Frequentist and Bayesian parameter estimation

Bachelor thesis of:

Philipp Schwartenbeck

Supervisor: Prof. Dr. Wolfgang Trutschnig

January 19, 2017

Contents

1 Introduction

The intention of this thesis is to investigate the process of parameter estimation from two different perspectives: the frequentist approach based on maximum likelihood estimation on the one hand, and the Bayesian approach based on inferring a posterior distribution over the entire parameter space on the other. The goal is not to contribute to a debate about the correctness or superiority of one of the two perspectives - there has been much controversy about this topic over the past 250 years without reaching a satisfactory conclusion. Rather, this thesis adopts Carnap's view arguing that both perspectives are in themselves consistent and comply with the basic axiomatic desiderata of probability theory (*Kolmogorov's axioms*) but (unfortunately) use the same term *probability* in a very different way (Carnap, 1959). I will describe and compare basic simulations and results for both point-wise and interval-based parameter estimation obtained from the two different perspectives. Furthermore, I will discuss important theoretical aspects of both theories, in particular those of Bayesian inference, which can be understood as an extension to maximum-likelihood estimation by incorporating prior knowledge. The most prominent feature of Bayesian inference is the use of a prior density, which regularises the maximum-likelihood solution in an estimation problem.

In the first section, I will introduce the required basics of probability theory to derive maximum-likelihood and Bayesian parameter estimation. Sections two and three discuss and compare parameter estimation under the two perspectives, where section two introduces point-wise estimation and section three compares frequentist confidence and Bayesian credibility intervals around point estimates. Section four deals with the most delicate aspect of Bayesian inference, namely the selection of a prior density. One major criticism of Bayesian approaches is the degree of arbitrariness in selecting a prior distribution. I will discuss the benefits and pitfalls of using priors to regularise maximum-likelihood and, in particular, I will elaborate on the most influential proposals for formal rules for prior specification and objective Bayesianism. Importantly, this also relates to the crucial issue of expressing the absence of prior knowledge (i.e., uninformative priors). The aspect of uninformative priors becomes particularly delicate in the case of unrestricted parameter spaces in estimation problems. Finally, I will illustrate a practical example in which two aspects of the use (and popularity in certain fields of science) of Bayesian inference are illustrated: using Bayesian inference to model a cognitive process and to infer individual parameters of subjects that account for their behavioural tendencies.

2 Probability theory and Bayes' theorem

We start with the most important definitions in probability theory (cf., Billingsley, 1979; Kallenberg, 2002; Klenke, 2008) that will lead to a formal definition of a likelihood function and the derivation of Bayes' rule.

2.1 Probability measure

Definition 2.1.1
A collection $\mathcal{A}$ of subsets of a set Ω is called **σ-Algebra** if it satisfies the following conditions:

1. $\Omega \in \mathcal{A}$
2. A $\in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
3. $A_1, A_2, A_3, ... \in \mathcal{A} \Rightarrow \bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{A}$

We call any set A $\in \mathcal{A}$ **measurable** and the tuple $(\Omega, \mathcal{A})$ a **measure space**. We will need the concept of a σ-algebra to arrive at a generic definition of a probability measure that includes cases where the power set $\mathbb{P}(\Omega)$ is uncountable.

Definition 2.1.2
Given a measure space $(\Omega, \mathcal{A})$ we call the function $\mathcal{P}$: $\mathcal{A} \to$ [0,1] a **probability measure** if it satisfies

1. $\mathcal{P}(\Omega) = 1$
2. $\mathcal{P}(\bigcup\limits_{n=1}^{\infty} A_n) = \sum\limits_{n=1}^{\infty} \mathcal{P}(A_n)$ for any pairwise disjoint sequence $A_1, A_2, A_3, ... \in \mathcal{A}$

   This attribute is called **σ-additivity**.

We call the triple $(\Omega, \mathcal{A}, \mathcal{P})$ a **probability space**.

A function $X : (\Omega, \mathcal{A}, \mathcal{P}) \to (\mathbb{R}, B(\mathbb{R}))$, where $B(\mathbb{R})$ is the Borel σ-algebra (the σ-algebra generated by all open sets in $\mathbb{R}$), is called a **random variable**.

Remark 2.1.3
A probability measure $\mathcal{P}(A)$ of an event A $\in \mathcal{A}$ is assumed to satisfy three **Kolmogorov axioms**. The second and the third axioms are points 1. and 2. of Definition 2.1.2, respectively. The first axiom states that $\mathcal{P}(A) \geqslant 0$ and $\mathcal{P}(A) \in \mathbb{R}$ for every $A \in \mathcal{A}$.

2.2 Sum and product rule

Before we can define Bayes' theorem and move on to maximum likelihood and Bayesian parameter estimation, we have to introduce the notion of a conditional probability:

Definition 2.2.1
In a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ with sets A,B $\in \mathcal{A}$ and $\mathcal{P}(B) \neq 0$, the **probability of A conditioned on B** is defined as:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}$$

It follows that $\mathcal{P}(A|B) = \mathcal{P}(A)$ in the case when A and B are independent, because then $\mathcal{P}(A \cap B) = \mathcal{P}(A) \cdot \mathcal{P}(B)$.

The notion of conditional probabilities lies at the heart of probability theory and, particularly, of parameter estimation based on both frequentist and Bayesian methods. Conditional probabilities also lead to the following central rules in probability theory:

Definition 2.2.2
In a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ with sets A,B $\in \mathcal{A}$, the **product rule** states that multiplying a conditional probability with the **marginal likelihood** of the conditioning variable yields the **joint distribution** of the two variables:

$$\begin{aligned}\mathcal{P}(A \cap B) &= \mathcal{P}(A|B) \cdot \mathcal{P}(B)\\ &= \mathcal{P}(B|A) \cdot \mathcal{P}(A)\\ &= \mathcal{P}(B \cap A)\end{aligned}$$

If $B_1, B_2, ...$ denotes a measurable partition of B, then the **sum rule** states that if we sum over all instances of the conditional variable, we are left with the marginal likelihood of the conditioned variable:

$$\mathcal{P}(A) = \sum_{i=1}^{\infty} \mathcal{P}(A \cap B_i)$$

Applying product rule, it follows that

$$\mathcal{P}(A) = \sum_{i=1}^{\infty} \mathcal{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathcal{P}(A|B_i)\mathcal{P}(B_i).$$

2.3 Bayes' rule

Having established the concept of a probability measure and the product and sum rule, we can move on to introduce Bayes' rule for measurable sets, before defining Bayes' theorem more generally for probability densities.

Theorem 2.3.1
In a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ with sets A,B $\in \mathcal{A}$ and $\mathcal{P}(B) \neq 0$, the probability of A conditioned on B can be written as:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A) \cdot \mathcal{P}(A)}{\mathcal{P}(B)}$$

This is called **Bayes' rule**.

The **proof** is very simple:

$$\begin{aligned}\mathcal{P}(A|B) \cdot \mathcal{P}(B) &= \mathcal{P}(A \cap B) \\ &= \mathcal{P}(B|A) \cdot \mathcal{P}(A) \\ \Leftrightarrow \mathcal{P}(A|B) &= \frac{\mathcal{P}(B|A) \cdot \mathcal{P}(A)}{\mathcal{P}(B)}\end{aligned}$$

■

If $A_1, A_2, ...$ denotes a measurable partition of $A \in \mathcal{A}$, it follows analogously that:

$$\mathcal{P}(A_i|B) = \frac{\mathcal{P}(B|A_i)\mathcal{P}(A_i)}{\sum_{i=1}^{\infty} \mathcal{P}(B|A_i)\mathcal{P}(A_i)}.$$

where

$$\sum_{i=1}^{\infty} \mathcal{P}(B|A_i)\mathcal{P}(A_i) = \mathcal{P}(B)$$

according to the law of total probability (cf., Bathke, 2013).

Finally, before we can define Bayes' theorem, we need to introduce the notion of a probability density function.

Definition 2.3.2
For a given random variable X in probability space $(\Omega, \mathcal{A}, \mathcal{P})$, the (cumulative) **distribution function** $F_X(x) : \mathbb{R} \to [0,1]$ is defined as (cf., Trutschnig, 2014)

$$F_X(x) := \mathcal{P}^X((-\infty, x]) = \mathcal{P}(\{\omega \in \Omega : X(\omega) \leq x\}).$$

If $\mathcal{P}^X$ is absolutely continuous, an integrable function $f : \mathbb{R} \to [0, \infty)$ with

$$\int_{\mathbb{R}} f(x)dx = 1$$

is called **probability density function**. In particular, we see that

$$F_X(x) = \mathcal{P}^X((-\infty, x]) = \int_{(-\infty,x]} f(t)dt$$

If X and Y are two continuous random variables, we can define the **conditional density function** as

$$f(x|y) = \frac{f(x,y)}{f(y)},$$

where

$$f(y) = \int f(x,y)dx$$

is called **marginal density function** - after marginalising out random variable X from the joint density function $f(x, y)$ (Felsenstein, 2008).

Bayes' Theorem 2.3.3
Based on the concept of a probability density function, we can now introduce a more general form of Bayes' theorem (Felsenstein, 2008):

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

The proof is analogous to the proof of Bayes' rule as introduced in Theorem 2.3.1 based on the definition of a conditional density function (Definition 2.3.2).

In the above theorem, $\pi(\theta)$ reflects a **prior density** over an unknown quantity or **parameter**, which will be discussed in much more detail in the following sections. In brief, a prior density reflects any existing a-priori knowledge about an unknown quantity. The second part of the numerator, $f(x|\theta)$, is called **likelihood function** and will be discussed in detail in the next sections. The term $\pi(\theta|x)$ is called **posterior density**, which can be understood as an update of the prior distribution after observing random variable X. Consequently, the posterior density has on average smaller variance and entropy than the prior density (Felsenstein, 2008). The denominator is called **evidence** or **marginal likelihood**, because θ is marginalised out and thus reduces the denominator to $\int f(x|\theta)\pi(\theta)d\theta = f(x)$. This integral often cannot be solved analytically and has to be approximated. Popular approximation methods are Markov Chain Monte Carlo (MCMC) or variational methods, which are, however, beyond the scope of this thesis.

Having introduced the most relevant concepts of probability theory as well as the concept of a likelihood function and Bayes' theorem, we can now move on to investigate their application in the domain of estimation problems.

3 Parameter estimation

Parameter estimation generally refers to the problem of inferring the true parameterisation of a distribution based on a limited amount of data (cf., Felsenstein, 2008; Koch, 2007). Popular examples are the estimation of the mean or variance of a given population or the probability of success, e.g. in a clinical test. Formally, when we have a random sample $X = (X_1, ..., X_n)$, where the random variables X_i follow a parametric distribution P under parameterisation θ from parameter space Θ, $X_i \sim (P_\theta)_{\theta \in \Theta}$, we may want to infer $\hat{\theta}_n$. This is the estimate of θ given $n \in \mathbb{N}$ realisations $x = (x_1, ..., x_n)$ of the random sample (Trutschnig, 2014).

In the following, I will discuss two approaches to estimate θ. First, I will introduce the frequentist solution called **maximum likelihood estimation**. The subsequent section discusses the Bayesian solution called **maximum a-posteriori** estimation. Note that in the beginning I will introduce these concepts in the context of point estimates, which are the 'best' estimates $\hat{\theta}_n$ for θ under the given approach. Afterwards, I will discuss the construction of intervals around point-estimates under both perspectives. As will be shown in the following, often maximum likelihood and maximum a-posteriori estimation will provide very similar results. The crucial difference between the two methods is the inclusion of prior knowledge in the Bayesian case, which can serve as an important factor for regularising the maximum likelihood solution and prevent *overfitting* (i.e., being to 'close' to the data) but at the same time can bias the estimation in an uncontrollable way. One crucial difference between the two approaches, however, lies in the definition of a parameter itself: under the frequentist perspective a parameter is an unknown constant whereas in the Bayesian perspective a parameter is treated as a random variable.

3.1 Maximum likelihood estimation

First, we need to introduce the likelihood-function:

Definition 3.1.1
We assume that random variable X follows a parameteric distribution P under parameterisation θ, $X \sim (P_\theta)_{\theta \in \Theta}$ and that P_θ is absolutely continuous with density f_θ. Further, $(X_1, ..., X_n)$ is a random sample and $x := (x_1, ..., x_n)$ a concrete realisation of that random sample. Then we call the function

$$\begin{aligned} \mathcal{L} : \Theta &\to [0, 1] \\ \theta &\mapsto f(x|\theta) := \prod_{i=1}^{n} f(x_i; \theta) \end{aligned}$$

the **likelihood function**.

Intuitively, the likelihood function can be understood as a mapping from possible parameter values $\theta \in \Theta$ (parameter space) to the probability of the observed data x under that parameterisation. Note that the likelihood function is not a probability density function and thus does not necessarily integrate to 1. We can now move on and define parameter estimation based on the likelihood function:

Definition 3.1.2
In the same setting as in Defintion 3.1.1, we define

$$\hat{\theta}_{MLE}(x) := \mathcal{L}(\hat{\theta}_n) = \underset{\theta \in \Theta}{sup}\ \mathcal{L}(\theta) = \underset{\theta \in \Theta}{argmax}\ f(x|\theta)$$

as the **maximum likelihood estimate** (**MLE**) of θ based on observations x. In general, such a θ does not necessarily always exist nor does it have to be one single value (in case of several maxima).

3.2 Maximum a-posteriori

In analogy to Defintion 2.3.2 and building on the concepts of Definition 3.1.1, we can now move on to define the Bayesian version of parameter estimation.

Definition 3.2.1
Bayesian parameter estimation rests on Bayes' theorem as defined above:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} = \frac{f(x|\theta)\pi(\theta)}{f(x)}$$

As noted above, $f(x|\theta) = \mathcal{L}(\theta)$ is the **likelihood function** and $\pi(\theta)$ is the prior density, which accounts for any **prior knowledge** we may have about the parameter space Θ. $f(x)$ is called the **evidence** or **marginal likelihood**, which serves as a normalisation factor to ensure that the **posterior distribution** $\pi(\theta|x)$ sums to one (cf., Kolmogorov axiom two). Thus, the main characteristic of the Bayesian account of parameter estimation can be reduced to

$$\pi(\theta|x) \propto f(x|\theta) \cdot \pi(\theta) = f(x,\theta)$$

The joint probability density $f(x,\theta)$ is called **generative model** and plays a central role in machine learning applications (Bishop, 2006), since it provides a probabilistic mapping from hidden states (latent variables) θ (e.g., the parameterisation of a distribution) to observables x (e.g., a random sample). This shows that a generative model is based on a likelihood function and a prior density and results from the application of the product rule. Applying Bayes' rule to compute a posterior density over θ can be understood as inverting the generative model to infer the most likely hidden states (parameterisation θ) underlying the observed data x.

Definition 3.2.2
Based on Definition 3.2.1, we can now define

$$\hat{\theta}_{MAP}(x) := \underset{\theta \in \Theta}{argmax}\ \pi(\theta|x) = \underset{\theta \in \Theta}{argmax}\ f(x|\theta) \cdot \pi(\theta)$$

as the **maximum a-posteriori estimate** (**MAP**) of θ based on observations x. As before, such a θ does not always exist nor does it have to be one single value.

Remark 3.2.3
Since the logarithm is a monotonic function, finding the maximum of $\pi(\theta|x)$ is equivalent to maximising $ln\pi(\theta|x)$, therefore:

$$\hat{\theta}_{MAP}(x) = \underset{\theta \in \Theta}{argmax}\ \pi(\theta|x) = \underset{\theta \in \Theta}{argmax}\ (lnf(x|\theta) + ln\pi(\theta))$$

This shows that in the case of a uniform prior on a bounded interval the ML and MAP estimates are identical, because $ln\pi(\theta)$ will be constant for all θ in Θ. This is an important result because this implies that MLE and MAP estimation will agree perfectly if the prior expresses a state of zero prior knowledge - a so called **uninformative prior**. As will be discussed in more detail in Section 5, the question whether it is possible to define truly uninformative priors for any given problem is an important topic of much current and past debate.

Remark 3.2.4
Note the close resemblance between $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MAP}$, where in fact the only difference between MLE and MAP is the inclusion of prior knowledge $\pi(\theta)$ into the estimation. Crucially, as will be shown in the following simulations, the influence of the prior on the estimation diminishes with increasing sample size $n \in \mathbb{N}$ of the random sample $(x_1, ..., x_n)$. Formally, this implies that the MLE and MAP estimate are asymptotically equivalent, i.e. $\lim\limits_{n \to \infty} (\hat{\theta}_{MAP} - \hat{\theta}_{MLE}) = 0$ (see below).

3.3 Simulation: estimating the probability of success

The standard introductory example to illustrate the difference between MLE and MAP is estimating the probability of success, for example the probability θ of observing heads when tossing an (unfair) coin. This is an artificially simple example with only one free parameter that is restricted to [0,1]. This implies that it is possible to express zero prior knowledge and that we can easily derive a formal rule for specifying the prior (although even in this simple example several rules for defining the prior density exist, as will be discussed in Section 5).

