STA421
Foundations of Bayesian Methodology
FS22

Małgorzata Roos

Department of Biostatistics at Epidemiology, Biostatistics and Prevention Institute,

University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland

malgorzata.roos@uzh.ch

May 31, 2022

Contents

Chapter 1

Lecture 1: Classical vs Bayesian paradigms and conditional probability

1.1 Overview of the lecture

Bayesian methods combine prior knowledge with observed data and are powerful tools for data analysis in many domains of science. However, underlying concepts, derivations, and computations can be challenging. This lecture reviews fundamental concepts of Bayesian methodology and provides an accessible introduction to theoretical and practical tools with medical applications. A successful participant will be able to apply Bayesian methods in other areas of research.

Probability calculus	Distributions	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ Likelihood $\times$ Prior		Likelihood
Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 1.1: Foundations of Bayesian Methodology: content of the lecture.

1.2 Overview of the individual project

Table 2 of Baeten et al. [2013] provides results of a Bayesian analysis of ASA20 responders at week 6 for Secukinumab and Placebo. This case-control study considers Ankylosing spondylitis in an experimental treatment with Secukinumab (monoclonal antibody) and uses historical controls. The primary binary endpoint ASAS20 indicates patients with a 20% response according to the Assessment of Spondylo Arthritis international Society criteria for improvement at week 6.

A classical clinical trial would for example use a 1:1 sampling with n=24 patients in the treatment group and n=24 patients in the placebo group. This Bayesian analysis uses a smaller number of patients. It applies a 4:1 study design with n=24 patients in the treatment group and only n=6 patients in the placebo group, but uses 8 similar historical placebo-controlled clinical trials to derive an informative prior for the placebo group instead.

Potential benefits of Bayesian analysis

- Reduces the number of placebo patients in the new trial
- Decreases costs
- Shortens trials duration (→ faster decision)
- Facilitates recruitment (→ faster decision)
- Can be more ethical in some situations

Secukinumab	**Placebo**
Sample size computation	
	Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)	Beta(11, 32)
Data	Data
Posterior (S)	Posterior (P)
Posterior probability of superiority	

Table 1.2: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

- An intermediate study can be conducted at any timepoint

Potential dangers of Bayesian analysis

- Posteriors hinge on the prior elicited for the placebo group
- The prior elicited for the placebo group depends on the prior for the between-study precision in a Bayesian meta-analysis

1.3 History

The history of both the Bayesian and the classical approaches to statistics is intertwined. This section reviews the most relevant historical facts.

Bayes

INDUCTIVE LOGIC
(θ) before $\longleftarrow$ after ($\mathbf{y}$)
before: possible, probable causes
after: effects, results

- James Bernoulli (1713)
- Reverend Thomas Bayes (1763)
- Laplace (1812)

Bayes Theorem
timeline $A \rightarrow B$:

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B]}$$

or

$$P[\theta \mid \mathbf{y}] = \frac{P[\mathbf{y} \mid \theta]P[\theta]}{P[\mathbf{y}]}$$

or

$$P[H \mid D] = \frac{P[D \mid H]P[H]}{P[D]}$$

– quantification of evidence

– Bayes factor

Classical

DEDUCTIVE LOGIC
(θ) before $\longrightarrow$ after ($\mathbf{y}$)
before: causes
after: results
general rules, promises (θ) lead to
certain results and conclusions ($\mathbf{y}$)

- Pearson, Galton - 1890, 1900
- Gosset, Fisher - 1910, 1920
- Pearson, Neyman - 1930

Likelihood
timeline $A \rightarrow B$:
Only interested in $P[B \mid A]$ or $P[\mathbf{y} \mid \theta]$

– 95% confidence intervals

– tests

– p-values

– statistical programs

1940 Physics
1950 MCMC Metropolis Hastings
1980 Gibbs Sampling
1990 WinBUGS
. . . OpenBUGS, JAGS, Stan, INLA,
Variational Bayes, bayesmeta

Nowadays, parallel usage of Bayesian and classical paradigms is quite common. See, for example, your individual project motivated by Baeten et al. [2013].

Note that the Bayes approach is also based on the likelihood. Therefore, all problems for classical inference such as uncertainty about the sampling model, randomness of the data (outliers) and model complexity propagate.

However, Bayes needs more work. For example, priors $P[\theta]$ must be elicitated from contextual information. Contextual information is usually provided by mean, standard deviation, minimum, maximum (range). A good understanding of properties of different distributions is necessary in order to define a correct $P[\theta]$ prior. See, for example, distributions zoo: Leemis and McQueston [2008]. Moreover, good communication with experts is necessary to get the correct information. Bayesian computation comprises:

- Conjugate analyses
- MCMC sampling: R, JAGS, OpenBUGS, Stan
- Bayesian numerical approximations: INLA, bayesmeta

Recommended reading: Bayarri and Berger [2004], Martin et al. [2020], and Johnson et al. [2022]. You can also check interactive visualizations Seeing Theory `http://students.brown.edu/seeing-theory/index.html`.

1.4 Probability calculus

The probability calculus is based on three axioms:

$P[A] \geq 0$

$P[A] = 1$ if A is true

$P[A \text{ or } B] = P[A \cup B] = P[A] + P[B]$ if $A \cap B = \emptyset$ and $P[A \text{ and } B] = P[A \cap B] = 0$ (events A and B are mutually exclusive).

There are several important properties of probabilities.

Conditional probability

$$P[A \mid B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{P[A \cap B]}{P[B]}, \tag{1.1}$$

given that $P[B] > 0$.

Two events A and B are called independent if the occurrence of B does not change the probability of A

$$P[A \mid B] = P[A]$$

and vice versa

$$P[B \mid A] = P[B].$$

Thus,

$$P[A \text{ and } B] = P[A]P[B].$$

Note that from Equation (1.1)

$$P[A \text{ and } B] = P[A \mid B]P[B] = P[B \mid A]P[A].$$

This observation leads to the **Bayes theorem**

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B]}. \tag{1.2}$$

Assume that event A has a disjoint, complementary event A^c such that $P[A]+P[A^c] = 1$. Conditional probabilities behave like ordinary probabilities, so that we have

$$P[A \mid B] + P[A^c \mid B] = 1.$$

This leads to the simplest version of the law of total probability:

$$P[B] = P[B \mid A]P[A] + P[B \mid A^c]P[A^c].$$

Therefore, the **Bayes theorem** from Equation (1.2) can be rewritten as

$$P[A \mid B] = \frac{P[B \mid A]P[A]}{P[B \mid A]P[A] + P[B \mid A^c]P[A^c]}. \tag{1.3}$$

For formulas applying to more than two events see Held and Sabanés Bové [2020, Sections A.1.1–A.1.2].

Note that there is a link between probability P and odds O:

$$O = \frac{P}{1-P}$$

and

$$P = \frac{O}{1+O}.$$

We can obtain the odds form for the Bayes theorem

$$\frac{P[A \mid B]}{P[A^c \mid B]} = \frac{P[B \mid A]}{P[B \mid A^c]}\frac{P[A]}{P[A^c]}, \tag{1.4}$$

by dividing Equation (1.2) by the same equation applied to the disjoint, complementary event A^c instead of A.

The ratio

$$\frac{P[A \mid B]}{P[A^c \mid B]} = \frac{P[A \mid B]}{1 - P[A \mid B]}$$

is called posterior odds, and

$$\frac{P[A]}{P[A^c]} = \frac{P[A]}{1 - P[A]}$$

is called prior odds, and the ratio

$$\frac{P[B \mid A]}{P[B \mid A^c]}$$

is the Bayes factor (likelihood ratio).

Remark: This Bayes factor is a measure of evidence [Held and Ott, 2018] of the null hypothesis H_0 against an alternative hypothesis H_A, when we replace events A and A^c in Equation (1.4) by H_0 and H_A.

Remark: One can also derive a conditional version of the Bayes theorem

$$P[A \mid B, I] = \frac{P[A \mid I] P[B \mid A, I]}{P[B \mid I]},$$

where I is an additional piece of information.

Recommended reading: Held and Sabanés Bové [2020]: Sections 6.1 and 6.2, A1, A1.1, A1.2, A2.1, A2.2, A2.3. See also Rouder and Morey [2019] for a deeper insight into the meaning of the Bayes theorem.

1.5 Example: Breast cancer and diagnostic tests

This section demonstrates on one medical example that $P[D^+ \mid T^+] \neq P[T^+ \mid D^+]$.

