadratic approximation illustrated for data with  $n = 30$  and  $\bar{x} = 0.7$ :

## Asymptotic normality of maximum likelihood estimates

Intuitively, it makes sense to associate large amount of curvature of the log-likelihood at the MLE with low variance of the MLE (and conversely, low amount of curvature with high variance).

From the above we see that for regular models:

- normality implies a quadratic log-likelihood,
- conversely, taking a quadratic approximation of the log-likelihood implies approximate normality, and
- in the quadratic approximation **the inverse observed Fisher information plays the role of the covariance** of the MLE.

This suggests the following theorem:

**Asymptotically, the MLE of the parameters of a regular model is normally distributed around the true parameter and with covariance equal to the inverse of the observed Fisher information:**

$$\hat{\theta}_{ML} \stackrel{d}{\sim} \underbrace{N}_{\text{multivariate normal}} \left( \underbrace{\theta_{\text{true}}}_{\text{mean vector}}, \underbrace{J_n(\hat{\theta}_{ML})^{-1}}_{\text{covariance matrix}} \right)$$

This theorem about the distributional properties of MLEs greatly enhances the usefulness of the method of maximum likelihood. It implies that in regular settings maximum likelihood is not just a method for obtaining point estimates but also provides estimates of their uncertainty.

## Remarks on the asymptotics

However, we need to clarify what “asymptotic” actually means in the context of the above theorem:

- 1) Primarily, it means to have sufficient sample size so that the log-likelihood  $\ell_n(\theta)$  is sufficiently well approximated by a quadratic function around  $\hat{\theta}_{ML}$ . The better the local quadratic approximation the better the normal approximation!
- 2) In a regular model with positive definite observed Fisher information matrix this is guaranteed for large sample size  $n \rightarrow \infty$  thanks to the central limit theorem).



- 3) However,  $n$  going to infinity is in fact not always required for the normal approximation to hold! Depending on the particular model a good local fit to a quadratic log-likelihood may be available also for finite  $n$ . As a trivial example, for the normal log-likelihood it is valid for any  $n$ .
- 4) In the other hand, in non-regular models (with nondifferentiable log-likelihood at the MLE and/or a singular Fisher information matrix) no amount of data, not even  $n \rightarrow \infty$ , will make the quadratic approximation work.

Remarks:

- The asymptotic normality of MLEs was first discussed in Fisher (1925)<sup>1</sup>
- The technical details of the above considerations are worked out in the theory of [locally asymptotically normal \(LAN\) models](#) pioneered in 1960 by [Lucien LeCam \(1924–2000\)](#).
- There are also methods to obtain higher-order (higher than quadratic and thus non-normal) asymptotic approximations. These relate to so-called [saddle point approximations](#).

### Information inequality and asymptotic optimal efficiency

Assume now that  $\hat{\theta}$  is an arbitrary and unbiased estimator for  $\theta$  and the underlying model  $P$  is regular with density  $p(x|\theta)$ .

[Harald Cramér \(1893–1985\)](#), [C. Radhakrishna Rao \(1920–2023\)](#) and others demonstrated in 1945 the so-called **information inequality**,

$$\text{Var}(\hat{\theta}) \geq (n\mathcal{I}_P(\theta_{\text{true}}))^{-1}$$

which puts a lower bound on the variance of an estimator for  $\theta$ . (Note for dimension  $d > 1$  this is a matrix inequality, meaning that the difference matrix is positive semidefinite). The information inequality is a consequence of the data processing inequality of the KL divergence and hence Fisher information (Chapter 4).

<sup>1</sup>Fisher R. A. 1925. *Theory of statistical estimation*. Math. Proc. Cambridge Philos. Soc. 22:700–725. <https://doi.org/10.1017/S0305004100009580>

For large sample size with  $n \rightarrow \infty$  and  $\hat{\theta}_{ML} \rightarrow \theta_{\text{true}}$  the observed Fisher information becomes  $J_n(\hat{\theta}_{ML}) \rightarrow n\mathcal{I}_P(\theta_{\text{true}})$  and therefore we can write the asymptotic distribution of  $\hat{\theta}_{ML}$  as

$$\hat{\theta}_{ML} \stackrel{a}{\sim} N\left(\theta_{\text{true}}, (n\mathcal{I}_P(\theta_{\text{true}}))^{-1}\right)$$

This means that for large  $n$  in regular models  $\hat{\theta}_{ML}$  achieves the lowest variance possible according to the Cramér-Rao information inequality. In other words, for large sample size maximum likelihood is optimally efficient and thus the best available estimator will in fact be the MLE!

However, as we will see later this does not hold for small sample size where it is indeed possible (and necessary) to improve over the MLE (e.g. via Bayesian estimation or regularisation).

### Non-regular models

For non-regular models the asymptotic normality does not hold. Instead, the (asymptotic) distribution of the MLE must be obtained in by other means.

**Example 10.2.** Distribution of MLE for upper bound  $\theta$  of the uniform distribution:

This continues Example 8.4 to find the distribution of  $\hat{\theta}_{ML}$ . Due to a discontinuity in the density at the MLE the observed Fisher information cannot be computed and the normal approximation for the distribution of  $\hat{\theta}_{ML}$  is not valid.

Nonetheless, one can still obtain the sampling distribution of  $\hat{\theta}_{ML} = x_{[n]}$ . However, *not* via asymptotic arguments but instead by understanding that  $x_{[n]}$  is an order statistic (see [https://en.wikipedia.org/wiki/Order\\_statistic](https://en.wikipedia.org/wiki/Order_statistic)) with the following properties:

$$x_{[n]} \sim \theta \text{ Beta}(n, 1) \quad \text{"n-th order statistic"}$$

$$\mathbb{E}(x_{[n]}) = \frac{n}{n+1} \theta$$

$$\text{Var}(x_{[n]}) = \frac{n}{(n+1)^2(n+2)} \theta^2 \approx \frac{\theta^2}{n^2}$$

Note that the variance decreases with  $\frac{1}{n^2}$  which is much faster than the usual  $\frac{1}{n}$  of an “efficient” estimator. Correspondingly,  $\hat{\theta}_{ML}$  is a so-called “super efficient” estimator.



## 10.2 Quantifying the uncertainty of maximum likelihood estimates

### Estimating the variance of MLEs

In the previous section we saw that MLEs are asymptotically normally distributed, with the inverse Fisher information (both expected and observed) linked to the asymptotic variance.

This leads to the question whether to use the observed Fisher information  $J_n(\hat{\theta}_{ML})$  or the expected Fisher information at the MLE  $n\mathcal{I}_P(\hat{\theta}_{ML})$  to estimate the variance of the MLE?

- Clearly, for  $n \rightarrow \infty$  both can be used interchangeably.
- However, they can be very different for finite  $n$  in particular for models that are not exponential families.
- Also normality may occur well before  $n$  goes to  $\infty$ .

Therefore one needs to choose between the two, considering also that

- the expected Fisher information at the MLE is the average curvature at the MLE, whereas the observed Fisher information is the actual observed curvature, and
- the observed Fisher information naturally occurs in the quadratic approximation of the log-likelihood.

All in all, the observed Fisher information as estimator of the variance is more appropriate as it is based on the actual observed data and also works for large  $n$  (in which case it yields the same result as using expected Fisher information):

$$\widehat{\text{Var}}(\hat{\theta}_{ML}) = J_n(\hat{\theta}_{ML})^{-1}$$

and its square-root as the estimate of the standard deviation

$$\widehat{\text{SD}}(\hat{\theta}_{ML}) = J_n(\hat{\theta}_{ML})^{-1/2}$$

Note that in the above we use *matrix inversion* and the (inverse) *matrix square root*.

The reasons for preferring observed Fisher information are made mathematically precise in a classic paper by Efron and Hinkley (1978)<sup>2</sup>

<sup>2</sup>Efron, B., and D. V. Hinkley. 1978. *Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information*. *Biometrika* 65:457–482. <https://doi.org/10.1093/biomet/65.3.457>

**Example 10.3.** Estimated variance and distribution of the MLE of the mean parameter for the normal distribution with known variance:

From Example 8.2 and Example 9.2 we know that

$$\hat{\mu}_{ML} = \bar{x}$$

and that the corresponding observed Fisher information at  $\hat{\mu}_{ML}$  is

$$J_n(\hat{\mu}_{ML}) = \frac{n}{\sigma^2}$$

The estimated variance of the MLE is therefore

$$\widehat{\text{Var}}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$$

and the corresponding asymptotic normal distribution is

$$\hat{\mu}_{ML} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Note that in this case the distribution is not asymptotic but is **exact**, i.e. valid also for small  $n$  (as long as the data  $x_i$  are actually from  $N(\mu, \sigma^2)$ !).

**Example 10.4.** Estimated variance and distribution of the MLE of a proportion:

From Example 8.1 and Example 9.1 we know the MLE

$$\hat{\theta}_{ML} = \bar{x} = \frac{k}{n}$$

and the corresponding observed Fisher information

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

The estimated variance of the MLE is therefore

$$\widehat{\text{Var}}(\hat{\theta}_{ML}) = \frac{\hat{\theta}_{ML}(1 - \hat{\theta}_{ML})}{n}$$

and the corresponding asymptotic normal distribution is

$$\hat{\theta}_{ML} \stackrel{a}{\sim} N\left(\theta, \frac{\hat{\theta}_{ML}(1 - \hat{\theta}_{ML})}{n}\right)$$



### Wald statistic

Centering the MLE  $\hat{\theta}_{ML}$  with  $\theta_0$  followed by standardising using the observed Fisher information yields the **Wald statistic** (named after [Abraham Wald, 1902–1950](#)):

$$t(\theta_0) = J_n(\hat{\theta}_{ML})^{1/2}(\hat{\theta}_{ML} - \theta_0)$$

The **squared Wald statistic** is a scalar defined as

$$\begin{aligned} t(\theta_0)^2 &= t(\theta_0)^T t(\theta_0) \\ &= (\hat{\theta}_{ML} - \theta_0)^T J_n(\hat{\theta}_{ML})(\hat{\theta}_{ML} - \theta_0) \end{aligned}$$

Both  $t(\theta_0)$  and  $t(\theta_0)^2$  are commonly called Wald statistics in the literature. In this text we explicitly use the qualifier “squared” if we refer to the latter.

We now assume that the true underlying parameter is  $\theta_0$ . Since the MLE is asymptotically normal the Wald statistic is asymptotically **standard normal** distributed:

$$\begin{aligned} t(\theta_0) &\stackrel{a}{\sim} N(0, I_d) \quad \text{for vector } \theta \\ t(\theta_0) &\stackrel{a}{\sim} N(0, 1) \quad \text{for scalar } \theta \end{aligned}$$

Correspondingly, the **squared** Wald statistic is chi-squared distributed:

$$\begin{aligned} t(\theta_0)^2 &\stackrel{a}{\sim} \chi_d^2 \quad \text{for vector } \theta \\ t(\theta_0)^2 &\stackrel{a}{\sim} \chi_1^2 \quad \text{for scalar } \theta \end{aligned}$$

The degree of freedom of the chi-squared distribution is the dimension  $d$  of the parameter vector  $\theta$ .

**Example 10.5.** Wald statistic for the mean parameter of a normal distribution with known variance:

We continue from Example 10.3. With  $\hat{\mu}_{ML} = \bar{x}$  and  $\widehat{\text{Var}}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$  and thus  $\widehat{\text{SD}}(\hat{\mu}_{ML}) = \frac{\sigma}{\sqrt{n}}$  we get as **Wald statistic**:

$$t(\mu_0) = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Note this is the one sample  $t$ -statistic with given  $\sigma$ . The **squared Wald statistic** is:

$$t(\mu_0)^2 = \frac{(\bar{x} - \mu_0)^2}{\sigma^2/n} \sim \chi_1^2$$

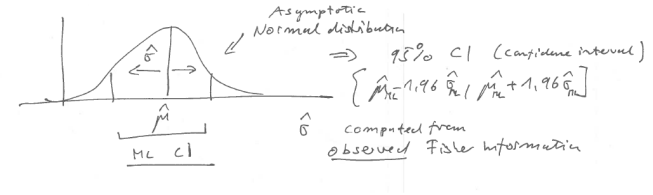


Figure 10.3: Construction of a 95% symmetric normal confidence interval for a maximum likelihood estimate.

Again, in this instance this is the exact distribution, not just the asymptotic one.

Using the Wald statistic or the squared Wald statistic we can test whether a particular  $\mu_0$  can be rejected as underlying true parameter, and we can also construct corresponding confidence intervals.

**Example 10.6.** Wald statistic for a proportion:

We continue from Example 10.4. With  $\hat{\theta}_{ML} = \bar{x}$  and  $\widehat{\text{Var}}(\hat{\theta}_{ML}) = \frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}$  and thus  $\widehat{\text{SD}}(\hat{\theta}_{ML}) = \sqrt{\frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}}$  we get as **Wald statistic**:

$$t(\theta_0) = \frac{\bar{x} - \theta_0}{\sqrt{\bar{x}(1-\bar{x})/n}} \stackrel{a}{\sim} N(0, 1)$$

The **squared Wald statistic** is:

$$t(\theta_0)^2 = n \frac{(\bar{x} - \theta_0)^2}{\bar{x}(1-\bar{x})} \stackrel{a}{\sim} \chi_1^2$$

## 10.3 Normal confidence intervals and tests

### Normal confidence intervals using the Wald statistic

See Section A.10 to review relevant background from year 1.

The asymptotic normality of MLEs derived from regular models enables us to construct a corresponding normal confidence interval (Figure 10.3). For example, to construct the asymptotic normal CI for the MLE of a scalar parameter  $\theta$  we use the MLE  $\hat{\theta}_{ML}$  as estimate of the mean and



its standard deviation  $\widehat{SD}(\hat{\theta}_{ML})$  computed from the observed Fisher information:

$$CI = [\hat{\theta}_{ML} \pm c_{\text{normal}} \widehat{SD}(\hat{\theta}_{ML})]$$

Here  $c_{\text{normal}}$  is a critical value for the standard-normal symmetric confidence interval chosen to achieve the desired nominal coverage. The critical values are computed using the inverse standard normal distribution function via  $c_{\text{normal}} = \Phi^{-1}\left(\frac{1+\kappa}{2}\right)$ . A list of critical values for the standard normal distribution is found in Table A.1. For example, for a CI with 95% coverage one uses the factor 1.96 so that

$$CI = [\hat{\theta}_{ML} \pm 1.96 \widehat{SD}(\hat{\theta}_{ML})]$$

The normal CI can be expressed using Wald statistic as follows:

$$CI = \{\theta_0 : |t(\theta_0)| < c_{\text{normal}}\}$$

Similary, it can also be expressed using the squared Wald statistic:

$$CI = \{\theta_0 : t(\theta_0)^2 < c_{\text{chisq}}\}$$

Note that this form facilitates the construction of normal confidence intervals for a parameter vector  $\theta_0$ .

A list of critical values for the chi-squared distribution with one degree of freedom is found in Table A.2.

The following lists contains the critical values resulting from the chi-squared distribution with degree of freedom  $m = 1$  for the three most common choices of coverage  $\kappa$  for a normal CI for a univariate parameter: For example, for a 95% interval the critical value equals 3.84 (which is the square of the critical value 1.96 for the standard normal).

**Example 10.7.** Normal confidence interval for the mean:

We continue from Example 10.3 and Example 10.5. Assume that we observe  $n = 25$  measurements with average  $\bar{x} = 10$ , from a normal with unknown mean and variance  $\sigma^2 = 4$ .

Then  $\hat{\mu}_{ML} = \bar{x} = 10$  and  $\widehat{SD}(\hat{\mu}_{ML}) = \sqrt{\frac{\sigma^2}{n}} = \frac{2}{5}$ .

The symmetric asymptotic normal CI for  $p$  with 95% coverage is given by  $\hat{\mu}_{ML} \pm 1.96 \widehat{SD}(\hat{\mu}_{ML})$  which for the present data results in the interval  $[9.216, 10.784]$ .

**Example 10.8.** Asymptotic normal confidence interval for a proportion:

We continue from Example 10.4 and Example 10.6. Assume we observe  $n = 30$  measurements with average  $\bar{x} = 0.7$ . Then  $\hat{\theta}_{ML} = \bar{x} = 0.7$  and  $\widehat{SD}(\hat{\theta}_{ML}) = \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \approx 0.084$ .

The symmetric asymptotic normal CI for  $p$  with 95% coverage is given by  $\hat{\theta}_{ML} \pm 1.96 \widehat{SD}(\hat{\theta}_{ML})$  which for the present data results in the interval  $[0.536, 0.864]$ .

### Normal tests using the Wald statistic

Finally, recall the **duality between confidence intervals and statistical tests**. Specifically, a confidence interval with coverage  $\kappa$  can be also used for testing as follows:

- for every  $\theta_0$  inside the CI the data do not allow to reject the hypothesis that  $\theta_0$  is the true parameter with significance level  $\alpha = 1 - \kappa$ .
- Conversely, all values  $\theta_0$  outside the CI can be rejected to be the true parameter with significance level  $\alpha = 1 - \kappa$ .

Hence, in order to test whether  $\theta_0$  is the true underlying parameter value we can compute the corresponding (squared) Wald statistic, find the desired critical value and then decide on rejection.

**Example 10.9.** Normal test for the mean:

We continue from Example 10.7.

We now consider two possible values ( $\mu_0 = 9.5$  and  $\mu_0 = 11$ ) as potentially true underlying mean parameter.

The value  $\mu_0 = 9.5$  lies inside the 95% confidence interval  $[9.216, 10.784]$ . This implies we cannot reject the hypothesis that this is the true underlying parameter on 5% significance level. In contrast,  $\mu_0 = 11$  is outside the confidence interval so we can indeed reject this value. In other words, data plus model exclude this value as a statistically implausible.

This can be verified more directly by computing the corresponding (squared) Wald statistics (see Example 10.5) and comparing them with the relevant critical values:

- $t(9.5)^2 = \frac{(10-9.5)^2}{4/25} = 1.56 < 3.84$  hence  $\mu_0 = 9.5$  cannot be rejected.



- $t(11)^2 = \frac{(10-11)^2}{4/25} = 6.25 > 3.84$  hence  $\mu_0 = 11$  can be rejected.

The squared Wald statistic at the boundaries of the confidence interval equals the critical value.

Note that this is the standard one-sample test of the mean, and that it is exact, not an approximation.

**Example 10.10.** Asymptotic normal test for a proportion:

We continue from Example 10.8.

We now consider two possible values ( $\theta_0 = 0.5$  and  $\theta_0 = 0.8$ ) as potentially true underlying proportion.

The value  $\theta_0 = 0.8$  lies inside the 95% confidence interval  $[0.536, 0.864]$ . This implies we cannot reject the hypothesis that this is the true underlying parameter on 5% significance level. In contrast,  $\theta_0 = 0.5$  is outside the confidence interval so we can indeed reject this value. In other words, data plus model exclude this value as statistically implausible.

This can be verified more directly by computing the corresponding (squared) Wald statistics (see Example 10.6) and comparing them with the relevant critical value (3.84 from chi-squared distribution for 5% significance level):

- $t(0.5)^2 = \frac{(0.7-0.5)^2}{0.084^2} = 5.71 > 3.84$  hence  $\theta_0 = 0.5$  can be rejected.
- $t(0.8)^2 = \frac{(0.7-0.8)^2}{0.084^2} = 1.43 < 3.84$  hence  $\theta_0 = 0.8$  cannot be rejected.

Note that the squared Wald statistic at the boundaries of the normal confidence interval is equal to the critical value.

## 11 Likelihood-based confidence interval and likelihood ratio

### 11.1 Likelihood-based confidence intervals and Wilks statistic

#### Wilks log-likelihood ratio statistic

Instead of relying on the normal resp. quadratic approximation, we can also use the log-likelihood directly to find **likelihood confidence intervals** (Figure 11.1).

Idea: find all  $\theta_0$  that have a log-likelihood that is almost as good as  $\ell_n(\hat{\theta}_{ML})$ .

$$CI = \{\theta_0 : \ell_n(\hat{\theta}_{ML}) - \ell_n(\theta_0) \leq \Delta\}$$

Here  $\Delta$  is our tolerated deviation from the maximum log-likelihood. We will see below how to determine a suitable  $\Delta$ .

The above leads naturally to the **Wilks log-likelihood ratio statistic**

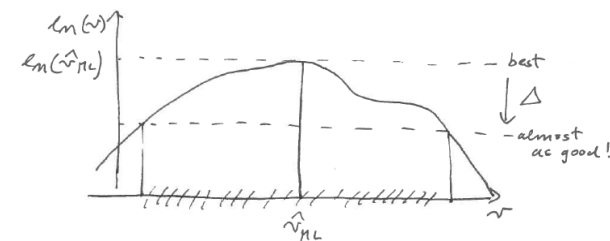


Figure 11.1: Construction of a likelihood-based confidence intervals.



$W(\theta_0)$  defined as:

$$\begin{aligned} W(\theta_0) &= 2 \log \left( \frac{L(\hat{\theta}_{ML}|D)}{L(\theta_0|D)} \right) \\ &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\theta_0)) \end{aligned}$$

With its help we can write the likelihood CI as follows:

$$CI = \{\theta_0 : W(\theta_0) \leq 2\Delta\}$$

The Wilks statistic is named after [Samuel S. Wilks \(1906–1964\)](#).

Advantages of using a likelihood-based CI:

- not restricted to be symmetric
- enables to construct multivariate CIs for parameter vector easily even in non-normal cases
- contains normal CI as special case

**Example 11.1.** The likelihood ratio statistic:

As alternative to the Wilks log-likelihood ratio statistic  $W(\theta_0)$  one may also use the **likelihood ratio** statistic

$$\Lambda(\theta_0) = \frac{L(\theta_0|D)}{L(\hat{\theta}_{ML}|D)}$$

The two statistics can be transformed into each other using

$$W(\theta_0) = -2 \log \Lambda(\theta_0)$$

and

$$\Lambda(\theta_0) = e^{-W(\theta_0)/2}$$

Hence large values of  $W(\theta_0)$  correspond to small values of  $\Lambda(\theta_0)$  and the other way round.

In this course we will only use  $W(\theta_0)$  as it is both easier to compute and its sampling distribution is easier to obtain.

**Example 11.2.** Wilks statistic for the mean parameter of a normal model:

The Wilks statistic is

$$W(\mu_0) = \frac{(\bar{x} - \mu_0)^2}{\sigma^2/n}$$

See Worksheet L2 for a derivation of the Wilks statistic from the normal log-likelihood function.

Note this is the same as the squared Wald statistic discussed in Example 10.5.