We can assume that we are running an experiment with success probability θ n times n-times, $n \in \mathbb{N}$, and have observed a random sample $(x_1, ..., x_n)$, where x_i=1 if flip i is heads and 0 otherwise. Thus, $n_h = \sum_{i=1}^{n} x_i$ gives us the number of observed heads and $n - n_h$ the number of observed tails. The likelihood function for this problem is a Bernoulli distribution:

$$f(x|\theta) = \theta^{n_h} \cdot (1-\theta)^{n-n_h}$$

Thus, the maximum likelihood estimate of θ becomes

$$\hat{\theta}_{MLE}(x) = \underset{\theta\in\Theta}{argmax}\ f(x|\theta) = \underset{\theta\in\Theta}{argmax}\ (\theta^{n_h} \cdot (1-\theta)^{n-n_h}).$$

Given that, as noted above, the logarithm is a monotonic function and thus maximising $f(x|\theta)$ is equivalent to maximising

$$lnf(x|\theta) = n_h \cdot ln\theta + (n - n_h) \cdot ln(1-\theta)$$

We can now easily derive the maximum-likelihood solution by hand:

$$\begin{aligned} 0 = \frac{\partial lnf(x|\hat{\theta})}{\partial\hat{\theta}} &= \frac{n_h}{\hat{\theta}} - \frac{n-n_h}{1-\hat{\theta}} \\ \Leftrightarrow n_h \cdot (1-\hat{\theta}) &= (n-n_h) \cdot \hat{\theta} \\ \Leftrightarrow n_h - n_h \cdot \hat{\theta} &= n \cdot \hat{\theta} - n_h \cdot \hat{\theta} \\ \Leftrightarrow \hat{\theta} &= \frac{n_h}{n} \end{aligned}$$

This result also holds true for the special cases $n_h = n$ and $n_h = 0$ (cf., Trutschnig, 2014, p 60). It follows that the maximum likelihood estimate of θ is simply the proportion of observed heads in the random sample:

$$\hat{\theta}_{MLE}(x) = \frac{n_h}{n}$$

To obtain the MAP-estimate, we need to define a prior over θ. The most prominent choice for $\pi(\theta)$ is a Beta distribution $\theta \sim B(\alpha, \beta)$, i.e.

$$\pi(\theta) = Beta(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$

where α and β are (scale and rate) parameters that determine the shape of the beta distribution and $B(\alpha, \beta) = \int_0^1 t^{\alpha-1} \cdot (1-t)^{\beta-1} dt$ is the beta function.

Remark 3.3.1
The choice of a Beta distribution has several advantages. First, it is a suitable distribution for a parameter that lies in the interval of [0,1]. Second, when choosing a Beta distribution as a prior we are in the rare situation where we can explicitly formulate zero prior knowledge, as will be shown when we compute the MAP estimate below. Finally, a Beta distribution is a conjugate prior for a Bernoulli likelihood. That means that the posterior follows the same distribution as the prior density, such that in our example the posterior distribution will be a Beta distribution, which we can **proof** easily:

We can simply observe the form of the posterior based on a Bernoulli likelihood function and a Beta prior density

$$\pi(\theta|x) = \frac{f(x|\theta) \cdot \pi(\theta)}{f(x)} = \frac{\theta^{n_h} \cdot (1-\theta)^{n-n_h} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}}{f(x) \cdot B(\alpha, \beta)}$$

and thus

$$\pi(\theta|x) \propto \theta^{\alpha+n_h-1} \cdot (1-\theta)^{\beta+n-n_h-1} = Beta(\theta; \alpha', \beta')$$

with $\alpha' = \alpha + n_h$ and $\beta' = \beta + n - n_h$.

■

This result shows that if the prior is conjugate, the update from prior to posterior density only updates the **hyperparameters** of the density function, i.e. the parameters that govern the shape of the density function (cf., Felsenstein, 2008 p36).

We can now derive the MAP estimate of θ just as before by noting that maximising $\pi(\theta|x)$ is equivalent to maximising

$$\begin{aligned} lnf(x|\theta) + ln\pi(\theta) = & n_h \cdot log(\theta) + (n - n_h) \cdot log(1-\theta) \\ & + (\alpha - 1) \cdot log(\theta) + (\beta - 1) \cdot log(1-\theta) - log(B(\alpha, \beta)) \end{aligned}$$

In consequence, we can now derive the MAP estimate in analogy to the ML estimate:

$$\begin{aligned} 0 = \frac{\partial log(\pi(\hat{\theta}|x))}{\partial \hat{\theta}} &= \frac{n_h}{\hat{\theta}} - \frac{n - n_h}{1 - \hat{\theta}} + \frac{\alpha - 1}{\hat{\theta}} - \frac{\beta - 1}{1 - \hat{\theta}} \\ \Leftrightarrow \hat{\theta} &= \frac{n_h + \alpha - 1}{n + \beta - 1 + \alpha - 1} \end{aligned}$$

and obtain

$$\hat{\theta}_{MAP}(x) = \frac{n_h + \alpha - 1}{n + \beta - 1 + \alpha - 1}$$

This is an important result, because it shows that when we choose a uniform prior distribution parameterised by $\alpha = 1 = \beta$, we find that $\hat{\theta}_{MLE}(x) = \hat{\theta}_{MAP}(x)$, which we can verify in a simulation easily (Figure 1). For simplicity here and in all subsequent simulations θ was discretised (with step-size 0.001). This does not affect the interpretation of the results in any way but simplifies the computation - particularly in the example of improper priors discussed in 5.5.2.

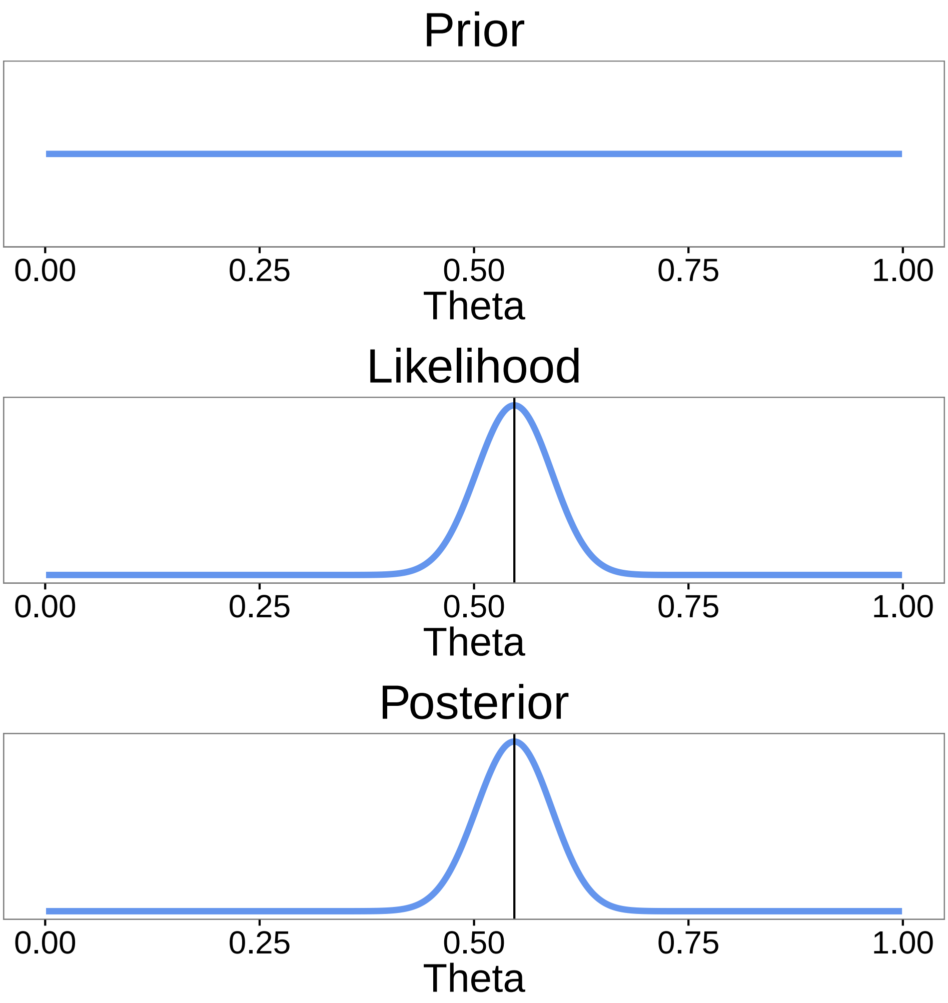

Figure 1. Results of simulating a random sample in a coin toss experiment using a Bernoulli likelihood under a true θ of 0.5 and a sample size of 128. When choosing a uniform prior,

we see that the maximum-likelihood estimate and the maximum a-posteriori estimate of θ are in perfect agreement (black vertical lines) and close to the true value of 0.5.

This also provides a straightforward and formal way of how we can define a prior density: α should reflect the number of heads (n_h) we have previously encountered and β reflects how often we have seen tails ($n - n_h$). Therefore, this is an example for which we find an objective procedure of how to specify the prior density. Furthermore, it is possible to express zero prior knowledge in terms of a uniform distribution.

From the above results we can also derive a second important result, namely the strong consistency of both the MLE and MAP estimate of θ.

Remark 3.3.2
According to the **strong law of large numbers** (SLLN), for a series of identically distributed, pairwise independent and integrable random variables in $(\Omega, \mathcal{A}, \mathcal{P})$ the sample mean of observations converges almost surely to the true expected value (in our example the probability of success θ), i.e. (Trutschnig, 2014 p 41):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \lim_{n\to\infty} \overline{X_n} = \mathbb{E}(X) = \theta \ [P]$$

From the SLLN we can directly derive the **strong consistency**, i.e. $\hat{\theta_n} \to \theta \ [P]$ (Trutschnig, 2014 p 55), for both estimators:

$$\lim_{n\to\infty} \hat{\theta}_{MLE}(x) = \lim_{n\to\infty} \frac{n_h}{n} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i}{n} = \lim_{n\to\infty} \overline{X_n} = \theta \ [P]$$

and

$$\lim_{n\to\infty} \hat{\theta}_{MAP}(x) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n + \alpha + \beta - 2} = \lim_{n\to\infty} \frac{\overline{X_n} + \frac{\alpha-1}{n}}{1 + \frac{\alpha+\beta-2}{n}} = \theta \ [P]$$

This is an important result, because it suggests that according to the SLLN both estimators are strongly consistent, i.e. converge almost surely to the true value of θ for every $\theta \in [0, 1]$. Practically, this means that the prior distribution is particularly influential in small sample sizes, but will be less important if the sample size increases (because both estimators have the same limit with increasing sample size). This can be illustrated if we simulate a prior far off the true theta ($\alpha = 22, \beta = 4$) for small (Figure 2) and large (Figure 3) sample sizes. We find that the maximum a-posteriori estimate is severely biased in a small sample but much closer to the maximum likelihood estimate in larger sample sizes.

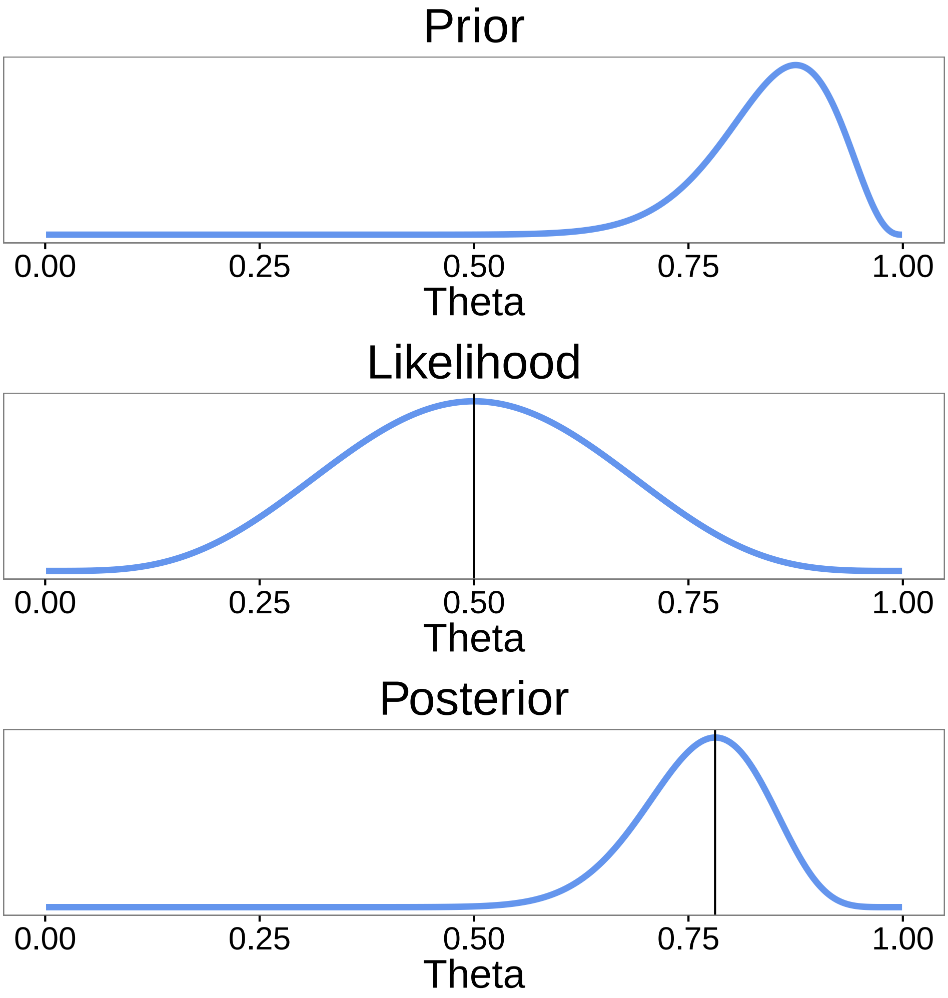

Figure 2. *In a small sample size of $n = 8$, the posterior is heavily influenced by a prior that is far off the true θ of 0.5 (prior density parameterised with $\alpha = 22, \beta = 4$).*

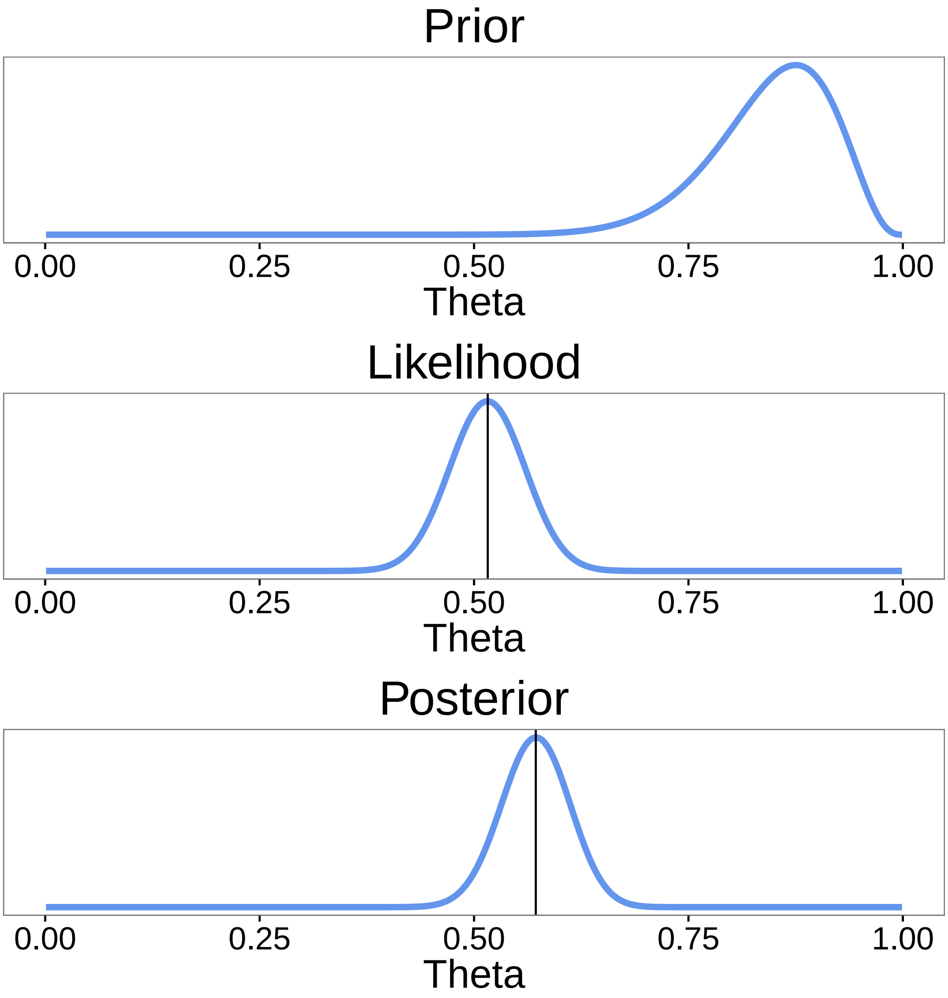

Figure 3. In a large sample size of $n = 128$, the posterior is much less influenced by a prior that is far off the true θ of 0.5 (prior density parameterised with $\alpha = 22, \beta = 4$).

In practice, however, the simulations above will rarely be very useful for realistic applications. On the one hand, we usually do not want to compute point-estimates but intervals that tell us something about the uncertainty in our estimates. On the other hand, we are often confronted with much more complicated problems, for example unrestricted parameter spaces where finding the correct prior density is less straightforward. These situations will be discussed in Sections 4 and 5, respectively.

4 Confidence intervals versus credibility intervals

When we infer parameters we usually want to measure our uncertainty in our estimate in addition to the most likely point estimate. In this instance the frequentist and Bayesian perspective diverge: while frequentists think of a parameter as an unknown constant around which we would like to compute a **confidence interval**, the Bayesian school treats the parameter as a random variable around which we can derive a **credibility** or **highest densitiy interval** from the posterior. In the following sections these two approaches will be defined and compared. I will start with the initial introductory example of a coin toss and will then move on to more interesting problems including unconstrained parameter spaces in Section 5.