Assumptions

Prevalence of breast cancer: $P[D^+] = 0.045$

Sensitivity: $P[T^+ \mid D^+] = 0.866$

Specificity: $P[T^- \mid D^-] = 0.968$

$$\text{Full bivariate distribution } P[T \cap D] = \begin{array}{c c} & \begin{array}{cc} D^+_\text{yes} & D^-_\text{no} \end{array} \\ \begin{array}{c} T^+_\text{yes} \\ T^-_\text{no} \end{array} & \begin{pmatrix} 0.03897 & 0.03056 \\ 0.00603 & 0.92444 \end{pmatrix} \end{array}$$

$$\text{marginal distribution for Test } P[T] = \begin{array}{c} T^+ \\ T^- \end{array} \begin{pmatrix} 0.06953 \\ 0.93047 \end{pmatrix}$$

$$\text{marginal distribution for Disease } P[D] = \begin{array}{c} D^+ \\ D^- \end{array} \begin{pmatrix} 0.045 \\ 0.955 \end{pmatrix}$$

$$P[D^- \mid T^-] = \frac{P[D^- \cap T^-]}{P[T^-]} = \frac{0.92444}{0.93047} = 0.994$$

$$P[D^+ \mid T^+] = \frac{P[D^+ \cap T^+]}{P[T^+]} = \frac{0.03897}{0.06953} = 0.56 \neq P[T^+ \mid D^+]$$

Remark: Another way of computing

$$P[D^+ \mid T^+] = \frac{P[D^+] P[T^+ \mid D^+]}{P[T^+]} = \frac{P[D^+] P[T^+ \mid D^+]}{P[D^+] P[T^+ \mid D^+] + P[D^-] P[T^+ \mid D^-]}$$

1.6 Overview of the classical statistic

The classical statistic is based on the likelihood (Figure 1.1).

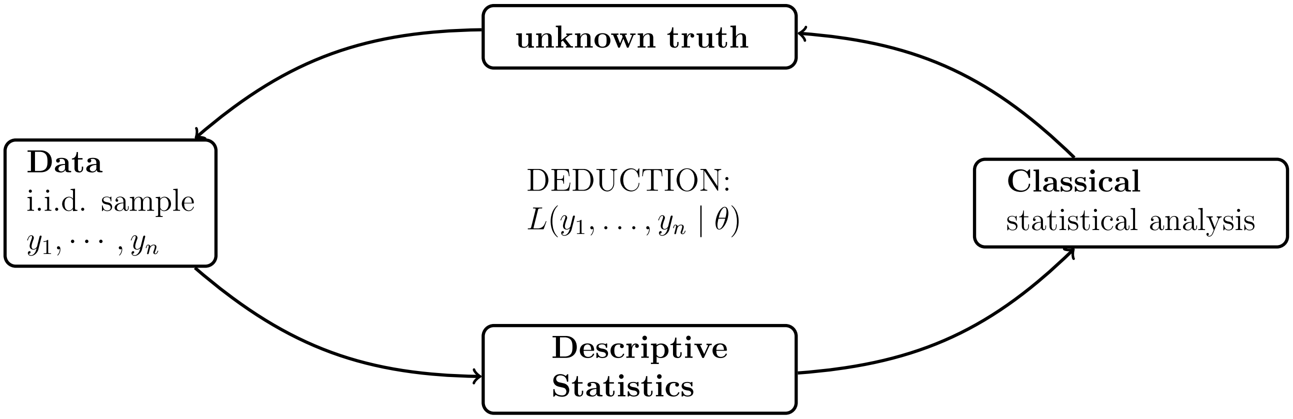


Figure 1.1: Overview of the classical statistic

1.6.1 Example: Primary outcome follows normal distribution

Data $Y \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$ and $\theta = \mu$.

Density $f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$.

Likelihood

$$\begin{aligned} L(y_1, \ldots, y_n \mid \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}, \end{aligned} \tag{1.5}$$

and the log-likelihood

$$\log L(y_1, \ldots, y_n \mid \mu) = n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

In order to derive an estimator of μ, compute

$$\frac{d \log L(y_1, \ldots, y_n \mid \mu)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \mu)(-1)\bigg|_{\mu=\hat{\mu}} = 0,$$

$$\sum_{i=1}^{n} y_i - n\hat{\mu} = 0.$$

Thus,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

One can also derive that $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2}$.

1.6.2 Example: Primary outcome follows Bernoulli distribution

Let $Y \overset{i.i.d.}{\sim} \mathrm{Be}(p)$ with $\theta = p$ and $P[Y = 0] = 1 - p$ and $P[Y = 1] = p$.

Density $f(y_i) = p^{y_i}(1-p)^{1-y_i}$.

Likelihood

$$\begin{aligned} L(y_1, \ldots, y_n \mid p) &= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i}, \end{aligned} \tag{1.6}$$

Log-likelihood

$$\log L(y_1, \ldots, y_n \mid p) = \sum_{i=1}^{n} y_i \log p + \left(n - \sum_{i=1}^{n} y_i\right) \log(1-p).$$

In order to derive an estimator of p, compute

$$\frac{d \log L(y_1, \ldots, y_n \mid p)}{dp} = \sum_{i=1}^{n} y_i \frac{1}{p} + \left(n - \sum_{i=1}^{n} y_i\right) \frac{1}{1-p}(-1)\Bigg|_{p=\hat{p}} = 0,$$

$$\sum_{i=1}^{n} y_i \frac{1}{\hat{p}} - \left(n - \sum_{i=1}^{n} y_i\right) \frac{1}{1-\hat{p}} = 0,$$

$$\sum_{i=1}^{n} y_i (1-\hat{p}) = \left(n - \sum_{i=1}^{n} y_i\right) \hat{p},$$

$$\sum_{i=1}^{n} y_i = n\hat{p},$$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

1.7 Overview of the Bayesian methodology

There is a consent that probability calculus leading to the Bayes formula in Equation (1.2) is objective. Bayesian methodology extends the classical approach based on the likelihood and considers

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

More specifically,

$$P[\theta \mid y_1, \ldots, y_n] \propto \mathrm{L}(y_1, \ldots, y_n \mid \theta) \times P[\theta].$$

Figure 1.2 provides an overview of the Bayesian methodology and its relation to the classical statistics.

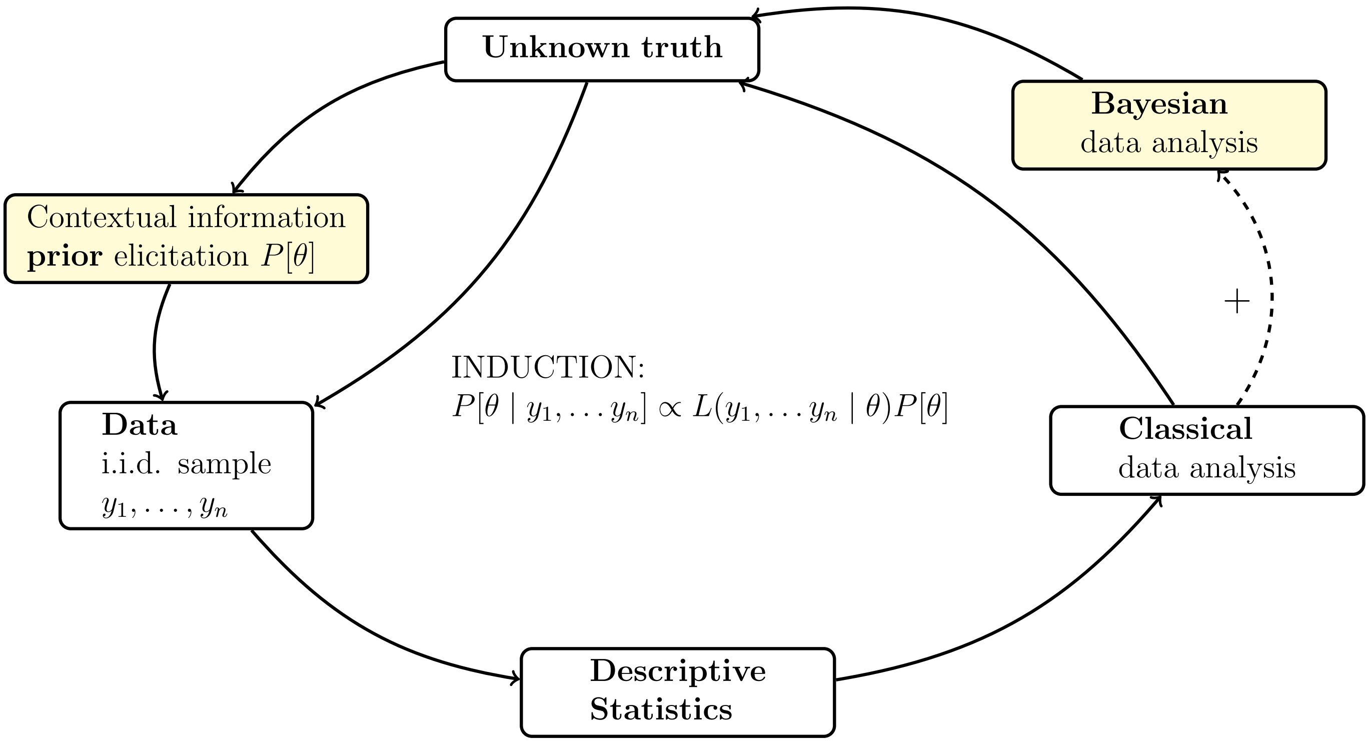


Figure 1.2: Overview of the Bayesian methodology. Fields with yellow background correspond to Bayes-specific steps.

1.7.1 Bayes factors and p-values

The Bayesian methodology enables an independent view of classical hypothesis testing. In applications, the estimation of the credibility of a conclusion expressed by the probability of H_0 given the data is usually of primary interest. The Bayes factor directly quantifies whether the data have increased or decreased the odds of H_0. Thus, Bayes factors facilitate direct conclusions about the probability of H_0 given the data, provided that both null H_0 and alternative H_1 hypotheses have been specified.