Comparing with Example 4.7 we see that in this case the Wilks statistic is essentially (apart from a scale factor  $2n$ ) the KL divergence between two normal distributions with different means and variance equal to  $\sigma^2$ :

$$W(\mu_0) = 2n D_{KL}(N(\hat{\mu}_{ML}, \sigma^2), N(\mu_0, \sigma^2))$$

**Example 11.3.** Wilks statistic for the proportion:

The log-likelihood function for the parameter  $\theta$  is (cf. Example 8.1)

$$\ell_n(\theta) = n(\bar{x} \log \theta + (1 - \bar{x}) \log(1 - \theta))$$

Hence the Wilks statistic is with  $\hat{\theta}_{ML} = \bar{x}$

$$\begin{aligned} W(\theta_0) &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\theta_0)) \\ &= 2n \left( \bar{x} \log \left( \frac{\bar{x}}{\theta_0} \right) + (1 - \bar{x}) \log \left( \frac{1 - \bar{x}}{1 - \theta_0} \right) \right) \end{aligned}$$

Comparing with Example 4.5 we see that in this case the Wilks statistic is essentially (apart from a scale factor  $2n$ ) the KL divergence between two Bernoulli distributions:

$$W(\theta_0) = 2n D_{KL}(\text{Ber}(\hat{\theta}_{ML}), \text{Ber}(\theta_0))$$

**Example 11.4.** ► Wilks statistic for the canonical parameter of an exponential family:

The log-likelihood for the canonical parameter  $\eta$  is (cf. Example 8.7)

$$\ell_n(\eta) = -na(\eta) + n\eta^T \overline{t(x)}$$

with

$$\hat{\eta}_{ML} = \mu_t^{-1}(\overline{t(x)})$$

Hence the Wilks statistic is

$$\begin{aligned} W(\eta_0) &= 2(\ell_n(\hat{\eta}_{ML}) - \ell_n(\eta_0)) \\ &= 2n \left( (\hat{\eta}_{ML} - \eta_0)^T \overline{t(x)} - (a(\hat{\eta}_{ML}) - a(\eta_0)) \right) \\ &= 2n \left( (\hat{\eta}_{ML} - \eta_0)^T \mu_t(\hat{\eta}_{ML}) - (a(\hat{\eta}_{ML}) - a(\eta_0)) \right) \end{aligned}$$



Comparing with Example 4.10 we see that in this case the Wilks statistic is essentially (apart from a scale factor  $2n$ ) the KL divergence between two members of the exponential family:

$$W(\eta_0) = 2n D_{\text{KL}}(P(\hat{\eta}_{ML}), P(\eta_0))$$

### Quadratic approximation of the Wilks statistic

Recall the *quadratic approximation* of the log-likelihood function  $\ell_n(\theta_0)$  (= second order Taylor series around the MLE  $\hat{\theta}_{ML}$ ):

$$\ell_n(\theta_0) \approx \ell_n(\hat{\theta}_{ML}) - \frac{1}{2}(\theta_0 - \hat{\theta}_{ML})^T J_n(\hat{\theta}_{ML})(\theta_0 - \hat{\theta}_{ML})$$

With this we can then approximate the Wilks statistic:

$$\begin{aligned} W(\theta_0) &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\theta_0)) \\ &\approx (\theta_0 - \hat{\theta}_{ML})^T J_n(\hat{\theta}_{ML})(\theta_0 - \hat{\theta}_{ML}) \\ &= t(\theta_0)^2 \end{aligned}$$

Thus the **quadratic approximation of the Wilks statistic yields the squared Wald statistic**.

Conversely, the Wilks statistic can be understood a generalisation of the squared Wald statistic.

**Example 11.5.** Quadratic approximation of the Wilks statistic for the mean parameter of a normal model (continued from Example 11.2):

The normal log-likelihood is already quadratic in the mean parameter (cf. Example 8.2). Correspondingly, the Wilks statistic is quadratic in the mean parameter as well. Hence in this particular case the quadratic “approximation” is in fact exact and the Wilks statistic and the squared Wald statistic are identical!

Correspondingly, confidence intervals and tests based on the Wilks statistic are identical to those obtained using the Wald statistic.

**Example 11.6.** Quadratic approximation of the Wilks statistic for a proportion (continued from Example 11.3):

The Wilks statistic is

$$W(\theta_0) = 2n \left( \bar{x} \log \left( \frac{\bar{x}}{\theta_0} \right) + (1 - \bar{x}) \log \left( \frac{1 - \bar{x}}{1 - \theta_0} \right) \right)$$

Computing Taylor series of second order (for  $\theta_0$  around  $\bar{x}$ ) yields the following approximations:

$$\log \left( \frac{\bar{x}}{\theta_0} \right) \approx -\frac{\theta_0 - \bar{x}}{\bar{x}} + \frac{(\theta_0 - \bar{x})^2}{2\bar{x}^2}$$

and

$$\log \left( \frac{1 - \bar{x}}{1 - \theta_0} \right) \approx \frac{\theta_0 - \bar{x}}{1 - \bar{x}} + \frac{(\theta_0 - \bar{x})^2}{2(1 - \bar{x})^2}$$

With the above we can approximate the Wilks statistic of the proportion as

$$\begin{aligned} W(\theta_0) &\approx 2n \left( -(\theta_0 - \bar{x}) + \frac{(\theta_0 - \bar{x})^2}{2\bar{x}} + (\theta_0 - \bar{x}) + \frac{(\theta_0 - \bar{x})^2}{2(1 - \bar{x})} \right) \\ &= n \left( \frac{(\theta_0 - \bar{x})^2}{\bar{x}} + \frac{(\theta_0 - \bar{x})^2}{(1 - \bar{x})} \right) \\ &= n \left( \frac{(\theta_0 - \bar{x})^2}{\bar{x}(1 - \bar{x})} \right) \\ &= t(\theta_0)^2. \end{aligned}$$

This verifies that the quadratic approximation of the Wilks statistic leads back to the squared Wald statistic of Example 10.6.

### Distribution of the Wilks statistic

The connection with the squared Wald statistic as quadratic approximation of the Wilks log-likelihood ratio statistic implies that both have asymptotically the same distribution.

Hence, under  $\theta_0$  the Wilks statistic is distributed asymptotically as

$$W(\theta_0) \stackrel{a}{\sim} \chi_d^2$$

where  $d$  is the number of parameters in  $\theta$ , i.e. the dimension of the model.

For scalar  $\theta$  (i.e. single parameter and  $d = 1$ ) this becomes

$$W(\theta_0) \stackrel{a}{\sim} \chi_1^2$$

This fact is known as **Wilks’ theorem**.



### Cutoff values for the likelihood CI

Table 11.1: Cutoff values for construction of likelihood confidence intervals for a single parameter.

coverage $\kappa$	$\Delta = \frac{c_{\text{chisq}}}{2}$
0.9	1.35
0.95	1.92
0.99	3.32

The asymptotic distribution for  $W$  is useful to choose a suitable  $\Delta$  for the likelihood CI noting that  $2\Delta = c_{\text{chisq}}$  where  $c_{\text{chisq}}$  is the critical value from Table A.2 for a specified coverage  $\kappa$ . This yields Table 11.1 valid for a scalar parameter.

Hence, in order to calibrate the likelihood interval we in effect compare it with a normal confidence interval.

**Example 11.7.** Likelihood confidence interval for a proportion:

We continue from Example 11.3, and as in Example 10.8 we assume we have data with  $n = 30$  and  $\bar{x} = 0.7$ .

This yields (via numerical root finding) as the 95% likelihood confidence interval the interval  $[0.524, 0.843]$ . It is similar but not identical to the corresponding asymptotic normal interval  $[0.536, 0.864]$  obtained in Example 10.8.

Figure 11.2 illustrates the relationship between the normal CI and the likelihood CI and also shows the role of the quadratic approximation (see also Example 10.1). Note that:

- the normal CI is symmetric around the MLE whereas the likelihood CI is not symmetric
- the normal CI is identical to the likelihood CI when using the quadratic approximation!

### Likelihood ratio test (LRT) using Wilks statistic

As in the normal case (with Wald statistic and normal CIs) one can also construct a test using the Wilks statistic:

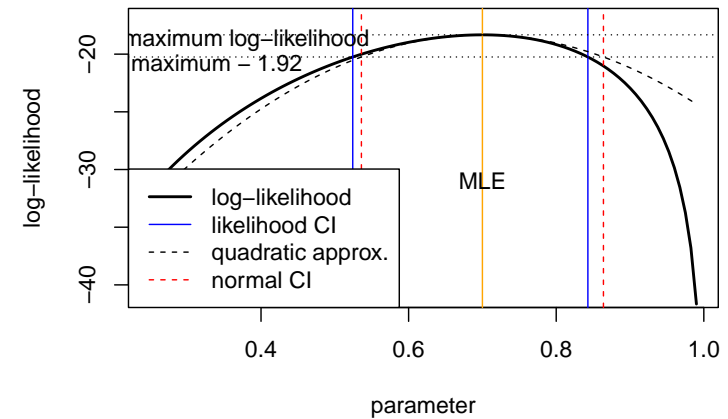


Figure 11.2: Likelihood-based CI and normal interval for the Bernoulli model.

$$\begin{array}{lll}
 H_0 : \theta = \theta_0 & \text{True model is } \theta_0 & \text{Null hypothesis} \\
 H_1 : \theta \neq \theta_0 & \text{True model is not } \theta_0 & \text{Alternative hypothesis}
 \end{array}$$

As test statistic we use the Wilks log likelihood ratio  $W(\theta_0)$ . Extreme values of this test statistic imply evidence against  $H_0$ .

Note that the null model is “simple” (= a single parameter value) whereas the alternative model is “composite” (= a set of parameter values).

#### Remarks:

- The composite alternative  $H_1$  is represented by a single point (the MLE).
- **Reject  $H_0$  for large values of  $W(\theta_0)$**
- under  $H_0$  and for large  $n$  the statistic  $W(\theta_0)$  is chi-squared distributed, i.e.  $W(\theta_0) \stackrel{a}{\sim} \chi_d^2$ . This allows to compute critical values (i.e thresholds to declared rejection under a given significance level) and also  $p$ -values corresponding to the observed test statistics.
- Models **outside** the CI are **rejected**



- Models **inside** the CI **cannot be rejected**, i.e. they can't be statistically distinguished from the best alternative model.

It can be shown that the likelihood ratio test to compare two simple models is optimal in the sense that for any given specified type I error (=probability of wrongly rejecting  $H_0$ , i.e. the significance level) it will maximise the power (=1- type II error, probability of correctly accepting  $H_1$ ). This is known as the **Neyman-Pearson theorem**.

As we have seen previously, the likelihood-based confidence interval differs from the the confidence interval based on the quadratic / normal approximation. Correspondingly, tests based on the log-likelihood ratio  $W(\theta_0)$  and on the squared Wald statistic  $t(\theta_0)^2$  will also yield different outcomes (e.g. rejection due to lying outside the confidence interval) even though both test statistics share the same asymptotic distribution and critical values.

**Example 11.8.** Likelihood test for a proportion:

We continue from Example 11.7 with 95% likelihood confidence interval [0.524, 0.843].

The value  $\theta_0 = 0.5$  is outside the CI and hence can be rejected whereas  $\theta_0 = 0.8$  is inside the CI and hence cannot be rejected on 5% significance level.

The Wilks statistic for  $\theta_0 = 0.5$  and  $\theta_0 = 0.8$  takes on the following values:

- $W(0.5) = 4.94 > 3.84$  hence  $\theta_0 = 0.5$  can be rejected.
- $W(0.8) = 1.69 < 3.84$  hence  $\theta_0 = 0.8$  cannot be rejected.

Note that the Wilks statistic at the boundaries of the likelihood confidence interval is equal to the critical value (3.84 corresponding to 5% significance level for a chi-squared distribution with 1 degree of freedom).

Compare also with the normal test for a proportion in Example 10.10.

### Likelihood ratio statistic and KL divergence

The likelihood ratio statistic is asymptotically linked to differences in the KL divergences of the two compared models with the underlying true model.

Assume that  $Q$  is the true (and unknown) data-generating model and that  $P(\theta)$  is a family of models and we would like to compare two candidate models  $P(\theta_A)$  and  $P(\theta_B)$  with parameters  $\theta_A$  and  $\theta_B$  on the basis of observed data  $D = \{x_1, \dots, x_n\}$ . The KL divergences  $D_{KL}(Q, P(\theta_A))$  and  $D_{KL}(Q, P(\theta_B))$  indicate how close each of the models fit the true  $Q$ . The difference of the two divergences is a way to measure the relative fit of the two models, and can be computed as

$$D_{KL}(Q, P(\theta_B)) - D_{KL}(Q, P(\theta_A)) = E_Q \log \frac{p(x|\theta_A)}{p(x|\theta_B)}$$

Replacing  $Q$  by the empirical distribution  $\hat{Q}_n$  leads to the large sample approximation (see also Example 4.3)

$$\begin{aligned} D_{KL}(Q, P(\theta_B)) - D_{KL}(Q, P(\theta_A)) \\ \approx D_{KL}(\hat{Q}_n, P(\theta_B)) - D_{KL}(\hat{Q}_n, P(\theta_A)) \\ = H(\hat{Q}_n, P(\theta_B)) - H(\hat{Q}_n, P(\theta_A)) \\ = (\ell_n(\theta_A) - \ell_n(\theta_B))/n \end{aligned}$$

Hence, the difference in the log-likelihoods provides an estimate of the difference in the KL divergence of the two models involved.

The Wilks log-likelihood ratio statistic

$$\begin{aligned} W(\theta_0) &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\theta_0)) \\ &= 2n(D_{KL}(\hat{Q}_n, P(\theta_0)) - D_{KL}(\hat{Q}_n, P(\hat{\theta}_{ML}))) \end{aligned}$$

thus compares the best-fit distribution with  $\hat{\theta}_{ML}$  as the parameter to the distribution with parameter  $\theta_0$ .

For *exponential families* the Wilks statistic can also be written in the form of the KL divergence:

$$W(\theta_0) = 2nD_{KL}(P(\hat{\theta}_{ML}), P(\theta_0))$$

This has been seen for specific examples in Example 11.2 and Example 11.3 and in Example 11.4 for exponential families in general.

This implies for exponential families the equality

$$\begin{aligned} D_{KL}(\hat{Q}_n, P(\theta_0)) &= D_{KL}(\hat{Q}_n, P(\hat{\theta}_{ML})) \\ &\quad + D_{KL}(P(\hat{\theta}_{ML}), P(\theta_0)) \end{aligned}$$

which is a variant of the **generalised Pythagorean theorem** for the KL divergence.



## 11.2 Generalised likelihood ratio test (GLRT)

Also known as **maximum likelihood ratio test (MLRT)**. The Generalised Likelihood Ratio Test (GLRT) works just like the standard likelihood ratio test with the difference that now the null model  $H_0$  is also a composite model.

$$\begin{aligned} H_0 : \theta \in \omega_0 \subset \Omega & \quad \text{True model lies in the restricted space} \\ H_1 : \theta \in \Omega & \quad \text{True model lies in the unrestricted space} \end{aligned}$$

Both  $H_0$  and  $H_1$  are now composite hypotheses.  $\Omega$  represents the unrestricted model space with dimension (=number of free parameters)  $d = |\Omega|$ . The constrained space  $\omega_0$  has degree of freedom  $d_0 = |\omega_0|$  with  $d_0 < d$ . Note that in the standard LRT the set  $\omega_0$  is a simple point with  $d_0 = 0$  as the null model is a simple distribution. Thus, LRT is contained in GLRT as special case!

The corresponding generalised (log) likelihood ratio statistic is given by

$$\begin{aligned} W &= 2 \log \left( \frac{L(\hat{\theta}_{ML}|D)}{L(\hat{\theta}_{ML}^0|D)} \right) \\ &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\hat{\theta}_{ML}^0)) \end{aligned}$$

and

$$\Lambda = \frac{\max_{\theta \in \omega_0} L(\theta|D)}{\max_{\theta \in \Omega} L(\theta|D)}$$

where  $L(\hat{\theta}_{ML}|D)$  is the maximised likelihood assuming the full model (with full parameter space  $\Omega$ ) and  $L(\hat{\theta}_{ML}^0|D)$  is the maximised likelihood for the restricted model (with restricted parameter space  $\omega_0$ ). Hence, to compute the GLRT test statistic we need to perform two optimisations, one for the full and another for the restricted model.

### Remarks:

- MLE in the restricted model space  $\omega_0$  is taken as a representative of  $H_0$ .
- The likelihood is **maximised in both numerator and denominator**.
- The restricted model is a special case of the full model (i.e. the two models are nested).

- The asymptotic distribution of  $W$  is chi-squared with degree of freedom depending on both  $d$  and  $d_0$ :

$$W \stackrel{a}{\sim} \chi_{d-d_0}^2$$

- This result is due to Wilks (1938)<sup>1</sup>. Note that it assumes that the true model is contained among the investigated models.
- If  $H_0$  is a simple hypothesis (i.e.  $d_0 = 0$ ) then the standard LRT (and corresponding CI) is recovered as special case of the GLRT.

### Example 11.9. GLRT example:

*Case-control study:* (e.g. “healthy” vs. “disease”)

we observe normal data  $D = \{x_1, \dots, x_n\}$  from two groups with sample size  $n_1$  and  $n_2$  (and  $n = n_1 + n_2$ ), with two different means  $\mu_1$  and  $\mu_2$  and common variance  $\sigma^2$ :

$$x_1, \dots, x_{n_1} \sim N(\mu_1, \sigma^2)$$

and

$$x_{n_1+1}, \dots, x_n \sim N(\mu_2, \sigma^2)$$

Question: are the two means  $\mu_1$  and  $\mu_2$  the same in the two groups?

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \quad (\text{with variance unknown, i.e. treated as nuisance parameter}) \\ H_1 : \mu_1 &\neq \mu_2 \end{aligned}$$

*Restricted and full models:*

$\omega_0$ : restricted model with two parameters  $\mu_0$  and  $\sigma_0^2$  (so that  $x_1, \dots, x_n \sim N(\mu_0, \sigma_0^2)$ ).

$\Omega$ : full model with three parameters  $\mu_1, \mu_2, \sigma^2$ .

*Corresponding log-likelihood functions:*

Restricted model  $\omega_0$ :

$$\ell_n(\mu_0, \sigma_0^2) = -\frac{n}{2} \log(\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2$$

<sup>1</sup>Wilks, S. S. 1938. *The large-sample distribution of the likelihood ratio for testing composite hypotheses*. Ann. Math. Statist. 9:60–62. <https://doi.org/10.1214/aoms/1177732360>



Full model  $\Omega$ :

$$\begin{aligned}\ell_n(\mu_1, \mu_2, \sigma^2) &= \left( -\frac{n_1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n_1} (x_i - \mu_1)^2 \right) + \\ &\quad \left( -\frac{n_2}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=n_1+1}^n (x_i - \mu_2)^2 \right) \\ &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \left( \sum_{i=1}^{n_1} (x_i - \mu_1)^2 + \sum_{i=n_1+1}^n (x_i - \mu_2)^2 \right)\end{aligned}$$

Corresponding MLEs:

$$\begin{aligned}\omega_0 : \quad \hat{\mu}_0 &= \frac{1}{n} \sum_{i=1}^n x_i & \hat{\sigma}_0^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_0)^2 \\ \Omega : \quad \hat{\mu}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} x_i & \hat{\sigma}^2 &= \frac{1}{n} \left\{ \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^n (x_i - \hat{\mu}_2)^2 \right\} \\ \hat{\mu}_2 &= \frac{1}{n_2} \sum_{i=n_1+1}^n x_i\end{aligned}$$

Note that the three estimated means are related by

$$\begin{aligned}\hat{\mu}_0 &= \frac{n_1}{n} \hat{\mu}_1 + \frac{n_2}{n} \hat{\mu}_2 \\ &= \hat{\pi}_1 \hat{\mu}_1 + \hat{\pi}_2 \hat{\mu}_2\end{aligned}$$

so the overall mean is the weighted average of the two individual group means.

Moreover, the two estimated variances are related by

$$\hat{\sigma}_0^2 = \hat{\pi}_1 \hat{\pi}_2 (\hat{\mu}_1 - \hat{\mu}_2)^2 + \hat{\sigma}^2$$

Note that this is an example of variance decomposition, where

- $\hat{\sigma}_0^2$  is the estimated total variance,
- $\hat{\pi}_1 \hat{\pi}_2 (\hat{\mu}_1 - \hat{\mu}_2)^2$  the estimated between-group (explained) variance, and
- $\hat{\sigma}^2$  is the estimated average within-group (unexplained) variance.

For the following we also note that

$$\begin{aligned}\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} &= \hat{\pi}_1 \hat{\pi}_2 \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\sigma}^2} + 1 \\ &= \frac{t_{\text{ML}}^2}{n} + 1 \\ &= \frac{t_{\text{UB}}^2}{n-2} + 1\end{aligned}$$

where

$$t_{\text{ML}} = \sqrt{n \hat{\pi}_1 \hat{\pi}_2} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}}$$

is the two sample  $t$ -statistic based on the ML variance estimate  $\hat{\sigma}^2$  and  $t_{\text{UB}} = t_{\text{ML}} \sqrt{\frac{n-2}{n}}$  is the conventional two sample  $t$ -statistic based on the unbiased variance estimate  $\hat{\sigma}_{\text{UB}}^2 = \frac{n}{n-2} \hat{\sigma}^2$  (see Section A.9).

Corresponding maximised log-likelihood:

Restricted model:

$$\ell_n(\hat{\mu}_0, \hat{\sigma}_0^2) = -\frac{n}{2} \log(\hat{\sigma}_0^2) - \frac{n}{2}$$

Full model:

$$\ell_n(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) = -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{n}{2}$$

Likelihood ratio statistic:

$$\begin{aligned}W &= 2 \left( \ell_n(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) - \ell_n(\hat{\mu}_0, \hat{\sigma}_0^2) \right) \\ &= n \log \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right) \\ &= n \log \left( \frac{t_{\text{ML}}^2}{n} + 1 \right) \\ &= n \log \left( \frac{t_{\text{UB}}^2}{n-2} + 1 \right)\end{aligned}$$



Thus, the log-likelihood ratio statistic  $W$  is a monotonic function (a one-to-one transformation!) of the (squared) two sample  $t$ -statistic!

*Asymptotic distribution:*

The degree of freedom of the full model is  $d = 3$  and that of the constrained model  $d_0 = 2$  so the generalised log likelihood ratio statistic  $W$  is distributed asymptotically as  $\chi_1^2$ . Hence, we reject the null model on 5% significance level for all  $W > 3.84$ .

*Other application of GLRTs*

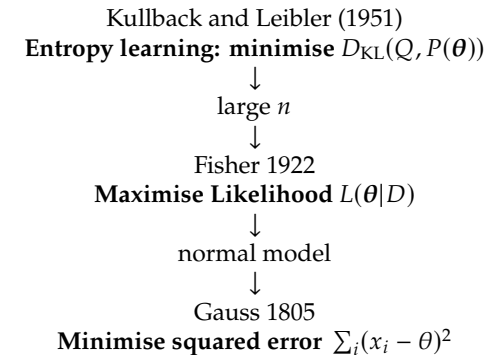
As shown above, the two sample  $t$  statistic can be derived as a likelihood ratio statistic.