4.1 Confidence intervals

Confidence intervals are the most commonly used intervals around estimators in frequentist statistics and are defined in the following way:

Definition 4.1.1
For a parameter estimate $\hat{\theta}$, an approximate **confidence interval** around θ is defined as:

$$CI = [\hat{\theta} \mp z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\mathbb{V}_{\hat{\theta}}}{n}}]$$

This approximation is based on the assumption that the parameter estimate reflects the true expected value. Approximate confidence intervals can be derived from the **central limit theorem**, which states that the distribution function of parameter estimates converges into the distribution function of a normal distribution with increasing n (Trutschnig, 2014). In the above definition, $x = \{x_1, ..., x_n\}$ is a random sample of size n, $\mathbb{V}_{\hat{\theta}}$ is the variance of the random sample under parameterisation $\hat{\theta}$ and $z_{1-\frac{\alpha}{2}}$ is the critical value of a normal distribution under significance level $(1-\alpha)$, which is usually set to 95 %. As noted above, from a frequentist perspective the parameter estimate is treated as an unknown constant and the lower and upper bound (L_n and U_n) of the confidence interval are random variables, respectively. Therefore, the correct interpretation of a (1-α) confidence interval is if we were to compute an infinite number of such confidence intervals around $\hat{\theta}$, they would include the true parameter θ in a proportion of (1-α) of all intervals, for example in 95% of all cases. A common misconception about confidence intervals is that they include the true parameter θ with probability (1-α). Formerly, a $(1-\alpha)$- confidence interval is defined as the interval between the random variables L_n and U_n, such that

$$P_\theta(L_n < \theta < U_n) = 1 - \alpha,$$

where P_θ is the probability based on the true parameter θ.

Example 4.1.2
If we take the above example of inferring the probability of success when we toss a coin, then the approximate confidence interval around $\hat{\theta}$ becomes:

$$[L_n, U_n] = [\hat{\theta} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\theta} \cdot (1-\hat{\theta})}{n}}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\theta} \cdot (1-\hat{\theta})}{n}}]$$

4.2 Highest density intervals

As noted above, from the perspective of Bayesian inference, the estimated parameter is a random variable that we can infer from the posterior distribution as obtained by Bayes' theorem. From this distribution we can extract both the maximum value $\hat{\theta}_{MAP}(x)$ but also a measure of variability or uncertainty around the maximum a-posteriori estimate. If there is very high certainty in a point estimate, then the obtained posterior distribution will have very low variance and vice versa. The most common measure of uncertainty around a MAP estimate is by defining the following interval:

Definition 4.2.1
The smallest interval of a posterior distribution, which includes the mode of the distribution as well as (1-α)$\cdot$100% of the mass, is called **(1-α)-credibility** or **highest density interval**. Note that in the case of a bi- or multi-modal posterior distribution, the highest density interval can be the union of several disjoint intervals.

It is important to note that several different definitions for credible and confidence intervals exist, such as the Pearson-Clopper or the Wald-type interval in the latter case. Discussing these intervals, however, goes beyond the scope of this thesis, which aims at comparing the two different conceptions of an interval and thus relies on the most popular definitions. For the following simulations it is important to keep in mind that unlike credible intervals, confidence intervals are not necessarily the smallest (freqeuentist) intervals around an estimate. In general, an increasing sample size will lead to a predominance of the likelihood in the posterior density. Thus, the highest density interval of the posterior will correspond to the highest density interval of the likelihood function. Note that this is not the same as the definition of a confidence interval, but there will be a close correspondence as illustrated in the simulations below.

4.3 Simulation: estimating the interval around the probability of success

We can now repeat the above simulations but in addition to estimating the probability of success, we can now also compute the confidence intervals around the ML-estimates and the highest density intervals around the MAP-estimates. In the following simulations α will always be set to 0.05.

Figure 4 shows a comparison of confidence intervals around ML estimates and credible intervals around MAP estimates under a uniform prior and a true θ of 0.5 ($n = 128$). We observe again that the MAP and ML estimates coincide, but also that the two intervals are very similar in these 20 simulations. For these simulations, the approximate confidence intervals as defined in Definition 4.1.1 were used. While there are other ways to compute (or approximate) confidence intervals, such as Pearson-Clopper intervals that tend to be slightly larger, these differences do not affect the general results of these comparative simulations. In consequence, small differences in the size of the intervals should not be over-interpreted.

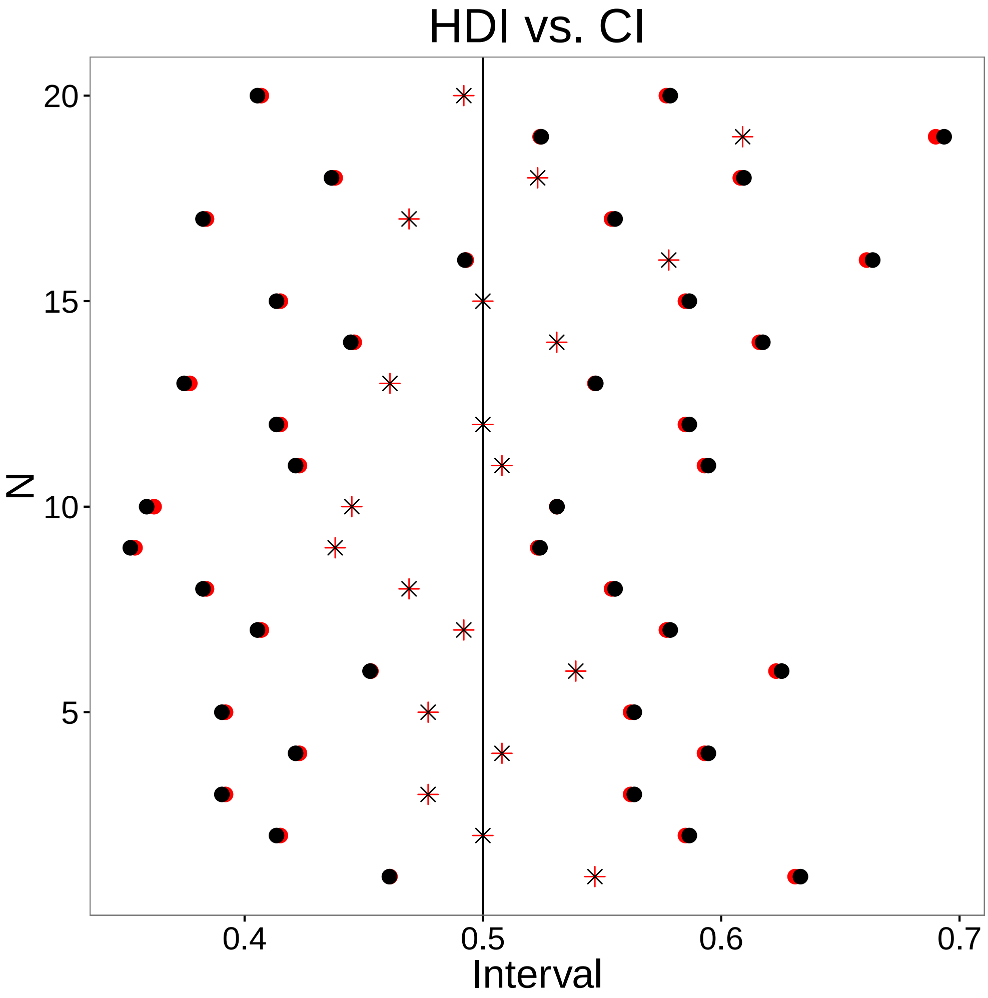

Figure 4. *Comparing 20 95% confidence intervals (black dots) around ML estimates (black asterisks) and credible intervals (red dots) around MAP estimates (red crosses) under a uniform prior and a true θ of 0.5 (black vertical line) in simulations with a sample size of* $n = 128$.

If we increase the number of simulations to 1000, we find that both intervals fail to capture the true parameter value about equally often and close to the expected proportion of failure of 0.05 or 5% (Figure 5). These simulations show that - in analogy to the point estimates - under a uniform prior the frequentist and Bayesian intervals will be highly consistent.

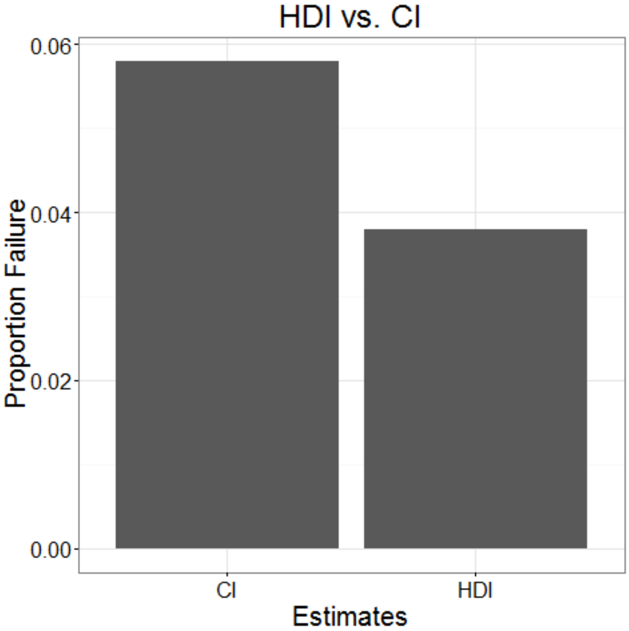

Figure 5. *Comparing the proportion of failure to capture the true θ of 0.5 in confidence intervals and credible intervals under a uniform prior (expected proportion of failure is 0.05, 1000 simulations using a sample size of 128). We see that both approaches have a very similar proportion of failure that is close to the expected value.*

We can now repeat our investigations above and test for the effects of a prior that is far off the true value. When we repeat the simulations with the biased prior introduced in Figure 2 and use a small sample size of $n = 8$, we find, unsurprisingly, that confidence intervals strongly outperform credible intervals (Figure 6).

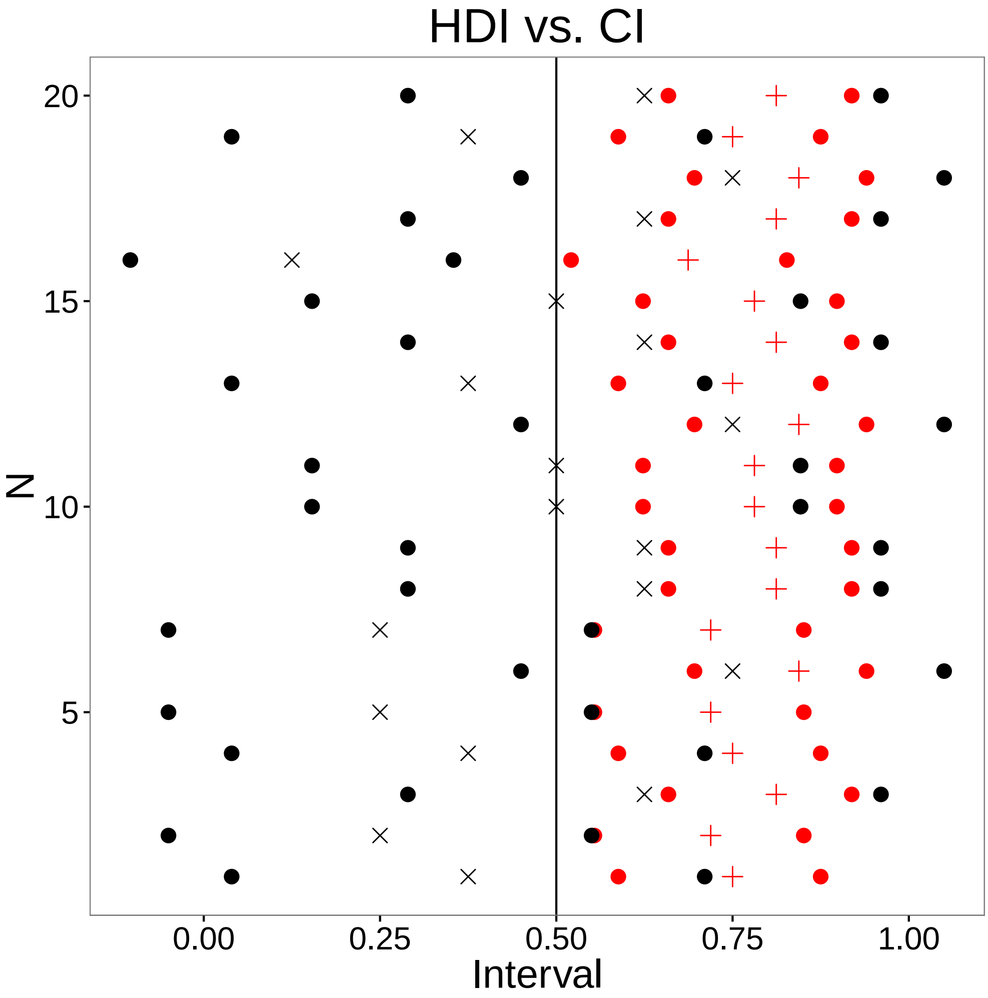

Figure 6. *Comparing 95% confidence intervals (black dots) around ML estimates (black asterisks) and credible intervals (red dots) around MAP estimates (red crosses) under a prior that is strongly biased towards higher values of θ as shown in Figures 2 and 3, using a small sample size of $n = 8$ and a true θ of 0.5 (black vertical line). As we would expect, we observe that credible intervals in contrast to confidence intervals fail to capture the true θ of 0.5*

Under these (extreme) conditions, we find that the credible interval fails to capture the true parameter every time in 1000 simulations (Figure 7).

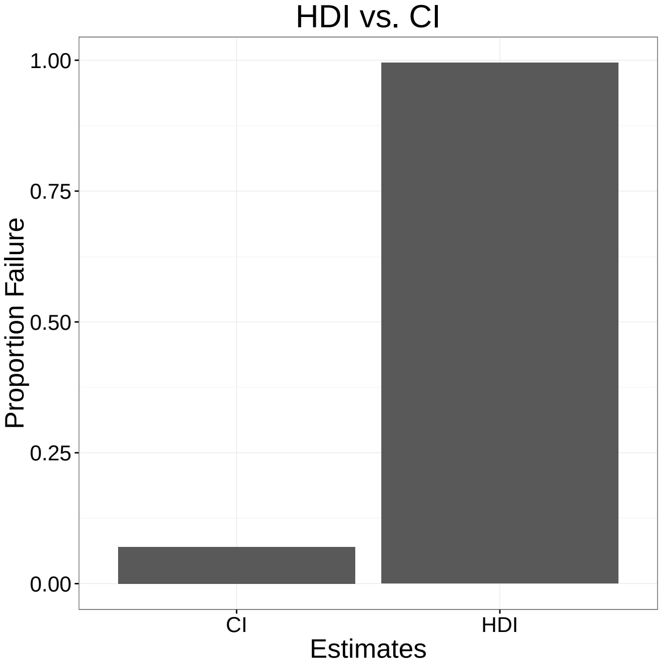

Figure 7. Comparing the proportion of failure to capture the true θ of 0.5 for confidence intervals and credible intervals under a prior that is strongly biased towards higher values of θ as shown in Figures 2 and 3, using a small sample size of $n = 8$ and a true θ of 0.5 (expected proportion of failure is 0.05).

However, when we increase the sample size from $n = 8$ to $n = 128$, we find that the credible interval is much less influenced by a 'malicious' prior as displayed in Figure 8. This result resonates with the results from Section 3.3, where it was shown that the MAP estimate is close to the ML estimate in large enough sample sizes, even if the prior is far off the true value of θ.

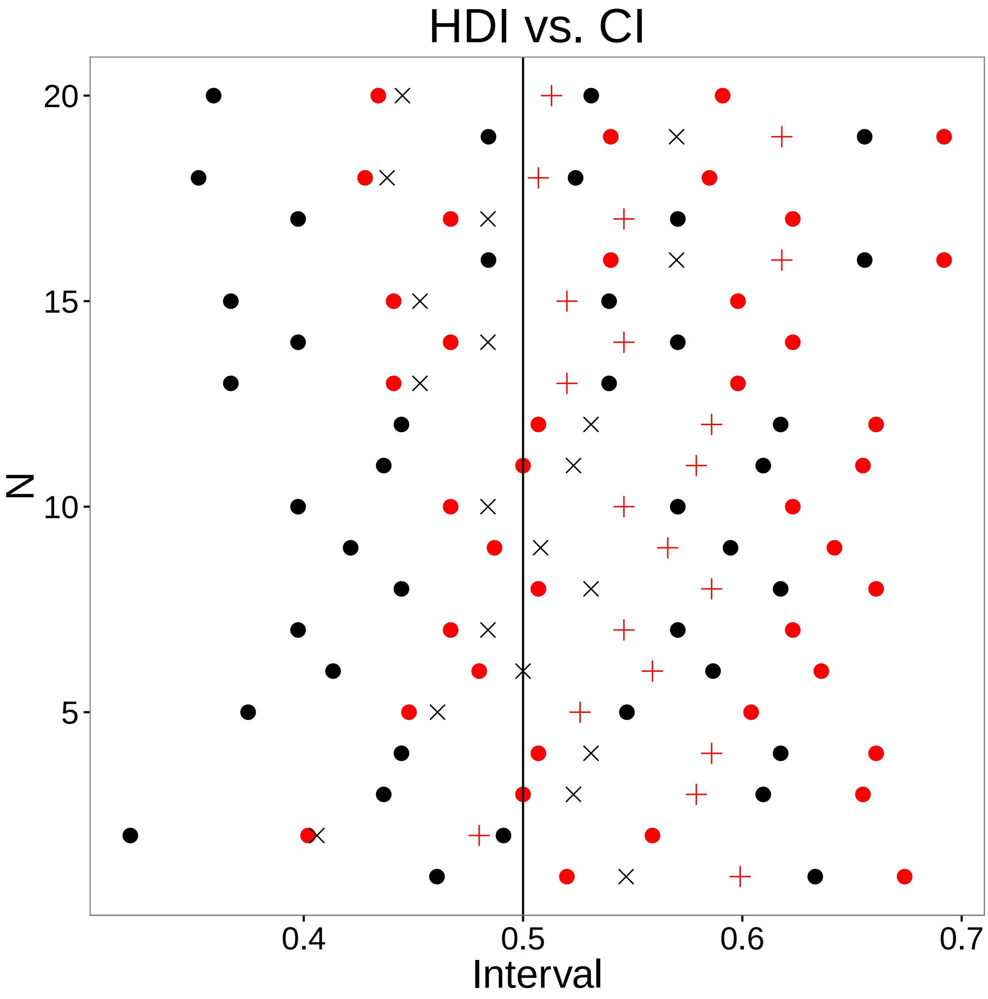

Figure 8. *Comparing 95% confidence intervals (black dots) around ML estimates (black asterisks) and credible intervals (red dots) around MAP estimates (red crosses) under a prior that is strongly biased towards higher values of θ as shown in Figures 2 and 3 but using a larger sample size of $n = 128$ and a true θ of 0.5 (black vertical line). We see that in larger sample sizes the posterior distribution is much less affected by 'malicious' priors, such that the credible capture the true parameter value much more often.*

We can see that in 100 simulations the proportion of failure to capture the true θ of credible intervals based on a prior that is far off the true parameter value is about four times higher than the proportion of failure for confidence intervals (Figure 9). Note that the proportion of failure for the credible interval showed a substantial decrease from 100% to 20% when increasing the sample size of the random sample.