On page 70, Held and Sabanés Bové [2020] define the p-value: the probability, under the assumption of the null hypothesis H_0, of obtaining a result equal to or more extreme than what was actually observed. A p-value is computed under the assumption that the null hypothesis H_0 is true. It does not allow for conclusions about the probability of H_0 given the data. A particular p-value can be obtained either for a large study with a small effect or for a small study with a large effect. Thus, the p-value does not say anything about the actual effect or evidence that such an effect exists.

Consider a significance test with a point null hypothesis $H_0 : \theta = \theta_0$. The alternative hypothesis can be either simple $H_1 : \theta = \theta_1 \neq \theta_0$ or composite $H_1 : \theta \neq \theta_0$. For a composite H_1 a prior distribution $f(\theta \mid H_1)$ must be specified.

Note that $P[H_1] = 1 - P[H_0]$ and $P[y] = f(y \mid H_0)P[H_0] + f(y \mid H_1)P[H_1]$. The Bayes formula for H_0

$$P[H_0 \mid y] = \frac{f(y \mid H_0)P[H_0]}{P[y]} \tag{1.7}$$

divided by the Bayes formula for H_1

$$P[H_1 \mid y] = \frac{f(y \mid H_1)P[H_1]}{P[y]} \tag{1.8}$$

render

$$\frac{P[H_0 \mid y]}{P[H_1 \mid y]} = BF_{01}(y)\frac{P[H_0]}{P[H_1]}, \tag{1.9}$$

where

$$BF_{01}(y) = \frac{f(y \mid H_0)}{f(y \mid H_1)}. \tag{1.10}$$

Note that the Bayes factor $BF_{01}(y)$ transforms the prior odds $P[H_0]/P[H_1]$ into posterior odds $P[H_0 \mid y]/P[H_1 \mid y]$ in the light of the data y. $BF_{01}(y)$ is a direct quantitative measure of how data y have increased or decreased the odds of H_0 and is referred to as the strength of evidence for or against H_0. The evidence against the null hypothesis H_0 is provided by small Bayes factors $BF_{01}(y) < 1$. The evidence in favor of the null hypothesis H_0 is provided by large Bayes factors $BF_{01}(y) > 1$. Table 2 of Held and Ott [2018] provides a categorization of Bayes factors $BF_{01}(y) \leq 1$ into levels of evidence against H_0: weak (1 to 1/3), moderate (1/3 to 1/10), substantial (1/10 to 1/30), strong (1/30 to 1/100), very strong (1/100 to 1/300), and decisive (< 1/300).

$BF_{01}(y)$ is the ratio of the likelihood $f(y \mid H_0) = f(y \mid \theta = \theta_0)$ of the observed data y under the null hypothesis H_0 and the marginal likelihood

$$f(y \mid H_1) = \int f(y \mid \theta)f(\theta \mid H_1)d\theta \tag{1.11}$$

under the alternative hypothesis H_1. Equation (1.11) is useful for composite alternative hypotheses H_1. It is the average likelihood $f(y \mid \theta)$ with respect to the prior distribution $f(\theta \mid H_1)$ for θ under the alternative H_1, which is called marginal likelihood (prior predictive distribution at the observed data). For a simple alternative, Equation (1.11) reduces to the likelihood $f(y \mid H_1) = f(y \mid \theta = \theta_1)$ and the $BF_{01}(y)$ reduces to a likelihood ratio.

Once we know $BF_{01}(y)$, we can solve the formula in Equation (1.9) for the posterior probability of H_0. Note that

$$\frac{P[H_0 \mid y]}{1 - P[H_0 \mid y]} = BF_{01}(y)\frac{P[H_0]}{P[H_1]}.$$

Thus,

$$P[H_0 \mid y] = \frac{BF_{01}(y)\frac{P[H_0]}{P[H_1]}}{1 + BF_{01}(y)\frac{P[H_0]}{P[H_1]}}. \tag{1.12}$$

Note that Bayes factors facilitate multiple hypothesis comparisons because they can be updated sequentially:

$$BF_{01}(y)BF_{12}(y) = \frac{f(y \mid H_0)}{f(y \mid H_1)}\frac{f(y \mid H_1)}{f(y \mid H_2)} = \frac{f(y \mid H_0)}{f(y \mid H_2)} = BF_{02}(y).$$

The minimum Bayes factor is the smallest Bayes factor within a certain class of alternative hypotheses. Minimum Bayes factors are very interesting because they quantify the maximal evidence of a p-value against a point H_0 within a certain class of alternative hypotheses.

The Bayesian approach provides a way of transforming p-values to direct measures of evidence against the null hypothesis expressed by Bayes factors. This transformation is called calibration. Held and Ott [2018] consider different transformations of p-values to minimum Bayes factors and show that minimum Bayes factors provide less evidence against the null hypothesis than the corresponding p-value might suggest. They also demonstrate that many techniques have been proposed to calibrate p-values and there is no consensus which calibration is the optimal one.
Recommended reading: Held and Sabanés Bové [2020] Sections 3.3 and 7.2.1, Goodman [1999b], Goodman [1999a], Held and Ott [2018], and `pCalibrate` package.

Example: Discuss `pCalibrate` to show the calibration of p-values by Bayes factors on the border between the classical and the full Bayes analysis.

1.7.2 Priors

The use of prior distributions for Bayesian analysis can be controversial. Therefore, a good understanding of different distributions is very important.

- Discussion of different distributions Leemis and McQueston [2008].
- Monte Carlo (MC) simulations vs true parameters (expectation and variance).
- The Change-of-Variables Formula Held and Sabanés Bové [2020, Section A.2.3].

  Assume a one-to-one and differentiable transformation $g(\dot{})$. Assume that the random variable Y with probability density function $f_Y(y)$ is a transformation of a continuous random variable X with probability density function $f_X(x)$, where g is a one-to-one and differentiable transformation and $Y = g(X)$. Then

  $$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{dg^{-1}(y)}{dy}\right|.$$

1.8 Worksheet 1

Probability calculus	**Distributions**	Change of variables formula
Priors	**MC sampling**	Asymptotics
Bayes		**Classical**
Posterior $\propto$ Likelihood $\times$ Prior		**Likelihood**
Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 1.3: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 1.

Secukinumab		**Placebo**
Classical Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		**Data**
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 1.4: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 2

Lecture 2: Conjugate Bayes, point estimates, and interval estimates

This chapter deals with the conjugate Bayes argument and the resulting posterior Bayesian estimates. The Bayes theorem from Equation (1.2) can be rewritten in terms of densities

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)},$$

where $f(y) = \int f(y \mid \theta) f(\theta) d\theta$ is called the marginal likelihood or the prior predictive distribution at the observed data y. A conjugate Bayes emerges when both the likelihood and the prior are based on distributions that allow for a direct multiplication of distributional cores. This leads to a simple rule of computation

$$f(\theta \mid y) \propto f(y \mid \theta) f(\theta),$$

which will be demonstrated on two examples.

This simple rule can be perceived in two different ways. First, we start with the prior $f(\theta)$ which is then modified by the current data contained in the likelihood $f(y \mid \theta)$ to get the posterior $f(\theta \mid y)$. Second, we start with the current data summarized by the likelihood $f(y \mid \theta)$ and this likelihood is modified by the prior $f(\theta)$ to get the posterior $f(\theta \mid y)$. Depending on application, the first or the second way of reading this simple rule is more useful, which makes the Bayesian methodology very practicable.

Recommended reading: Held and Sabanés Bové [2020] Sections 6.3.1 and 6.4 and Hartnack and Roos [2021]. Note that Table 6.2 of Held and Sabanés Bové [2020] is very relevant. It shows which posterior distributions can be derived analytically given a likelihood and a conjugate prior.

Remark: Rouder and Morey [2019] discuss in detail the ratio form of the Bayes theorem

$$\frac{f(\theta \mid y)}{f(\theta)} = \frac{f(y \mid \theta)}{f(y)}$$

in terms of updating factors. Whereas the factor to the left denotes the strength of evidence form the data about θ, the factor to the right denotes the gain in predictive accuracy for θ, i.e. how well the data are predicted when conditioned on the value of θ relative to the marginal prediction. Thus, the strength of evidence is the relative predictive accuracy.

2.1 Binary data: Vision correction

We consider an example of vision correction. In a class with $20 - 30$ years old students, 16 participants out of 22 required vision correction. What is the true probability π that young students need vision correction?

We know from the classical statistics that the best practice to analyze these data is to provide both a point estimate $\hat{\pi} = 16/22 = 0.727$ and an interval estimate in form of a 95% confidence interval for the true probability π (95%CI(π)) [Hartnack and Roos, 2021].

```
library(DescTools)

BinomCI(x = 16, n = 22, conf.level = 0.95, method = "wilson")

##            est    lwr.ci    upr.ci
## [1,] 0.7272727 0.5184827 0.8684924
```

The interpretation of a classical 95%CI is based on a repeated execution of the same experiment. A confidence interval with confidence $1 - \alpha$ provides limits T_l and T_u such that for the parameter of interest θ

$$P[T_l \leq \theta \leq T_u] = 1 - \alpha$$

holds. Classical confidence intervals aim to uniformly provide a prespecified coverage probability conditionally on any single point in parameter space.