More generally, it turns out that many other commonly used familiar statistical tests and test statistics can be interpreted as GLRTs. This shows the wide applicability of this procedure.

## 12 Optimality properties and conclusion

### 12.1 Properties of maximum likelihood encountered so far

1. MLE is a special case of KL divergence minimisation *valid for large samples*.
2. MLE can be seen as generalisation of least squares (and conversely, least squares is a special case of ML).



3. Given a model, derivation of the MLE is basically automatic (only optimisation required)!
4. MLEs are **consistent**, i.e. if the true underlying model  $Q$  is contained in the set of specified candidates models  $P(\theta)$  then the MLE will converge to the true model corresponding to parameter  $\theta_{\text{true}}$ .
5. Correspondingly, **MLEs are asymptotically unbiased**.
6. However, MLEs are *not* necessarily unbiased in finite samples (e.g. the MLE of the variance parameter in the normal distribution).



7. The achieved maximum likelihood is invariant under transformations of parameters.
8. In regular situations (when local quadratic approximation is possible) MLEs are **asymptotically normally distributed**, with the asymptotic variance determined by the inverse observed Fisher information.
9. In regular situations and for large sample size MLEs are **asymptotically optimally efficient** (Cramér-Rao theorem): For large samples the MLE achieves the lowest possible variance possible in an estimator — this is the so-called Cramér-Rao lower bound. The variance decreases to zero with  $n \rightarrow \infty$  typically with rate  $1/n$ .
10. The likelihood ratio can be used to construct optimal tests (in the sense of the Neyman-Pearson theorem).

## 12.2 Summarising data and the concept of (minimal) sufficiency

### Induced partitioning of data space and likelihood equivalence

Every sufficient statistic  $t(D)$  induces a partitioning of the space of data sets by clustering all hypothetical outcomes for which the statistic  $t(D)$  assumes the same value  $t$ :

$$\mathcal{X}_t = \{D : t(D) = t\}$$

The **data sets in  $\mathcal{X}_t$  are equivalent in terms of the sufficient statistic  $t(D)$** . Note that this implies that  $t(D)$  is not a 1:1 transformation of  $D$ . Instead of  $n$  data points  $x_1, \dots, x_n$  as few as one or two summaries (such as empirical mean and variance) may be sufficient to fully convey all the information in the data about the model parameters. Thus, transforming data  $D$  using a sufficient statistic  $t(D)$  may result in substantial **data reduction**.

Two data sets  $D_1$  and  $D_2$  for which the ratio of the corresponding likelihoods  $L(\theta|D_1)/L(\theta|D_2)$  does not depend on  $\theta$  (so the two likelihoods are proportional to each other by a constant) are called **likelihood equivalent** because a likelihood-based procedure to learn about  $\theta$  will draw identical conclusions from  $D_1$  and  $D_2$ . For data sets  $D_1, D_2 \in \mathcal{X}_t$  which

are equivalent with respect to a sufficient statistic  $T$  it follows directly from the Fisher-Pearson factorisation

$$L(\theta|D) = h(t(D), \theta) k(D)$$

that the ratio

$$L(\theta|D_1)/L(\theta|D_2) = k(D_1)/k(D_2)$$

and thus is constant with regard to  $\theta$ . As a result, all **data sets in  $\mathcal{X}_t$  are likelihood equivalent**. However, the converse is not true: depending on the sufficient statistics there usually will be many likelihood equivalent data sets that are not part of the same set  $\mathcal{X}_t$ .

### Minimal sufficient statistics

Of particular interest is therefore to find those sufficient statistics that achieve the coarsest partitioning of the sample space and thus may allow the highest data reduction. Specifically, a **minimal sufficient statistic** is a sufficient statistic for which all likelihood equivalent data sets also are equivalent under this statistic.

Therefore, to check whether a sufficient statistic  $t(D)$  is minimal sufficient we need to verify whether for any two likelihood equivalent data sets  $D_1$  and  $D_2$  it also follows that  $t(D_1) = t(D_2)$ . If this holds true then  $T$  is a minimal sufficient statistic.

An equivalent non-operational definition is that a minimal sufficient statistic  $t(D)$  is a sufficient statistic that can be computed from any other sufficient statistic  $S(D)$ . This follows from the above directly: assume any sufficient statistic  $S(D)$ , this defines a corresponding set  $\mathcal{X}_s$  of likelihood equivalent data sets. By implication any  $D_1, D_2 \in \mathcal{X}_s$  will necessarily also be in  $\mathcal{X}_t$ , thus whenever  $S(D_1) = S(D_2)$  we also have  $t(D_1) = t(D_2)$ , and therefore  $t(D_1)$  is a function of  $S(D_1)$ .

A trivial but **important example of a minimal sufficient statistic is the likelihood function itself** since by definition it can be computed from any set of sufficient statistics. Thus the likelihood function  $L(\theta)$  captures all information about  $\theta$  that is available in the data. In other words, it provides an *optimal summary* of the observed data with regard to a model. Note that in Bayesian statistics (to be discussed in Part 2 of the module) the likelihood function is used as proxy/summary of the data.



**Example: normal distribution**

**Example 12.1.** Sufficient statistics for the parameters of the normal distribution:

The normal model  $N(\mu, \sigma^2)$  with parameter vector  $\theta = (\mu, \sigma^2)^T$  and log-likelihood

$$\ell_n(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

One possible set of minimal sufficient statistics for  $\theta$  are  $\bar{x}$  and  $\overline{x^2}$ , and with these we can rewrite the log-likelihood function without any reference to the original data  $x_1, \dots, x_n$  as follows

$$\ell_n(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{n}{2\sigma^2} (\overline{x^2} - 2\bar{x}\mu + \mu^2)$$

An alternative set of minimal sufficient statistics for  $\theta$  consists of  $s^2 = \overline{x^2} - \bar{x}^2 = \sigma_{ML}^2$  and  $\bar{x} = \hat{\mu}_{ML}$ . The log-likelihood written in terms of  $s^2$  and  $\bar{x}$  is

$$\ell_n(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{n}{2\sigma^2} (s^2 + (\bar{x} - \mu)^2)$$

Note that in this example the dimension of the parameter vector  $\theta$  equals the dimension of the minimal sufficient statistic, and furthermore, that the MLEs of the parameters are in fact minimal sufficient!

**MLEs of parameters of an exponential family are minimal sufficient statistics**

The conclusion from Example 12.1 holds true more generally: **in an exponential family model** (such as the normal distribution as particular important case) **the MLEs of the parameters are minimal sufficient statistics**. Thus, there will typically be substantial dimension reduction from the raw data to the sufficient statistics.

However, outside exponential families the MLE is not necessarily a minimal sufficient statistic, and may not even be a sufficient statistic. This is because **a (minimal) sufficient statistic of the same dimension as the parameters does not always exist**. A classic example is the Cauchy distribution for which the minimal sufficient statistics are the ordered observations, thus the MLE of the parameters do not constitute sufficient

statistics, let alone minimal sufficient statistics. However, the MLE is of course still a function of the minimal sufficient statistic.

In summary, the likelihood function acts as perfect data summariser (i.e. as minimal sufficient statistic), and in exponential families (e.g. normal distribution) the MLEs of the parameters  $\hat{\theta}_{ML}$  are minimal sufficient.

Finally, while sufficiency is clearly a useful concept for data reduction one needs to keep in mind that this is always in reference to a specific model. Therefore, unless one strongly believes in a certain model it is generally a good idea to keep (and not discard!) the original data.

**12.3 Concluding remarks on maximum likelihood****Application of KL divergence in statistics**

In statistics the typical roles of the distribution  $Q$  and  $P$  in the KL divergence  $D_{KL}(Q, P)$  are:

- $Q$  is the (unknown) underlying true model for the data-generating process
- $P$  is the approximating model (typically a parametric distribution family)

Optimising (i.e. minimising) the KL divergence with regard to  $P$  amounts to *approximation* and optimising with regard to  $Q$  to *imputation*.

In previous chapters we have seen how the KL divergence leads to maximum likelihood (via minimum empirical risk) and also allows to choose distribution families (via maximum entropy). Later we will also see how KL divergence is linked to Bayesian learning.

Since the KL divergence is not symmetric there two distinct ways to minimise the divergence between a fixed  $F_0$  and the family  $F(\theta)$  (see Figure 12.1). minimising the parameter  $\theta$  in  $D_{KL}(\hat{F}_0, F(\theta))$  ("forward KL") and in  $D_{KL}(F(\theta), \hat{F}_0)$  ("backward KL").

Each way has different properties:



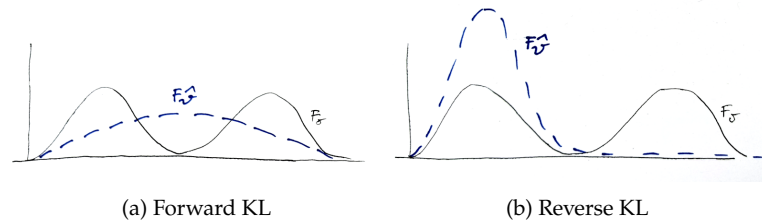


Figure 12.1: Illustration of (a) forward KL and b) reverse KL optimisation.

a) **forward KL, approximation KL**:  $\min_{\theta} D_{\text{KL}}(F_0, F(\theta))$ 

Here we keep the first argument fixed and minimise KL by changing the second argument. This is also called an “M (Moment) projection” or, alternatively, “reverse I-projection”. It has a **zero avoiding** property:  $f(x|\theta) > 0$  whenever  $f_0(x) > 0$ .

This procedure is mean-seeking and inclusive, i.e. when there are multiple modes in the density of  $F_0$  a fitted unimodal density  $F(\hat{\theta})$  will seek to cover all modes (**mass covering** property).

Maximum likelihood is based on “forward KL”.

b) **reverse KL, inference KL**:  $\min_{\theta} D_{\text{KL}}(F(\theta), F_0)$ 

Here we keep the second argument fixed and minimise KL by changing the first argument. This is also called an “I (Information) projection”. It has a **zero forcing** property:  $f(x|\theta) = 0$  whenever  $f_0(x) = 0$ .

This procedure is mode-seeking and exclusive, i.e. when there are multiple modes in the density of  $F_0$  a fitted unimodal density  $F(\hat{\theta})$  will seek out one mode to the exclusion of the others (**mode attracting** property).

Bayesian updating and variational Bayes approximations use “reverse KL”.

**What happens if  $n$  is small?**

From the long list of optimality properties of ML it is clear that for large sample size  $n$  the best estimator will typically be the MLE.

However, for **small sample size** it is indeed possible (and necessary) to **improve over the MLE** (e.g. via Bayesian estimation or regularisation). Some of these ideas will be discussed in Part II.

- Likelihood will *overfit*!

Alternative methods need to be used:

- regularised/penalised likelihood
- Bayesian methods

which are essentially two sides of the same coin.

Classic example of a simple non-ML estimator that is better than the MLE: **Stein’s example / Stein paradox** (C. Stein, 1955):

- Problem setting: estimation of the mean in multivariate case
- Maximum likelihood estimation breaks down!  $\rightarrow$  average (=MLE) is worse in terms of MSE than Stein estimator.
- For small  $n$  the asymptotic distributions for the MLE and for the LRT are not accurate, so for inference in these situations the distributions may need to be obtained by simulation (e.g. parametric or nonparametric bootstrap).

**Model selection**

- CI are sets of models that are not statistically distinguishable from the best ML model
- in doubt, choose the simplest model compatible with data
- better prediction, avoids overfitting
- Useful for model exploration and model building.
- Note that, by construction, the model with more parameters always has a higher likelihood, implying likelihood favours complex models
- Complex model may overfit!
- For comparison of models penalised likelihood or Bayesian approaches may be necessary
- Model selection in small samples and high dimension is challenging



- Recall that the aim in statistics is **not** about rejecting models (this is easy as for large sample size any model will be rejected!)
- Instead, the aim is model building, i.e. to find a model that **explains the data well** and that **predicts well**!
- Typically, this will not be the best-fit ML model, but rather a simpler model that is close enough to the best / most complex model.

## Part IV

# Bayesian statistics



## 13 Conditioning and Bayes rule

In this chapter we review conditional probabilities. Conditional probability is essential for Bayesian statistical modelling.

### 13.1 Conditional probability

We consider two random variables  $x$  and  $y$  and assume a **joint density** (or joint PMF)  $p(x, y)$ . By definition  $\int_{x,y} p(x, y) dx dy = 1$ .

The **marginal densities** for the individual random variables  $x$  and  $y$  are given by  $p(x) = \int_y p(x, y) dy$  and  $p(y) = \int_x p(x, y) dx$ , respectively. Therefore, the marginal densities are derived from the joint density by integrating over all possible states of the variable that is being excluded. As necessary for any density, the marginal densities also integrate to one, i.e.  $\int_x p(x) dx = 1$  and  $\int_y p(y) dy = 1$ .

As alternative to integrating out a random variable in the joint density  $p(x, y)$  we may wish to keep it fixed at some value. For instance we may want to keep  $y$  fixed at  $y_0$ . In this case  $p(x, y = y_0)$  is proportional to the **conditional density** (or PMF) given by the ratio

$$p(x|y = y_0) = \frac{p(x, y = y_0)}{p(y = y_0)}$$

In this formula the denominator  $p(y = y_0) = \int_x p(x, y = y_0) dx$  ensures that  $\int_x p(x|y = y_0) dx = 1$ , thus it renormalises  $p(x, y = y_0)$  so that it becomes a density integrating to one. To simplify notation, the particular value on which a variable is conditioned is often left out, and we just write  $p(x|y)$ .

### 13.2 Bayes' theorem

[Thomas Bayes \(1701-1761\)](#) was the first to state [Bayes' theorem](#) on conditional probabilities.

Using the definition of conditional probabilities we see that the joint density can be written as the product of marginal and conditional density in two different ways:

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

This directly leads to **Bayes' theorem**:

$$p(x|y) = p(y|x) \frac{p(x)}{p(y)}$$

This rule relates the two possible conditional densities (or conditional probability mass functions) for the random variables  $x$  and  $y$ , enabling the change of the order of conditioning.

Bayes's theorem was [published in 1763](#) only after his death by [Richard Price \(1723-1791\)](#):

[Pierre-Simon Laplace](#) independently published Bayes' theorem in 1774 and he was in fact the first to routinely apply it to statistical calculations.

### 13.3 Conditional mean and variance

The conditional distribution  $P_{x|y}$  with density  $p(x|y)$  has mean  $E(x|y)$  and variance  $\text{Var}(x|y)$ . These are called the **conditional mean** and **conditional variance**, respectively.

The conditional mean  $E(x|y)$  is also denoted by  $E\left(P_{x|y}\right)$  and  $E_{P_{x|y}}(x)$ . It is obtained by calculating

$$E(x|y) = E\left(P_{x|y}\right) = E_{P_{x|y}}(x) = \begin{cases} \sum_x p(x|y) x & \text{discrete case} \\ \int_x p(x|y) x dx & \text{continuous case} \end{cases}$$

The conditional variance  $\text{Var}(x|y)$  is also denoted by  $\text{Var}\left(P_{x|y}\right)$  and  $\text{Var}_{P_{x|y}}(x)$ .



$$\text{Var}(x|y) = \text{Var}\left(P_{x|y}\right) = \text{Var}_{P_{x|y}}(x) = E_{P_{x|y}}((x - E_{P_{x|y}}(x))^2) = E((x - E(x|y))^2|y)$$

The **law of total expectation** links the means of the marginal distribution and of the conditional distributions, stating that

$$\begin{aligned} E(x) &= E(E(x|y)) \\ &= E_{P_y} E_{P_{x|y}}(x) \\ &= E_{P_{x,y}}(x) \end{aligned}$$

Hence, the total mean (left side) is the weighted average (outer expectation) of the various conditional means (inner expectation).

Similarly, the **law of total variance** states that

$$\begin{aligned} \text{Var}(x) &= \text{Var}(E(x|y)) + E(\text{Var}(x|y)) \\ &= \text{Var}_{P_y} E_{P_{x|y}}(x) + E_{P_y} \text{Var}_{P_{x|y}}(x) \end{aligned}$$

The total variance (left side) decomposes into the “explained” variance or “between-group” variance (first term on right side) and the “unexplained” variance or “mean within group” variance (second term on the right side). Again, the outer expectations are with regard to  $P_y$  and the inner expectations with regard to  $P_{x|y}$ .

**Example 13.1.** Mean and variance of a mixture model:

Assume  $K$  groups indicated by a discrete variable  $y \in \{1, 2, \dots, K\}$  with probability  $p(y) = \pi_y$ . In each group the observations  $x$  follow a density  $p(x|y)$  with conditional mean  $E(x|y) = \mu_y$  and conditional variance  $\text{Var}(x|y) = \sigma_y^2$ . The joint density for  $x$  and  $y$  is  $p(x, y) = \pi_y p(x|y)$ . The marginal density for  $x$  is  $p(x) = \sum_{y=1}^K \pi_y p(x|y)$ . This is called a mixture model.

The total mean  $E(x) = \mu_0$  is equal to

$$\mu_0 = \sum_{y=1}^K \pi_y \mu_y$$

The total variance  $\text{Var}(x) = \sigma_0^2$  is equal to

$$\sigma_0^2 = \sum_{y=1}^K \pi_y (\mu_y - \mu_0)^2 + \sum_{y=1}^K \pi_y \sigma_y^2$$

## 13.4 Conditional entropy and entropy chain rules

Similar to mean and variance one can also define conditional versions of entropies. These lead to decomposition of the joint entropy that are also known as **entropy chain rules**.

### Conditional entropy

For the entropy of the joint distribution we find that

$$\begin{aligned} H(P_{x,y}) &= -E_{P_{x,y}} \log p(x, y) \\ &= -E_{P_x} E_{P_{y|x}} (\log p(x) + \log p(y|x)) \\ &= -E_{P_x} \log p(x) - E_{P_x} E_{P_{y|x}} \log p(y|x) \\ &= H(P_x) + H(P_{y|x}) \end{aligned}$$

thus it decomposes into the entropy of the marginal distribution and the **conditional entropy** defined as

$$H(P_{y|x}) = -E_{P_x} E_{P_{y|x}} \log p(y|x)$$

Note that to simplify notation by convention the expectation  $E_{P_x}$  over the variable  $x$  that we condition on ( $x$ ) is implicitly assumed.

### Conditional cross-entropy

Similarly, for the cross-entropy we get

$$\begin{aligned} H(Q_{x,y}, P_{x,y}) &= -E_{Q_{x,y}} \log p(x, y) \\ &= -E_{Q_x} E_{Q_{y|x}} \log (p(x) p(y|x)) \\ &= -E_{Q_x} \log p(x) - E_{Q_x} E_{Q_{y|x}} \log p(y|x) \\ &= H(Q_x, P_x) + H(Q_{y|x}, P_{y|x}) \end{aligned}$$

where the **conditional cross-entropy** is defined as

$$H(Q_{y|x}, P_{y|x}) = -E_{Q_x} E_{Q_{y|x}} \log p(y|x)$$

Note again the implicit expectation  $E_{Q_x}$  over  $x$  implied in this notation.



### Conditional KL divergence

The KL divergence between the joint distributions can be decomposed as follows:

$$\begin{aligned}
 D_{\text{KL}}(Q_{x,y}, P_{x,y}) &= E_{Q_{x,y}} \log \left( \frac{q(x, y)}{p(x, y)} \right) \\
 &= E_{Q_x} E_{Q_{y|x}} \log \left( \frac{q(x)q(y|x)}{p(x)p(y|x)} \right) \\
 &= E_{Q_x} \log \left( \frac{q(x)}{p(x)} \right) + E_{Q_x} E_{Q_{y|x}} \log \left( \frac{q(y|x)}{p(y|x)} \right) \\
 &= D_{\text{KL}}(Q_x, P_x) + D_{\text{KL}}(Q_{y|x}, P_{y|x})
 \end{aligned}$$

with the **conditional KL divergence** defined as

$$D_{\text{KL}}(Q_{y|x}, P_{y|x}) = E_{Q_x} E_{Q_{y|x}} \log \left( \frac{q(y|x)}{p(y|x)} \right)$$

(again the expectation  $E_{Q_x}$  is usually dropped for convenience). The conditional KL divergence can also be computed from the conditional (cross-)entropies by the familiar relationship

$$D_{\text{KL}}(Q_{y|x}, P_{y|x}) = H(Q_{y|x}, P_{y|x}) - H(Q_{y|x})$$

### Conditional entropy with prior measure

Entropy with prior measure is the negative of the KL divergence.

Hence:

$$\begin{aligned}
 G(Q_{x,y}, P_{x,y}) &= G(Q_x, P_x) + G(Q_{y|x}, P_{y|x}) \\
 &= G(Q_y, P_y) + G(Q_{x|y}, P_{x|y})
 \end{aligned}$$

## 13.5 Entropy bounds for the marginal variables

The chain rule for KL divergence directly shows that

$$\begin{aligned}
 \underbrace{D_{\text{KL}}(Q_{x,y}, P_{x,y})}_{\text{upper bound}} &= D_{\text{KL}}(Q_x, P_x) + \underbrace{D_{\text{KL}}(Q_{y|x}, P_{y|x})}_{\geq 0} \\
 &\geq D_{\text{KL}}(Q_x, P_x)
 \end{aligned}$$

This means that the KL divergence between the joint distributions forms an **upper bound for the KL divergence between the marginal distributions**, with the difference given by the conditional KL divergence  $D_{\text{KL}}(Q_{y|x}, P_{y|x})$ .

Equivalently, we can state an **upper bound for the marginal cross-entropy**:

$$\begin{aligned}
 \underbrace{H(Q_{x,y}, P_{x,y}) - H(Q_{y|x})}_{\text{upper bound}} &= H(Q_x, P_x) + \underbrace{D_{\text{KL}}(Q_{y|x}, P_{y|x})}_{\geq 0} \\
 &\geq H(Q_x, P_x)
 \end{aligned}$$

Instead of an upper bound we may as well express this as **lower bound for the negative marginal cross-entropy**

$$\begin{aligned}
 -H(Q_x, P_x) &= \underbrace{-H(Q_x Q_{y|x}, P_{x,y}) + H(Q_{y|x})}_{\text{lower bound}} + \underbrace{D_{\text{KL}}(Q_{y|x}, P_{y|x})}_{\geq 0} \\
 &\geq F(Q_x, Q_{y|x}, P_{x,y})
 \end{aligned}$$

Since entropy and KL divergence is closely linked with maximum likelihood the above bounds play a major role in statistical learning of models with unobserved latent variables (here  $y$ ). They form the basis of important methods such as the EM algorithm as well as of variational Bayes.



## 14 Models with latent variables and missing data

### 14.1 Complete data log-likelihood versus observed data log-likelihood

It is frequently the case that we need to employ models where not all variables are observable and the corresponding data are missing.