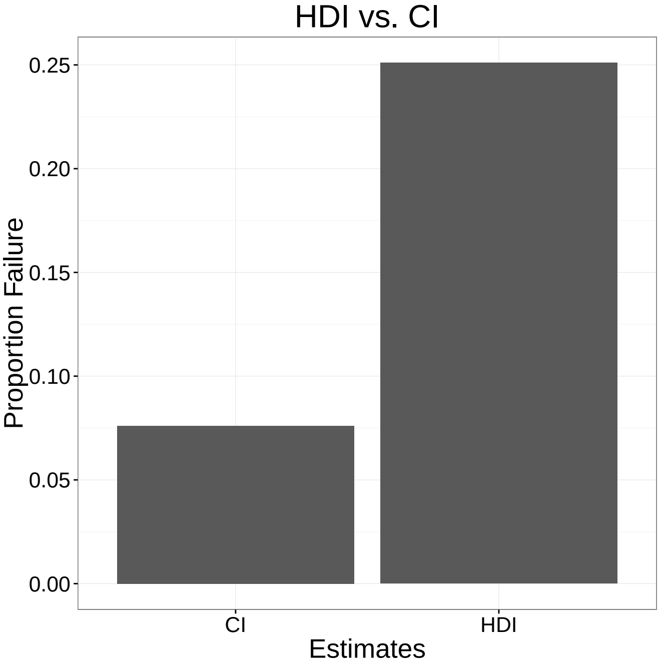

Figure 9. Comparing the proportion of failure to capture the true θ of 0.5 for confidence intervals and credible intervals under a prior that is strongly biased towards higher values of θ as shown in Figures 2 and 3 but using a large sample size of $n = 128$ (expected proportion of failure is 0.05). We observe that the proportion of failure of Bayesian intervals is about four times higher than the proportion of failure of frequentist intervals.

We have now discussed the process of both point-wise and interval/based parameter estimation under a frequentist and Bayesian approach. However, so far we have used a highly simplified example of estimating only one parameter in a restricted parameter space between 0 and 1. Much more relevant for investigating the practical differences between frequentist and Bayesian parameter estimation, however, are more complicated problems and situations when the parameter space is unrestricted, such as when we want to estimate the mean or variance of a Gaussian. In the previous simulations I illustrated a problem for which there exists a straightforward formal rule about ho to specify the prior density. Here, the formal rule was using a beta distribution parameterised by scale and rate parameters *alpha* and *beta*, which reflect the previously encountered number of successes and failures in Bernoulli trials. Importantly, this formal rule also accounts for the case of a complete absence of prior knowledge, i.e. the case of no previous encounters of successes of failures, by using a uniform distribution between 0 and 1. In the next section I will discuss several attempts to specify more general formal rules for defining prior densities, including problems with unrestricted parameter spaces, which leads to the discussion of the general role of priors in Bayesian inference. In the past, many formal rules for specifying the prior and thus developing an 'objective' Bayesianism have been suggested, where the Jeffreys' prior and Jaynes' maximum entropy principle are amongst the most influential proposals. However, as we will see, none of the proposed formal rules are truly capable of accounting for zero prior knowledge in an unrestricted parameter space and hence will always induce some sense of bias in parameter estimation.

5 The role of priors in Bayesian inference

In this section I will illustrate the most controversial and perhaps most interesting difference between Bayesian inference and MLE: the role of prior densities. It is easy to show that Bayesian inference can outperform frequentist parameter estimation as soon as the estimation problem becomes more complex (e.g., hierarchical estimation or a larger number of free parameters with high inter-dependencies) or the sample size becomes small, as will be shown below. In these instances, Bayesian inference can outperform a frequentist approach because it allows for the incorporation of prior knowledge. However, the critical question then becomes how to specify the prior density to provide an accurate account of the prior state of knowledge. This is a particularly critical question in situations where we do not have any prior knowledge at all. The role of the prior density is to bias or regularise the maximum likelihood estimation and trade-off current observations with previous knowledge. In the absence of any prior knowledge we need to find a prior that expresses zero knowledge and thus leaves the estimate unbiased, resulting in Remark 3.2.3. As we will see below, despite several previous attempts there is no general rule for specifying the prior density that truly accounts for the absence of prior knowledge and thus allows for an unbiased parameter estimate.

5.1 Regularising maximum likelihood

The functional role of priors in Bayesian inference can be understood as a regularisation of maximum likelihood estimation. This can become particularly useful in problems with only a small number of observations or in more complex problems that require hierarchical modelling (for a detailed discussion see Gelman, 2013), which explains its popularity in fields such as machine learning (Bishop, 2006). A popular criticism of maximum likelihood estimation is a tendency to 'overfit' the data (cf., Bishop, 2006, p. 10), which in this context can be understood as being too close to the observed data as illustrated in Figure 10.

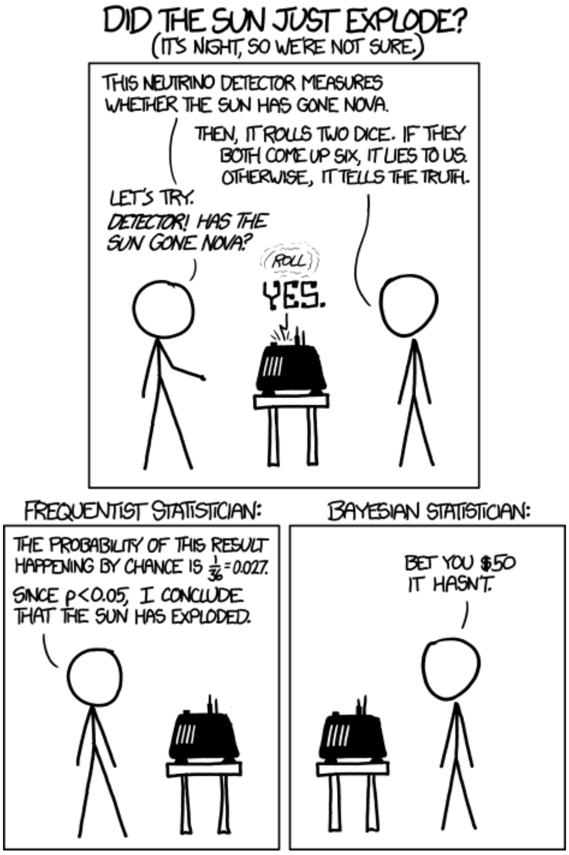
DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Figure 10. A common criticism of maximum likelihood (frequentist) parameter estimation is the fact that the estimation is too heavily influenced by observed data, particularly in a small sample size (retrieved from https://xkcd.com/1132/)

More formally, we can imagine an experiment as simulated above with a probability of success $\theta \in [0, 1]$, in which we observe three successes (n_h) in a row. The maximum likelihood solution for the underlying probability of success would be

$$\hat{\theta}_{MLE}(x) = \frac{n_h}{n} = \frac{3}{3} = 1$$

However, even if we use a strongly biased prior as in the examples before with $\alpha = 22, \beta = 4$, we obtain a more moderate estimate of θ (see Figure 11), namely

$$\hat{\theta}_{MAP}(x) = \frac{n_h + \alpha - 1}{n + \beta - 1 + \alpha - 1} = \frac{24}{27} = 0.89$$

If we would use a more uninformative (i.e., symmetrical and with higher variance) prior with $\alpha = 4, \beta = 4$, we would obtain a much more moderate estimate of

$$\hat{\theta}_{MAP}(x) = \frac{n_h + \alpha - 1}{n + \beta - 1 + \alpha - 1} = \frac{2}{3}$$

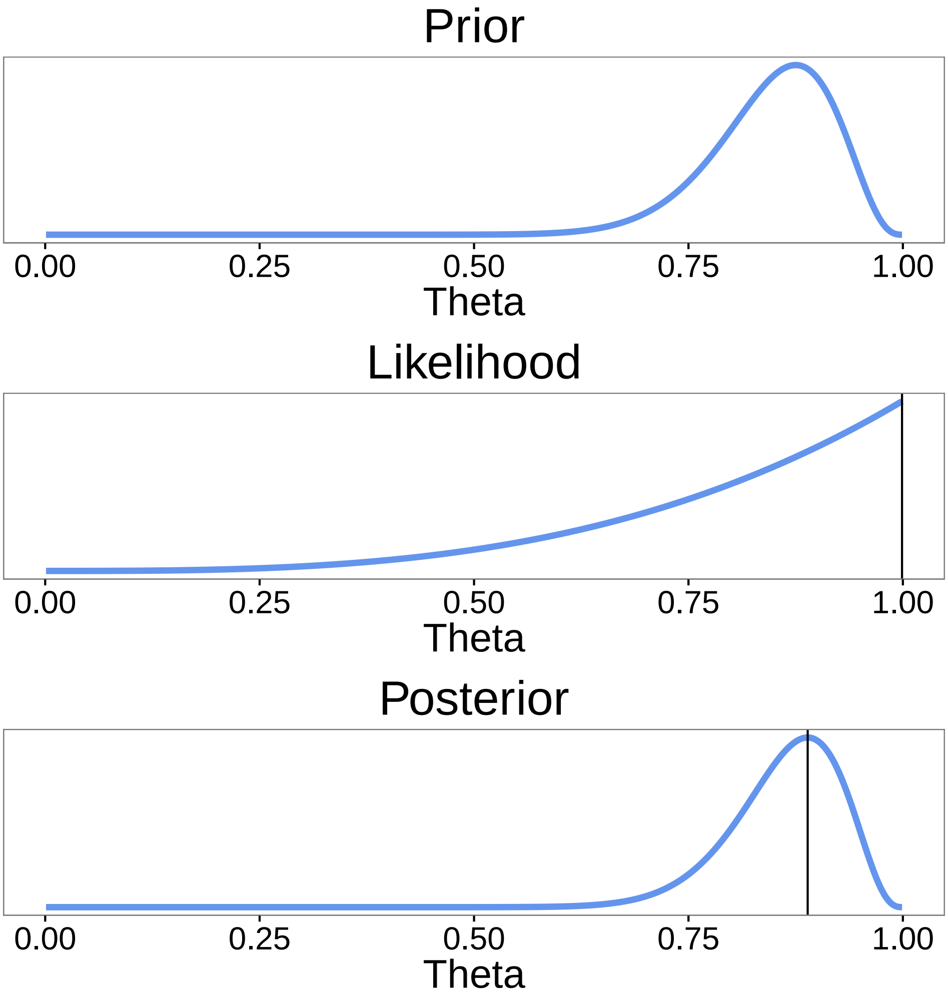

Figure 11. *Even when we use a strongly biased prior, we find that the maximum a-posteriori estimate of θ is much less biased by an extreme sample (3 heads in a row) than a maximum likelihood solution, which provides an estimate of $\hat{\theta}_{MLE}(x) = 1$. Note that this effect is particularly strong in small sample sizes. These problems can be connected to other statistical phenomena that have attracted much recent interest, such as the 'replication crisis' or the 'winner's curse' (e.g., Button et al., 2013)*

This example demonstrates that it can be useful to use an informative prior. The most important question then, however, is how to specify an appropriate prior. One of the main criticisms of Bayesian methods is the degree of arbitrariness in defining a prior density, or, equivalently, a lack of objectivism (Efron, 1986). In the following, I will discuss some of the most important formal rules for specifying a prior density that have been proposed in the past, attempting to create an 'objective Bayesianism', and test their practicability in simulated examples.

5.2 Formal rules and uninformative priors

The development of formal rules to define a prior, which are often referred to as *reference priors* (Bernardo, 1979; Berger, 2004), has been a topic of much interest over the past decades (cf., Kass & Wassermann, 1996). In general, within the 'Bayesian school' one can dissociate a subjectivist and an objectivist tradition. The subjectivist school (de Finetti, 1972; Savage 1954) has proposed that the notion of a prior probability is inherently subjective and will always represent an observer's subjective degree of belief. In contrast, the objective tradition in Bayesian inference argues that, given the same prior knowledge, everybody should converge to the same prior density used for Bayesian inference (Jaynes, 1968). The most influential proponents of objective Bayesianism were Harold Jeffreys and Edwin Jaynes. Both of them have proposed a formal rule for specifying a prior, namely the Jeffreys' prior and the maximum entropy principle, which will be discussed below. A critical question for a generic formal account of prior specification is not only how one converges to the same prior given the same knowledge, but also whether it is possible to express true ignorance. While in the above example of a coin toss a uniform prior will express zero prior knowledge and thus will lead to the unbiased result of Remark 3.2.3, the use of a uniform prior becomes impossible in problems where the parameter space is unrestricted. Before discussing the solutions proposed by Jeffreys and Jaynes, I will mention two general desiderata that have been proposed for the formal definition of a prior, based on the following definition by Kass & Wassermann (1996):

Definition 5.2.1
In a finite parameter space, we refer to the **partitioning paradox** in a situation where the prior probabilities for single events change in case we partition the parameter-space (see example below). In a finite or infinite parameter space, a prior is called **parameterisation invariant** if the information encoded in the prior does not depend on the parameterisation of the prior density.

Example 5.2.2
In the most simple example of the partition problem, we can think of a parameter space consisting of two values, $\Theta = \{\theta_1, \theta_2\}$, in which case assigning a uniform, 'uninformative prior' would give us $\pi(\theta_1) = \pi(\theta_2) = \frac{1}{2}$. If we now imagine a second parameter space $\Phi = \{\phi_1, \phi_2, \phi_3\}$, and a mapping between Θ and Φ that states that $\theta_1 = \phi_1$ (other relations arbitrary), then a uniform prior on Φ takes the values $\mu(\phi_1) = \mu(\phi_2) = \mu(\phi_3) = \frac{1}{3}$. We end up with the paradox that $\pi(\theta_1) = \frac{1}{2} \neq \frac{1}{3} = \mu(\phi_1)$, whereas $\theta_1 = \phi_1$ by design.

Remark 5.2.3
The partitioning paradox is a special (finite-set) case of the lack of parameterisation invariance (Kass & Wassermann, 1996). This can be seen easily by assuming a mapping $g : \Phi \rightarrow \Theta$, which can be understood as a re-parameterisation from Φ to Θ. A prior μ on Φ then induces a prior π on Θ, such that $\pi(\theta) = \mu(g^{-1}(\theta))$. In Example 5.2.2, we would obtain $g(\omega_1) = \theta_1$. In general, if μ is uniform, then e^{μ} or $ln\mu$ will not be uniform and thus not invariant to re-parameterisation. This is one reason speaking against the general use of uniform priors as a reference prior for zero prior knowledge in an infinite or unbounded parameter space (see below).

In practice, we are are looking for formal rules for prior specification that are immune to the partitioning problem and that are parameterisation invariant.

5.3 Uniform prior

In the absence of any prior knowledge, both Thomas Bayes and Pierre-Simon Laplace proposed to assign equal probability to each event, also known as the *principle of insufficient reason* (Bayes, 1763; Laplace, 1812). However, this procedure, resulting in a uniform prior and the result of Remark 3.2.3 is limited because it fails to provide priors that are parameterisation invariant in many cases (e.g., unbounded parameter spaces). Furthermore, as soon as the parameter space becomes uncountable it is not straightforward how to determine a uniform prior without obtaining an *improper prior* (see below).

5.4 Jeffreys' prior

Harold Jeffreys was perhaps the most influential individual in attempting to develop an 'objective Bayesianism' (Kass & Wassermann, 1996; Jeffreys, 1939; Ghosh, 2011). While he argued that there cannot only be one logically correct prior to express ignorance, he initially proposed a prior density that is constant (an *improper prior* for unrestricted parameter spaces). Eventually, in a publication in 1946 (Jeffreys, 1946), he proposed the following general approach for defining a prior density:

Definition 5.3.1
We call

$$I(\theta) = \mathbb{E}_\theta \left(\frac{\partial ln \mathcal{L}(x;\theta)}{\partial\theta} \right)^2$$

the **Fisher information**, where $\mathcal{L}(x;\theta)$ is the likelihood as defined in Definition 3.1.1. Then, the **Jeffreys prior** is defined as

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

(or $\pi(\theta) \propto \sqrt{det(I(\theta))}$ in the multi-parameter case).

The motivation for Jeffreys proposal was that the Fisher-Information can be understood as the amount of information about θ that is provided by the distribution (Syversveen, 1998) and that the corresponding prior has the desired property of invariance to parameter transformation. The inverse of the Fisher information also provides a lower bound on the variance of a parameter estimate (cf., Trutschnig, 2014, p 57). Here, however, the Fisher information is understood more generally as the information that random variable X carries about parameter θ (Lehmann and Casella, 1998, p 115).

Remark 5.3.2
The invariance to parameter transformation for the Jeffreys prior follows directly from the change of variables theorem (cf., Kass and Wasserman, 1996). If we have a re-parameterisation $\theta \mapsto \gamma$, then $\pi(\gamma) \propto \sqrt{I(\gamma)}$ can be derived directly from $\pi(\theta) \propto \sqrt{I(\theta)}$.