Theoretical interpretation: For repeated random samples from a distribution with unknown parameter θ, a $(1 - \alpha)\dot{1}00\%$ confidence interval will cover θ in $(1 - \alpha)\dot{1}00\%$ of all cases (Held and Sabanés Bové [2020] on page 57 in Section 3.2.2).

Practical interpretation: In the vision correction example for the random sample at hand, we can say that the 95%CI for the true probability π of vision correction is 95%CI(π) (0.518, 0.868). Although we do not know whether this 95%CI(π) covers the true probability π of vision correction, we know that the procedure used for computation of the 95%CI(π) has a desirable characteristic. For repeated independent random samples drawn from the distribution of vision correction with the unkown true parameter π, 95%CI(π) intervals cover the true probability π in approximately 95% of cases.

2.2 Bayes analysis of binary data

2.2.1 Beta prior

A prior expresses contextual information or knowledge in form of a probability distribution. The Beta distribution is based on the observation that the integral $\int_0^1 u^{x-1}(1-u)^{y-1}du$ exists. This integral is called the Beta function $B(x, y)$.

- Beta function

$$\int_0^1 u^{x-1}(1-u)^{y-1}du = B(x, y),$$

where

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

and

$$\Gamma(x) = (x-1)!$$

- Beta distribution

  Let

$$f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{2.1}$$

be the density of the Beta(α,β) distribution with two shape parameters α and β. Then

$$\frac{1}{B(\alpha,\beta)} \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp = 1.$$

Beta distribution is very flexible. It can attain different forms which can be symmetric and asymmetric. Elicitation of shape parameters α and β by moments matching is a convenient way to define a Beta prior (See your individual project Part 2B).

$$X \sim \text{Beta}(\alpha,\beta)$$
$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta}$$
$$\text{Var}X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mathbb{E}(X)(1-\mathbb{E}X)}{(\alpha+\beta+1)}.$$

Prior effective sample size

$$\text{PriESS} \approx \frac{1}{\text{Var}(X)} \approx \frac{\mathbb{E}(X)(1-\mathbb{E}X) - \text{Var}(X)}{\text{Var}(X)} = \alpha+\beta \tag{2.2}$$

Therefore, it is convenient to think about $\alpha+\beta$ as a prior sample size. This number informs us about the weight of the prior.

Example Vision correction:
We consider three Beta priors, which are depicted in Figure 2.1.

- skeptical prior $\alpha+\beta = 0.5+0.5 = 1$ (equivalent to 1 observation)
- neutral prior $\alpha+\beta = 1+1 = 2$ (equivalent to 2 observations)
- enthusiastic prior $\alpha+\beta = 12+12 = 24$ (equivalent to more observations than in the final sample of 22)

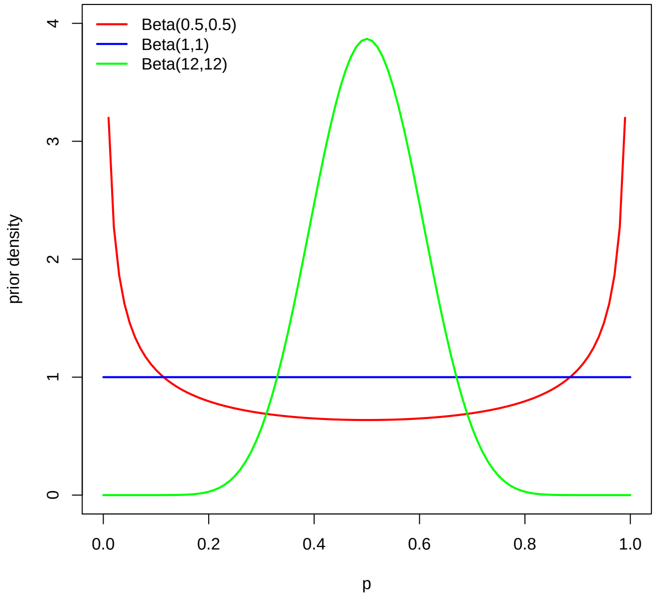


Figure 2.1: Beta priors in the vision correction example.

2.2.2 Likelihood

We assume that each binary observation y_i is a realization of independent identically distributed ($i.i.d.$) random variables, which follow the Bernoulli Be(p) distribution:

$$y_i \overset{i.i.d.}{\sim} \text{Be}(p) = \begin{cases} 1, & \text{with } p \\ 0, & \text{with } 1-p \end{cases} = p^{y_i}(1-p)^{1-y_i}, \quad i = 1, \cdots, n.$$

The likelihood is equal

$$L(y_1, \cdots, y_n \mid p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} = p^{n\bar{y}}(1-p)^{n-n\bar{y}}, \qquad (2.3)$$

where $n\bar{y} = n(\frac{1}{n}\sum_{i=1}^n y_i) = \sum_{i=1}^n y_i$ is the number of binary observations attaining value 1 in the sample of n observations. This likelihood is proportional to the binomial likelihood.

2.2.3 Posterior distribution

We begin with the computation of the posterior distribution with all constants. Note that due to conjugacy, the numerator of the Bayes formula, which is the multiplication of the likelihood $L(y_1, \cdots, y_n \mid p)$ from Equation (2.3) and the Beta(α, β) prior with density in Equation (2.1), is equal:

$$L(y_1, \cdots, y_n \mid p) f(p) = p^{n\bar{y}}(1-p)^{n-n\bar{y}} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} = \frac{1}{B(\alpha,\beta)} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}.$$

Therefore,

$$\begin{aligned}
f(p \mid y_1, \cdots, y_n) &= \frac{L(y_1, \cdots, y_n \mid p) f(p)}{\int_0^1 L(y_1, \cdots, y_n \mid p) f(p) dp} \\
&= \frac{\frac{1}{B(\alpha,\beta)} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}}{\frac{1}{B(\alpha,\beta)} B(\alpha+n\bar{y}, \beta+n-n\bar{y}) \underbrace{\int_0^1 \frac{1}{B(\alpha+n\bar{y}, \beta+n-n\bar{y})} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1} dp}_{=1}} \\
&= \frac{1}{B(\alpha+n\bar{y}, \beta+n-n\bar{y})} p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}.
\end{aligned} \tag{2.4}$$

Thus, $p \mid y_1, \cdots, y_n \sim \text{Beta}(\alpha+n\bar{y}, \beta+n-n\bar{y})$.

Alternatively, we can identify the posterior distribution based on cores of distributions:

$$\begin{aligned}
f(p \mid y_1, \cdots, y_n) &\propto L(y_1, \cdots, y_n \mid p) f(p) \\
&= \underbrace{p^{n\bar{y}}(1-p)^{n-n\bar{y}}}_{\text{likelihood kernel}} \underbrace{p^{\alpha-1}(1-p)^{\beta-1}}_{\text{prior kernel}} \\
&= \underbrace{p^{\alpha+n\bar{y}-1}(1-p)^{\beta+n-n\bar{y}-1}}_{\text{kernel of the posterior distribution Beta}(\alpha+n\bar{y}, \beta+n-n\bar{y})}.
\end{aligned} \tag{2.5}$$

Again, $p \mid y_1, \cdots, y_n \sim \text{Beta}(\alpha+n\bar{y}, \beta+n-n\bar{y})$.

The expectation of the posterior distribution is equal:

$$\begin{aligned}
\mathbb{E}(p \mid y_1, \cdots, y_n) &= \frac{\alpha+n\bar{y}}{\alpha+n\bar{y}+\beta+n-n\bar{y}} = \frac{\alpha+n\bar{y}}{\alpha+\beta+n} \\
&= \frac{\alpha}{\alpha+\beta+n} + \frac{n\bar{y}}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n} \underbrace{\frac{\alpha}{\alpha+\beta}}_{\mathbb{E}(\text{prior})} + \big(1 - \frac{\alpha+\beta}{\alpha+\beta+n}\big) \underbrace{\bar{y}.}_{\text{MLE}}
\end{aligned} \tag{2.6}$$

It is a weighted average of the prior mean $\frac{\alpha}{\alpha+\beta}$ and the ML estimate $\bar{y}$. The relative prior sample size $\frac{\alpha+\beta}{\alpha+\beta+n}$ quantifies the weight of the prior mean in the posterior expectation. Note that this quantity decreases to 0 with data sample size increasing to ∞.

Posterior effective sample size (PostESS) of a $\text{Beta}(\alpha+n\bar{y}, \beta+n-n\bar{y})$ distribution is approximatively equal:

$$PostESS \approx \overbrace{\frac{1}{\text{Var}(p \mid y_1, \cdots, y_n)}}^{\text{posterior precision}} \approx \alpha+n\bar{y}+\beta+n-n\bar{y} = \alpha+\beta+n. \tag{2.7}$$

Note the analogy to the prior effective sample size.

Example: Vision correction

Data: $n\bar{y} = 16$ participants out of $n = 22$ required vision correction. Note that $n - n\bar{y} = 6$ participants did not require any vision correction. Figure 2.2 shows the standardized likelihood, the Beta(0.5, 0.5) prior and the resulting Beta(16.5, 6.5) posterior distributions.