For example consider two random variables  $x$  and  $y$  with a joint density

$$p(x, y|\theta)$$

and parameters  $\theta$ . If we observe data  $D_x = \{x_1, \dots, x_n\}$  and  $D_y = \{y_1, \dots, y_n\}$  for  $n$  samples we can use the **complete data log-likelihood**

$$\ell(\theta|D_x, D_y) = \sum_{i=1}^n \log p(x_i, y_i|\theta)$$

to estimate  $\theta$ . Recall that

$$\ell(\theta|D_x, D_y) = -nH(\hat{Q}_{x,y}, P_{x,y|\theta})$$

where  $\hat{Q}_{x,y}$  is the empirical joint distribution based on both  $D_x$  and  $D_y$  and  $P_{x,y|\theta}$  the joint model, so maximising the complete data log-likelihood minimises the cross-entropy  $H(\hat{Q}_{x,y}, P_{x,y|\theta})$ .

Now assume that  $y$  is not observable and hence is a so-called **latent variable**. Then we don't have observations  $D_y$  and therefore cannot use the complete data likelihood. Instead, for maximum likelihood estimation with missing data we need to use the **observed data log-likelihood**.

From the joint density we obtain the marginal density for  $x$  by integrating out the unobserved variable  $y$ :

$$p(x|\theta) = \int_y p(x, y|\theta) dy$$

Using the marginal model we then compute the **observed data log-likelihood**

$$\ell(\theta|D_x) = \sum_{i=1}^n \log p(x_i|\theta) = \sum_{i=1}^n \log \int_y p(x_i, y|\theta) dy$$

Note that only the data  $D_x$  are used.

Maximum likelihood estimation based on the marginal model proceeds as usual by maximising the corresponding observed data likelihood function which is

$$\ell(\theta|D_x) = -nH(\hat{Q}_x, P_{x|\theta})$$

where  $\hat{Q}_x$  is the empirical distribution based only on  $D_x$  and  $P_{x|\theta}$  is the model family. Hence, maximising the observed data log-likelihood minimises the cross-entropy  $H(\hat{Q}_x, P_{x|\theta})$ .

**Example 14.1.** Two group normal mixture model:

Assume we have two groups labelled by  $y = 1$  and  $y = 2$  (thus the variable  $y$  is discrete). The data  $x$  observed in each group are normal with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively. The probability of group 1 is  $\pi_1 = p$  and the probability of group 2 is  $\pi_2 = 1 - p$ . The density of the joint model for  $x$  and  $y$  is

$$p(x, y|\theta) = \pi_y N(x|\mu_y, \sigma_y^2)$$

The model parameters are  $\theta = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$  and they can be inferred from the complete data comprised of  $D_x = \{x_1, \dots, x_n\}$  and the group allocations  $D_y = \{y_1, \dots, y_n\}$  of each sample using the complete data log-likelihood

$$\ell(\theta|D_x, D_y) = \sum_{i=1}^n \log \pi_{y_i} + \sum_{i=1}^n \log N(x_i|\mu_{y_i}, \sigma_{y_i}^2)$$

However, typically we do not know the class allocation  $y$  and thus we need to use the marginal model for  $x$  alone which has density

$$\begin{aligned} p(x|\theta) &= \sum_{y=1}^2 \pi_y N(x|\mu_y, \sigma_y^2) \\ &= pN(x|\mu_1, \sigma_1^2) + (1-p)N(x|\mu_2, \sigma_2^2) \end{aligned}$$



This is an example of a **two-component mixture model**. The corresponding observed data log-likelihood is

$$\ell(\theta|D_x) = \sum_{i=1}^n \log \sum_{y=1}^2 \pi_y N(x|\mu_y, \sigma_y^2)$$

Note that the form of the observed data log-likelihood is more complex than that of the complete data log-likelihood because it contains the logarithm of a sum that cannot be simplified. It is used to estimate the model parameters  $\theta$  from  $D_x$  without requiring knowledge of the class allocations  $D_y$ .

**Example 14.2.** Alternative computation of the observed data likelihood:

An alternative way to arrive at the observed data likelihood is to marginalise the complete data likelihood.

$$L(\theta|D_x, D_y) = \prod_{i=1}^n p(x_i, y_i|\theta)$$

and

$$L(\theta|D_x) = \int_{y_1, \dots, y_n} \prod_{i=1}^n p(x_i, y_i|\theta) dy_1 \dots dy_n$$

The integration (sum) and the multiplication can be interchanged as per [Generalised Distributive Law](#) leading to

$$L(\theta|D_x) = \prod_{i=1}^n \int_y p(x_i, y|\theta) dy$$

which is the same as constructing the likelihood from the marginal density.

## 14.2 Estimation of the unobservable latent states using Bayes theorem

After estimating the marginal model it is straightforward to obtain a probabilistic prediction about the state of the latent variables  $y_1, \dots, y_n$ . Since

$$p(x, y|\theta) = p(x|\theta) p(y|x, \theta) = p(y|\theta) p(x|y, \theta)$$

given an estimate  $\hat{\theta}$  we are able to compute for each observation  $x_i$

$$p(y_i|x_i, \hat{\theta}) = \frac{p(x_i, y_i|\hat{\theta})}{p(x_i|\hat{\theta})} = \frac{p(y_i|\hat{\theta}) p(x_i|y_i, \hat{\theta})}{p(x_i|\hat{\theta})}$$

the probabilities / densities of all states of  $y_i$  (note this an application of Bayes' theorem).

**Example 14.3.** Latent states of two group normal mixture model:

Continuing from Example 14.1 above we assume the marginal model has been fitted with parameter values  $\hat{\theta} = (\hat{p}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2)^T$ . Then for each sample  $x_i$  we can get probabilistic prediction about group association of each sample by

$$p(y_i|x_i, \hat{\theta}) = \frac{\hat{\pi}_{y_i} N(x_i|\hat{\mu}_{y_i}, \hat{\sigma}_{y_i}^2)}{\hat{p} N(x_i|\hat{\mu}_1, \hat{\sigma}_1^2) + (1 - \hat{p}) N(x_i|\hat{\mu}_2, \hat{\sigma}_2^2)}$$

## 14.3 EM Algorithm

Computing and maximising the observed data log-likelihood can be difficult because of the integration over the unobserved variable (or summation in case of a discrete latent variable). In contrast, the complete data log-likelihood function may be easier to compute.

The widely used **EM algorithm**, formally described by Dempster and others (1977) but also used before, addresses this problem and maximises the observed data log-likelihood indirectly in an iterative procedure comprising two steps:

- 1) First ("E" step), the missing data  $D_y$  is imputed using Bayes' theorem. This provides probabilities ("soft allocations") for each possible state of the latent variable.
- 2) Subsequently ("M" step), the expected complete data log-likelihood function is computed, where the expectation is taken with regard to the distribution over the latent states, and it is maximised with regard to  $\theta$  to estimate the model parameters.

The EM algorithm leads to the exact same estimates as if the observed data log-likelihood would be optimised directly. Therefore the EM algorithm is in fact *not* an approximation, it is just a different way to find the MLEs.



The EM algorithm and application to clustering is discussed in more detail in the module [MATH38161 Multivariate Statistics and Machine Learning](#).

In a nutshell, the justification for the EM algorithm follows from the entropy chain rules and the corresponding bounds, such as  $D_{\text{KL}}(Q_{x,y}, P_{x,y}) \geq D_{\text{KL}}(Q_x, P_x)$  (see previous chapter). Given observed data for  $x$  we know the empirical distribution  $\hat{Q}_x$ . Hence, by minimising  $D_{\text{KL}}(\hat{Q}_x Q_{y|x}, P_{x,y}^\theta)$  iteratively

- 1) with regard to  $Q_{y|x}$  (“E” step) and
- 2) with regard to the parameters  $\theta$  of  $P_{x,y}^\theta$  (“M” step”)

one minimises  $D_{\text{KL}}(\hat{Q}_x, P_x^\theta)$  with regard to the parameters of  $P_x^\theta$ .

Interestingly, in the “E” step the first argument of the KL divergence is optimised (“I” projection) and in the “M” step the second argument (“M” projection).

Alternatively, instead of bounding the marginal KL divergence one can also either minimise the upper bound of the cross-entropy or maximise the lower bound of the negative cross-entropy. All of these three procedures yield the same EM algorithm.

Note that the optimisation of the entropy bound in the “E” step requires variational calculus since the argument is a distribution! The EM algorithm is therefore in fact a special case of a **variational Bayes algorithm** since it not only provides estimates of  $\theta$  but also yields the distribution of the latent states by means of the calculus of variations.

Finally, the previous discussion illustrates that we can gain insights into unobservable states using Bayes’ theorem. By applying the same principle to parameters and models, we arrive at the Bayesian approach to statistical learning.

## 15 Essentials of Bayesian statistics

### 15.1 Principle of Bayesian learning

#### From prior to posterior distribution

Bayesian statistical learning applies Bayes’ theorem to update our state of knowledge about a parameter in the light of data.

Ingredients:

- $\theta$  parameter(s) of interest, unknown and fixed.
- prior distribution with density  $p(\theta)$  describing the *uncertainty* (not randomness!) about  $\theta$
- data-generating process  $p(x|\theta)$

Note the **model underlying the Bayesian approach is the joint distribution**

$$p(\theta, x) = p(\theta)p(x|\theta)$$

as both a prior distribution over the parameters as well as a data-generating process have to be specified.

Question: new information in the form of a new observation  $x$  arrives - how does the uncertainty about  $\theta$  change?

Answer: use Bayes’ theorem to **update the prior density to the posterior density** (see Figure 15.1).

$$\underbrace{p(\theta|x)}_{\text{posterior}} = \underbrace{p(\theta)}_{\text{prior}} \frac{p(x|\theta)}{p(x)}$$



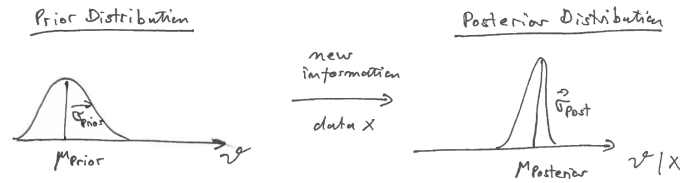


Figure 15.1: Bayesian learning by updating the prior distribution to the posterior distribution.

For the denominator in Bayes formula we need to compute  $p(x)$ . This is obtained by

$$\begin{aligned} p(x) &= \int_{\theta} p(x, \theta) d\theta \\ &= \int_{\theta} p(x|\theta)p(\theta) d\theta \end{aligned}$$

i.e. by marginalisation of the parameter  $\theta$  from the joint distribution of  $\theta$  and  $x$ . (For discrete  $\theta$  replace the integral by a sum). Depending on the context this quantity is either called the

- **normalisation constant** as it ensures that the posterior density  $p(\theta|x)$  integrates to one.
- **prior predictive density** of the data  $x$  given the model  $M$  before seeing any data. To emphasise the implicit conditioning on a model we may write  $p(x|M)$ . Since all parameters have been integrated out  $M$  in fact refers to a model *class*.
- **marginal likelihood** of the underlying **model** (class)  $M$  given data  $x$ . To emphasise this may write  $L(M|x)$ . Sometimes it is also called **model likelihood**.

### Zero forcing property

It is easy to see that if in Bayes rule the prior density/probability is zero for some parameter value  $\theta$  then the posterior density/probability will remain at zero for that  $\theta$ , regardless of any data collected. This **zero-forcing property** of the Bayes update rule has been called **Cromwell's rule** by [Dennis Lindley \(1923–2013\)](#). Therefore, assigning prior density/probability 0 to an event should be avoided.

Note that this implies that assigning prior probability 1 should be avoided, too.

### Bayesian update and likelihood

After *independent and identically distributed* (iid) data  $D = \{x_1, \dots, x_n\}$  have been observed the Bayesian posterior is computed by

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{p(\theta)}_{\text{prior}} \frac{L(\theta|D)}{p(D)}$$

involving the likelihood  $L(\theta|D) = \prod_{i=1}^n p(x_i|\theta)$  and the marginal likelihood  $p(D) = \int_{\theta} p(\theta)L(\theta|D)d\theta$  with  $\theta$  integrated out.

The marginal likelihood serves as a standardising factor so that the posterior density for  $\theta$  integrates to 1:

$$\int_{\theta} p(\theta|D)d\theta = \frac{1}{p(D)} \int_{\theta} p(\theta)L(\theta|D)d\theta = 1$$

Unfortunately, the integral to compute the marginal likelihood is typically analytically intractable and requires numerical integration and/or approximation.

Comparing likelihood and Bayes procedures note that

- conducting a Bayesian statistical analysis requires integration respectively averaging (to compute the marginal likelihood)
- in contrast to a likelihood analysis that requires optimisation (to find the maximum likelihood).

### Sequential updates

Note that the Bayesian update procedure can be repeated again and again: we can use the posterior as our new prior and then update it with further data. Thus, we may also update the posterior density sequentially, with the data points  $x_1, \dots, x_n$  arriving one after the other, by computing first  $p(\theta|x_1)$ , then  $p(\theta|x_1, x_2)$  and so on until we reach  $p(\theta|x_1, \dots, x_n) = p(\theta|D)$ .

For example, for the first update we have

$$p(\theta|x_1) = p(\theta) \frac{p(x_1|\theta)}{p(x_1)}$$



with  $p(x_1) = \int_{\theta} p(x_1|\theta)p(\theta)d\theta$ .

The second update yields

$$\begin{aligned} p(\theta|x_1, x_2) &= p(\theta|x_1) \frac{p(x_2|\theta, x_1)}{p(x_2|x_1)} \\ &= p(\theta|x_1) \frac{p(x_2|\theta)}{p(x_2|x_1)} \\ &= p(\theta) \frac{p(x_1|\theta)p(x_2|\theta)}{p(x_1)p(x_2|x_1)} \end{aligned}$$

with  $p(x_2|x_1) = \int_{\theta} p(x_2|\theta)p(\theta|x_1)d\theta$ .

Note that *given*  $\theta$  the  $x_i$  are independent, hence  $p(x_2|\theta, x_1) = p(x_2|\theta)$  and the joint distribution given  $\theta$  is  $p(x_1, x_2|\theta) = p(x_1|\theta)p(x_2|\theta)$ . In contrast, the joint distribution with  $\theta$  integrated out is  $p(x_1, x_2) = p(x_1)p(x_2|x_1) \neq p(x_1)p(x_2)$  so without specification of  $\theta$  out the  $x_i$  are dependent.

The final step is

$$p(\theta|D) = p(\theta|x_1, \dots, x_n) = p(\theta) \frac{\prod_{i=1}^n p(x_i|\theta)}{p(D)}$$

with the marginal likelihood factorising into

$$p(D) = \prod_{i=1}^n p(x_i|x_{<i})$$

with

$$p(x_i|x_{<i}) = \int_{\theta} p(x_i|\theta)p(\theta|x_{<i})d\theta$$

The last factor is the **posterior predictive density** of the new data  $x_i$  after seeing data  $x_1, \dots, x_{i-1}$  (given the model class  $M$ ).

Intuitively, we can understand why the probability of the new  $x_i$  depends on the previously observed data points. This is because the uncertainty about the model parameter  $\theta$  depends on how many data points we have already observed, and typically with more data the uncertainty about  $\theta$  decreases. Therefore the marginal likelihood  $p(D)$  is *not* simply the product of the marginal densities  $p(x_i)$  at each  $x_i$  but instead the product of the conditional densities  $p(x_i|x_{<i})$ .

Only when the parameter is fully known, and there is no uncertainty about  $\theta$ , the observations  $x_i$  are independent. This leads back to the standard likelihood where we condition on a particular  $\theta$  and the likelihood is the product  $p(D|\theta) = \prod_{i=1}^n p(x_i|\theta)$ .

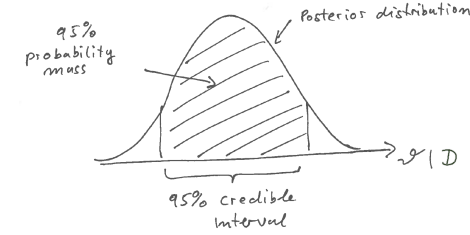


Figure 15.2: Bayesian credible interval based on the posterior distribution.

## Summaries of posterior distributions and credible intervals

**The Bayesian estimate is the full complete posterior distribution!**

However, it is useful to summarise aspects of the posterior distribution:

- Posterior mean  $E(\theta|D)$
- Posterior variance  $\text{Var}(\theta|D)$
- Posterior mode etc.

In particular the mean of the posterior distribution is often taken as a *Bayesian point estimate*.

The posterior distribution also allows to define **credible regions** or **credible intervals**. These are the **Bayesian equivalent to confidence intervals** and are constructed by finding the areas of highest probability mass (say 95%) in the posterior distribution (Figure 15.2).

Bayesian credible intervals, unlike frequentist confidence intervals, are thus very easy to interpret as they simply correspond to the part of the parameter space in which we can find the parameter with a given specified probability. In contrast, in frequentist statistics it does not make sense to assign a probability to a parameter value.

Note that there are typically many credible intervals with the given specified coverage  $\alpha$  (say 95%). Therefore, we may need further criteria to construct these intervals.

For univariate parameter  $\theta$  a **two-sided equal-tail credible interval** is obtained by finding the corresponding lower  $1 - \alpha/2$  and upper  $\alpha/2$  quantiles. Typically this type of credible interval is easy to compute. However, note that the density values at the left and right boundary



points of such an interval are typically different. Also this does not generalise well to a multivariate parameter  $\theta$ .

As alternative, a **highest posterior density (HPD)** credible interval of coverage  $\alpha$  is found by identifying the shortest interval (i.e. with smallest support) for the given  $\alpha$  probability mass. Any point within an HPD credible interval has higher density than a point outside the HPD credible interval. Correspondingly, the density at the boundary of an HPD credible interval is constant taking on the same value everywhere along the boundary.

A Bayesian HPD credible interval is constructed in a similar fashion as a likelihood-based confidence interval, starting from the mode of the posterior density and then looking for a common threshold value for the density to define the boundary of the credible interval. When the posterior density has multiple modes the HPD interval may be disjoint. HPD intervals are also well defined for multivariate  $\theta$  with the boundaries given by the contour lines of the posterior density resulting from the threshold value.

In the Worksheet B1 examples for both types of credible intervals are given and compared visually.

## Practical application of Bayes statistics on the computer

As we have seen Bayesian learning is *conceptually straightforward*:

- 1) Specify prior uncertainty  $p(\theta)$  about the parameters of interest  $\theta$ .
- 2) Specify the data-generating process for a specified parameter:  $p(x|\theta)$ .
- 3) Apply Bayes' theorem to update prior uncertainty in the light of the new data.

In practice, however, computing the posterior distribution can be *computationally very demanding*, especially for complex models.

For this reason specialised software packages have been developed for computational Bayesian modelling, for example:

- Bayesian statistics in R: <https://cran.r-project.org/web/views/Bayesian.html>
- Stan probabilistic programming language (interfaces with R, Python, Julia and other languages) — <https://mc-stan.org>

- Bayesian statistics in Python: [PyMC](#) using [PyTensor](#) as backend, [NumPyro](#) using [JAX](#) as backend, [TensorFlow Probability on JAX](#) using [JAX](#) as backend, [Pyro](#) using [PyTorch](#) as backend, [TensorFlow Probability](#) using [Tensorflow](#) as backend.
- Bayesian statistics in Julia: [Turing.jl](#)
- Bayesian hierarchical modelling with [BUGS](#), [JAGS](#) and [NIMBLE](#).

In addition to numerical procedures to sample from the posterior distribution there are also many procedures aiming to approximate the Bayesian posterior, employing the [Laplace approximation](#), integrated nested Laplace approximation (INLA), [variational Bayes](#) etc.

## 15.2 Some background on Bayesian statistics

### Bayesian interpretation of probability

#### What makes you “Bayesian”?

If you use Bayes' theorem are you therefore automatically a Bayesian? No!!

Bayes' theorem is a mathematical fact from probability theory. Hence, Bayes' theorem is valid for everyone, whichever form for statistical learning you are subscribing (such as frequentist ideas, likelihood methods, entropy learning, Bayesian learning).

As we discuss now the key difference between Bayesian and frequentist statistical learning lies in the differences in *interpretation of probability*, not in the mathematical formalism for probability (which includes Bayes' theorem).

#### Mathematics of probability

The mathematics of probability in its modern foundation was developed by [Andrey Kolmogorov \(1903–1987\)](#). In this book [Foundations of the Theory of Probability \(1933\)](#) he establishes probability in terms of set theory / measure theory. This theory provides a coherent mathematical framework to work with probabilities.

However, Kolmogorov's theory does *not* provide an interpretation of probability!



→ The Kolmogorov framework is the basis for both the frequentist and the Bayesian interpretation of probability.

### Interpretations of probability

Essentially, there are two major commonly used interpretation of probability in statistics - the **frequentist interpretation** and the **Bayesian interpretation**.

#### A: Frequentist interpretation

Probability is interpreted as a relative frequency (of an event in a long-running series of identically repeated experiments, ideally based on infinitely many trials)

Note that the law of large numbers provides justification to interpret long-running relative frequencies as probabilities. However, the converse assumption — and the key frequentist assumption! — that all probabilities have a frequentist interpretation does *not* follow from the law of large numbers.

By definition, the frequentist view of probability is very restrictive. For example, frequentist probability cannot be used to describe events that occur only a single time. Frequentist probability thus can only be applied asymptotically, for large samples.

The frequentist view of probability is also called the *ontological view* of probability as it assumes that probability “exists” independently of an observer and our knowledge.

In turn, ontological probability implies actual inherent randomness which, as we have discussed already in the introduction, amounts only to a small fraction of settings in which probability is used.

As a result, in classical frequentist statistics there are constructs other than probability to assess the uncertainty about parameters (e.g. confidence intervals, sampling distributions).

#### B: Bayesian probability

“Probability does not exist” — famous quote by [Bruno de Finetti \(1906–1985\)](#), a Bayesian statistician.

What does this mean?

Probability is a **description of the state of knowledge** and of **uncertainty**.

Probability is thus an *epistemological quantity* that is assigned and that changes with more information rather than something that is an inherent property.

Note that this notion of probability does not require any repeated experiments. The Bayesian interpretation of probability is valid regardless of sample size or the number or repetitions of an experiment.

Epistemological probability can be applied both to single repetitions and to large number of trials, in the latter case (by means of the law of large numbers) it agrees numerically with frequentist probability.

**Hence, the key difference between frequentist and Bayesian approaches is not the use of Bayes’ theorem. Rather it is whether you consider probability as ontological (frequentist) or epistemological entity (Bayesian).**

As a result, in Bayesian statistics to assess the uncertainty about parameters one can use probability directly (in the form of prior and posterior distributions).