Proof: (single-parameter case)
The result follows directly from the following calculation:

$$\begin{aligned}
\pi(\gamma) = \pi(\theta) \cdot \left|\frac{\partial\theta}{\partial\gamma}\right| &\propto \sqrt{I(\theta)\left(\frac{\partial\theta}{\partial\gamma}\right)^2} \\
&= \sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial ln\mathcal{L}}{\partial\theta}\right)^2\right]\left(\frac{\partial\theta}{\partial\gamma}\right)^2} \\
&= \sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial ln\mathcal{L}}{\partial\theta}\frac{\partial\theta}{\partial\gamma}\right)^2\right]} \\
&= \sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial ln\mathcal{L}}{\partial\gamma}\right)^2\right]} \\
&= \sqrt{I(\gamma)}
\end{aligned}$$

■

5.5 Simulating the Jeffreys' prior

To compute the Jeffreys prior, one simply needs the Fisher information based on the likelihood function for a given problem. In the following, I will illustrate the use of a Jeffreys prior in the experiment with success probability θ as illustrated above as well as in a problem with an unrestricted parameter space (estimating the mean of a normal distribution).

5.5.1 Bernoulli distribution

In a Bernoulli trial, the Jeffreys prior over θ is

$$\pi(\theta) \propto \sqrt{I(\theta)} = \sqrt{\mathbb{E}(\frac{n_h}{\theta} - \frac{n-n_h}{1-\theta})^2} = \sqrt{\theta \cdot (\frac{1}{\theta} - \frac{0}{1-\theta})^2 + (1-\theta)\cdot(\frac{0}{\theta} - \frac{1}{1-\theta})^2} = \frac{1}{\sqrt{\theta\cdot(1-\theta)}}$$

where we use the fact that

$$\frac{\partial}{\partial\theta}(n_h \cdot log(\theta) + (n-n_h)\cdot log(1-\theta)) = \frac{n_h}{\theta} - \frac{n-n_h}{1-\theta}$$

This shows that the Jeffreys prior does not depend on sample size n and the Fisher information can be computed based on a sample size of 1 (cf., Trutschnig, 2014 p 57). As can be seen in Figure 12, the Jeffreys prior assigns more mass on the extremes of the parameter space $\Theta = [0, 1]$. Thus, this prior is biased towards categorical (0 or 1) decisions, such as the presence or absence of a disease, rather than values in between, such as a coin toss of a fair coin. As Figure 12 shows, this also implies that the Jeffreys prior falls short of the same 'overfitting' problem illustrated in 5.1.

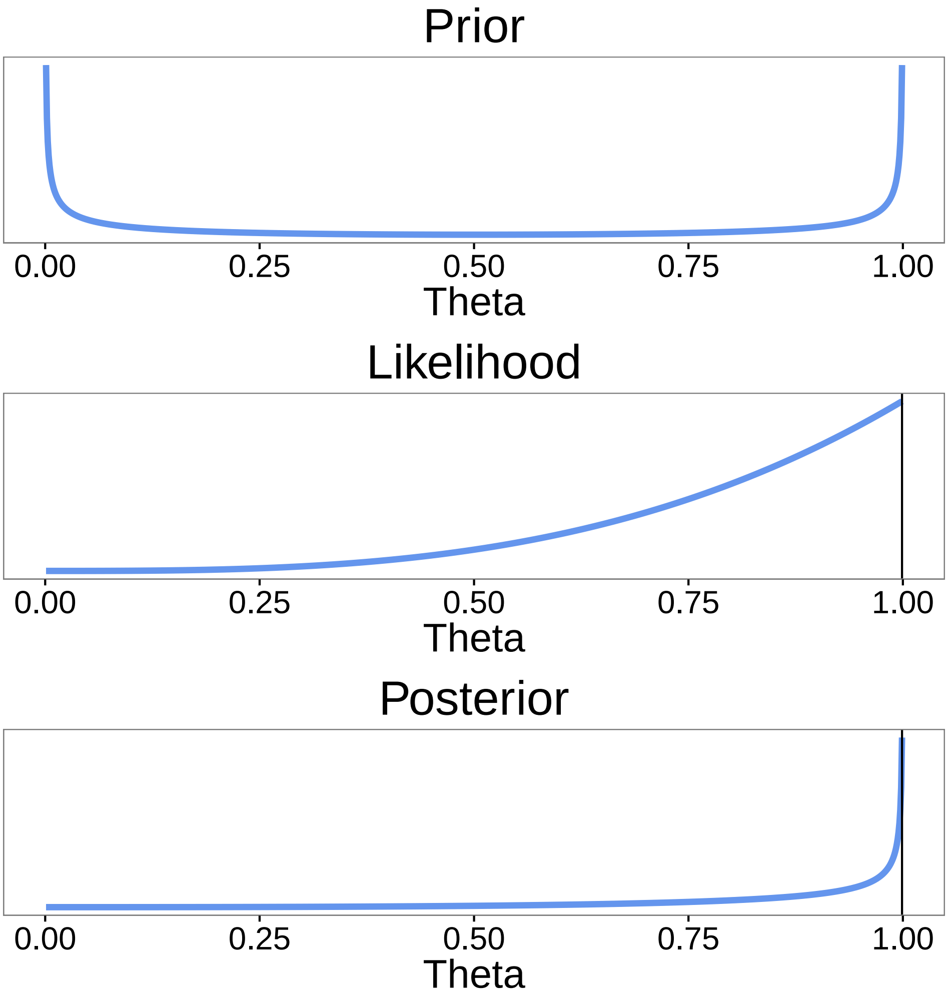

Figure 12. *The Jeffreys prior for a coin toss experiment puts high probability mass on the extreme values of* Θ*, thus providing a similar result as the maximum likelihood solution if the observe 3 successes in a row.*

Figure 13 depicts a close correspondence between confidence intervals and credibility intervals based on the Jeffreys prior (1000 simulations based on a sample size of 128).

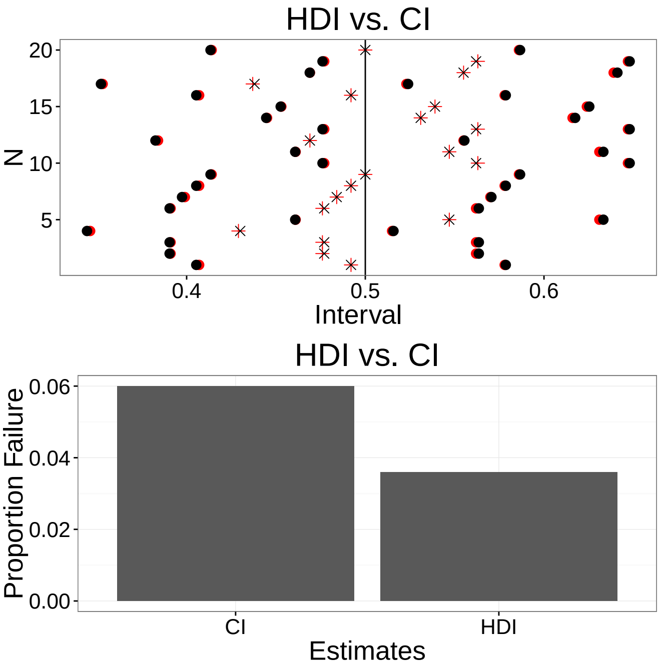

Figure 13. *Same setup as in the previous figures. Simulations show a close correspondence between confidence and credibility intervals (top) and a similar proportion of failure to include the true value (bottom) if the credibility intervals are based on the Jeffreys prior (sample size of 128, bottom row: 1000 simulated experiments).*

Others (e.g., Smith, in press) have argued that instead of Jeffreys' prior $\pi(\theta) = \dfrac{1}{\theta \cdot (1-\theta)}$ should be used, which is sometimes called the Haldane prior. This prior, however, is improper and does not integrate to 1.

5.5.2 Normal distribution

We can now turn to a more interesting example with an unbounded parameter space, namely estimating the mean μ of a normal distribution with a known standard deviation σ. Based on the likelihood function for the normal distribution $\mathcal{L}(x;\mu) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\dfrac{x-\theta}{2\sigma^2}}$ we obtain

$$\frac{\partial ln\mathcal{L}(x;\mu)}{\partial\theta} = \frac{x-\theta}{\sigma^2}$$

and thus for the Jeffreys prior on the mean of a normal distribution

$$\pi(\mu) \propto \sqrt{I(\mu)} = \sqrt{\mathbb{E}(\frac{x-\theta}{\sigma^2})^2} = \sqrt{\frac{1}{\sigma^2}\mathbb{E}(x-\mu)^2} = \sqrt{\frac{\sigma^2}{\sigma^4}} = \frac{1}{\sigma}$$

We see that this prior is constant in $(-\infty, \infty)$ and thus the integral over this prior does not exist. This is called an improper prior, which requires some sort of approximation. While it has been argued that the use of improper priors can still lead to proper posteriors (Kass and Wasserman, 1996; Berger, 2004), the exact method of how to work with an improper prior is less clear. As noted before and analogously to the previous simulations, for simplicity θ was discretised (with step-size 0.001) in the following computations. This does not affect the interpretation of the results in any way but simplifies the computation, particularly when dealing with an improper prior (because the marginal likelihood can be computed as a sum and not as an integral).Then, we obtain a very similar result as in Figure 1, such that the prior will not bias the estimation and we obtain the same results for MLE and MAP estimation (see Figure 14).

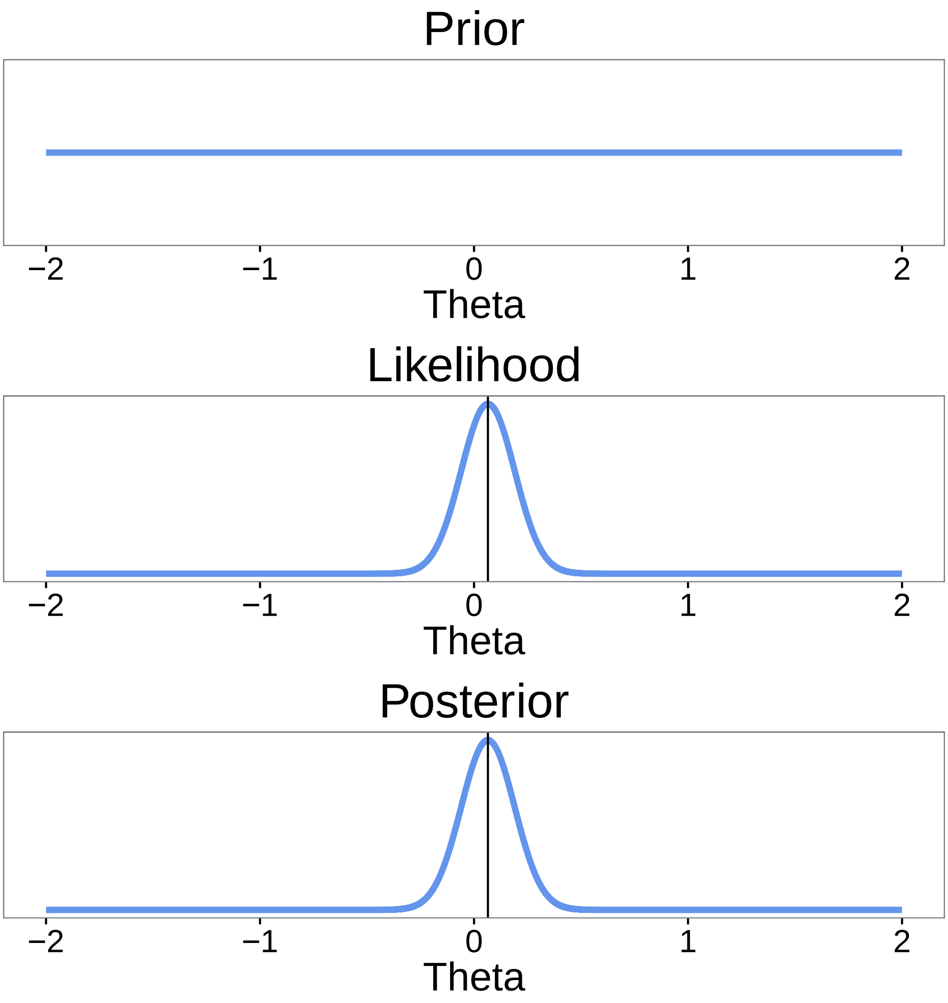

Figure 14. Using the Jeffreys prior to estimate the mean of a normal distribution, we find that we obtain a very similar result under MLE and MAP estimation (simulated with a true mean of 0.5 and a standard deviation of 2, i.e. $X \sim \mathcal{N}(0.5, 2)$*, and a sample size of 128).*

Concerning the intervals, we observe that the credibility intervals become much narrower than the confidence intervals, and show a much higher proportion of failures to capture the true value in 100 simulations (Figure 15).

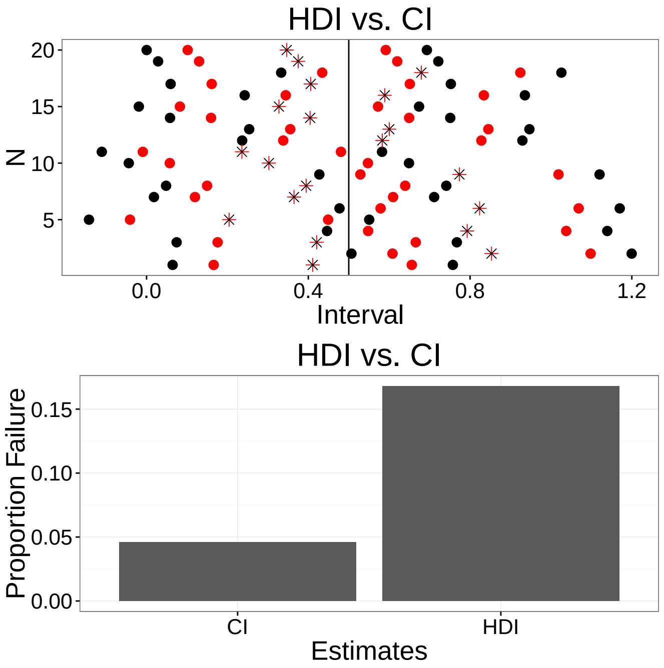

Figure 15. The credibility intervals become narrower (top) and show a substantially higher proportion of failures to capture the true mean (bottom) of a normal distribution when we use the Jeffreys prior (same setup as in Figure 14, sample size of 128, bottom row: 1000 simulated experiments).

5.6 Maximum entropy principle

While Jeffreys' proposal was arguably the most influential in defining formal rules for reference priors, a second approach is the maximum entropy principle formulated by Edwin Jaynes (Jaynes, 1968). Jaynes argued that probabilities should not be understood as relative frequencies but rather as a degree of belief and, furthermore, tried to ground probability theory in fundamental logic (Jaynes, 2003; Friedman and Shimony, 1971). The underlying proposal of the maximum entropy principle is then about how to specify probabilities with little or no available knowledge: Jaynes suggested to choose a prior that has maximal entropy whilst incorporating all available knowledge (thus, in the absence of any knowledge, the prior will be uniform), where entropy is defined as $H = -\sum_{i=1}^{n} p_i \cdot log\ p_i$ in the discrete case (Shannon, 1948). For illustrative purposes, only the discrete case will be discussed (for relevant literature discussing extensions to continuous problems and critique of this approach see Remark 5.6.3).

The maximum entropy prior can be obtained by (cf., Kass & Wassermann, 1996)

$$\pi(\theta_i) = exp[\sum_{j=1}^{m} \lambda_j f_j(x)]$$

where $f_j(x)$ specifies the constraints in terms of prior knowledge (see examples below), λ_j are constants to ensure that $\pi(\Theta)$ sums to one and m reflects the number of constraints.

Example 5.6.1
In a situation with zero prior knowledge except for the number of n discrete events, the maximum entropy prior can be obtainted by solving the Lagrangian

$$L \equiv \max_{p_i}[-\sum_{i=1}^{n} p_i \; log \; p_i - (\lambda_0 - 1)(\sum_{i=1}^{n} p_i - 1)]$$

We obtain $p_i = e^{-\lambda_0}$ and $\lambda_0 = log \; n$, such that the maximum entropy prior is uniform with $\pi(\theta_i) = \frac{1}{n}$. This is a classical result, such that in finite, discrete problems a uniform distribution has highest entropy and thus highest uncertainty about possible outcomes.

Example 5.6.2
In a situation where X is a discrete random variable with three possible values, $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, and an expected value of $E(X) = 2.5$, the maximum entropy prior can be obtained by solving the Lagrangian

$$L \equiv \max_{p_i}[-\sum_{i=1}^{3} p_i \; log \; p_i - (\lambda_0 - 1)(\sum_{i=1}^{3} p_i - 1) - \lambda_1(\sum_{i=1}^{3} p_i x_i - 2.5)]$$

We obtain $p_i = e^{-\lambda_0 - x_i \lambda_i}$ and $\lambda_0 = 2.987$, $\lambda_1 = -0.834$. Hence, the maximum entropy prior is $\pi(\theta_1) = 0.116$, $\pi(\theta_2) = 0.268$, $\pi(\theta_3) = 0.616$. Note that assuming a mean of 2 would lead again to a uniform prior (examples adapted from Shroeder, 2016).

Remark 5.6.3
While the maximum entropy priniple has been highly influential and successfully applied in many different fields, such as image processing and time-series analysis (Kass & Wasserman, 1996), there are several problems with the maximum entropy principle in practical use for classical parameter estimation. For example, as can be seen easily from the discrete form of the maximum entropy solution, this method is also subject to the partitioning problem as described above (Kass & Wasserman, 1996) and is therefore not invariant to parameter transformations (Seidenfeld, 1987). Furthermore, applying maximum entropy to continuous problems can become quite difficult, because entropy of a density π needs to be measured with respect to a base measure μ (Jaynes 1968), such that the entropy becomes $H = -\int \pi log \; \pi \; d\mu$. Finding such a base measure, however, can be difficult and be essentially as hard as choosing the prior in the first place (Kass & Wasserman, 1996). Seidenfeld (1987) provides a more detailed discussion and criticism of using the maximum entropy principle for defining priors in parameter estimation.