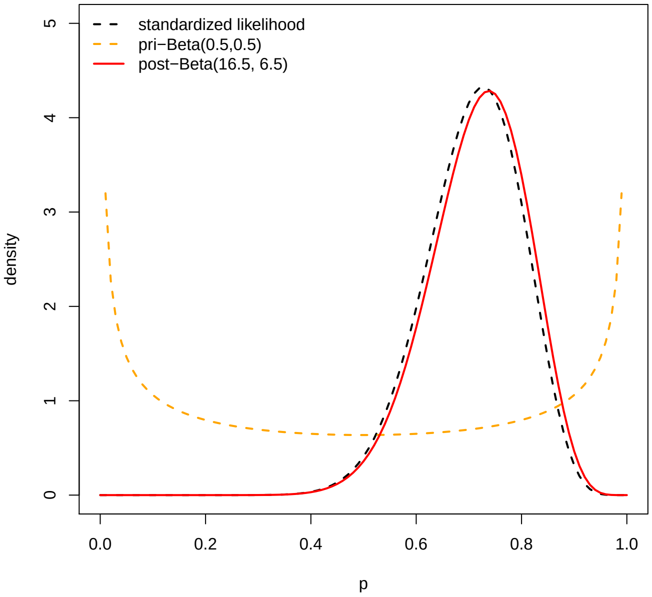


Figure 2.2: Likelihood, prior Beta(0.5, 0.5), posterior Beta(16.5, 6.5) in the vision correction example with $n\bar{y} = 16$ and $n = 22$.

Posterior distributions can be summarized by point and interval estimates. For example, for the Beta(16.5, 6.5) posterior we get the posterior mean $16.5/(16.5 + 6.5) = 0.717$. We can also compute an equi-tailed 95% credible interval (95%CrI):

```
qbeta(p = c(0.025, 0.975), shape1 = 16.5, shape2 = 6.5)

## [1] 0.5217688 0.8772947
```

The **interpretation of a 95%CrI** differs from that of confidence intervals. We can state that the posterior probability of vision correction p lies between 0.521 and 0.877 with probability 95%, when a Beta(0.5, 0.5) prior is assumed. This result addresses the actual question more directly and can be interpreted intuitively. Bayesian credible intervals account for the prior distribution and provide a coverage on average over the prior. They directly relate to the knowledge about the parameter after considering the data at hand.

For different priors, we get

- The skeptical prior Beta(0.5, 0.5) leads to a Beta(16.5, 6.5) posterior.
- The neutral prior Beta(1, 1) leads to a Beta(17, 7) posterior.
- The enthusiastic prior Beta(12, 12) leads to a Beta(28, 18) posterior.

These posterior distributions are shown in Figure 2.3.

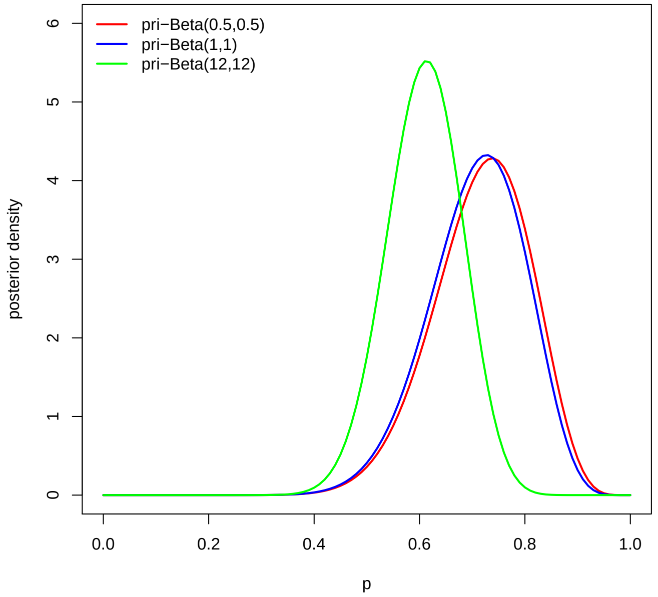


Figure 2.3: Posteriors in the vision correction example obtained for three different priors: Beta(0.5, 0.5), Beta(1, 1), and Beta(12, 12).

Note that posterior results depend on the prior. For the Beta(28, 18) posterior we get the posterior mean $28/(28 + 18) = 0.609$ and the equi-tailed 95%CrI:

```
qbeta(p = c(0.025, 0.975), shape1 = 28, shape2 = 18)

## [1] 0.4654101 0.7430241
```

2.3 Bayes analysis of normal data

Assume that $y_1, \ldots, y_n$ are realizations (observations) generated by *i.i.d.* random variables which follow a $\mathrm{N}(m, \kappa^{-1})$ distribution. Assume that the prior of m follows a $\mathrm{N}(\mu, \lambda^{-1})$ distribution, where κ, μ and λ are fixed (known) constants. We derive the posterior distribution of m given that $y_1, \ldots, y_n$ have been observed.

According to the Bayes formula in Equation (1.2), which has been rewritten in terms of densities, we get:

$$f(m \mid y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n \mid m) f(m)}{\int_{-\infty}^{\infty} f(y_1, \ldots, y_n \mid m) f(m) dm}. \tag{2.8}$$

The denominator is known as the marginal likelihood:

$$\int_{-\infty}^{\infty} f(y_1, \ldots, y_n \mid m) f(m) dm = \int_{-\infty}^{\infty} f(y_1, \ldots, y_n, m) dm = f(y_1, \ldots, y_n).$$

We derive the posterior based on kernels of distributions and use:

$$\underbrace{f(m \mid y_1, \ldots, y_n)}_{\text{posterior}} \propto \underbrace{f(y_1, \ldots, y_n \mid m)}_{\text{likelihood}} \underbrace{f(m).}_{\text{prior}} \tag{2.9}$$

We combine the likelihood

$$f(y_1, \ldots, y_n \mid m) = \left(\frac{\kappa}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i - m)^2\right\}$$

and the prior

$$f(m) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(m - \mu)^2\right)$$

and, following Equation (2.9), we get

$$f(m \mid y_1, \ldots, y_n) \propto \exp\left\{-\frac{\kappa}{2}\sum_{i=1}^{n}(y_i - m)^2 - \frac{\lambda}{2}(m - \mu)^2\right\}.$$

At this stage, formulas for combining quadratic forms (Held and Sabanés Bové [2020] Section B.1.5) can be applied to show that

$$f(m \mid y_1, \ldots, y_n) \propto \exp\left\{-\frac{(n\kappa + \lambda)}{2}\left(m - \frac{\kappa n\bar{y} + \lambda\mu}{n\kappa + \lambda}\right)^2\right\}.$$

Thus, we get

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n\bar{y} + \lambda\mu}{n\kappa + \lambda}, (n\kappa + \lambda)^{-1}\right).$$

Note that the expectation of the posterior can be rewritten as

$$\frac{\lambda}{n\kappa + \lambda}\mu + \big(1 - \frac{\lambda}{n\kappa + \lambda}\big)\bar{y}.$$

The weight $\frac{\lambda}{n\kappa+\lambda}$ of the prior mean in the posterior expectation decreases to 0 with data sample size increasing to ∞. Moreover, the variance of the posterior distribution $(n\kappa + \lambda)^{-1}$ is smaller than the variance λ^{-1} of the prior distribution, because the precision is larger. The posterior variance decreases to 0 with data sample size increasing to ∞. In addition, the prior effective sample size is equal $PriESS \approx \lambda$ and the posterior effective sample size $PostESS \approx n\kappa + \lambda$ is larger than $PriESS$.

2.4 Point estimates

Bayesian point estimates such as mean, mode, and median have a deeper decision-theoretic meaning. They minimize an expected loss with respect to the posterior distribution. For example, the posterior mean minimizes the quadratic loss function $l(a, \theta) = (a - \theta)^2$, because the first derivative with respect to a of $\mathbb{E}(l(a, \theta) \mid y) = \int (a - \theta)^2 f(\theta \mid y) d\theta$ set to 0 results in $a = \int \theta f(\theta \mid y) d\theta = \mathbb{E}(\theta \mid y)$. For more details on point estimates and loss functions see Held and Sabanés Bové [2020, Section 6.4].

2.5 Credible intervals

There are at least two ways to compute Bayesian credible intervals $(1-\alpha)100\%\mathrm{CrI}(\theta)$:

(a) equi-tailed credible intervals

(b) highest posterior density (HPD) intervals.

These Bayesian credible intervals allow for direct probability statements. However, they have different properties.

An equi-tailed $(1-\alpha)$ credible interval has $\frac{\alpha}{2}$, $1-\frac{\alpha}{2}$ quantiles of $\pi(\theta \mid y)$ at its endpoints. One discards equal amounts of posterior probability on either side of the interval. An equi-tailed credible interval is:

- Intuitively straightforward
- Easy to compute from MC and MCMC samples
- Has a nice invariance property

Let h be a monotone function (can be non-linear). A $(1-\alpha)$ equi-tailed credible interval for $h(\theta \mid y)$ can be obtained by applying $h()$ to the endpoints of the $(1-\alpha)$ equi-tailed credible interval for $\theta \mid y$. There is no concern about the chosen scale for inference. In fact, this property applies to all quantiles (also median) of the posterior.