### Historical developments

- Bayesian statistics is named after [Thomas Bayes](#) (1701-1761). His paper<sup>1</sup> introducing the famous theorem was published only after his death (1763).
- [Pierre-Simon Laplace](#) (1749-1827) was the first to practically use Bayes’ theorem for statistical calculations, and he also independently discovered Bayes’ theorem in 1774<sup>2</sup>
- This activity was then called “*inverse probability*” and not “Bayesian statistics”.
- Between 1900 and 1940 classical mathematical statistics was developed and the field was heavily influenced and dominated by [R.A. Fisher](#) (who invented likelihood theory and ANOVA, among other things - he was also working in biology and was professor of genetics). Fisher was very much opposed to Bayesian statistics.

<sup>1</sup>Bayes, T. 1763. *An essay towards solving a problem in the doctrine of chances*. The Philosophical Transactions 53:370–418. <https://doi.org/10.1098/rstl.1763.0053>

<sup>2</sup>Laplace, P.-S. 1774. *Mémoire sur la probabilité de causes par les événements*. Mémoires de mathématique et de physique, présentés à l’Académie Royale des sciences par divers savants et lus dans ses assemblées. Paris, Imprimerie Royale, pp. 621–657.



- 1931 [Bruno de Finetti](#) publishes his “[representation theorem](#)”. This shows that the joint distribution of a sequence of exchangeable events (i.e. where the ordering can be permuted) can be represented by a mixture distribution that can be constructed via Bayes’ theorem. (Note that exchangeability is a weaker condition than i.i.d.) This theorem is often used as a justification of Bayesian statistics (along with the so-called Dutch book argument, also by de Finetti).
- 1933 publication of [Andrey Kolmogorov](#)’s book on probability theory.
- 1946 Cox theorem by [Richard T. Cox \(1898–1991\)](#): the aim to generalise classical logic from TRUE/FALSE statements to continuous measures of uncertainty inevitably leads to probability theory and Bayesian learning! This justification of Bayesian statistics was later popularised by [Edwin T. Jaynes \(1922–1998\)](#) in various books (1959, 2003).
- 1955 Stein Paradox - [Charles M. Stein \(1920–2016\)](#) publishes a paper on the Stein estimator — an estimator of the mean that dominates the ML estimator (i.e. the sample average). The Stein estimator is better in terms of MSE than the ML estimator, which was very puzzling at that time but it is easy to understand from a Bayesian perspective.
- Only from the 1950s the use of the term “Bayesian statistics” became prevalent — see Fienberg (2006)<sup>3</sup>

Due to advances in personal computing from 1970 onwards Bayesian learning has become more pervasive!

- Computers allow to do the complex (numerical) calculations needed in Bayesian statistics .
- Metropolis-Hastings algorithm published in 1970 (which allows to sample from a posterior distribution without explicitly computing the marginal likelihood).
- Development of regularised estimation techniques such as penalised likelihood in regression (e.g. ridge regression 1970).
- penalised likelihood via KL divergence for model selection (Akaike 1973).
- A lot of work on interpreting Stein estimators as empirical Bayes estimators (Efron and Morris 1975)

<sup>3</sup>Fienberg, S. E. 2006. *When did Bayesian inference become “Bayesian”?* Bayesian Analysis 1:1–40. <https://doi.org/10.1214/06-BA101>

- regularisation originally was only meant to make singular systems/matrices invertible, but then it turned out regularisation has also a Bayesian interpretation.
- Reference priors (Bernardo 1979) proposed as default priors for models with multiple parameters.
- The EM algorithm (published in 1977) uses Bayes theorem for imputing the distribution of the latent variables.

Another boost was in the 1990/2000s when in science (e.g. genomics) many complex and high-dimensional data set were becoming the norm, not the exception.

- Classical statistical methods cannot be used in this setting (overfitting!) so new methods were developed for high-dimensional data analysis, many with a direct link to Bayesian statistics
- 1996 lasso (L1 regularised) regression invented by [Robert Tibshirani](#).
- Machine learning methods for nonparametric and extremely highly parametric models (neural network) require either explicit or implicit regularisation.
- Many Bayesians in this field, many using [variational Bayes techniques](#) which may be viewed as generalisation of the EM algorithm and are also linked to methods used in statistical physics.



# 16 Bayesian learning in practice

In this chapter we discuss how three basic problems, namely how to estimate a proportion, the mean and the variance in a Bayesian framework.

## 16.1 Estimating a proportion using the beta-binomial model

### Binomial likelihood

In order to apply Bayes' theorem we first need to find a suitable likelihood. We use the Bernoulli model as in Example 8.1:

Repeated Bernoulli experiment (binomial model):

Bernoulli data-generating process:

$$x \sim \text{Ber}(\theta)$$

- $x \in \{0, 1\}$  (e.g. "success" vs. "failure")
- The "success" is indicated by outcome  $x = 1$  and the "failure" by  $x = 0$
- Parameter:  $\theta$  is the probability of "success"
- probability mass function (PMF):  $\Pr(x = 1) = \theta$ ,  $\Pr(x = 0) = 1 - \theta$
- Mean:  $E(x) = \theta$
- Variance  $\text{Var}(x) = \theta(1 - \theta)$

Binomial model  $\text{Bin}(n, \theta)$  (sum of  $n$  Bernoulli experiments):

- $y \in \{0, 1, \dots, n\} = \sum_{i=1}^n x_i$
- Mean:  $E(y) = n\theta$
- Variance:  $\text{Var}(y) = n\theta(1 - \theta)$
- Mean of standardised  $y$ :  $E(y/n) = \theta$
- Variance of standardised  $y$ :  $\text{Var}(y/n) = \frac{\theta(1-\theta)}{n}$

Maximum likelihood estimate of  $\theta$ :

### 16.1 Estimating a proportion using the beta-binomial model

- We conduct  $n$  Bernoulli trials and observe data  $D = \{x_1, \dots, x_n\}$  with average  $\bar{x}$  and  $n_1$  successes and  $n_2 = n - n_1$  failures.
- Binomial likelihood:

$$L(\theta|D) = \binom{n}{n_1} \theta^{n_1} (1 - \theta)^{n_2}$$

Note that the binomial coefficient arises as the ordering of the  $x_i$  is irrelevant but it may be discarded as it does not contain the parameter  $\theta$ .

- From Example 8.1 we know that the maximum likelihood estimate of the proportion  $\theta$  is the frequency

$$\hat{\theta}_{ML} = \frac{n_1}{n} = \bar{x}$$

Thus, the MLE  $\hat{\theta}_{ML}$  can be expressed as an average (of the individual data points). This seemingly trivial fact is important for Bayesian estimation of  $\theta$  using linear shrinkage, as will become evident below.

### Beta distribution

A random variable with support  $x \in [0, 1]$  is often described by a **beta distribution**

$$x \sim \text{Beta}(\alpha_1, \alpha_2)$$

with parameters  $\alpha_1 \geq 0$  and  $\alpha_2 \geq 0$ .

For  $\alpha_1 = \alpha_2 = 1$  the beta distribution reduces to the uniform distribution  $\text{Unif}(0, 1)$ .

The density of the beta distribution is

$$p(x|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} x^{\alpha_1-1} (1-x)^{\alpha_2-1}$$

where  $B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$  is the beta function.

The mean is

$$E(x) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

and the variance is

$$\text{Var}(x) = \frac{\mu(1-\mu)}{\alpha_1 + \alpha_2 + 1}$$



Instead of the parameters  $\alpha_1$  and  $\alpha_2$  it is often convenient to consider the mean parametrisation

$$x \sim \text{Beta}(\alpha_1 = m\mu, \alpha_2 = m(1 - \mu))$$

with  $\mu = \alpha_1/m$  as mean parameter and  $m = \alpha_1 + \alpha_2 \geq 0$  as concentration parameter.

In mean parametrisation the mean is

$$E(x) = \mu$$

and the variance

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

See the [Probability and Distribution Refresher notes](#) for further properties of the beta distribution and related distributions (such as the Dirichlet distribution as its multivariate version).

## Beta prior distribution

In Bayesian statistics the model is comprised the data-generating process (as for maximum likelihood) but we also require the specification of a prior distribution over the parameters. Therefore, we need to **explicitly model our prior uncertainty about  $\theta$** .

The parameter  $\theta$  has support  $[0, 1]$ . Therefore it is natural to use a **beta distribution**  $\text{Beta}(\alpha_1, \alpha_2)$  as prior for  $\theta$ . We will see below that the beta distribution is a natural choice as a prior in conjunction with a binomial likelihood.

The parameters of a prior (here  $\alpha_1 \geq 0$  and  $\alpha_2 \geq 0$ ) are also known as the **hyperparameters** of the model to distinguish them from the parameters of the likelihood function (here  $\theta$ ).

We write for the prior distribution

$$\theta \sim \text{Beta}(\alpha_1, \alpha_2)$$

with density

$$p(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \theta^{\alpha_1-1} (1 - \theta)^{\alpha_2-1}$$

and prior mean

$$E(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

It is important that this does not actually mean that  $\theta$  is random. It only means that we model the uncertainty about  $\theta$  using a beta-distributed random variable. The flexibility of the beta distribution allows to accommodate a large variety of possible scenarios for our prior knowledge using just two parameters.

Note the mean and variance of the beta prior and the mean and variance of the standardised binomial variable  $y/n$  have the same form. This is further indication that the binomial likelihood and the beta prior are well matched — see the discussion below about “conjugate priors”.

## Computing the posterior distribution

After observing data  $D = \{x_1, \dots, x_n\}$  with  $n_1$  “successes” and  $n_2 = n - n_1$  “failures” we can compute the posterior density over  $\theta$  using Bayes’ theorem:

$$p(\theta|D) = \frac{p(\theta)L(\theta|D)}{p(D)}$$

Applying Bayes’ theorem results in the posterior distribution:

$$\theta|D \sim \text{Beta}(\alpha_1 + n_1, \alpha_2 + n_2)$$

with density

$$p(\theta|D) = \frac{1}{B(\alpha_1 + n_1, \alpha_2 + n_2)} \theta^{\alpha_1+n_1-1} (1 - \theta)^{\alpha_2+n_2-1}$$

and posterior mean

$$E(\theta|D) = \frac{\alpha_1 + n_1}{\alpha_1 + \alpha_2 + n_1 + n_2}$$

For a proof see Worksheet B1.

Thus, when updating from the prior to the posterior the parameters of the beta distribution are updated in the following simple fashion:

- $\alpha_1 \longrightarrow \alpha_1 + n_1$
- $\alpha_2 \longrightarrow \alpha_2 + n_2$



## Update from prior to posterior in terms of mean parametrisation

It is instructive to consider the update in terms of mean and concentration parameters.

- The prior concentration parameter  $m$  is set to  $k_0 = \alpha_1 + \alpha_2$
- The prior mean parameter  $\mu$  is set to  $\mu_0 = \alpha_1/k_0$ .

The prior mean is therefore

$$E(\theta) = \mu_0$$

and the prior variance

$$\text{Var}(\theta) = \frac{\mu_0(1 - \mu_0)}{k_0 + 1}$$

The posterior mean and concentration parameters are then as follows:

- The concentration parameter  $m$  is updated to  $k_1 = k_0 + n$
- The mean parameter  $\mu$  is updated to

$$\mu_1 = \frac{\alpha_1 + n_1}{k_1}$$

This can be written as

$$\begin{aligned} \mu_1 &= \frac{\alpha_1}{k_1} + \frac{n_1}{k_1} \\ &= \frac{k_0}{k_1} \frac{\alpha_1}{k_0} + \frac{n}{k_1} \frac{n_1}{n} \\ &= \lambda \mu_0 + (1 - \lambda) \hat{\theta}_{ML} \end{aligned}$$

with  $\lambda = \frac{k_0}{k_1}$ . Hence,  $\mu_1$  is a convex combination of the prior mean and the MLE.

Therefore, the posterior mean is

$$E(\theta|D) = \mu_1$$

and the posterior variance is

$$\text{Var}(\theta|D) = \frac{\mu_1(1 - \mu_1)}{k_1 + 1}$$

Thus the prior to posterior update in mean parametrisation is:

- $k_0 \longrightarrow k_1 = k_0 + n$
- $\mu_0 \longrightarrow \mu_1 = \lambda \mu_0 + (1 - \lambda) \hat{\theta}_{ML}$  with  $\lambda = k_0/k_1$

## 16.2 Properties of Bayesian learning

The beta-binomial model, even though it is one of the simplest possible models, already allows to observe a number of important features and properties of Bayesian learning. Many of these apply also to other models as we will see later.

### Prior acting as pseudodata

In the expression for the mean and variance you can see that the concentration parameter  $k_0 = \alpha_1 + \alpha_2$  behaves like an implicit sample size connected with the prior information about  $\theta$ .

Specifically,  $\alpha_1$  and  $\alpha_2$  act as **pseudocounts** that influence both the posterior mean and the posterior variance, exactly in the same way as conventional observations.

For example, the larger  $k_0$  (and thus the larger  $\alpha_1$  and  $\alpha_2$ ) the smaller is the posterior variance, with variance decreasing proportional to the inverse of  $k_0$ . If the prior is highly concentrated, i.e. if it has low variance and large precision (=inverse variance) then the implicit data size  $k_0$  is large. Conversely, if the prior has large variance, then the prior is vague and the implicit data size  $k_0$  is small.

Hence, a prior has the same effect as if one would add data — but without actually adding data! This is precisely this why a prior acts as a regulariser and prevents overfitting, because it increases the effective sample size.

Another interpretation is that a prior summarises data that may have been available previously as observations.

### Linear shrinkage of mean

In the beta-binomial model the **posterior mean is a convex combination (i.e. the weighted average) of the ML estimate and the prior mean** as can be seen from the update formula

$$\mu_1 = \lambda \mu_0 + (1 - \lambda) \hat{\theta}_{ML}$$

with weight  $\lambda \in [0, 1]$

$$\lambda = \frac{k_0}{k_1}.$$



Thus, the posterior mean  $\mu_1$  is a linearly adjusted  $\hat{\theta}_{ML}$ . The factor  $\lambda$  is called the **shrinkage intensity** — note that this is the ratio of the “prior sample size” ( $k_0$ ) and the “effective total sample size” ( $k_1$ ).

1. This adjustment of the MLE is called *shrinkage*, because the  $\hat{\theta}_{ML}$  is “shrunk” towards the prior mean  $\mu_0$  (which is often called the “target”, and sometimes the target is zero, and then the terminology “shrinking” makes most sense).
2. If the shrinkage intensity is zero ( $\lambda = 0$ ) then the ML point estimator is recovered. This happens when  $\alpha_1 = 0$  and  $\alpha_2 = 0$  or for  $n \rightarrow \infty$ .

Remark: using maximum likelihood to estimate  $\theta$  (for moderate or small  $n$ ) is the same as Bayesian posterior mean estimation using the beta-binomial model with prior  $\alpha_1 = 0$  and  $\alpha_2 = 0$ . This prior is extremely “u-shaped” and the implicit prior for the ML estimation. Would you use such a prior intentionally?

3. If the shrinkage intensity is large ( $\lambda \rightarrow 1$ ) then the posterior mean corresponds to the prior. This happens if  $n = 0$  or if  $k_0$  is very large (implying that the prior is sharply concentrated around the prior mean).
4. Since the ML estimate  $\hat{\theta}_{ML}$  is unbiased the Bayesian point estimate is biased (for finite  $n$ !). And the bias is induced by the prior mean deviating from the true mean. This is also true more generally as Bayesian learning typically produces biased estimators (but asymptotically they will be unbiased like in ML).
5. The fact that the posterior mean is a linear combination of the MLE and the prior mean is not a coincidence. In fact, this is true for all distributions that are exponential families, see e.g. Diaconis and Ylvisaker (1979)<sup>1</sup>. Crucially, exponential families can always be parametrised such that the corresponding MLEs are expressed as averages of functions of the data (more technically: for an exponential family the MLE of the expectation parameters is given as the average of the canonical statistics). In conjunction with a particular type of prior (conjugate priors, always existing for exponential families, see below) this allows to write the update from the prior to posterior mean as a linear adjustment of the MLE.
6. Furthermore, it is possible (and indeed quite useful for computational reasons!) to formulate Bayes learning assuming only first and second moments (i.e. without full distributions) and in terms

<sup>1</sup>Diaconis, P., and D Ylvisaker. 1979. *Conjugate Priors for Exponential Families*. Ann. Statist. 7:269–281. <https://doi.org/10.1214/aos/1176344611>

of linear shrinkage, see e.g. Hartigan (1969)<sup>2</sup>. The resulting theory is called “Bayes linear statistics” (Goldstein and Wooff, 2007)<sup>3</sup>.

## Conjugacy of prior and posterior distribution

In the beta-binomial model for estimating the proportion  $\theta$  the choice of the **beta distribution as prior distribution** along with the binomial likelihood resulted in having the **beta distribution as posterior distribution** as well.

If the prior and posterior belong to the same distributional family the prior is called a **conjugate prior**. This will be the case if the prior has the same functional form as the likelihood. Therefore one also says that the prior is conjugate for the likelihood.

It can be shown that conjugate priors exist for all likelihood functions that are based on data-generating models that are exponential families.

In the beta-binomial model the likelihood is based on the binomial distribution and has the following form (only terms depending on the parameter  $\theta$  are shown):

$$\theta^n (1 - \theta)^{n_2}$$

The form of the beta prior is (again, only showing terms depending on  $\theta$ ):

$$\theta^{\alpha_1-1} (1 - \theta)^{\alpha_2-1}$$

Since the posterior is proportional to the product of prior and likelihood the posterior will have exactly the same form as the prior:

$$\theta^{\alpha_1+n_1-1} (1 - \theta)^{\alpha_2+n_2-1}$$

Choosing the prior distribution from a family conjugate for the likelihood greatly simplifies Bayesian analysis since the Bayes formula can then be written in form of an update formula for the parameters of the beta distribution:

$$\begin{aligned}\alpha_1 &\rightarrow \alpha_1 + n_1 = \alpha_1 + n \hat{\theta}_{ML} \\ \alpha_2 &\rightarrow \alpha_2 + n_2 = \alpha_2 + n(1 - \hat{\theta}_{ML})\end{aligned}$$

Thus, conjugate prior distributions are very convenient choices. However, in their application it must be ensured that the prior distribution is flexible

<sup>2</sup>Hartigan, J. A. 1969. *Linear Bayesian methods*. J. Roy. Statist. Soc. B 31:446–454 <https://doi.org/10.1111/j.2517-6161.1969.tb00804.x>

<sup>3</sup>Goldstein, M., and D. Wooff. 2007. *Bayes Linear Statistics: Theory and Methods*. Wiley. <https://doi.org/10.1002/9780470065662>



enough to encapsulate all prior information that may be available. In cases where this is not the case alternative priors should be used (and most likely this will then require to compute the posterior distribution numerically rather than analytically).

### Large sample limits of mean and variance

If  $n$  is large and  $n \gg \alpha, \beta$  then  $\lambda \rightarrow 0$  and hence the posterior mean and variance become asymptotically

$$E(\theta|D) \stackrel{a}{=} \frac{n_1}{n} = \hat{\theta}_{ML}$$

and

$$\text{Var}(\theta|D) \stackrel{a}{=} \frac{\hat{\theta}_{ML}(1 - \hat{\theta}_{ML})}{n}$$

Thus, if the sample size is large then the Bayes' estimator turns into the ML estimator! Specifically, the posterior mean becomes the ML point estimate, and the posterior variance is equal to the asymptotic variance computed via the observed Fisher information.

Thus, for large  $n$  the data dominate and any details about the prior (such as the settings of the hyperparameters  $\alpha_1$  and  $\alpha_2$ ) become irrelevant!

### Asymptotic normality of the posterior distribution

Also known as **Bayesian Central Limit Theorem (CLT)**.

Under some regularity conditions (such as regular likelihood and positive prior probability for all parameter values, finite number of parameters, etc.) for large sample size the Bayesian posterior distribution converges to a normal distribution centred around the MLE and with the variance of the MLE:

$$\text{for large } n: p(\theta|D) \rightarrow N(\hat{\theta}_{ML}, \text{Var}(\hat{\theta}_{ML}))$$

So not only are the posterior mean and variance converging to the MLE and the variance of the MLE for large sample size, but also the posterior distribution itself converges to the sampling distribution!

This holds generally in many regular cases, not just in the simple case above.

The Bayesian CLT is generally known as the **Bernstein-von Mises theorem** (who discovered it at around 1920–30), but special cases were already known by Laplace.

In the Worksheet B1 the asymptotic convergence of the posterior distribution to a normal distribution is demonstrated graphically.

### Posterior variance for finite $n$

From the Bayesian posterior we can obtain a Bayesian point estimate for the proportion  $\theta$  by computing the posterior mean

$$E(\theta|D) = \frac{\alpha_1 + n_1}{k_1} = \hat{\theta}_{\text{Bayes}}$$

along with the posterior variance

$$\text{Var}(\theta|D) = \frac{\hat{\theta}_{\text{Bayes}}(1 - \hat{\theta}_{\text{Bayes}})}{k_1 + 1}$$

Asymptotically for large  $n$  the posterior mean becomes the maximum likelihood estimate (MLE), and the posterior variance becomes the asymptotic variance of the MLE. Thus, for large  $n$  the Bayesian point estimate will be indistinguishable from the MLE and shares its favourable properties.

In addition, for finite sample size the posterior variance will typically be *smaller* than both the asymptotic posterior variance (for large  $n$ ) and the prior variance, showing that combining the information available in the prior and in the data leads to a more efficient estimate.

## 16.3 Estimating the mean using the normal-normal model

### Normal likelihood

As in Example 8.2 where we estimated the mean parameter by maximum likelihood we assume as data-generating model the normal distribution with unknown mean  $\mu$  and known variance  $\sigma^2$ :

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



We observe  $n$  samples  $D = \{x_1, \dots, x_n\}$ . This yields using maximum likelihood the estimate  $\hat{\mu}_{ML} = \bar{x}$ .

We note that the MLE  $\hat{\mu}_{ML}$  is expressed as an average of the data points, which is what enables the linear shrinkage seen below.

### Normal prior distribution

The **normal distribution is the conjugate distribution for the mean parameter of a normal likelihood**, so if we use a normal prior then posterior for  $\mu$  is normal as well.

To model the uncertainty about  $\mu$  we use the normal distribution in the form  $N(\mu, \sigma^2/k)$  with a mean parameter  $\mu$  and a concentration parameter  $k > 0$  (remember that  $\sigma^2$  is given and is also used in the likelihood).

Specifically, we use as normal **prior distribution** for the mean

$$\mu \sim N\left(\mu_0, \frac{\sigma^2}{k_0}\right)$$

- The prior concentration parameter is set to  $k_0$
- The prior mean parameter is set to  $\mu_0$

Hence the prior mean is

$$E(\mu) = \mu_0$$

and the prior variance

$$\text{Var}(\mu) = \frac{\sigma^2}{k_0}$$

where the concentration parameter  $k_0$  corresponds the implied sample size of the prior. Note that  $k_0$  does not need to be an integer value.