6 Conclusion

The aim of this work was to provide a comparison of frequentist and Bayesian parameter estimation. Both ways will provide the same result when we express a state of zero knowledge in the prior density, which, however, appears to be impossible in many situations, such as in unrestricted parameter spaces. Bayesian maximum a-posteriori estimation has the advantage of regularising the maximum likelihood estimation, which may provide faster or in fact better results and can be particularly important in more complex problems or small sample sizes. The critical question, however, is how to translate any existing state of knowledge into a prior density, which can be particularly difficult in problems with many parameters or an unrestricted parameter space. So far, it appears that none of the proposed rules for defining objective prior densities are capable of expressing a state of zero prior knowledge except in some very specific (and practically quite unrealistic) instances (Dawid, 1997; Efron, 1986). Particularly in problems with unrestricted parameter spaces many of the proposed formal rules are difficult to apply. The tentative conclusion of this thesis, therefore, is to be aware of the advantages and disadvantages of frequentist and Bayesian estimation routines, as different problems may be best addressed with different (frequentist or Bayesian) approaches.

7 A practical example

Finally, in this section I will to provide a worked example of applied Bayesian methods in the domain of cognitive neuroscience. Here, the contribution of Bayesian theory is twofold: on the one hand we are faced with the problem of estimating individual parameters given observed behaviour from participants, which should account for individual characteristics in a cognitive process. This aspect relates to the theoretical part of this thesis and the arguments discussed above, leaving us with the question of whether we want to perform Bayesian or frequentist parameter estimation. On the other hand I will describe a model of a cognitive process based on sequential Bayesian belief updating. Here, the emphasis is not on the methods of the statistical analysis but rather on providing a model that might be informative about how a particular cognitive process is implemented in brain function. In the following, I will provide an overview over the experiment, the Bayesian model and parameter estimation and illustrate a model-based analyses of physiological data.

7.1 Experimental setup

In this study we investigated how subjects perform inference on the current context to make adaptive choices in a decision-making (gambling) task. The structure of the task is illustrated in Figure 16 and required subjects to decide whether they wanted to accept or reject a gamble. If they decided to accept the gamble, they could either win or lose a monetary amount (20 cents) in a trial, which eventually determined the subject's payment after the experiment. If they decided to reject the offer, they would not win or lose anything but still observe the outcome (i.e., whether they would have won or lost had they accepted the offer). Each offer was represented by the presentation of a particular shape and a tone, and each possible combination of shapes and tones was presented equally often.

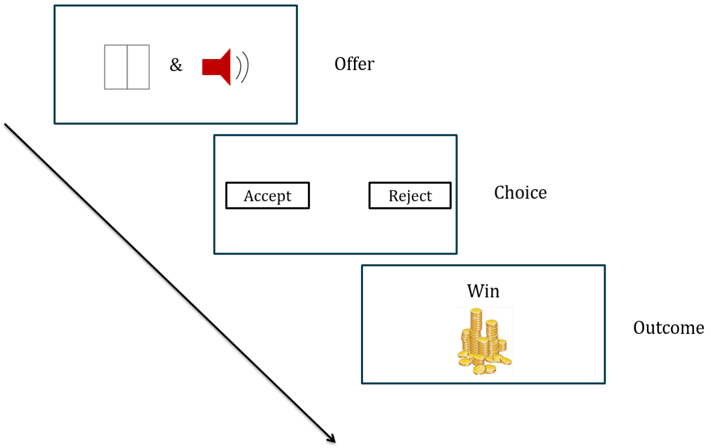

Figure 16. Experimental design: subjects had to decide whether to accept or reject an offer consisting of a shape and tone. After their decision they observed an outcome, which was a monetary win or loss.

Subjects learned prior to performing the experiment that there was one 'good' and one 'bad' shape and tone, where 'good' means being predictive of a win and 'bad' means being predictive of a loss (see Figure 17) and were trained on the identity of these stimuli. Furthermore, these predictions were not perfectly valid but only true in 85% of the cases. An 'offer' or 'gamble' was then defined as one out of four possible combinations of the auditory and visual cues (good tone and shape, good tone and bad shape, bad tone and good shape, bad tone and shape).

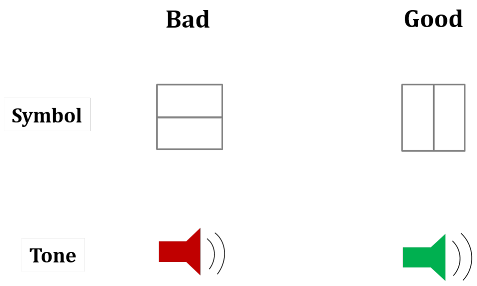

Figure 17. Subjects learned that one shape and one tone was predictive of a loss with 85% validity ('bad') and one shape and one tone was predictive of a win with 85% validity ('good')

Importantly, to decide whether to accept or reject an offer, subjects had to infer the current context of a trial, where subjects were instructed that the context would change about two to three times per session. Subjects could be in one out of three different contexts, which determined which stimulus feature to pay attention to. Subjects could be in a context where only shape is relevant, only tone is relevant or both dimensions are relevant. The corresponding choice rules are depicted in Figure 18, where in each context cues predicting a loss should be rejected and cues predicting a win should be accepted. Importantly, subjects had to infer the current context by observing the cue-outcome contingencies. For example, in Figure 16 subjects observe a 'good' shape and a 'bad' tone followed by a win. Observing this sequence makes the context shape more likely than context tone or conjunction, because only context shape predicted a win in this trial. Based on their inference on the current context, subjects had to infer their choice (accept or reject). For example, if subjects observe the offer of Figure 16 and believe they are in context shape they should accept the gamble and reject otherwise.

To model the subjects' inference processes in that task, we used a Bayesian model of sequential belief updating, which will be illustrated below.

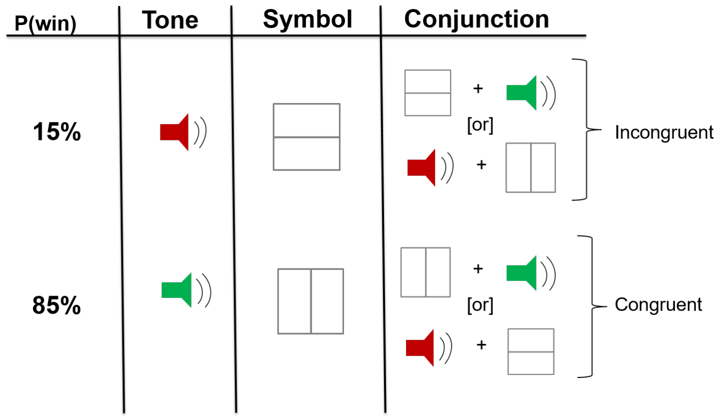

Figure 18. Subjects could be in one out of three different contexts. The current context implied which feature of the offer to pay attention to: either only the tone was relevant, only the shape was relevant or both dimensions were relevant.

7.2 A Bayesian model for belief updating

We assume that subjects hold a prior belief about the current context at any given trial and update their belief according to Bayes' rule based on new observations. At the first trial, we assume that subjects hold a (maximum entropy) uniform prior, i.e. $\pi(\theta_i) = 1/3$ for every $i \in \{1, 2, 3\}$.

Thereby $\theta_1, \theta_2, \theta_3$ refer to the three different contexts, respectively. To account for possible reversals in the current context, we multiply the prior at any trial with matrix E, where E is defined as

$$E = \begin{bmatrix} 0.95 & \frac{0.05}{2} & \frac{0.05}{2} \\ \frac{0.05}{2} & 0.95 & \frac{0.05}{2} \\ \frac{0.05}{2} & \frac{0.05}{2} & 0.95 \end{bmatrix}$$

This matrix accounts for possible reversals in contexts. The matrix can be read as a $\frac{57}{60} = 0.95$ probability that there is no shift of context at a given trial, whereas there is a $\frac{(1-\frac{57}{60})}{2} = 0.025$ probability that there was a shift to one of the other two contexts. These probabilities correspond to the true contingencies of a session, which involved three shifts in 60 trials. Effectively, the computational role of the E-matrix is simply to reduce the current belief about being in a particular context, which is important for detecting context reversals.

Finally, to apply Bayes' theorem we need to define the likelihood function of a given observation in a trial. An observation is comprised by the the particular offer (shape and tone combination) and the outcome (win or loss), where the likelihood should reflect the probability of observing a particular outcome given the observed offer under the three different contexts. Thus, we can define three different likelihood functions for the three different contexts:

$$\mathcal{L}^1 = \begin{bmatrix} 0.85 & 0.85 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.85 & 0.85 \end{bmatrix}$$

$$\mathcal{L}^2 = \begin{bmatrix} 0.85 & 0.15 & 0.85 & 0.15 \\ 0.15 & 0.85 & 0.15 & 0.85 \end{bmatrix}$$

$$\mathcal{L}^3 = \begin{bmatrix} 0.85 & 0.15 & 0.15 & 0.85 \\ 0.15 & 0.85 & 0.85 & 0.15 \end{bmatrix}$$

Where $\mathcal{L}^1, \mathcal{L}^2, \mathcal{L}^3$ refer to the contexts tone, shape and conjunction, the rows of the likelihood functions refer to the outcomes (1 = win, 2 = loss) and the columns refer to the four different stimulus combinations (1 = good tone/ good shape, 2 = good tone/ bad shape, 3 = bad tone/ good shape, 4 = bad tone/ bad shape).

We can now model trial-by-trial belief updates as

$$\pi(\theta_i|x) = \frac{\pi(\theta_i) \cdot \mathcal{L}^i_x}{\sum_i \pi(\theta_i) \cdot \mathcal{L}^i_x}$$

where x is a tuple that refers to a particular cue-outcome association: $x = (c, o)$, $o \in \{1, 2\}$ and $c \in \{1, 2, 3, 4\}$ (note that here '·' refers to the dot product). Then, the posterior belief about the context at a given trial is used as a prior at the subsequent trial, where the process described above repeats.

Given this simple model of belief updating, we can simulate a full session consisting of 60 trials, as illustrated in Figure 19.

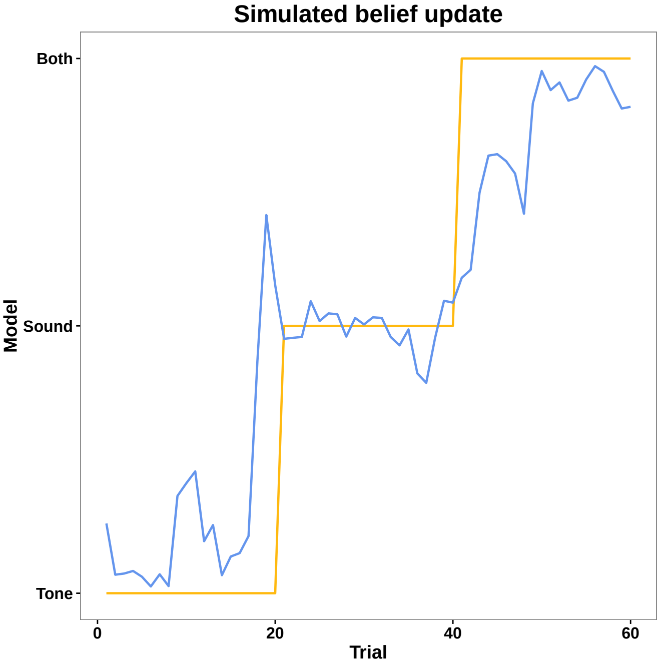

Figure 19. Simulated session consisting of 60 trials, where a subject starts in the context tone being relevant, then moves to context shape and finally ends up in context conjunction (yellow line). The blue line reflects the the agent's simulated prior belief about the current context. We see that the Bayesian belief update model tracks the true state of the task reasonably well (first 4 trials discarded to allow for initial belief formation).

Finally, we can also predict choice behaviour based on this model, which will become important for parameter estimation in the next session. A simple way to predict choice behaviour is by defining stimulus specific choice rules for each context and then predict behaviour with respect to the agent's current belief about the context. We can define an action-rule matrix in analogy to the likelihood-functions above:

$$\mathcal{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

where the columns refer to the observed stimuli in analogy to the likelihood-functions above and the rows refer to contexts tone, shape and conjunction, respectively. An entry of 1 refers to action 'accept', whereas 0 refers to 'reject'. The predicted action at any given trial then simply becomes

$$\hat{a} = \pi(\theta_i)\mathcal{A}_{1-3,s} = \pi(\theta_i) \begin{bmatrix} \mathcal{A}_{1,s} \\ \mathcal{A}_{2,s} \\ \mathcal{A}_{3,s} \end{bmatrix}$$

This is called Bayesian model averaging, because the action under each context (i.e., model) is averaged according to the posterior belief (from the previous trial) about each context, with $\hat{a} \in [0, 1]$

7.3 Parameter estimation

From having specified the model of belief updates, we can now investigate individual differences in the belief updating by estimating individual parameters in the model given observed behaviour from data collected in the experiment. Note the important difference between the actual Bayesian model of belief updating, where subjects are assumed to infer (i.e., estimate) the hidden state concerning the current context, and inferring individual parameters of that model that describe the specific choice behaviour of subjects. One crucial difference in these two estimation problems is the relevant data for specifying the likelihood function - while in the Bayesian belief updating model, the relevant data to estimate the current context are the observed offers and outcomes of the trials, whereas to estimate individual parameters of that model the relevant data are the observed choices from the actual subjects of the experiment.

Here, we will estimate two different parameters: r, which stands for the probability of no reversal at a given trial in the E matrix, and v, which reflects the validity of cues as specified in the likelihood functions, such that every 0.85 entry will be replaced with v and every 0.15 entry will be replaced by $1 - v$. Thus, the E-matrix now becomes

$$E = \begin{bmatrix} r & \frac{1-r}{2} & \frac{1-r}{2} \\ \frac{1-r}{2} & r & \frac{1-r}{2} \\ \frac{1-r}{2} & \frac{1-r}{2} & r \end{bmatrix}$$

with $r \in [0, 1]$. If $r = 1$, E will be an identity matrix and thus leave $\pi(\theta_i)$ unchanged. Intuitively, r can be understood as the probability that there is no reversal in contexts at a given trial. Thus, the individual estimate of r for each subject given her behaviour should reflect her belief about the volatility in the given experiment. Furthermore, the likelihood functions become

$$\mathcal{L}^1 = \begin{bmatrix} v & v & 1-v & 1-v \\ 1-v & 1-v & v & v \end{bmatrix}$$

$$\mathcal{L}^2 = \begin{bmatrix} v & 1-v & v & 1-v \\ 1-v & v & 1-v & v \end{bmatrix}$$

$$\mathcal{L}^3 = \begin{bmatrix} v & 1-v & 1-v & v \\ 1-v & v & v & 1-v \end{bmatrix}$$

Where $v \in [0, 1]$, which intuitively can be understood as how well the cues predict the outcomes. Obtaining individual estimates for v should reflect the subject's belief about the reliability of a stimulus.

To estimate these parameters, we need to define a likelihood function based on the subjects' behaviour. This is less straightforward than in the examples above, because the likelihood function needs to take the specifics of each trial into account to predict the choice of an agent (therefore we cannot use one particular, fixed distribution to predict the choices over all trials). One common way to define the likelihood function is

$$\mathcal{L} = \sum_{i=1}^{T} log(\hat{a_i} \cdot a_i) + log((1 - \hat{a_i}) \cdot (1 - a_i))$$

where T stands for the number of trials in total, $\hat{a_i} \in [0, 1]$ reflects the predicted action at any given trial i and $a_i \in \{0, 1\}$ the observed action by a participant (0 = reject, 1 = accept). Thus, $\mathcal{L}$ will be negative with an upper bound of 0 and maximum likelihood estimation simply has to find a parameter setting that maximises $\mathcal{L}$ (or, equivalently, minimises $-\mathcal{L}$).

To perform maximum a-posteriori estimation, we simply need to add the log-likelihood of each parameter value under the prior distribution as illustrated in Remark 3.2.3. This can be done separately for r and v, which both are bounded by 0 and 1. Thus, we can use a beta distribution as a prior for both parameters, and simply add the probability density function of a given parameter value under the given prior to $\mathcal{L}$.

7.3.1 Results

The subsequent analyses were performed with MATLAB, using the fminsearch function to find the minumum of $-\mathcal{L}$ (or its MAP equivalent) given a set of parameter configurations (this is a standard function used in parameter estimation based on a simplex search method (Lagarias et al.,1998) to perform unconstrained, nonlinear optimisation and find the minimum of a function).

First, we can try to replicate the equivalence of MAP and MLE when we use a uniform (beta) prior over parameters r and v. Figure 20 shows the parameter estimates for 13 subjects who performed the task described above. Each of these subjects performed three sessions of the task and thus 180 trials in total (except three subjects who only performed two sessions). As expected, we find that the ML and MAP estimates are equivalent for both parameters, which are close to their original values of $r = 0.95$ and $v = 0.85$.