Remark: The invariance property doesn't hold for expectations. In fact, there is functional non-invariance for a non-linear function $g()$:

$$\mathbb{E}(g(Y)) = \int_\Omega g(u)dP_Y \neq g(\mathbb{E}(Y)).$$

For example, if g is convex then $\mathbb{E}(g(Y)) \geq g(\mathbb{E}(Y))$ and if g is concave $\mathbb{E}(g(Y)) \leq g(\mathbb{E}(Y))$. See Jensen's inequality in [Held and Sabanés Bové, 2020] page 354, Section A.3.7.

The highest posterior probability HPD interval fulfills $\mathbb{P}[\theta : f(\theta \mid y) > c] = 1-\alpha$ and provides the shortest possible interval. For symmetric and unimodal posteriors, it coincides with $(1-\alpha)$ equi-tailed credible intervals. But for bimodal or multimodal posteriors this correspondence does not hold. Note that the invariance property for transformation $h()$ does not hold any more. For more details on Bayesian HPD credible intervals see Held and Sabanés Bové [2020, Section 6.4].

Remark: Sequential step by step vs pooled data in one step.
Assume a sequence of three measurements y_1, y_2, y_3. Then,

$$\begin{aligned}\mathbb{P}[\theta \mid y_1, y_2, y_3, I] &\propto \mathbb{P}[y_3 \mid \theta, y_1, y_2, I]\mathbb{P}[\theta \mid y_1, y_2, I] \\ &\propto \mathbb{P}[y_3 \mid \theta, y_1, y_2, I]\mathbb{P}[y_2 \mid \theta, y_1, I]\underbrace{\mathbb{P}[y_1 \mid \theta, I]\overbrace{\mathbb{P}[\theta \mid I]}^{\text{prior}}}_{\propto\ \mathbb{P}[\theta \mid y_1, I]} \\ &\propto \underbrace{\prod_{i=1}^{3}\mathbb{P}[y_i \mid \theta, I]}_{\text{pooled likelihood}}\mathbb{P}[\theta \mid I].\end{aligned}$$

Example:

Step by step

1.1 Prior Beta(α, β)

1.2 Data y_1

1.3 Posterior $(\alpha + y_1, \beta + 1 - y_1)$

2.1 Prior Beta$(\alpha + y_1, \beta + 1 - y_1)$

2.2 Data y_2

2.3 Posterior
Beta$(\alpha + y_1 + y_2, \beta + 2 - (y_1 + y_2))$

Pooled

a Prior Beta(α, β)

b Data y_1, y_2

c Posterior
Beta$(\alpha + y_1 + y_2, \beta + 2 - (y_1 + y_2))$

Note that $y_1 + y_2$ is a sufficient statistic with respect to the Binomial likelihood "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter (p)".

2.6 Worksheet 2

Probability calculus	**Distributions**	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ **Likelihood** $\times$ **Prior**		Likelihood
Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 2.1: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 2.

Secukinumab		**Placebo**
Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		**Beta(11, 32)**
Data (S)		**Data (P)**
	Classical analysis	
Posterior (S)		**Posterior (P)**
	Posterior probability of superiority	

Table 2.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 3

Lecture 3: Predictive distributions, asymptotics, and Monte Carlo simulations

This chapter further explores the strength and the beauty of the Bayes formula from Equation (1.2) rewritten in terms of densities:

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)}.$$

Bayesian analysis can be conducted at several stages.

Prior stage:

- prior distribution
- prior predictive distribution

Intermediate stage:

- intermediate posterior distribution
- intermediate posterior predictive distribution

Final stage:

- final posterior distribution
- final posterior predictive distribution

This chapter deals with prior predictive and posterior predictive distributions, independent Monte Carlo sampling, and Bayesian asymptotics.

Recommended reading: Held and Sabanés Bové [2020] Sections 9.3.1, 9.3.2, 8.3.1, and 6.6. See also Spiegelhalter et al. [1994] for an extensive and nuanced discussion of Bayesian approaches to randomized clinical trials.

3.1 Predictive distributions for binary data

We use notation from Section 2.2 and extend the argument to predictive distributions for binary data.

3.1.1 Prior predictive distribution

$$f(y_1,\ldots,y_k) = \int_0^1 f(y_1,\ldots,y_k,p)dp = \int_0^1 \underbrace{f(y_1,\ldots,y_k \mid p)}_{\text{Binomial likelihood}} \underbrace{f(p)}_{prior} dp$$

Let

$$\bar{y}^{(k)} = \frac{1}{k}\sum_{i=n+1}^{n+k} y_i$$

If

$$n = 0 \implies \bar{y}^{(k)} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

$$\begin{aligned}
&\int_0^1 \binom{k}{k\bar{y}^{(k)}} p^{k\bar{y}^{(k)}}(1-p)^{k-k\bar{y}^{(k)}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}dp \\
&= \binom{k}{k\bar{y}^{(k)}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k\bar{y}^{(k)})\Gamma(\beta+k-k\bar{y}^{(k)})}{\Gamma(\alpha+\beta+k)} \underbrace{\int_0^1 \frac{1}{\mathrm{B}(\alpha+k\bar{y}^{(k)})} p^{\alpha+k\bar{y}^{(k)}-1}(1-p)^{\beta+k-k\bar{y}^{(k)}-1}dp}_{=1} \\
&= \binom{k}{k\bar{y}^{(k)}} \frac{\mathrm{B}(\alpha+k\bar{y}^{(k)}, \beta+k-k\bar{y}^{(k)})}{\mathrm{B}(\alpha,\beta)}
\end{aligned} \tag{3.1}$$

Remark For $k = 1$, we get a Bernoulli distribution for one future observation y

$$\begin{aligned}
\binom{1}{y}\frac{\mathrm{B}(\alpha+y,\beta+1-y)}{\mathrm{B}(\alpha,\beta)} &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(\beta+1-y)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{1}{\alpha+\beta}\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)}\frac{\Gamma(\beta+1-y)}{\Gamma(\beta)} \\
&= \begin{cases} \frac{\alpha}{\alpha+\beta}, & \text{if } y = 1 \\ \frac{\beta}{\alpha+\beta}, & \text{if } y = 0 \end{cases}.
\end{aligned} \tag{3.2}$$

Figure 3.1 demonstrates that the prior predictive distribution for binary data shows the probability of the number of events in a future sample based on k observations.

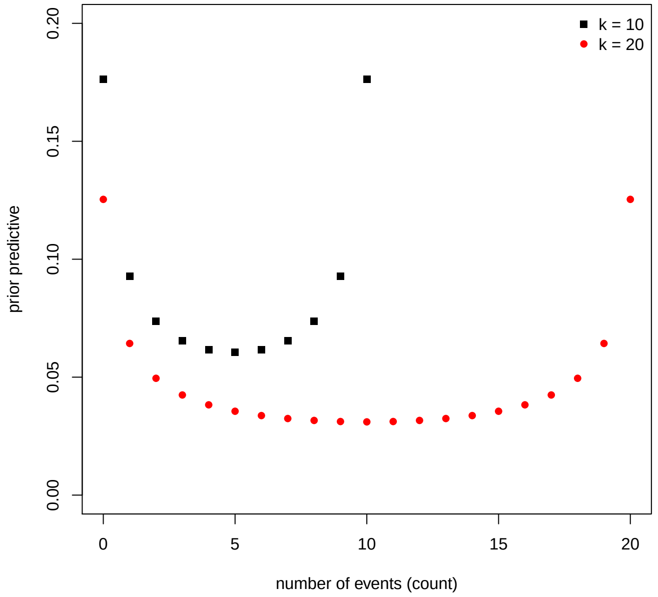


Figure 3.1: Prior predictive distribution for k future binary observations given a Beta(0.5, 0.5) prior.

3.1.2 Posterior predictive distribution

$$\begin{aligned} f(\underbrace{y_{n+1},\dots,y_{n+k}}_{\text{future observations}} \mid \underbrace{y_1,\dots,y_n}_{\text{known obserations}}) &= \int_0^1 f(y_{n+1},\dots,y_{n+k},p \mid y_1,\dots,y_n)dp \\ &\overset{\text{i.i.d. sample}}{=} \int_0^1 f(y_{n+1},\dots,y_{n+k} \mid p)\,\underbrace{f(p \mid y_1,\dots,y_n)}_{\text{posterior distribution}}\,dp \\ &= \text{compute again}\dots \\ &= \text{or use the conjugacy and the formula derived for} \\ &\text{the prior predictive distribution} \end{aligned} \tag{3.3}$$

Denote $\bar{y}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Recall that

$$p \mid y_1,\dots,y_n \sim \text{Beta}(\alpha + n\bar{y}^{(n)}, \beta + n + n\bar{y}^{(n)}).$$

Thus,

$$f(\underbrace{y_{n+1},\dots,y_{n+k}}_{\text{future observations}} \mid \underbrace{y_1,\dots,y_n}_{\text{known obserations}}) = \binom{k}{k\bar{y}^{(k)}} \frac{\mathrm{B}(\alpha + n\bar{y}^{(n)} + k\bar{y}^{(k)}, \beta + n - n\bar{y}^{(n)} + k - k\bar{y}^{(k)})}{\mathrm{B}(\alpha + n\bar{y}^{(n)}, \beta + n - n\bar{y}^{(n)})}. \tag{3.4}$$