### Normal posterior distribution

After observing data  $D$  the **posterior distribution** is also normal with updated parameters  $\mu = \mu_1$  and  $k_1$

$$\mu|D \sim N\left(\mu_1, \frac{\sigma^2}{k_1}\right)$$

- The posterior concentration parameter is updated to  $k_1 = k_0 + n$

- The posterior mean parameter is updated to

$$\mu_1 = \lambda\mu_0 + (1 - \lambda)\hat{\mu}_{ML}$$

with  $\lambda = \frac{k_0}{k_1}$ . This can be seen as linear shrinkage of  $\hat{\mu}_{ML}$  towards the prior mean  $\mu_0$ .

(For a proof see Worksheet B2.)

The posterior mean is

$$E(\mu|D) = \mu_1$$

and the posterior variance is

$$\text{Var}(\mu|D) = \frac{\sigma^2}{k_1}$$

### Large sample asymptotics

For  $n$  large and  $n \gg k_0$  the shrinkage intensity  $\lambda \rightarrow 0$  and  $k_1 \rightarrow n$ . As a result

$$E(\mu|D) \stackrel{a}{=} \hat{\mu}_{ML}$$

$$\text{Var}(\mu|D) \stackrel{a}{=} \frac{\sigma^2}{n}$$

i.e. we recover the MLE and its asymptotic variance!

Note that for finite  $n$  the posterior variance  $\frac{\sigma^2}{n+k_0}$  is smaller than both the asymptotic variance  $\frac{\sigma^2}{n}$  of the MLE and the prior variance  $\frac{\sigma^2}{k_0}$ .

## 16.4 Estimating the variance using the Inverse-Wishart-normal model

### Normal likelihood

As data-generating model we use normal distribution

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



with unknown variance  $\sigma^2$  and known mean  $\mu$ . This yields as maximum likelihood estimate for the variance

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Note that, again, the MLE is an average (of a quadratic function of the individual data points). This enables linear shrinkage of the MLE as seen below.

### Inverse-Wishart prior distribution

To model the uncertainty about the variance we use the inverse-gamma distribution, also known as the univariate inverse-Wishart distribution (refer to the [Probability and Distribution Refresher notes](#) for details of this distribution). The IG resp. univariate IW distribution are identical apart from parametrisation. They are conjugate for the variance parameter in the normal likelihood, hence both the prior and the posterior distribution are also inverse-Gamma resp. inverse-Wishart.

In the following we use the Wishart parametrisation, hence we call this an inverse-Wishart prior, and the whole model inverse-Wishart-normal model. However, this model is also often called inverse-gamma-normal model if the inverse-gamma distribution and corresponding parametrisation is used as prior.

Specifically, as prior distribution for  $\sigma^2$  we use the IW distribution in mean parametrisation:

$$\sigma^2 \sim \text{IWis}(\psi = \kappa_0 \sigma_0^2, k = \kappa_0 + 2)$$

- The prior concentration parameter is  $\kappa_0$
- The prior mean parameter is  $\sigma_0^2$

The corresponding prior mean is

$$E(\sigma^2) = \sigma_0^2$$

and the prior variance is

$$\text{Var}(\sigma^2) = \frac{2\sigma_0^4}{\kappa_0 - 2}$$

(note that  $\kappa_0 > 2$  is required for the variance to exist)

### Inverse-Wishart posterior distribution

After observing  $D = \{x_1, \dots, x_n\}$  the posterior distribution is also IW with updated parameters:

$$\sigma^2 | D \sim \text{IWis}(\psi = \kappa_1 \sigma_1^2, k = \kappa_1 + 2)$$

- The posterior concentration parameter is updated to  $\kappa_1 = \kappa_0 + n$
- The posterior mean parameter update follows the standard linear shrinkage rule:

$$\sigma_1^2 = \lambda \sigma_0^2 + (1 - \lambda) \hat{\sigma}_{ML}^2$$

$$\text{with } \lambda = \frac{\kappa_0}{\kappa_1}.$$

The posterior mean is

$$E(\sigma^2 | D) = \sigma_1^2$$

and the posterior variance

$$\text{Var}(\sigma^2 | D) = \frac{2\sigma_1^4}{\kappa_1 - 2}$$

### Large sample asymptotics

For large sample size  $n$  with  $n \gg \kappa_0$  the shrinkage intensity vanishes ( $\lambda \rightarrow 0$ ) and therefore  $\sigma_1^2 \rightarrow \hat{\sigma}_{ML}^2$ . We also find that  $\kappa_1 - 2 \rightarrow n$ .

This results in the asymptotic posterior mean

$$E(\sigma^2 | D) \stackrel{a}{=} \hat{\sigma}_{ML}^2$$

and the asymptotic posterior variance

$$\text{Var}(\sigma^2 | D) \stackrel{a}{=} \frac{2(\hat{\sigma}_{ML}^2)^2}{n}$$

Thus we recover the MLE of  $\sigma^2$  and its asymptotic variance.

### Other equivalent update rules

Above the update rule from prior to posterior IW distribution is stated for the mean parametrisation:

- $\kappa_0 \rightarrow \kappa_1 = \kappa_0 + n$



- $\sigma_0^2 \rightarrow \sigma_1^2 = \lambda \sigma_0^2 + (1 - \lambda) \widehat{\sigma}_{ML}^2$  with  $\lambda = \frac{\kappa_0}{\kappa_1}$

This has the advantage that the mean of the IW distribution is updated directly, and that the prior and posterior variance is also straightforward to compute.

The same update rule can also be expressed in terms of the other parametrisations. In terms of the conventional parameters  $\alpha$  and  $\beta$  of the IG distribution the update rule is

- $\alpha_0 \rightarrow \alpha_1 = \alpha_0 + \frac{n}{2}$
- $\beta_0 \rightarrow \beta_1 = \beta_0 + \frac{n}{2} \widehat{\sigma}_{ML}^2 = \beta_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2$

For the parameters  $\psi$  and  $k$  of the univariate inverse Wishart distribution the update rule is

- $k_0 \rightarrow k_1 = k_0 + n$
- $\psi_0 \rightarrow \psi_1 = \psi_0 + n \widehat{\sigma}_{ML}^2 = \psi_0 + \sum_{i=1}^n (x_i - \mu)^2$

For the parameters  $\tau^2$  and  $\nu$  of the scaled inverse chi-squared distribution the update rule is

- $\nu_0 \rightarrow \nu_1 = \nu_0 + n$
- $\tau_0^2 \rightarrow \tau_1^2 = \frac{\nu_0}{\nu_1} \tau_0^2 + \frac{n}{\nu_1} \widehat{\sigma}_{ML}^2$

(See Worksheet B2 for proof of the equivalence of all the above update rules.)

## 16.5 Estimating the precision using the Wishart-normal model

### MLE of the precision

Instead of estimating the variance  $\sigma^2$  we may wish to estimate the precision  $w = 1/\sigma^2$ , i.e. the inverse of the variance.

As above the data-generating model is a normal distribution

$$x \sim N(\mu, 1/w)$$

with unknown precision  $w$  and known mean  $\mu$ . This yields as maximum likelihood estimate (easily derived thanks to the invariance principle!)

$$\widehat{w}_{ML} = \frac{1}{\widehat{\sigma}_{ML}^2} = \frac{1}{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

Crucially, the MLE of the precision  $w$  is not an average itself (instead, it is a function of an average). Consequently, as shown below, the posterior mean of  $w$  cannot be expressed as a linear adjustment of the MLE.

### Wishart prior

For modelling the variance we have used an inverse-Wishart resp. inverse-Gamma distribution for the prior and posterior distributions. Thus, in order to model the precision we therefore now use a univariate Wishart resp. gamma distribution. Recall that these two distributions are identical apart from their parametrisation.

Specifically, we use the Wishart distribution in the mean parametrisation (see [Probability and Distribution Refresher notes](#) for details):

$$w \sim \text{Wis} \left( s^2 = \frac{w_0}{k_0}, k = k_0 \right)$$

- The prior concentration parameter is set to  $k_0$
- The prior mean parameter is set to  $w_0$

The corresponding prior mean of the precision is

$$E(w) = w_0$$

and the prior variance is

$$\text{Var}(w) = \frac{2w_0^2}{k_0}$$

### Wishart posterior

After observing  $D = \{x_1, \dots, x_n\}$  the posterior distribution is also Wishart with updated parameters:

$$w|D \sim \text{Wis} \left( s^2 = \frac{w_1}{k_1}, k = k_1 \right)$$

- The posterior concentration parameter is updated to  $k_1 = k_0 + n$



- The posterior mean parameter update follows the rule:

$$\frac{1}{w_1} = \lambda \frac{1}{w_0} + (1 - \lambda) \frac{1}{\hat{w}_{ML}}$$

with  $\lambda = \frac{k_0}{k_1}$ . Crucially, the linear update is applied to the inverse of the precision but **not** to the precision itself. This is because the MLE of the precision parameter cannot be expressed as an average.

The posterior mean is

$$E(w|D) = w_1$$

and the posterior variance

$$\text{Var}(w|D) = \frac{2w_1^2}{k_1}$$

Equivalent update rules directly for the parameters of a corresponding gamma distribution  $\text{Gam}(\alpha = k/2, \beta = 1/(2s^2))$  with mean  $\alpha/\beta = ks^2$  are given as follows. The shape parameter  $\alpha$  is updated according to

$$\alpha_1 = \alpha_0 + \frac{n}{2}$$

and the rate parameter  $\beta$  according to

$$\beta_1 = \beta_0 + \frac{n}{2} \hat{\sigma}_{ML}^2$$

This is the form you will find most often in textbooks. While elegant in terms of parameter updates for the gamma prior it obscures the fact the mean of the precision is not linearly updated.

## 17 Bayesian model comparison

### 17.1 Marginal likelihood as model likelihood

#### Simple and composite models

In the introduction of the Bayesian learning we already encountered the marginal likelihood  $p(D|M)$  of a model class  $M$  in the denominator of Bayes' rule:

$$p(\theta|D, M) = \frac{p(\theta|M)p(D|\theta, M)}{p(D|M)}$$

Computing this marginal likelihood is different for simple and composite models.

A model is called “simple” if it directly corresponds to a specific distribution, say, a normal distribution with fixed mean and variance, or a binomial distribution with a given probability for the two classes. Thus, a simple model is a point in the model space described by the parameters of a distribution family (e.g.  $\mu$  and  $\sigma^2$  for the normal family  $N(\mu, \sigma^2)$ ). For a simple model  $M$  the density  $p(D|M)$  corresponds to standard likelihood of  $M$  and there are no free parameters.

On the other hand, a model is “composite” if it is composed of simple models. This can be a finite set, or it can be comprised of infinite number of simple models. Thus a composite model represent a model class. For example, a normal distribution with a given mean but unspecified variance, or a binomial model with unspecified class probability, is a composite model.

If  $M$  is a composite model, with the underlying simple models indexed by a parameter  $\theta$ , the likelihood of the model is obtained by marginalisation over  $\theta$ :

$$\begin{aligned} p(D|M) &= \int_{\theta} p(D|\theta, M)p(\theta|M)d\theta \\ &= \int_{\theta} p(D, \theta|M)d\theta \end{aligned}$$



i.e. we *integrate* over all parameter values  $\theta$ .

If the distribution over the parameter  $\theta$  of a model is strongly concentrated around a specific value  $\theta_0$  then the composite model degenerates to a simple point model, and the marginal likelihood becomes the likelihood of the parameter  $\theta_0$  under that model.

**Example 17.1.** Beta-binomial distribution:

Assume that likelihood is binomial with mean parameter  $\theta$ . If  $\theta$  follows a Beta distribution then the marginal likelihood with  $\theta$  integrated out is the [beta-binomial distribution](#) (see also Worksheet B2). This is an example of a [compound probability distribution](#).

### Log-marginal likelihood as penalised maximum log-likelihood

By rearranging Bayes' rule we see that

$$\log p(D|M) = \log p(D|\theta, M) - \log \frac{p(\theta|D, M)}{p(\theta|M)}$$

The above is valid for all  $\theta$ .

Assuming concentration of the posterior around the MLE  $\hat{\theta}_{ML}$  we will have  $p(\hat{\theta}_{ML}|D, M) > p(\hat{\theta}_{ML}|M)$  and thus

$$\log p(D|M) = \underbrace{\log p(D|\hat{\theta}_{ML}, M)}_{\text{maximum log-likelihood}} - \underbrace{\log \frac{p(\hat{\theta}_{ML}|D, M)}{p(\hat{\theta}_{ML}|M)}}_{\text{penalty} > 0}$$

Therefore, the log-marginal likelihood is essentially a penalised version of the maximum log-likelihood, and the penalty depends on the concentration of the posterior around the MLE.

### Model complexity and Occams razor

Intriguingly, the penalty implicit in the log-marginal likelihood is linked to the complexity of the model, in particular to the number of parameters of  $M$ . We will see this directly in the Schwarz approximation of the log-marginal likelihood discussed below.

Thus, the averaging over  $\theta$  in the marginal likelihood has the effect of automatically penalising complex models. Therefore, when comparing models using the marginal likelihood a complex model may be ranked below simpler models. In contrast, when selecting a model by comparing maximum likelihood directly the model with the highest number of parameters always wins over simpler models. Hence, the penalisation implicit in the marginal likelihood prevents overfitting that occurs with maximum likelihood.

The principle of preferring a less complex model is called **Occam's razor** or the **law of parsimony**.

When choosing models a simpler model is often preferable over a more complex model, because the simpler model is typically better suited to both explaining the currently observed data as well as future data, whereas a complex model will typically only excel in fitting the current data but will perform poorly in prediction.

## 17.2 The Bayes factor for comparing two models

### Definition of the Bayes factor

The **Bayes factor** is the ratio of the likelihoods of the two models:

$$B_{12} = \frac{p(D|M_1)}{p(D|M_2)}$$

The **log-Bayes factor**  $\log B_{12}$  is also called the **weight of evidence** for  $M_1$  over  $M_2$ .

### Bayes theorem in terms of the Bayes factor

We would like to compare two models  $M_1$  and  $M_2$ . Before seeing data  $D$  we can check their **Prior odds** (= ratio of prior probabilities of the models  $M_1$  and  $M_2$ ):

$$\frac{\Pr(M_1)}{\Pr(M_2)}$$



After seeing data  $D = \{x_1, \dots, x_n\}$  we arrive at the **Posterior odds** (= ratio of posterior probabilities):

$$\frac{\Pr(M_1|D)}{\Pr(M_2|D)}$$

Using Bayes Theorem  $\Pr(M_i|D) = \Pr(M_i) \frac{p(D|M_i)}{p(D)}$  we can rewrite the posterior odds as

$$\underbrace{\frac{\Pr(M_1|D)}{\Pr(M_2|D)}}_{\text{posterior odds}} = \underbrace{\frac{p(D|M_1)}{p(D|M_2)}}_{\text{Bayes factor } B_{12}} \underbrace{\frac{\Pr(M_1)}{\Pr(M_2)}}_{\text{prior odds}}$$

The **Bayes factor** is the multiplicative factor that updates the prior odds to the posterior odds.

On the log scale we see that

$$\log\text{-posterior odds} = \text{weight of evidence} + \log\text{-prior odds}$$

### Interpretive scales for the Bayes factor

Table 17.1: Scale for the Bayes factor according to Jeffreys (1961).

$B_{12}$	$\log B_{12}$	evidence in favour of $M_1$ versus $M_2$
$> 100$	$> 4.6$	decisive
10 to 100	2.3 to 4.6	strong
3.2 to 10	1.16 to 2.3	substantial
1 to 3.2	0 to 1.16	not worth more than a bare mention

Following Harold Jeffreys (1961) <sup>1</sup> one may interpret the strength of the Bayes factor as listed in Table 17.1.

<sup>1</sup>Jeffreys, H. *Theory of Probability*. 3rd ed. Oxford University Press.

Table 17.2: Scale for the Bayes factor according to Kass and Raftery (1995).

$B_{12}$	$\log B_{12}$	evidence in favour of $M_1$ versus $M_2$
$> 150$	$> 5$	very strong
20 to 150	3 to 5	strong
3 to 20	1 to 3	positive
1 to 3	0 to 1	not worth more than a bare mention

More recently, Kass and Raftery (1995) <sup>2</sup> proposed to use the a slightly modified scale (Table 17.2).

### Bayes factor versus likelihood ratio

If both  $M_1$  and  $M_2$  are simple models then the **Bayes factor is identical to the likelihood ratio** of the two models.

However, if one of the two models is composite then the Bayes factor and the generalised likelihood ratio differ: In the Bayes factor the representative of a composite model is the **model average** of the simple models indexed by  $\theta$ , with weights taken from the prior distribution over the simple models contained in  $M$ . In contrast, in the generalised likelihood ratio statistic the representative of a composite model is chosen by *maximisation*.

Thus, **for composite models**, the **Bayes factor does not equal the corresponding generalised likelihood ratio statistic**. In fact, the key difference is that the Bayes factor is a penalised version of the likelihood ratio, with the penalty depending on the difference in complexity (number of parameters) of the two models

## 17.3 Approximate computations

The marginal likelihood and the Bayes factor can be difficult to compute in practice. Therefore, a number of approximations have been developed. The most important is the so-called Schwarz (1978) approximation of the log-marginal likelihood. It is used to approximate the log-Bayes factor and also yields the BIC (Bayesian information criterion) which can be interpreted as penalised maximum likelihood.

<sup>2</sup>Kass, R.E., and A.E. Raftery. 1995. *Bayes factors*. JASA 90:773–795. <https://doi.org/10.1080/01621459.1995.10476572>



## Schwarz (1978) approximation of log-marginal likelihood

The logarithm of the marginal likelihood of a model can be approximated following Schwarz (1978)<sup>3</sup> as follows:

$$\log p(D|M) \approx \ell_n^M(\hat{\theta}_{ML}^M) - \frac{1}{2}d_M \log n$$

where  $d_M$  is the dimension of the model  $M$  (number of parameters in  $\theta$  belonging to  $M$ ) and  $n$  is the sample size and  $\hat{\theta}_{ML}^M$  is the MLE. For a simple model  $d_M = 0$  so then there is no approximation as in this case the marginal likelihood equals the likelihood.

The above formula can be obtained by quadratic approximation of the likelihood **assuming large  $n$**  and assuming that the prior is locally uniform around the MLE. The Schwarz (1978) approximation is therefore a special case of a [Laplace approximation](#).

Note that the approximation is the maximum log-likelihood minus a penalty that depends on the model complexity (as measured by dimension  $d$ ), hence this is an example of penalised ML! Also note that the distribution over the parameter  $\theta$  is not required in the approximation.

## Bayesian information criterion (BIC)

The BIC (Bayesian information criterion) of the model  $M$  is the approximated log-marginal likelihood times the factor -2:

$$BIC(M) = -2\ell_n^M(\hat{\theta}_{ML}^M) + d_M \log n$$

Thus, when comparing models one aims to maximise the marginal likelihood or, as approximation, minimise the BIC.

The reason for the factor “-2” is simply to have a quantity that is on the same scale as the Wilks log likelihood ratio. Some people / software packages also use the factor “2”.

<sup>3</sup>Schwarz, G. 1978. *Estimating the dimension of a model*. Ann. Statist. 6:461–464. <https://doi.org/10.1214/aos/1176344136>

## Approximating the weight of evidence (log-Bayes factor) with BIC

Using BIC (twice) the log-Bayes factor can be approximated as

$$\begin{aligned} 2 \log B_{12} &\approx -BIC(M_1) + BIC(M_2) \\ &= 2 \left( \ell_n^{M_1}(\hat{\theta}_{ML}^{M_1}) - \ell_n^{M_2}(\hat{\theta}_{ML}^{M_2}) \right) - \log(n)(d_{M_1} - d_{M_2}) \end{aligned}$$

i.e. it is the penalised log-likelihood ratio of model  $M_1$  vs.  $M_2$ .

## 17.4 Bayesian testing using false discovery rates

We introduce False Discovery Rates (FDR) as a Bayesian method to distinguish a null model from an alternative model. This is closely linked with classical frequentist multiple testing procedures.

### Setup for testing a null model $H_0$ versus an alternative model $H_A$

We consider two models:

$H_0$  : null model, with density  $f_0(x)$  and distribution  $F_0(x)$

$H_A$  : alternative model, with density  $f_A(x)$  and distribution  $F_A(x)$

Aim: given observations  $x_1, \dots, x_n$  we would like to decide for each  $x_i$  whether it belongs to  $H_0$  or  $H_A$ .

This is done by a critical decision threshold  $x_c$ : if  $x_i > x_c$  then  $x_i$  is called “significant” and otherwise called “not significant”.

In classical statistics one of the the most widely used approach to find the decision threshold is by computing  $p$ -values from the  $x_i$  (this uses only the null model but not the alternative model), and then thresholding the  $p$ -values at a certain level (say 5%). If  $n$  is large then often the test is modified by adjusting the  $p$ -values or the threshold (e.g. if Bonferroni correction).

Note that this procedure ignores any information we may have about the alternative model!



## Test errors

### True and false positives and negatives

For any decision threshold  $x_c$  we can distinguish the following errors:

- False positives (FP), “false alarm”, type I error:  $x_i$  belongs to null but is called “significant”
- False negative (FN), “miss”, type II error:  $x_i$  belongs to alternative, but is called “not significant”

In addition we have:

- True positives (TP), “hits”: belongs to alternative and is called “significant”
- True negatives (TN), “correct rejections”: belongs to null and is called “not significant”

### Specificity and Sensitivity

From counts of TP, TN, FN, FP we can derive further quantities:

- False Positive Rate FPR, false alarm rate, type I error probability:  $FPR = \alpha_I = \frac{FP}{TN+FP} = 1 - TNR$
- False Negative Rate FNR, miss rate, type II error probability:  $FNR = \alpha_{II} = \frac{FN}{TP+FN} = 1 - TPR$
- True Negative Rate TNR, **specificity**:  $TNR = \frac{TN}{TN+FP} = 1 - FPR = 1 - \alpha_I$
- True Positive Rate TPR, **sensitivity, power**, recall:  $TPR = \frac{TP}{TP+FN} = 1 - FNR = 1 - \alpha_{II}$
- Accuracy:  $ACC = \frac{TP+TN}{TP+TN+FP+FN}$

Another common way to choose the decision threshold  $x_d$  in classical statistics is to balance sensitivity/power vs. specificity (maximising both power and specificity, or equivalently, minimising both false positive and false negative rates). ROC curves plot TPR/sensitivity vs. FPR = 1-specificity.

## FDR and FNDR

It is possible to link the above with the observed counts of TP, FP, TN, FN:

- False Discovery Rate (FDR):  $FDR = \frac{FP}{FP+TP}$
- False Nondiscovery Rate (FNDR):  $FNDR = \frac{FN}{TN+FN}$
- Positive predictive value (PPV), True Discovery Rate (TDR), precision:  $PPV = \frac{TP}{FP+TP} = 1 - FDR$
- Negative predictive value (NPV):  $NPV = \frac{TN}{TN+FN} = 1 - FNDR$

In order to choose the decision threshold it is natural to balance FDR and FNDR (or PPV and NPV), by minimising both FDR and FNDR or maximising both PPV and NPV.