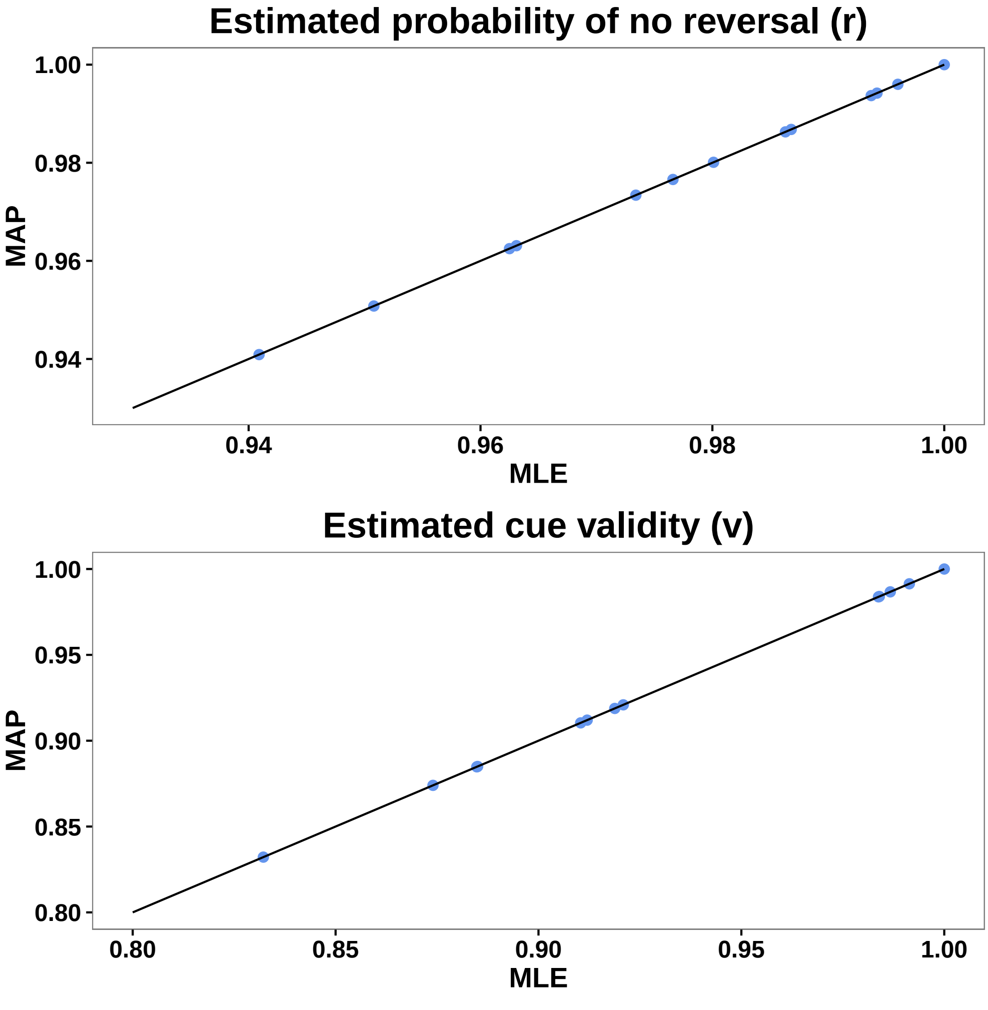

Figure 20. Using a uniform (beta) prior, we find that the parameter estimates for both the probability of no reversal (r) and the cue validity (v) are identical under ML and MAP estimation. Furthermore, we see that the estimates are close to the true values of $r = 0.95$ *and* $v = 0.85$. *(Black line: identity)*

Figure 21 shows that the model based on the ML estimates provides a fairly high explained variance of observed behaviour (R^2), defined as the squared (Spearman) correlation coefficient between predicted ($\hat{a}$) and observed (a) behaviour.

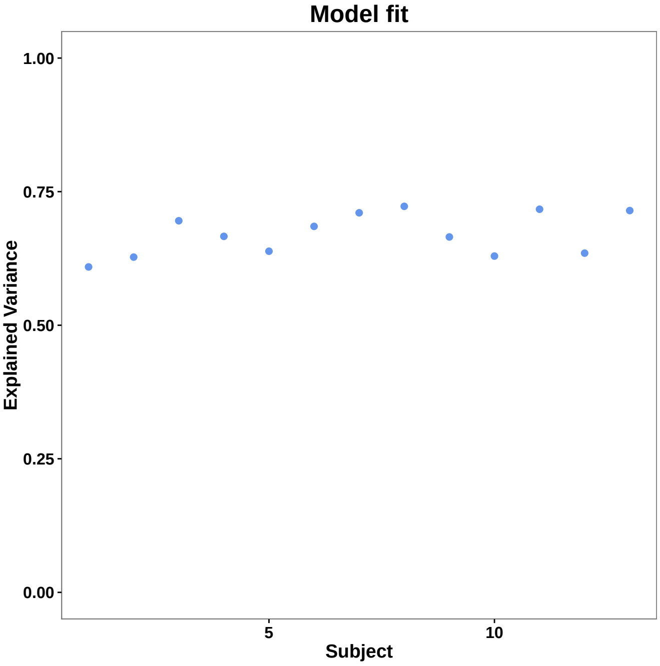

Figure 21. Explained variance of observed behaviour by the Bayesian model based on the maximum likelihood parameters for each subject.

We can now move on and explore the role of the prior on r and v. In this case we do have prior knowledge about the two parameters, because we know their true values in the experiment. The subjects were trained intensely on these contingencies prior to the experiment and were communicated explicitly to the participants. Thus, we can explore what happens if we put higher probability mass on higher parameter values that are closer to the true values of the parameters, namely $r = 0.95$ and $v = 0.85$. More specifically, we can define the beta priors in a way that their expected value will be the true parameter values, where the expected value of a beta-distribution is $\frac{\alpha}{\alpha+\beta}$. Figure 22 shows the difference between ML and MAP estimates when we specify the priors on r and v such that their expected value reflects the true value of these parameters with very little variance ($r : \alpha = 95, \beta = 5$, $v : \alpha = 85, \beta = 15$). We see that even under such a parameterisation with very little variance the ML and MAP estimates are in close correspondence (even though ML estimation tends to produces slightly larger parameter estimates). This behaviour is due to the fact that in this example, ML provides a very good and accurate estimation of the individual parameters and does not run into overfitting issues as discussed in Section 5.1. This can be explained by the use of a relatively simple behavioural model and the small number of parameters that are estimated. Furthermore, given that for almost all subjects we have 180 data points it is likely that the prior has only weak influence on the estimation. In more complex models, for example when we try to estimate a hierarchical model to infer a group effect or when we estimate a larger number of parameters with higher dependencies, it might be more important to regularise the ML solution by using a prior on each parameter.

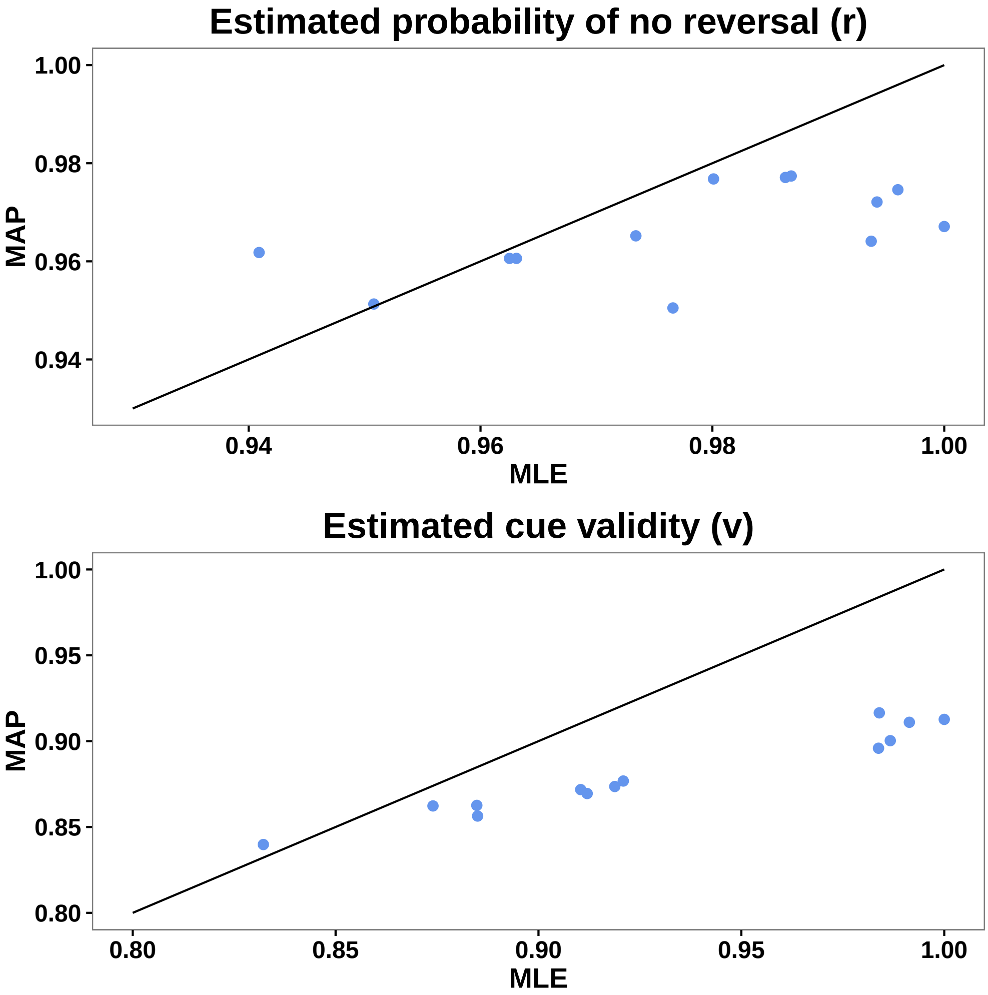

Figure 22. *Even when using highly precise priors on the two parameters (*$r : \alpha = 95, \beta = 5$*,* $v : \alpha = 85, \beta = 15$*), we find a close correspondence between ML and MAP estimates (even though ML estimates produce slightly higher parameter estimates than the corresponding MAP estimates). This behaviour is due to the fact that ML estimation works reasonably well in this example due to a relatively simple model and only two parameters that are estimated as well as a large sample size for each subject.*

We also see that ML and MAP estimation provide very similar model fits in terms of explained variance of observed behaviour as illustrated in Figure 23.

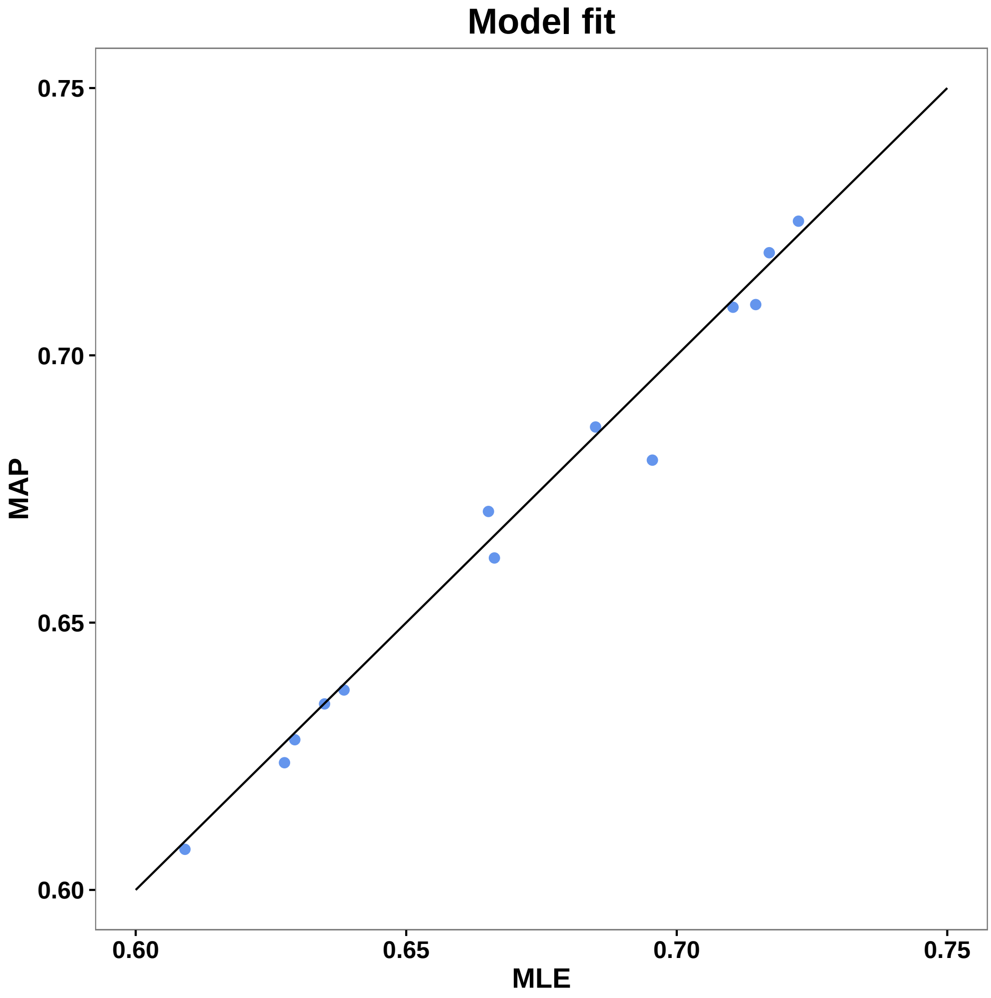

Figure 23. We see that the model fits calculated from ML and MAP estimates are very similar even when we use highly informative priors. We find that MLE model fits are generally marginally higher than MAP model fits, which is a tendency that becomes more pronounced when we use more complex models or smaller sample sizes.

Finally, we can also illustrate the effect of 'wrong' priors on the parameter estimates. Figure 24 shows that parameter estimates differ strongly when we reverse the parameterisation of the priors ($r : \alpha = 5, \beta = 95$, $v : \alpha = 15, \beta = 85$), resulting in a much lower model fit for MAP estimates as illustrated in Figure 25.

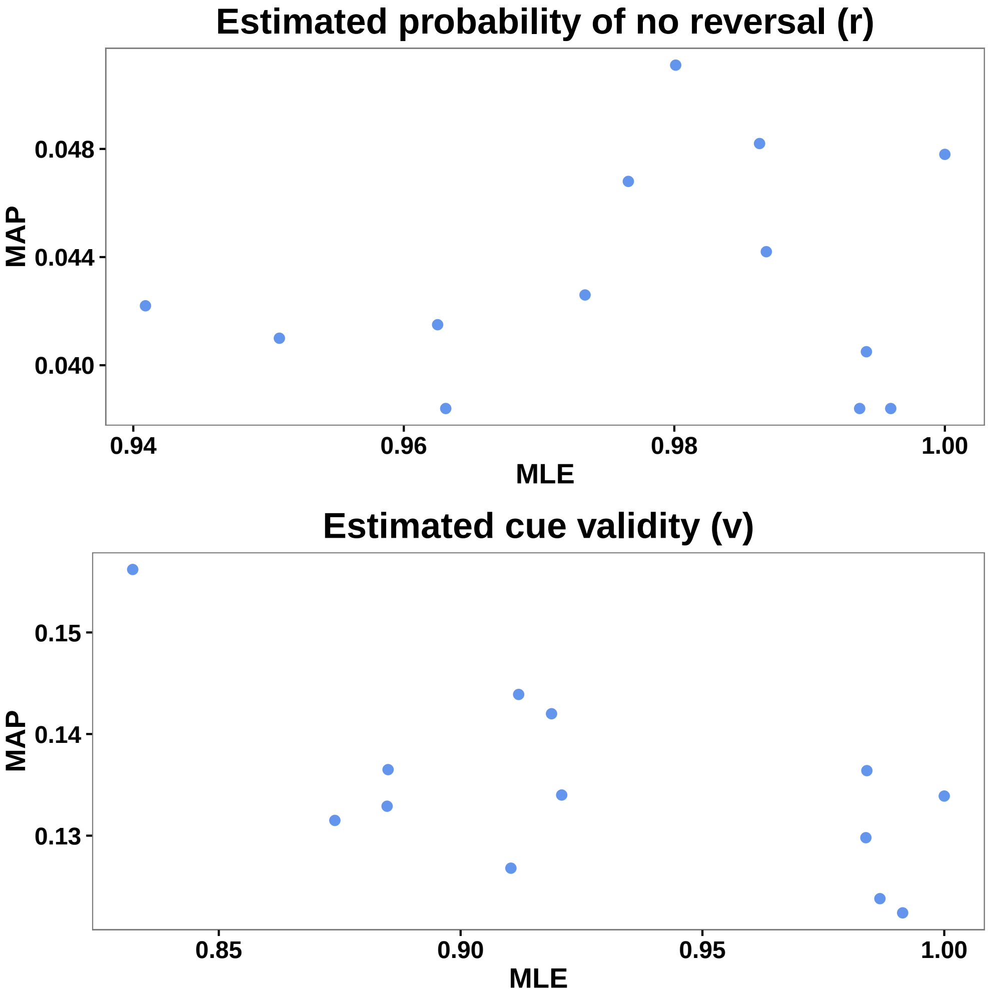

Figure 24. *When we use highly precise ’malicious’ priors on the two parameters* ($r : \alpha = 5, \beta = 95$, $v : \alpha = 15, \beta = 85$)*, we find a pronounced difference between ML and MAP estimates, such that MAP estimates are far off the true parameter values.*

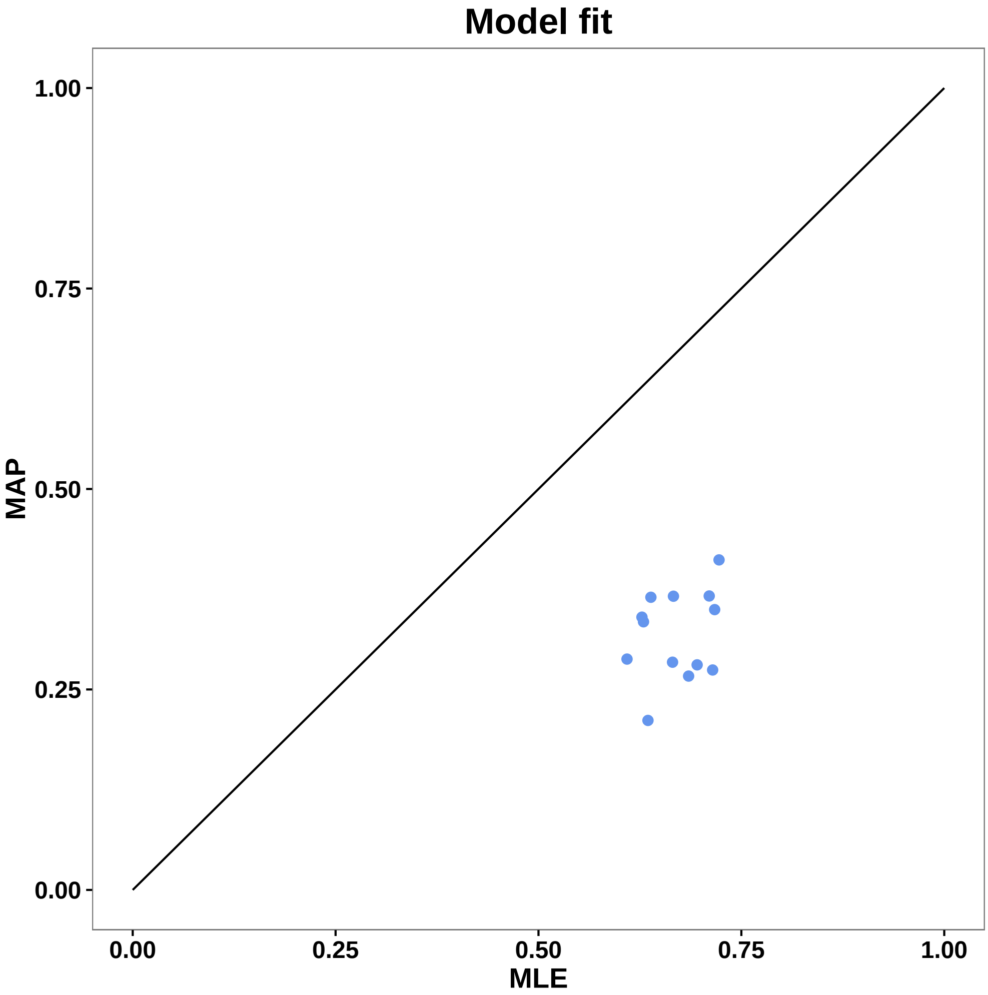

Figure 25. Under precise 'malicious' priors, we find that ML estimation provides a much higher model fit that MAP estimation.