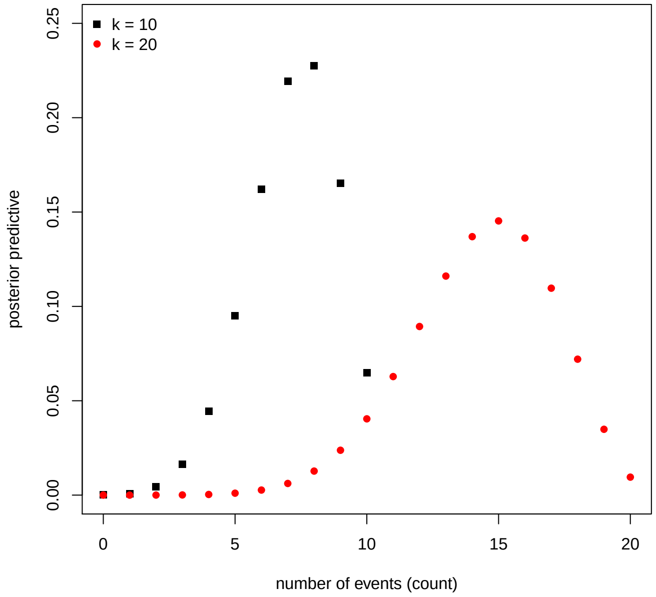


Figure 3.2: Posterior predictive distribution for k future observations given already observed data $n\bar{y}^{(n)} = 16$ in $n = 22$ observations and prior Beta(0.5, 0.5).

Remark With $k = 1$ we get a Bernoulli distribution for the posterior predictive distribution of one future observation y_{n+1} given $y_1, \dots, y_n$.

$$\text{Be}\left(\frac{\alpha+n\bar{y}^{(n)}}{\alpha+\beta+n}\right) = \begin{cases} \frac{\alpha+n\bar{y}^{(n)}}{\alpha+\beta+n}, & \text{if } y_{n+1} = 1, \\ \frac{\beta+n-n\bar{y}^{(n)}}{\alpha+\beta+n}, & \text{if } y_{n+1} = 0. \end{cases}$$

Example: Vision correction

Data: $n\bar{y}^{(n)} = 16$ participants out of $n = 22$ required vision correction. Note that $n - n\bar{y}^{(n)} = 6$ participants did not require any vision correction. Data $n\bar{y}^{(n)} = 16$ and $n = 22$ combined with the prior Beta(0.5, 0.5) result in a posterior Beta(16.5, 6.5), which reflects our current knowledge. Figure 3.2 shows the posterior predictive distribution, which is based on this current posterior knowledge, for the number of events in future samples based on $k = 10$ and $k = 20$ observations.

This posterior predictive distribution can be used to provide prediction intervals [Hartnack and Roos, 2021].

```
PI(x = 16, n = 22, a = 0.5, b = 0.5, k = 10, conf.level = 0.95)

## lower upper
##     4    10

PI(x = 16, n = 22, a = 0.5, b = 0.5, k = 20, conf.level = 0.95)

## lower upper
##     9    19
```

3.2 Predictive distributions for normal data

This section is a continuation of the argument presented in Section 2.3. Assume that $y_1, \ldots, y_n$ are i.i.d. observations generated by a sampling distribution $\mathrm{N}(m, \kappa^{-1})$ and the prior for m follows a $\mathrm{N}(\mu, \lambda^{-1})$ distribution, where κ, μ and λ are fixed constants.

3.2.1 Prior predictive distribution

We derive analytically the prior predictive distribution for one future observation y assuming that no observations have been collected yet. The prior predictive distribution is defined by

$$f(y) = \int_{-\infty}^{\infty} f(y \mid \theta) f(\theta) d\theta, \tag{3.5}$$

where $f(\theta)$ is the prior and $f(y \mid \theta)$ is the likelihood. Thus,

$$\begin{aligned}
f(y) &= \int_{-\infty}^{\infty} f(y \mid m) f(m) dm \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\kappa^{-1}}} \exp\left(-\frac{\kappa(y-m)^2}{2}\right) \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(-\frac{\lambda(m-\mu)^2}{2}\right) dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\kappa(y^2+m^2-2my)+\lambda(m^2+\mu^2-2m\mu)\right)\right\} dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\kappa y^2+\lambda\mu^2-\frac{(\kappa y+\lambda\mu)^2}{\kappa+\lambda}+(\kappa+\lambda)\left(m-\frac{\kappa y+\lambda\mu}{\kappa+\lambda}\right)^2\right)\right\} dm \\
&= \frac{\sqrt{\kappa\lambda}}{2\pi} \exp\left\{-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right\} \sqrt{\frac{2\pi}{\kappa+\lambda}} \underbrace{\int_{-\infty}^{\infty} \sqrt{\frac{\kappa+\lambda}{2\pi}} \exp\left\{-\frac{\kappa+\lambda}{2}\left(m-\frac{\kappa y+\lambda\mu}{\kappa+\lambda}\right)^2\right\} dm}_{=1}
\end{aligned} \tag{3.6}$$

According to the property of probability density functions, the integral at the right hand side above is equal to 1. The remaining part of the equation reads:

$$f(y) = \sqrt{\frac{\kappa\lambda}{2\pi(\kappa+\lambda)}} \exp\left\{-\frac{\kappa\lambda(y-\mu)^2}{2(\kappa+\lambda)}\right\} = \sqrt{\frac{1}{2\pi\left(\frac{1}{\lambda}+\frac{1}{\kappa}\right)}} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\left(\frac{1}{\lambda}+\frac{1}{\kappa}\right)}\right\}$$

.

Hence, the prior predictive distribution of one future observation y is $\mathrm{N}\left(\mu, \lambda^{-1}+\kappa^{-1}\right)$.

3.2.2 Posterior predictive distribution

Derive analytically the posterior predictive distribution for one future observation y_{n+1} given $y_1, \ldots, y_n$ have been observed. According to the definition of the posterior predictive distribution in [Held and Sabanés Bové, 2020] Section 9.3.1, we have

$$\begin{aligned} f(y_{n+1} \mid y_1, \ldots, y_n) &= \int_{-\infty}^{\infty} f(y_{n+1}, m \mid y_1, \ldots, y_n) dm \\ &= \int_{-\infty}^{\infty} f(y_{n+1} \mid m, y_1, \ldots, y_n) f(m \mid y_1, \ldots, y_n) dm \\ &\overbrace{=}^{\text{cond. ind.}} \int_{-\infty}^{\infty} \underbrace{f(y_{n+1} \mid m)}_{\text{likelihood}} \underbrace{f(m \mid y_1, \ldots, y_n)}_{\text{posterior density}} dm \end{aligned} \tag{3.7}$$

We have

$$f(y_{n+1} \mid m) = \sqrt{\frac{\kappa}{2\pi}} \exp\left(\frac{\kappa(y_{n+1} - m)^2}{2}\right)$$

We have already derived the posterior distribution in Section 2.3

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n\bar{y} + \lambda\mu}{n\kappa + \lambda}, \frac{1}{n\kappa + \lambda}\right).$$

Denote

$$\mu_{\text{post}} = \frac{\kappa n\bar{y} + \lambda\mu}{n\kappa + \lambda}$$

and

$$\lambda_{\text{post}} = n\kappa + \lambda.$$

Hence,

$$f(y_{n+1} \mid y_1, \ldots, y_n) = \int_{-\infty}^{\infty} \sqrt{\frac{\kappa}{2\pi}} \exp\left(-\frac{\kappa}{2}(y_{n+1} - m)^2\right) \sqrt{\frac{\lambda_{\text{post}}}{2\pi}} \exp\left(-\frac{\lambda_{\text{post}}}{2}(m - \mu_{\text{post}})^2\right) dm.$$

Following the argument for the prior predictive distribution in Section 3.2.1, we obtain the posterior predictive distribution:

$$y_{n+1} \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\mu_{\text{post}}, \lambda_{\text{post}}^{-1} + \kappa^{-1}\right).$$

Summary of results for the posterior distribution, the prior predictive distribution, and the posterior predictive distribution for normal observations. Frequently, these distributions are parametrized by variances or standard deviations instead of precisions.

Prerequisites:

Data: $y_1, y_2, \ldots, y_n \sim \mathrm{N}(m, \sigma^2) = \mathrm{N}(m, \kappa^{-1})$ with $\kappa = 1/\sigma^2$.

Prior: $m \sim \mathrm{N}(\mu, \tau^2) = \mathrm{N}(\mu, \lambda^{-1})$ with $\lambda = 1/\tau^2$.

Posterior distribution:

Posterior distribution parametrized by precisions:

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda}, (n\kappa + \lambda)^{-1}\right). \tag{3.8}$$

Posterior distribution parametrized by variances:

$$m \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, (\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1}\right) = \mathrm{N}\left(\frac{\tau^2 n\bar{y} + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right). \tag{3.9}$$

Note that posterior variance is smaller than the prior variance, because the precision gets larger. The impact of the prior decreases with increasing sample size.

Prior predictive distribution:

Prior predictive distribution parametrized by precisions:

$$y \sim \mathrm{N}\left(\mu, \lambda^{-1} + \kappa^{-1}\right) \tag{3.10}$$

Prior predictive distribution parametrized by variances:

$$y \sim \mathrm{N}\left(\mu, \sigma^2 + \tau^2\right) \tag{3.11}$$

The variance of the prior predictive distribution is larger than the variance of the prior and the data alone.