In machine learning it is common to use “precision-recall plots” that plot precision (=PPV, TDR) vs. recall (=power, sensitivity).

## Bayesian perspective

### Two component mixture model

In the Bayesian perspective the problem of choosing the decision threshold is related to computing the posterior probability

$$\Pr(H_0|x_i),$$

i.e. probability of the null model given the observation  $x_i$ , or equivalently computing

$$\Pr(H_A|x_i) = 1 - \Pr(H_0|x_i)$$

the probability of the alternative model given the observation  $x_i$ .

This is done by assuming a mixture model

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_A(x)$$

where  $\pi_0 = \Pr(H_0)$  is the prior probability of  $H_0$  and.  $\pi_A = 1 - \pi_0 = \Pr(H_A)$  the prior probability of  $H_A$ .

Note that the weights  $\pi_0$  can in fact be estimated from the observations by fitting the mixture distribution to the observations  $x_1, \dots, x_n$  (so it is effectively an empirical Bayes method where the prior is informed by the data).



**Local FDR**

The posterior probability of the null model given a data point is then given by

$$\Pr(H_0|x_i) = \frac{\pi_0 f_0(x_i)}{f(x_i)} = LFDR(x_i)$$

This quantity is also known as the **local FDR** or **local False Discovery Rate**.

In the given one-sided setup the local FDR is large (close to 1) for small  $x$ , and will become close to 0 for large  $x$ . A common decision rule is given by thresholding local false discovery rates: if  $LFDR(x_i) < 0.1$  the  $x_i$  is called significant.

**q-values**

In correspondence to  $p$ -values one can also define tail-area based false discovery rates:

$$Fdr(x_i) = \Pr(H_0|X > x_i) = \frac{\pi_0 F_0(x_i)}{F(x_i)}$$

These are called **q-values**, or simply **False Discovery Rates (FDR)**. Intriguingly, these also have a frequentist interpretation as adjusted  $p$ -values (using a Benjamini-Hochberg adjustment procedure).

**Software**

There are a number of R packages to compute (local) FDR values:

For example:

- [locfdr](#)
- [qvalue](#)
- [fdrtool](#)

and many more.

Using FDR values for screening is especially useful in high-dimensional settings (e.g. when analysing genomic and other high-throughput data).

FDR values have both a Bayesian as well as frequentist interpretation, providing further evidence that good classical statistical methods do have a Bayesian interpretation.



# 18 Choosing priors in Bayesian analysis

## 18.1 Choosing a prior

### Prior as part of the model

It is **essential in a Bayesian analysis to specify your prior uncertainty about the model parameters**. Note that this is simply **part of the modelling process!** Thus in a Bayesian approach the data analyst needs to be more explicit about all modelling assumptions.

Typically, when choosing a suitable prior distribution we consider the overall form (shape and domain) of the distribution as well as its key characteristics such as the mean and variance. As we have learned the precision (inverse variance) of the prior may often be viewed as implied sample size.

For large sample size  $n$  the posterior mean converges to the maximum likelihood estimate (and the posterior distribution to normal distribution centered around the MLE), so for large  $n$  we may ignore specifying a prior.

However, for small  $n$  it is essential that a prior is specified. In non-Bayesian approaches this prior is still there but it is either implicit (maximum likelihood estimation) or specified via a penalty (penalised maximum likelihood estimation).

### Some guidelines

So the question remains what are good ways to choose a prior? Two useful ways are:

1. Use a weakly informative prior. This means that you do have an idea (even if only vague) about the suitable values of the parameter of interest, and you use a corresponding prior (for example with moderate variance) to model the uncertainty. This acknowledges that there are no uninformative priors and but also aims that the prior does not dominate the likelihood (i.e. the data). The result is a weakly regularised estimator. Note that it is often desirable that the prior adds information (if only a little) so that it can act as a regulariser.
2. Empirical Bayes methods can often be used to determine one or all of the hyperparameters (i.e. the parameters in the prior) from the observed data. There are several ways to do this, one of them is to tune the shrinkage parameter  $\lambda$  to achieve minimum MSE. We discuss this further below.

Furthermore, there also exist many proposals advocating so-called “uninformative priors” or “objective priors”. However, there are no actually uninformative priors, since a prior distribution that looks uninformative (i.e. “flat”) in one coordinate system can be informative in another — this is a simple consequence of the rule for transformation of probability densities. As a result, often the suggested objective priors are in fact improper, i.e. are not actually probability distributions!

## 18.2 Default priors or uninformative priors

Objective or for default priors are attempts 1) to automatise specification of a prior and 2) to find uninformative priors.

### Jeffreys prior

The most well-known non-informative prior is given by a proposal by [Harold Jeffreys \(1891–1989\)](#) in 1946<sup>1</sup>.

Specifically, this prior is constructed from the Fisher information of the data generating model used in the likelihood and thus promises automatic construction of objective uninformative priors via

$$p(\theta) \propto \sqrt{\det I(\theta)}$$

<sup>1</sup>Jeffreys, H. 1946. *An invariant form for the prior probability in estimation problems*. Proc. Roy. Soc. A **186**:453–461. <https://doi.org/10.1098/rspa.1946.0056>



The reasoning underlying this prior is its **invariance against transformation of the coordinate system of the parameters**.

For the Beta-Binomial model the Jeffreys prior corresponds to  $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ . Note this is not the uniform distribution but a U-shaped prior.

For the normal-normal model it corresponds to the flat improper prior  $p(\mu) = 1$ .

For the Inverse-Wishart-normal model the Jeffreys prior is the improper prior  $p(\sigma^2) = \frac{1}{\sigma^2}$ .

This already illustrates the main problem with this type of prior – namely that it often is improper, i.e. the prior distribution is not actually a probability distribution (i.e. the density does not integrate to 1).

Another issue is that Jeffreys priors are usually not conjugate which complicates the update from the prior to the posterior.

Furthermore, if there are multiple parameters ( $\theta$  is a vector) then Jeffreys priors do not usually lead to sensible priors.

## Reference priors

An alternative to Jeffreys priors are the so-called **reference priors** developed by Bernardo (1979)<sup>2</sup>. This type of priors aims to choose the prior such that there is maximal “correlation” between the data and the parameter. More precisely, the mutual information between  $\theta$  and  $x$  is maximised (i.e. the the expected KL divergence between the posterior and prior distribution). The underlying motivation is that the data and parameters should be maximally linked (thereby minimising the influence of the prior).

For univariate settings the reference priors are identical to Jeffreys priors. However, reference prior also provide reasonable priors in multivariate settings.

In both Jeffreys’ and the reference prior approach the choice of prior is by expectation over the data, i.e. not for the specific data set at hand (this can be seen both as a positive and negative!).

<sup>2</sup>Bernardo, J. M. 1979. *Reference posterior distributions for Bayesian inference (with discussion)*. JRSS B 41:113–147. <https://doi.org/10.1111/j.2517-6161.1979.tb01066.x>

## 18.3 Empirical Bayes

In empirical Bayes the data analyst specifies a family of prior distribution (say a beta distribution with two parameters), and then the data at hand are used to find an optimal choice for the hyper-parameters (hence the name “empirical”). Thus the hyper-parameters are not specified but themselves estimated.

### Type II maximum likelihood

In particular, assuming data  $D$ , a likelihood  $p(D|\theta)$  for some model with parameters  $\theta$  as well as a prior  $p(\theta|\lambda)$  for  $\theta$  with hyper-parameter  $\lambda$  the marginal likelihood now depends on  $\lambda$ :

$$p(D|\lambda) = \int_{\theta} p(D|\theta)p(\theta|\lambda)d\theta$$

We can therefore use maximum (marginal) likelihood find optimal values of  $\lambda$  given the data.

Since maximum-likelihood is used in a second level step (the hyper-parameters) this type of empirical Bayes is also often called “type II maximum likelihood”.

### Shrinkage estimation using empirical risk minimisation

An alternative (but related) way to estimate hyper-parameters is by minimising the empirical risk.

In the examples for Bayesian estimation that we have considered so far the posterior mean of the parameter of interest was obtained by linear shrinkage

$$\hat{\theta}_{\text{shrink}} = E(\theta|D) = \lambda\theta_0 + (1 - \lambda)\hat{\theta}_{\text{ML}}$$

of the MLE  $\hat{\theta}_{\text{ML}}$  towards the prior mean  $\theta_0$ , with shrinkage intensity  $\lambda = \frac{k_0}{k_1}$  determined by the ratio of the prior and posterior concentration parameters  $k_0$  and  $k_1$ .

The resulting point estimate  $\hat{\theta}_{\text{shrink}}$  is called *shrinkage estimate* and is a convex combination of  $\theta_0$  and  $\hat{\theta}_{\text{ML}}$ . The prior mean  $\theta_0$  is also called the “target”.



The hyperparameter in this setting is  $k_0$  (linked to the precision of the prior) and or equivalently the shrinkage intensity  $\lambda$ .

An optimal value for  $\lambda$  can be obtained by minimising the mean squared error of the estimator  $\hat{\theta}_{\text{shrink}}$ .

In particular, by construction, the target  $\theta_0$  has low or even zero variance but non-vanishing and potentially large bias, whereas the MLE  $\hat{\theta}_{\text{ML}}$  will have low or zero bias but a substantial variance. By combining these two estimators with opposite properties the aim is to achieve a *bias-variance tradeoff* so that the resulting estimator  $\hat{\theta}_{\text{shrink}}$  has lower MSE than either  $\theta_0$  and  $\hat{\theta}_{\text{ML}}$ .

Specifically, the aim is to find

$$\lambda^* = \arg \min_{\lambda} E \left( (\theta - \hat{\theta}_{\text{shrink}})^2 \right)$$

It turns out that this can be minimised without knowing the actual true value of  $\theta$  and the result for an unbiased  $\hat{\theta}_{\text{ML}}$  is

$$\lambda^* = \frac{\text{Var}(\hat{\theta}_{\text{ML}})}{E((\hat{\theta}_{\text{ML}} - \theta_0)^2)}$$

Hence, the shrinkage intensity will be small if the variance of the MLE is small and/or if the target and the MLE differ substantially. On the other hand, if the variance of the MLE is large and/or the target is close to the MLE the shrinkage intensity will be large.

Choosing the shrinkage parameter by optimising expected risk (here mean squared error) is also a form empirical Bayes.

**Example 18.1.** James-Stein estimator:

Empirical risk minimisation to estimate the shrinkage parameter of the normal-normal model for a single observation yields the James-Stein estimator (1955).

Specifically, James and Stein propose the following estimate for the multivariate mean  $\mu$  of using a single sample  $x$  drawn from the multivariate normal  $N_d(\mu, I)$ :

$$\hat{\mu}_{JS} = \left( 1 - \frac{d-2}{\|x\|^2} \right) x$$

Here, we recognise  $\hat{\mu}_{ML} = x$ ,  $\mu_0 = 0$  and shrinkage intensity  $\lambda^* = \frac{d-2}{\|x\|^2}$ .

Efron and Morris (1972) and Lindley and Smith (1972) later generalised the James-Stein estimator to the case of multiple observations  $x_1, \dots, x_n$  and target  $\mu_0$ , yielding an empirical Bayes estimate of  $\mu$  based on the normal-normal model.



## 19 Optimality properties and summary

### 19.1 Bayesian statistics in a nutshell

- Bayesian statistics explicitly models the uncertainty about the parameters of interest by probability
- In the light of new evidence (observed data) the uncertainty is updated, i.e. the prior distribution is combined via Bayes rule with the likelihood to form the posterior distribution
- If the posterior distribution is in same family as the prior  $\rightarrow$  conjugate prior.
- In an exponential family the Bayesian update of the mean is always expressible as linear shrinkage of the MLE.
- For large sample size the posterior mean becomes maximum likelihood estimator and the prior plays no role.
- Conversely, for small sample size if no data is available the posterior stays close the prior.

#### Advantages

- Adding prior information has regularisation properties. This is very important in more complex models with many parameters, e.g., in the estimation of a covariance matrix (to avoid singularity).
- Improves small-sample accuracy (e.g. MSE)
- Bayesian estimators tend to perform better than MLEs - this is not surprising as they use the observed data plus the extra information available in the prior.
- Bayesian credible intervals are conceptually much more simple than frequentist confidence intervals.

### Frequentist properties of Bayesian estimators

A Bayesian point estimator (e.g. the posterior mean) can also be assessed by its frequentist properties.

- First, by construction due to introducing a prior the Bayesian estimator will be biased for finite  $n$  even if the MLE is unbiased.
- Second, intriguingly it turns out that the sampling variance of the Bayes point estimator (not to be confused with the posterior variance!) can be smaller than the variance of the MLE. This depends on the choice of the shrinkage parameter  $\lambda$  that also determines the posterior variance.

As a result, Bayesian estimators may have smaller MSE (=squared bias + variance) than the ML estimator for finite  $n$ .

In statistical decision theory this is called the theorem of **admissibility of Bayes rules**. It states that under mild conditions every admissible estimation rule (i.e. one that dominates all other estimators with regard to some expected loss, such as the MSE) is in fact a Bayes estimator with some prior.

Unfortunately, this theorem does not tell which prior is needed to achieve optimality, however an optimal estimator can often be found by tuning the hyperparameters.

### Specifying the prior — problem or advantage?

In Bayesian statistics the data analyst needs to be very explicit about the modelling assumptions:

**Model = data-generating process (likelihood) + prior uncertainty (prior distribution)**

Note that alternative statistical methods can often be interpreted as Bayesian methods assuming a specific *implicit* prior!

For example, likelihood estimation for the binomial model is equivalent to Bayes estimation using the Beta-Binomial model with a Beta(0,0) prior (=Haldane prior).

However, when choosing a prior explicitly for this model, interestingly most analysts would rather use a flat prior Beta(1, 1) (=Laplace prior) with implicit sample size  $k_0 = 2$  or a transformation-invariant prior Beta(1/2, 1/2) (=Jeffreys prior) with implicit sample size  $k_0 = 1$  rather than the Haldane prior!



→ be aware about the implicit priors!!

Better to acknowledge that a prior is being used (even if implicit!)  
Being specific about all your assumptions is enforced by the Bayesian approach.

Specifying a prior is thus best understood as an intrinsic part of model specification. It helps to improve inference and it may only be ignored if there is lots of data.

## 19.2 Optimality of Bayesian inference

The optimality of Bayesian model making use of full model specification (likelihood plus prior) can be shown from a number of different perspectives. Correspondingly, there are many theorems that prove (or at least indicate) this optimality:

- 1) **Richard Cox's theorem**: generalising classical logic invariably leads to Bayesian inference.
- 2) **de Finetti's representation theorem**: joint distribution of exchangeable observations can always be expressed as weighted mixture over a prior distribution for the parameter of the model. This implies the existence of the prior distribution and the requirement of a Bayesian approach.
- 3) **Frequentist decision theory**: all admissible decision rules are Bayes rules!
- 4) Entropy perspective: The posterior density (a function!) is obtained as a result of optimising an entropy criterion. Bayesian updating may thus be viewed as a *variational optimisation problem*. Specifically, Bayes theorem is the minimal update when new information arrives in form of observations (see below).

Remark: there exist a number of further (often somewhat esoteric) suggestions for propagating uncertainty such as "fuzzy logic", imprecise probabilities, etc. These contradict Bayesian learning and are thus in direct violation of the above theorems.

## 19.3 Connection with entropy learning

The *Bayesian update rule* is a very general form of learning when the *new information arrives in the form of data*. But actually there is an even more general principle of which the Bayesian update rule is just a special case: the **principle of minimal information update** (e.g. Jaynes 1959, 2003) or **principle of minimum information discrimination (MDI)** (Kullback 1959).

It can be summarised as follows: **Change your beliefs only as much as necessary to be coherent with new evidence!**

Under this principle of "inertia of beliefs" when new information arrives the uncertainty about a parameter is only minimally adjusted, only as much as needed to account for the new information. To implement this principle KL divergence is a natural measure to quantify the change of the underlying beliefs. This is known as **entropy learning**.

The Bayes rule emerges a special case of entropy learning:

- The KL divergence between the joint posterior  $Q_{x,\theta}$  and joint prior distribution  $P_{x,\theta}$  is computed, with the posterior distribution  $Q_{\theta|x}$  as free parameter.
- The conditional distribution  $Q_{\theta|x}$  is found by minimising the KL divergence  $D_{KL}(Q_{x,\theta}, P_{x,\theta})$ .
- The optimal solution to this **variational optimisation problem** is given by Bayes' rule!

This application of the KL divergence is an example of **reverse KL optimisation** (aka *I*-projection, see Part I of the notes). Intriguingly, this explains the zero forcing property of Bayes' rule (because that this is a general property of an *I*-projection).

Applying entropy learning therefore includes Bayesian learning as special case:

- 1) If information arrives in form of data → update prior by Bayes' theorem (Bayesian learning).

Interestingly, entropy learning will lead to other update rules for other types of information:

- 2) If information arrives in the form of another distribution → update using R. Jeffrey's rule of conditioning (1965).



- 3) If the information is presented in the form of constraints  $\rightarrow$  Kullback's principle of minimum MDI (1959), E. T. Jaynes maximum entropy (MaxEnt) principle (1957).

This shows (again) how fundamentally important KL divergence is in statistics. It not only leads to likelihood inference (via forward KL) but also to Bayesian learning, as well as to other forms of information updating (via reverse KL).

Furthermore, in Bayesian statistics the KL divergence is useful to choose priors (e.g. reference priors) and it also helps in (Bayesian) experimental design to quantify the information provided by an experiment.

## 19.4 Conclusion

Bayesian statistics offers a coherent framework for statistical learning from data, with methods for

- estimation
- testing
- model building

There are a number of theorems that show that “optimal” estimators (defined in various ways) are all Bayesian.

It is conceptually very simple — but can be computationally very involved!

It provides a coherent generalisation of classical TRUE/FALSE logic (and therefore does not suffer from some of the inconsistencies prevalent in frequentist statistics).

Bayesian statistics is a non-asymptotic theory, it works for any sample size. Asymptotically (large  $n$ ) it is consistent and converges to the true model (like ML!). But Bayesian reasoning can also be applied to events that take place only once — no assumption of hypothetical infinitely many repetitions as in frequentist statistics is needed.

Moreover, many classical (frequentist) procedures may be viewed as *approximations* to Bayesian methods and estimators, so using classical approaches in the correct application domain is perfectly in line with the Bayesian framework.

Bayesian estimation and inference also automatically regularises (via the prior) which is important for complex models and when there is the problem of overfitting.

## Bibliography

- Akaike, H. 1985. “Prediction and Entropy.” In *A Celebration of Statistics*, edited by A. C. Atkinson and S. E. Fienberg, 1–24. Springer. [https://doi.org/10.1007/978-1-4613-8560-8\\_1](https://doi.org/10.1007/978-1-4613-8560-8_1).
- Aldrich, J. 1997. “R. A. Fisher and the Making of Maximum Likelihood 1912–1922.” *Statist. Sci.* 12: 162–76. <https://doi.org/10.1214/ss/1030037906>.
- Amari, S.-I. 2016. *Information Geometry and Its Applications*. Springer. <https://doi.org/10.1007/978-4-431-55978-8>.
- Bernardo, J. M. 1979. “Expected Information as Expected Utility.” *Ann. Stat.* 7: 686–90. <https://doi.org/10.1214/aos/1176344689>.
- Dawid, A. P., and M. Musio. 2014. “Theory and Applications of Proper Scoring Rules.” *METRON* 72: 169–83. <https://doi.org/10.1007/s40300-014-0039-y>.
- Diaconis, P., and B. Skyrms. 2018. *Ten Great Ideas about Chance*. Princeton University Press.
- Efron, B. 2022. *Exponential Families in Theory and Practice*. Cambridge University Press. <https://doi.org/10.1017/9781108773157>.
- Fisher, R. A. 1922. “On the Mathematical Foundations of Theoretical Statistics.” *Phil. Trans. R. Soc. A* 222: 309–68. <https://doi.org/10.1098/rsta.1922.0009>.
- Gelman, A., J. B. Carlin, H. A. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. CRC Press.
- Good, I. J. 1952. “Rational Decisions.” *J. R. Statist. Soc. B* 14: 107–14. <https://doi.org/10.1111/j.2517-6161.1952.tb00104.x>.
- . 1979. “Studies in the History of Probability and Statistics. XXXVII A. M. Turing’s Statistical Work in World War II.” *Ann. Math. Statist.* 66: 393–96. <https://doi.org/10.1093/biomet/66.2.393>.
- Hand, D. J. 2015. “From Evidence to Understanding: A Commentary on Fisher (1922) ‘on the Mathematical Foundations of Theoretical Statistics’.” *Phil. Trans. R. Soc. A* 373: 20140252. <https://doi.org/10.1098/rsta.2014.0252>.
- Heard, N. 2021. *An Introduction to Bayesian Inference, Methods and Computation*. Springer.
- Held, L., and D. S. Bové. 2020. *Applied Statistical Inference: Likelihood and Bayes*. Second. Springer.



- Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511790423>.
- Kullback, S. 1959. *Information Theory and Statistics*. Wiley.
- Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *Ann. Math. Statist.* 22: 79–86. <https://doi.org/10.1214/aoms/1177729694>.
- Lemons, D. S. 2013. *A Student's Guide to Entropy*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511984556>.
- McGrayne, S. B. 2011. *The Theory That Would Not Die*. Yale University Press.
- Sahu, S. K. 2024. *Introduction to Probability, Statistics & R*. Springer. <https://doi.org/10.1007/978-3-031-37865-2>.
- Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell Syst. Tech. J.* 27: 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- Stiegler, S. M. 2007. "The Epic Story of Maximum Likelihood." *Statist. Sci.* 22: 598–620. <https://doi.org/10.1214/07-STS249>.
- Winkler, R. L. 1996. "Scoring Rules and the Evaluation of Probabilities (with Discussion)." *Test* 5: 1–69. <https://doi.org/10.1007/BF02562681>.

## A Statistics refresher

Below you find a brief overview over some relevant concepts in statistics that you should be familiar with from earlier modules.

### A.1 Data and statistics as functions of data

Broadly, by "data" we refer to quantitative observations and measurements collected in experiments.

We denote the observed data by  $D = \{x_1, \dots, x_n\}$  where  $n$  denotes the number of data points (the **sample size**). Each data point can be scalar or a multivariate quantity.

Generally, a **statistic**  $t(D)$  is function of the observed data  $D$ . The statistic  $t(D)$  can be of any type and value (scalar, vector, matrix etc. — even a function).  $t(D)$  is called a *summary statistic* if it describes important aspects of the data such as location (e.g. the average  $\text{avg}(D) = \bar{x}$ , the median) or scale (e.g. standard deviation, interquartile range).

### A.2 Statistical learning

The aim in statistics, and by extension in data science and machine learning, is to use data to learn about and better understand the world. It is a key feature of statistics to employ probabilistic models for that purpose.