7.4 Model-based analysis of physiological data

Finally, we can now use this model and individual parameter estimates to investigate the physiological basis of (Bayesian) belief updates. In this experiment, we recorded the pupil dilation of every subject during every trial. Pupil dilation has been associated with belief updates and the encoding of unexpectedness, strongly associated with the neuromodulator norepinephrine and the activity in a brain region called locus coeruleus (e.g. Nassar et al., 2012). The exact nature of the signal and its computational meaning, however, has remained unclear. We can now use different update signals that we can retrieve from the computational model to investigate the nature of the unexpectedness signal as reflected in the pupil dilation. Given that ML estimation was found to produce good and reliable estimates, the following analyses will be based on these estimates.

Figure 26 shows that we indeed found the classical unexpectedness response in the pupil dilation in our subjects. We computed the mean pupil dilation for expected and unexpected outcomes in the period of 800 to 1400 ms after outcome-onset (a classical time window to analyse a change in pupil dilation). Expected outcomes were defined as a correspondence between a subject's decision and an outcome, i.e. an accepted win or a rejected loss. Trials in which the subject's decision did not correspond to the outcome were defined as unexpected. As Figure 26 shows, we found that the pupil dilation was significantly higher in unexpected compared to expected trials (paired t-test: $t(12) = -6.646, p < 0.01$, expected: $mean = 1795.862, sd = 628.135$, unexpected: $mean = 1906.619, sd = 675.893$)

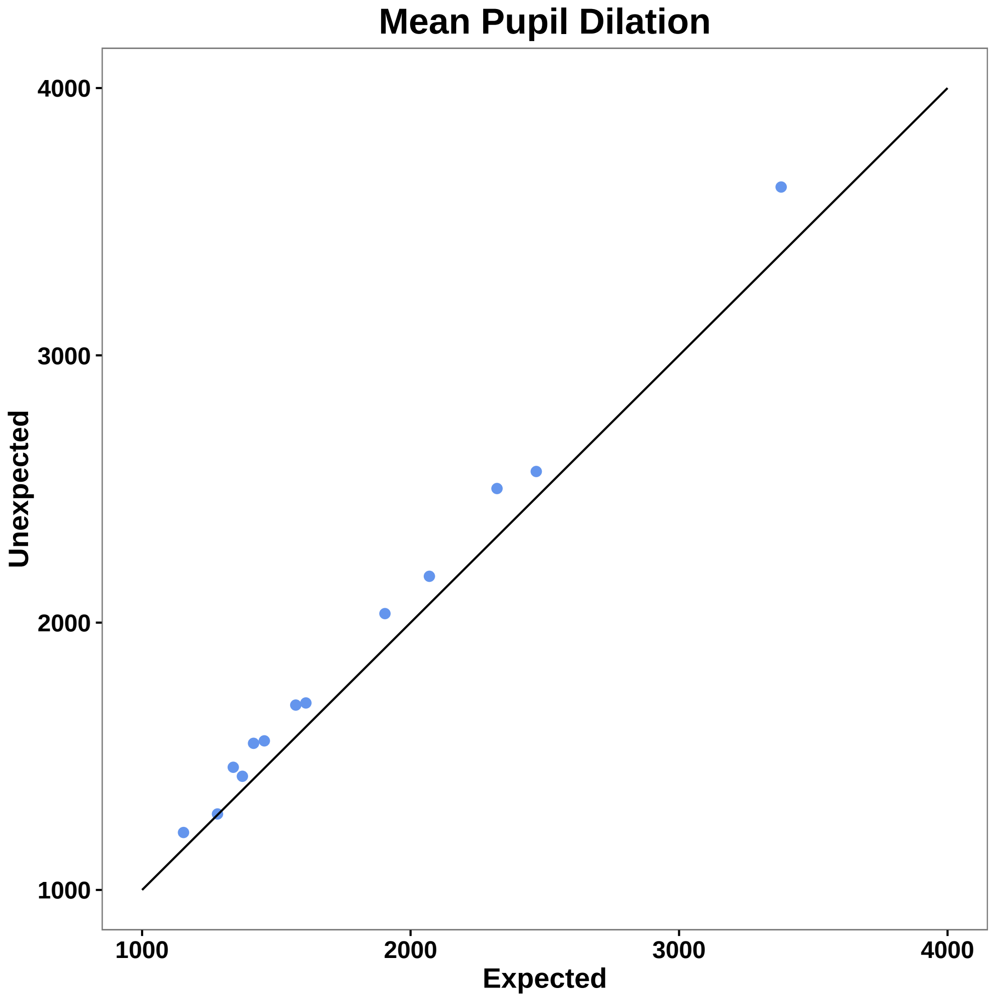

Figure 26. For all participants, we found higher average pupil dilation for unexpected compared to expected outcomes. The black line depicts the identity.

Having replicated a standard result in pupillometry, we can now use the Bayesian model of belief updating to investigate the computational nature of the increased pupil response to unexpected outcomes. In particular, recent work has highlighted a distinction between the pure unexpectedness of a stimulus and actual belief updates (O'Reilly et al., 2013;

Schwartenbeck, FitzGerald & Dolan, 2016). The pure unexpectedness can be defined as information-theoretic surprise (*surprisal*):

$$-ln\ \mathcal{P}(o|\pi(\theta), c)$$

where o refers to the outcome in a trial (win or loss), c to the observed cue and $\pi(\theta)$ to the beliefs about the hidden state (context). This quantitiy reflects the pure unexpectedness of an event, but it does not reflect how much an agent actually shifts its beliefs about the true context. The actual shifts in beliefs can be defined as Kullback-Leibler divergence from prior to posterior beliefs in a trial:

$$KL[\pi(\theta_i|x)\ ||\ \pi(\theta_i)] = \sum_i \pi(\theta_i|x) \cdot ln\left(\frac{\pi(\theta_i|x)}{\pi(\theta_i)}\right)$$

This divergence quantifies how much the prior density changes according to observations, and reflects how meaningful an (unexpected) observation is with respect to an agent's beliefs about the current context of a task, and thus how much these beliefs are updated on a trial-by-trial basis. The present task allows for a separate investigation of surprise and the Kullback-Leibler divergence and its physiological implementation (the average correlation between these two signals was $r = 0.58$). In particular, we can now derive trial-by-trial model-based regressors for information-theoretic surprise and belief updates and investigate which of the two predicts changes in pupil dilation. Figure 27 shows the individual regression coefficients for surprise and belief updates as predictors for pupil dilation. We found that belief updates in general had a much stronger and positive effect on pupil dilation, such that whenever there were stronger belief updates pupil dilation increased (one-sample t-test for regression coefficients against 0: $t(12) = 5.018, p < 0.01$). In contrast, we found a weaker and inconsistent effect for information-theoretic surprise (one-sample t-test for regression coefficients against 0: $t(12) = 0.019, p = 0.876$). In a direct comparison, we found that coefficients for belief updates had a stronger impact on the pupil dilation than coefficients for information-theoretic surprise ($t(12) = 4.936, p < 0.01$). Importantly, this suggests that physiological responses to unexpected outcomes as signalled by changes of the pupil diameter reflect actual model-updates, and are therefore sensitive to a subject's cognitive representation of the task, as opposed to the pure unexpectedness of an observation.

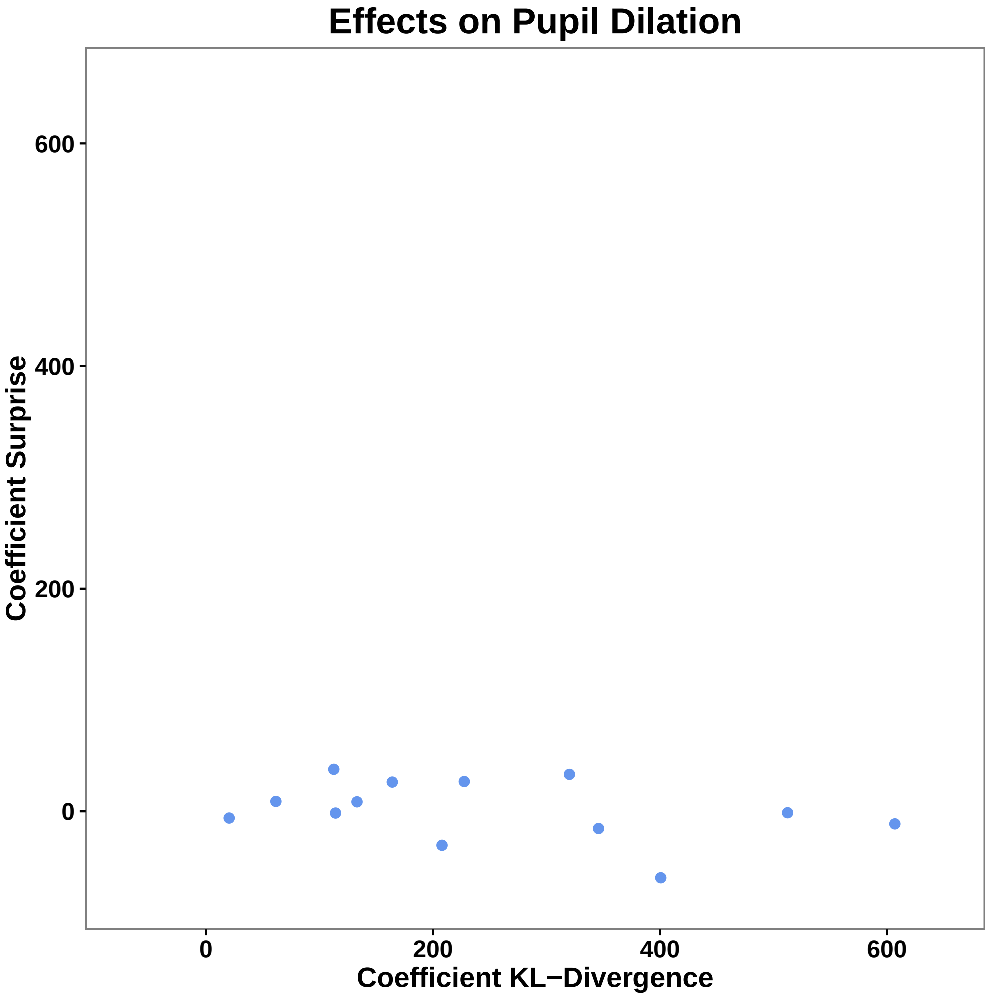

Figure 27. Individual coefficients for surprise and belief updates (Kullback-Leibler divergence) on pupil dilation. We found that belief updates exerted strong and consistently positive effects on pupil dilation, whereas surprise exerted weaker and inconsistent effects.

7.5 Summary

The intention of this chapter was to provide an insight into the use of Bayesian inference in an applied setting, namely a study in the domain of cognitive neuroscience. This practical example has illustrated the use of Bayesian inference in two ways: modelling a cognitive process as Bayesian belief updating and Bayesian parameter estimation. In this example, there was no advantage of using maximum a-posteriori estimation over maximum likelihood. We found that ML estimation produced reasonable parameter estimates and acceptable model-fits. In more complex problems, for example hierarchical models or problems with more free parameters and higher dependencies, it might be advantageous to use maximum a-posteriori estimates. On the other hand, we found that our Bayesian model of belief updating provided an accurate account of observed behaviour from 13 subjects. Particularly, this model allowed us to dissociate different update signals and test for their neural implementation. The results of our model/based analysis suggest that pupil dilation, a physiological response strongly associated with the processing of unexpected events, reflect actual belief updates (and thus the meaningful information content imparted by a stimulus) as opposed to the pure unexpectedness of a stimulus defined as information-theoretic surprise.

8 References

Bathke, A. (2013). Skriptum Stochastische Modellbildung, Universität Salzburg.

Bayes, T. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London, 53, 370–418.

Berger, J. (2004). The Case for Objective Bayesian Analysis. Bayesian Analysis, 1(1), 1–17.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society. Series B, Vol. 41, No. 2 (1979), pp. 113-147. Retrieved from http://www.jstor.org/stable/2985028

Billingsley, P. (1979). Probability and Measure. New York, Toronto, London: John Wiley and Sons.

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14. http://doi.org/10.1038/nrn3475

Carnap, R., & Stegmüller, W. (1959). Die beiden Wahrscheinlichkeitsbegriffe. In Induktive Logik und Wahrscheinlichkeit (pp. 12–37). Vienna: Springer Vienna.

Dawid, A. P. (1997). Comments on Non-informative priors do not exist. Journal of Statistical Planning and Inference, 65, 159–189.

de Finetti, B. (1974). Bayesianism: Its Unifying Role for Both the Foundations and Applications of Statistics. International Statistical Review, 42(2), 117. http://doi.org/10.2307/1403075

Efron, B. (2012). Why Isn't Everyone a Bayesian? The American Statistician.

Friedman, K., & Shimony, A. (1971). Jaynes's maximum entropy prescription and probability theory. Journal of Statistical Physics.

Felsenstein, K. (2008). Skriptum Mathematische Statistik.

Gelman, A. (2013). Bayesian data analysis.

Ghosh, M. (2011). Objective Priors: An Introduction for Frequentists. Statistical Science, 26(2), 187–202. http://doi.org/10.1214/10-STS338

Jaynes, E. T. (1968). Prior Probabilities. IEEE Transactions On Systems Science and Cybernetics, 4(3), 227–241.

Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

Jeffreys, H. (1939). Theory of probability. Clarendon Press.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 186(1007), 453–461. http://doi.org/10.1098/rspa.1946.0056

Kallenberg, O. (2002). Foundations of Modern Probability, 2nd ed. Springer Series in Statistics.

Kass, R. E., & Wasserman, L. (2012). The Selection of Prior Distributions by Formal Rules. Journal of the American Statistical Association.

Klenke, A. (2008). Probability Theory - a comprehensive course. Springer

Koch, K. R. (2007). Introduction to Bayesian Statistics. Springer

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. SIAM Journal on Optimization, 9(1), 112–147. http://doi.org/10.1137/S1052623496303470

Laplace, P. S. (1812). Théorie analytique des probabilités.

Lehmann, E.L., & Casella, G. (1998). Theory of Point Estimation. Springer

Letham, B., & Rudin, C. (n.d.). Probabilistic Modeling and Bayesian Analysis. University Script.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. Nature Neuroscience, 15(7), 1040–1046. http://doi.org/10.1038/nn.3130

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proceedings of the National Academy of Sciences of the United States of America, 110(38), E3660–E3669. http://doi.org/10.1073/pnas.1305373110

Savage, L. J. (1956). The Foundations of Statistics. Journal of the American Statistical Association, 51(276), 657. http://doi.org/10.2307/2281495

Schroeder, D. (2016). Accounting, managerial experimentation and causal effects.

Schwartenbeck, P., FitzGerald, T. H. B., & Dolan, R. (2015). Neural signals encoding shifts in beliefs. NeuroImage, 125, 578–586. http://doi.org/10.1016/j.neuroimage.2015.10.067

Seidenfeld, T. (1987). Entropy and Uncertainty. In Entropy and Uncertainty (pp. 259–287). Dordrecht: Springer Netherlands.

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423. http://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Smith, G. P. (2016). Expressing Prior Ignorance of a Probability Parameter.

Syversveen, A. R. (1998). Noninformative Bayesian Priors. Interpretation And Problems With Construction And Applications.

Trutschnig, W. (2014). Rohversion Skriptum zur Vorlesung Statistik. http://www.trutschnig.net