Let denote data models by  $p(x|\theta)$  where  $\theta$  represents the parameters of the model. Often (but not always)  $\theta$  can be interpreted as or is associated with some manifest property of the model. If there is only a single parameter we write  $\theta$  (scalar parameter). If we wish to highlight that there are multiple parameters we write  $\boldsymbol{\theta}$  (in bold type).

Specifically, our aim is to identify the best model(s) for the data in order to both



- explain the current data, and
- to enable good prediction of future data.

By choosing a sufficiently complex model the first aim (to explain the observed data) is often easily achieved. However, if the model is too complex it can fail to address the second aim (to predict unseen data well). Thus, when choosing a model we would like to avoid both the problem of **underfitting** (i.e. choosing an overly simplistic model) as well **overfitting** (i.e. choosing an overly complex model). Finding this balance will also help with interpreting the fitted model.

Typically, we focus the analysis to a specific model family with a some parameter  $\theta$ .

An **estimator** for  $\theta$  is a function  $\hat{\theta}(D)$  of the data that maps the data (input) to an informed guess (output) about  $\theta$ .

- A **point estimator** provides a single number for each parameter
- An **interval estimator** provides a set of possible values for each parameter.

Interval estimators can be linked to the concept of testing specified values for a parameter. Specifically a confidence interval contains all parameter values that are not significantly different from the best parameter.

### A.3 Sampling properties of a point estimator

A point estimator  $\hat{\theta}$  depends on the data, hence it exhibits **sampling variation**, i.e. estimate will be different for a new set of observations.

Thus  $\hat{\theta}$  can be seen as a random variable, and its distribution is called **sampling distribution** (across different experiments).

Properties of this distribution can be used to evaluate how far the estimator deviates (on average across different experiments) from the true value:

$$\begin{aligned}
 \text{Bias:} \quad \text{Bias}(\hat{\theta}) &= E(\hat{\theta}) - \theta \\
 \text{Variance:} \quad \text{Var}(\hat{\theta}) &= E\left((\hat{\theta} - E(\hat{\theta}))^2\right) \\
 \text{Mean squared error:} \quad \text{MSE}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\
 &= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
 \end{aligned}$$

The last identity about MSE follows from  $E(x^2) = \text{Var}(x) + E(x)^2$ .

At first sight it seems desirable to focus on unbiased (for finite sample size  $n$ ) estimators. However, requiring strict unbiasedness is not always a good idea. In many situations it is better to accept some bias in an estimator in order to achieve a smaller variance and an overall smaller MSE. This is called **bias-variance tradeoff** — as more bias is traded for smaller variance (or, conversely, less bias is traded for higher variance). This is also related to the above mentioned problems of underfitting (large bias) and overfitting (large variance).

### A.4 Efficiency and consistency of an estimator

Typically, Bias, Var and MSE all decrease with increasing sample size so that with more data  $n \rightarrow \infty$  the errors become smaller and smaller.

**Efficiency:** An estimator  $\hat{\theta}_A$  is said to more efficient than estimator  $\hat{\theta}_B$  if for same sample size  $n$  it has smaller error (e.g. MSE) than the competing estimator.

The typical rate of decrease in variance of a good estimator is  $\frac{1}{n}$  and the rate of decrease in the standard deviation is  $\frac{1}{\sqrt{n}}$ . Note that this implies that to get one digit more accuracy in an estimate (standard deviation decreasing by factor of 10) we need 100 times more data!

**Consistency:**  $\hat{\theta}$  is called consistent if

$$\text{MSE}(\hat{\theta}) \rightarrow 0 \text{ with } n \rightarrow \infty$$

Consistency is an essential yet relatively weak requirement for any reasonable estimator. Among all consistent estimators we generally choose those that are most **efficient**, meaning that they exhibit the smallest variance and/or MSE for a given finite  $n$ .

Consistency implies that, given an infinite amount of data, the true model can be accurately identified, provided that the model class includes the actual data-generating model. If the model class does not encompass the true model, strict consistency cannot be attained. Nevertheless, our goal remains to choose a model that is closest to the true model and approximates it as best as possible.



## A.5 Empirical distribution

Suppose we observe data  $D = \{x_1, \dots, x_n\}$  with each  $x_i \sim Q$  sampled independently and identically.

The **empirical distribution** based on data  $D$  can be written as the mixture

$$\hat{Q}_n = \frac{1}{n} \sum_{i=1}^n \delta(x_i)$$

where  $\delta(\mu)$  is the delta distribution, a *continuous* distribution with mean  $\mu$  representing a point mass at  $\mu$ . The “hat” is a reminder that  $\hat{Q}_n$  is a function of the observed data  $D$  with the subscript  $n$  indicating the sample size.

The corresponding **empirical pdf** is

$$\hat{q}_n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x|x_i)$$

This is a generalised function satisfying  $\hat{q}_n(x) = 0$  for all  $x \neq x_i$  with infinite spikes at the data points  $x_i$  and still integrating to one.

The **empirical cumulative distribution function** (ecdf) is given by

$$\hat{Q}_n(x) = \frac{1}{n} \sum_{i=1}^n [x_i \leq x]$$

where  $[A]$  is the indicator function in Iverson bracket notation which equals 1 if  $A$  is true and 0 otherwise. Thus  $\hat{Q}_n(x)$  counts how many observations  $x_i \in D$  are smaller or equal than  $x$  and then standardises by the total number of samples  $n$ . The empirical distribution function is monotonically non-decreasing from 0 to 1 in discrete steps.

The mean of the empirical distribution is the **empirical mean** and is given by the sample average:

$$\begin{aligned} E_{\hat{Q}_n}(x) &= E(\hat{Q}_n) \\ &= \frac{1}{n} \sum_{i=1}^n E(\delta(x_i)) \\ &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \end{aligned}$$

In R the empirical distribution function is computed by `ecdf()`.

## A.6 Law of large numbers

The **law of large numbers**, discovered by [Jacob Bernoulli \(1655-1705\)](#), asserts that, if the mean exists, the sample average will converge to the mean as the sample size  $n$  becomes large,

$$\frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \rightarrow E(x)$$

or equivalently,

$$E(\hat{Q}_n) \rightarrow E(Q)$$

Therefore, when the mean is defined, it can be approximated by the empirical mean for sufficiently large values of  $n$ .

A variant of the law of large numbers is that the empirical distribution  $\hat{Q}_n$  converges strongly (almost surely) to the underlying distribution  $Q$  as  $n \rightarrow \infty$ :

$$\hat{Q}_n \xrightarrow{a.s.} Q$$

The [Glivenko-Cantelli theorem](#) additionally asserts that the convergence is uniform.

As a result, with  $n \rightarrow \infty$  there's also convergence of the average of a function of the observed samples to the corresponding expectation of the function of the random variable:

$$E_{\hat{Q}_n}(h(x)) = \frac{1}{n} \sum_{i=1}^n h(x_i) \rightarrow E_Q(h(x))$$

Moreover, the law of large numbers provides a **justification for interpreting large-sample limits of frequencies as probabilities**. However, **the converse** assumption that all probabilities can be interpreted in frequentist manner **does not follow** from the law of large numbers or from the axioms of probability.

Finally, it is worth pointing out that the law of large number is only concerned with the asymptotic regime, finite sample behaviour (small  $n$ ) can be quite different from the large- $n$  case.



## A.7 Empirical estimators

Thanks to the law of large numbers, we may use the empirical distribution  $\hat{Q}_n$  based on data  $D$  as a consistent estimator of the unknown true distribution  $Q$ .

This motivates **empirical estimators**. If a quantity of interest  $\theta$  is a functional  $\theta = g(Q)$  of the true distribution  $Q$ , its **empirical estimator**  $\hat{\theta} = g(\hat{Q}_n)$  is obtained by replacing the true distribution with the empirical distribution  $\hat{Q}_n$ .

An example is the mean  $E_Q(x)$  with regard to  $Q$ . The **empirical mean** is the expectation with regard to the empirical distribution which equals the **average of the samples**:

$$\hat{E}(x) = \hat{\mu} = E_{\hat{Q}_n}(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

Similarly, other empirical estimators can be constructed simply by replacing the expectation in the definition of the quantity of interest by the sample average. For example, the **empirical variance** with unknown mean is given by

$$\widehat{\text{Var}}(x) = \hat{\sigma}^2 = E_{\hat{Q}_n}((x - \hat{\mu})^2) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Note the factor  $1/n$  before the summation sign. We can also write the empirical variance in terms of  $\overline{x^2} = \frac{1}{n} \sum_{k=1}^n x_k^2$  as

$$\widehat{\text{Var}}(x) = \overline{x^2} - \bar{x}^2$$

By construction, as a result of the strong convergence of  $\hat{Q}_n$  to  $Q$  empirical estimators are consistent, with their MSE, variance and bias all decreasing to zero with large sample size  $n$ . However, for finite sample size they do have a finite variance and may also be biased.

For example, the empirical variance given above is biased with  $\text{Bias}(\hat{\sigma}^2) = -\sigma^2/n$ . Note this bias decreases with  $n$ . An unbiased estimator can be obtained by rescaling the empirical estimator by the factor  $n/(n-1)$ :

$$\hat{\sigma}_{\text{UB}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The empirical estimators for the mean and variance can also be obtained for random vectors  $\mathbf{x}$ . In this case the data  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  is comprised of  $n$  vector-valued observations.

For the mean get

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k = \bar{\mathbf{x}}$$

and for the covariance

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T$$

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

With  $\overline{\mathbf{x}\mathbf{x}^T} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T$  we can also write

$$\hat{\Sigma} = \overline{\mathbf{x}\mathbf{x}^T} - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$$

## A.8 Sampling distribution of mean and variance estimators for normal data

If the underlying distribution family of  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  is known we can often obtain the exact distribution of an estimator.

For example, assuming normal distribution  $x_i \sim N(\mu, \sigma^2)$  we can derive the sampling distribution for the empirical mean and variance:

- The empirical estimator of the mean parameter  $\mu$  is given by  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ . Under the normal assumption the distribution of  $\hat{\mu}$  is

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Thus  $E(\hat{\mu}) = \mu$  and  $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$ . The estimate  $\hat{\mu}$  is unbiased as  $E(\hat{\mu}) - \mu = 0$ . The mean squared error of  $\hat{\mu}$  is  $\text{MSE}(\hat{\mu}) = \frac{\sigma^2}{n}$ .

- The empirical variance  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$  for normal data follows a one-dimensional Wishart distribution

$$\hat{\sigma}^2 \sim \text{Wis}\left(s^2 = \frac{\sigma^2}{n}, k = n - 1\right)$$



Thus,  $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$  and  $\text{Var}(\hat{\sigma}_{\text{ML}}^2) = \frac{2(n-1)}{n^2}\sigma^4$ . The estimate  $\hat{\sigma}^2$  is biased since  $E(\hat{\sigma}_{\text{ML}}^2) - \sigma^2 = -\frac{1}{n}\sigma^2$ . The mean squared error is  $\text{MSE}(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4$ .

- The unbiased variance estimate  $\hat{\sigma}_{\text{UB}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  for normal data follows a one-dimensional Wishart distribution

$$\hat{\sigma}_{\text{UB}}^2 \sim \text{Wis}\left(s^2 = \frac{\sigma^2}{n-1}, k = n-1\right)$$

Thus,  $E(\hat{\sigma}_{\text{UB}}^2) = \sigma^2$  and  $\text{Var}(\hat{\sigma}_{\text{UB}}^2) = \frac{2}{n-1}\sigma^4$ . The estimate  $\hat{\sigma}_{\text{ML}}^2$  is unbiased since  $E(\hat{\sigma}_{\text{UB}}^2) - \sigma^2 = 0$ . The mean squared error is  $\text{MSE}(\hat{\sigma}_{\text{UB}}^2) = \frac{2}{n-1}\sigma^4$ .

Interestingly, for any  $n > 1$  we find that  $\text{Var}(\hat{\sigma}_{\text{UB}}^2) < \text{Var}(\hat{\sigma}_{\text{ML}}^2)$  and  $\text{MSE}(\hat{\sigma}_{\text{UB}}^2) < \text{MSE}(\hat{\sigma}_{\text{ML}}^2)$  so that the biased empirical estimator has both lower variance and lower mean squared error than the unbiased estimator.

## A.9 *t*-statistics

### One sample *t*-statistic

Suppose we observe  $n$  independent data points  $x_1, \dots, x_n \sim N(\mu, \sigma^2)$ . Then the average  $\bar{x} = \sum_{i=1}^n x_i$  is distributed as  $\bar{x} \sim N(\mu, \sigma^2/n)$  and correspondingly

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

Note that  $z$  uses the *known variance*  $\sigma^2$ .

If the variance is unknown and is estimated by the *unbiased variance*

$$s_{\text{UB}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

then one arrives at the one sample *t*-statistic

$$t_{\text{UB}} = \frac{\bar{x} - \mu}{\sqrt{s_{\text{UB}}^2/n}} \sim t_{n-1}.$$

It is distributed according to a Student's *t*-distribution with  $n-1$  degrees of freedom, with mean 0 for  $n > 2$  and variance  $(n-1)/(n-3)$  for  $n > 3$ .

If instead of the unbiased estimate the *empirical variance* (i.e. the maximum likelihood estimate, ML)

$$s_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n} s_{\text{UB}}^2$$

is used then this leads to a slightly different statistic

$$t_{\text{ML}} = \frac{\bar{x} - \mu}{\sqrt{s_{\text{ML}}^2/n}} = \sqrt{\frac{n}{n-1}} t_{\text{UB}}$$

with

$$t_{\text{ML}} \sim t_{n-1}\left(0, \tau^2 = \frac{n}{n-1}\right)$$

Thus,  $t_{\text{ML}}$  follows a location-scale *t*-distribution, with mean 0 for  $n > 2$  and variance  $n/(n-3)$  for  $n > 3$ .

### Two sample *t*-statistic with common variance

Now suppose we observe normal data  $D = \{x_1, \dots, x_n\}$  from two groups with sample size  $n_1$  and  $n_2$  (and  $n = n_1 + n_2$ ) with two different means  $\mu_1$  and  $\mu_2$  and common variance  $\sigma^2$ :

$$x_1, \dots, x_{n_1} \sim N(\mu_1, \sigma^2)$$

and

$$x_{n_1+1}, \dots, x_n \sim N(\mu_2, \sigma^2)$$

Then  $\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$  and  $\hat{\mu}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^n x_i$  are the sample averages within each group.

The common variance  $\sigma^2$  may be estimated either by the *unbiased estimate*

$$s_{\text{UB}}^2 = \frac{1}{n-2} \left( \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^n (x_i - \hat{\mu}_2)^2 \right)$$

(note the factor  $n-2$ ) or by the *empirical estimate* (ML)

$$s_{\text{ML}}^2 = \frac{1}{n} \left( \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^n (x_i - \hat{\mu}_2)^2 \right) = \frac{n-2}{n} s_{\text{UB}}^2$$



The estimator for the common variance is often referred to as *pooled variance estimate* as information is pooled from two groups to obtain the estimate.

Using the unbiased pooled variance estimate the two sample  $t$ -statistic is given by

$$t_{UB} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{UB}^2}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{n}{n_1 n_2}\right) s_{UB}^2}}$$

In terms of empirical frequencies  $\hat{\pi}_1 = \frac{n_1}{n}$  and  $\hat{\pi}_2 = \frac{n_2}{n}$  it can also be written as

$$t_{UB} = \sqrt{n} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{\hat{\pi}_1} + \frac{1}{\hat{\pi}_2}\right) s_{UB}^2}} = \sqrt{n \hat{\pi}_1 \hat{\pi}_2} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s_{UB}^2}}$$

The two sample  $t$ -statistic is distributed as

$$t_{UB} \sim t_{n-2}$$

i.e. according to a Student's  $t$ -distribution with  $n - 2$  degrees of freedom, with mean 0 for  $n > 3$  and variance  $(n - 2)/(n - 4)$  for  $n > 4$ . Large values of the two sample  $t$ -statistic indicates that there are indeed two groups rather than just one.

The two sample  $t$ -statistic using the empirical (ML) pooled estimate of the variance is

$$\begin{aligned} t_{ML} &= \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{ML}^2}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{n}{n_1 n_2}\right) s_{ML}^2}} \\ &= \sqrt{n} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{\hat{\pi}_1} + \frac{1}{\hat{\pi}_2}\right) s_{ML}^2}} = \sqrt{n \hat{\pi}_1 \hat{\pi}_2} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s_{ML}^2}} \\ &= \sqrt{\frac{n}{n-2}} t_{UB} \end{aligned}$$

with

$$t_{ML} \sim t_{n-2} \left(0, \tau^2 = \frac{n}{n-2}\right)$$

Thus,  $t_{ML}$  follows a location-scale  $t$ -distribution, with mean 0 for  $n > 3$  and variance  $n/(n - 4)$  for  $n > 4$ .

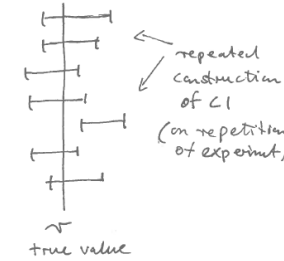


Figure A.1: Coverage property of confidence intervals.

## A.10 Confidence intervals

### General concept

A **confidence interval** (CI) is an **interval estimate**  $\widehat{CI}(x_1, \dots, x_n)$  that depends on data and has a random (sampling) variation as well as a **frequentist** interpretation.

A key property of a confidence interval is its **coverage probability**

$$\kappa = \Pr(\theta \in \widehat{CI})$$

which describes how often — in repeated identical experiments — the estimated confidence interval overlaps the true parameter value  $\theta$ , i.e. how often it will “cover”  $\theta$  (see Figure A.1). In the above  $\theta$  is fixed and  $\widehat{CI}$  is random. Note that  $\kappa$  is explicitly **not** the probability that the true value is contained in a specific instance of an estimated confidence interval. Specifically, any particular confidence interval either covers  $\theta$  or it doesn't.

For example, a coverage probability  $\kappa = 0.95$  (95%) implies that in 95 out of 100 repeated experiments the corresponding estimated confidence interval will contain the (unknown) true value.

It is trivial to create a confidence with high coverage, simply by assuming a wide interval. Therefore, a useful confidence interval must be both **compact** and have **high coverage**.

Finally, there is also a direct relationship between confidence intervals and statistical testing procedures. Specifically, a confidence interval can be interpreted as the set of parameter values that cannot be rejected. The



complement  $\alpha = 1 - \kappa$  is called the **rejection probability** or **significance level**.

### Symmetric normal confidence interval

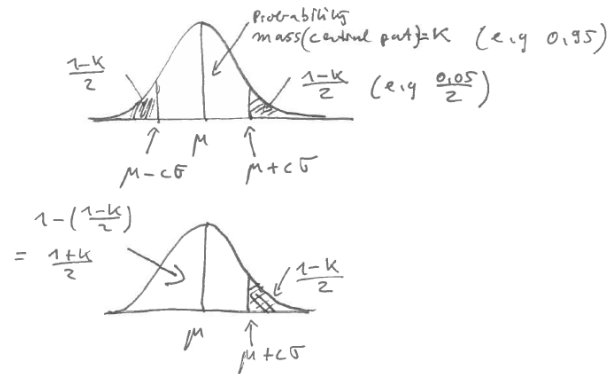


Figure A.2: Construction of a symmetric two-sided normal confidence interval.

For a normally distributed univariate random variable it is straightforward to construct a symmetric two-sided confidence interval with a given desired coverage  $\kappa$  (Figure A.2). The confidence interval corresponds to the central part of the density and contains probability mass  $\kappa = 1 - \alpha$  whereas both tails each contain mass  $(1 - \kappa)/2 = \alpha/2$  and correspond together to the rejection region.

A **symmetric normal confidence interval** with nominal coverage  $\kappa$  for

- a scalar parameter  $\theta$
- with normally distributed estimate  $\hat{\theta} \sim N(\theta, \sigma^2)$

is given by

$$\widehat{\text{CI}} = [\hat{\theta} \pm c\sigma]$$

where the critical value  $c$  is chosen to achieve the desired coverage probability

$$\kappa = \Pr(\hat{\theta} - c\sigma \leq \theta \leq \hat{\theta} + c\sigma)$$

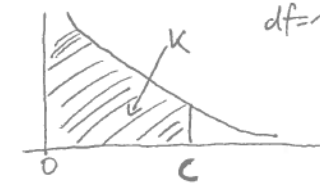


Figure A.3: Construction of a one-sided confidence interval based on a chi-squared distribution with one degree of freedom.

The critical value  $c$  is obtained as the  $(1 + \kappa)/2 = 1 - \alpha/2$  quantile  $z_{(1+\kappa)/2} = z_{1-\alpha/2}$  of the standard normal distribution  $N(0, 1)$  so that

$$c = \Phi^{-1}\left(\frac{1 + \kappa}{2}\right) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

where where  $\Phi(c)$  is the cumulative distribution function (CDF) of the standard normal  $N(0, 1)$  distribution. Its inverse  $\Phi^{-1}$  is the standard normal quantile function.

Table A.1: Critical values for the standard normal distribution.

Coverage $\kappa$	Critical value $c$	Quantile $z_{(1+\kappa)/2}$
0.90	1.6449	$z_{0.95}$
0.95	1.9600	$z_{0.975}$
0.99	2.5758	$z_{0.995}$

Table A.1 lists the critical values  $c$  for the three most commonly used values of  $\kappa$ . It is useful to memorise these values as they are used frequently.

### One-sided confidence interval based on the chi-squared distribution

For a chi-squared distributed statistic commonly a one-sided confidence interval of the form  $[0, c]$  is used with nominal coverage probability  $\kappa = \Pr(x \leq c)$  (see Figure A.3). The right tail contains  $1 - \kappa = \alpha$  probability mass.



We obtain the critical value  $c$  as the  $\kappa = 1 - \alpha$  quantile of the chi-squared distribution by using the quantile function, i.e. by inverting the CDF of the chi-squared distribution.

Table A.2: Critical values for the chi-squared distribution with one degree of freedom.

Coverage $\kappa$	Critical value $c$	Quantile $x_\kappa$
0.90	2.7055	$x_{.90}$
0.95	3.8415	$x_{.95}$
0.99	6.6349	$x_{.99}$

Table A.2 lists the critical values for the three most common choices of the coverage probability  $\kappa$  for a chi-squared distribution with one degree of freedom. Note that these critical values are the squared values of the corresponding thresholds in Table A.1 (with small discrepancies due to rounding).

## B Further study

In this module we can only touch the surface of likelihood and Bayes inference. As a starting point for further reading the following text books are recommended.

### B.1 Recommended reading

- Held and Bové (2020) *Applied Statistical Inference: Likelihood and Bayes (2nd edition)*. Springer.
- Sahu (2024) *Introduction to Probability, Statistics & R*. Springer

### B.2 Additional references

- Heard (2021) *An Introduction to Bayesian Inference, Methods and Computation*. Springer.
- Gelman et al. (2014) *Bayesian data analysis (3rd edition)*. CRC Press.