Posterior predictive distribution:

Posterior predictive distribution parametrized by precisions:

$$y_{n+1} \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\kappa n \bar{y} + \lambda \mu}{n\kappa + \lambda}, (n\kappa + \lambda)^{-1} + \kappa^{-1}\right) \tag{3.12}$$

Posterior predictive distribution parametrized by variances:

$$y_{n+1} \mid y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\tau^2 n\bar{y} + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2} + \sigma^2\right) \tag{3.13}$$

The variance of the posterior predictive distribution is larger than the variance of the data alone. The impact of the prior decreases with increasing sample size.

Remark: The Bayes theorem from Equation (1.2) can be rewritten in terms of densities

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)}, \tag{3.14}$$

where

$$f(y) = \int_{-\infty}^{\infty} f(y \mid \theta) f(\theta) d\theta. \tag{3.15}$$

The prior predictive distribution $f(y)$ evaluated at observed data is called marginal likelihood. Note that Equation (3.14) can be rewritten to get

$$f(y) = \frac{f(y \mid \theta) f(\theta)}{f(\theta \mid y)}. \tag{3.16}$$

Thus, Equation (3.16) provides an alternative way to compute $f(y)$ defined in Equation (3.15) within the conjugate Bayes framework. Note that for this computation all constants play an important role.

Remark: An alternative proof of the prior predictive distribution for $Y \mid m \sim \mathrm{N}(m, \sigma^2)$ with $m \sim \mathrm{N}(\mu, \tau^2)$ is by rules of iterated expectation and total variance ([Spiegelhalter et al., 2002, p. 17, 84], [Held and Sabanés Bové, 2020, Section A.3.4]).

$$\mathbb{E}(Y) = \mathbb{E}_m[\mathbb{E}_Y(Y \mid m)] = \mathbb{E}_m(m) = \mu$$

$$\mathrm{Var}(Y) = \mathrm{Var}_m[\mathbb{E}_Y(Y \mid m)] + \mathbb{E}_m[\mathrm{Var}_Y(Y \mid m)] = \mathrm{Var}_m[m] + \mathbb{E}_m[\sigma^2] = \tau^2 + \sigma^2 = \lambda^{-1} + \kappa^{-1}.$$

Therefore,

$$Y \sim \mathrm{N}\left(\mu, \lambda^{-1} + \kappa^{-1}\right)$$

For predictions, we add variances and the uncertainty increases.

3.3 Bayesian asymptotics

Consider a model with n independent identically distributed (*i.i.d.*) random variables $Y_i, i = 1, \ldots, n$ and the resulting likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$, where $\theta \in \Theta$ and Θ is an open set in $\mathbb{R}^p$. Denote the maximum likelihood estimator (MLE) by $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$, the ordinary unit Fisher information of one observation Y_i by

$$I^*(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(Y_i \mid \theta)$$

and the expected unit Fisher information of one observation by

$$J^*(\theta) = -\mathbb{E}_\theta\Big(\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(Y_i \mid \theta)\Big).$$

We denote the counterparts of the full random sample by $J_{1:n}(\theta) = nJ^*(\theta)$ and $I_{1:n}(\theta) = nI^*(\theta)$ [Held and Sabanés Bové, 2020, Section 4.2.3, p.97] and use their informal notation.

Under regularity conditions (see Held and Sabanés Bové [2020] Definition 4.1 on p. 80), if θ_0 is a fixed true parameter, then

$$\hat{\theta}_n \overset{\text{approx}}{\approx} \mathrm{N}(\theta_0, J_{1:n}(\theta_0)^{-1}) = \mathrm{N}(\theta_0, \frac{1}{n}J^*(\theta_0)^{-1}).$$

The proof of Bayes asymptotic combines a Taylor expansion and the computation of the posterior in the normal model discussed in Section 2.3. This asymptotic says that, under regularity conditions, for any smooth prior which is strictly positive in a neighborhood of θ_0

1.
$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \mathrm{N}(\hat{\theta}_n, I_{1:n}(\hat{\theta}_n)^{-1}) = \mathrm{N}(\hat{\theta}_n, \frac{1}{n}I^*(\hat{\theta}_n)^{-1}), \tag{3.17}$$

where $I_{1:n}(\hat{\theta}_n)$ is the observed Fisher information of the whole sample and $I^*(\hat{\theta}_n)$ is the observed unit Fisher information of one observation. This result holds irrespective of the distributional form of the prior. Therefore, the influence of the prior disappears asymptotically and the posterior is concentrated in a $\sqrt{1/n}$ neighborhood of the MLE. See Held and Sabanés Bové [2020] Section 6.6.2 on page 206 for further details.

Note that there is a difference in what is considered fixed and what is random in classical and Bayesian asymptotics. For example, in Equation (3.17), $\theta \mid y_1, \ldots, y_n$ is a sequence of posterior distributions, which depend on n, and θ is a random variable. On the right hand side of Equation (3.17), the MLE $\hat{\theta}_n$ is a function of observations $y_i, i = 1, \ldots, n$ that are fixed.

In addition to Equation (3.17), there are three other approximations to this asymptotic result.

2. The observed Fisher information can be replaced by the expected Fisher information.
$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \mathrm{N}(\hat{\theta}_n, J_{1:n}(\hat{\theta}_n)^{-1}), \tag{3.18}$$

3. With posterior mode $\mathrm{Mod}(\theta \mid y_1, \ldots, y_n)$ and the negative curvature $C_{1:n}^{-1}$:
$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \mathrm{N}(\mathrm{Mod}(\theta \mid y_1, \ldots, y_n), C_{1:n}^{-1}), \tag{3.19}$$

4. With posterior expectation $\mathrm{E}(\theta \mid y_1, \ldots, y_n)$ and posterior covariance $\mathrm{Cov}(\theta \mid y_1, \ldots, y_n)$:
$$\theta \mid y_1, \ldots, y_n \overset{\text{approx}}{\approx} \mathrm{N}(\mathrm{E}(\theta \mid y_1, \ldots, y_n), \mathrm{Cov}(\theta \mid y_1, \ldots, y_n)), \tag{3.20}$$

Formulas provided in Example 6.29 of Held and Sabanés Bové [2020] on page 208 can be applied to vision correction data. Figure 3.3 shows that asymptotic approximations 1–4 in Equations (3.17)–(3.20) work well for the Beta(16.5, 6.5) posterior obtained for 16 participants out of $n = 22$ who need vision correction combined with the Beta(0.5, 0.5) prior.

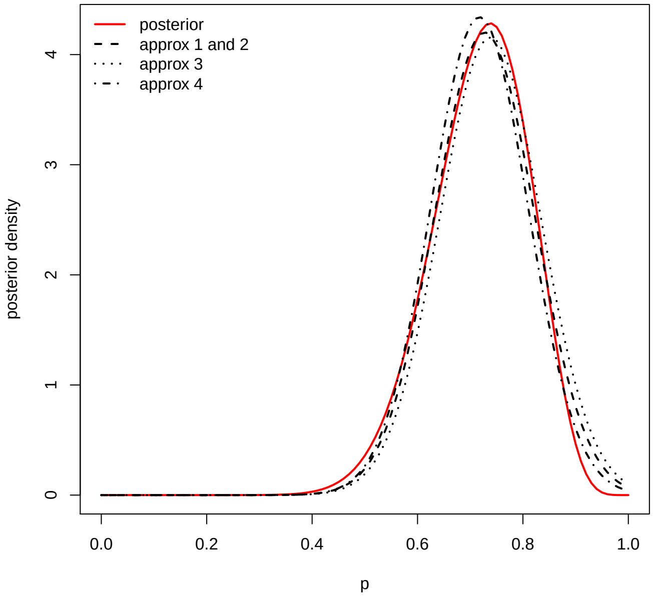


Figure 3.3: Asymptotic approximations of the posterior Beta(16.5, 6.5) obtained for vision correction data $n\bar{y} = 16$ and $n = 22$ combined with the Beta(0.5, 0.5) prior.

3.4 Monte Carlo simulations

The basis of the Monte Carlo (MC) method is the ability to generate independent identically distributed ($i.i.d.$) realizations from a uniform distribution in the unit interval $[0, 1]$. Inverse transform sampling takes this sample and uses it to generate $i.i.d.$ realizations of another random variable, provided that we have an inverse of the cumulative distribution function of this random variable. This approach works, because $P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x)$.

Example Exponential distribution: $F(x) = 1 - \exp(-\lambda x)$ for $x \geq 0$. We can solve $F(x) = 1 - \exp(-\lambda x) = u$ for x and obtain $x = F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. Thus, if we have an $i.i.d.$ sample from a uniform distribution (left panel of Figure 3.4), we can provide a sample from the exponential distribution (right panel of Figure 3.4).

The individual project in Worksheet 3 shows an interesting and in practice highly relevant application of the MC simulation. MC simulation is applied to posteriors in two different groups. Comparison of both samples directly demonstrates differences between both groups. The comparison can be expressed by differences, risk ratios, odds ratios or other measures. Based on these measures the posterior probability of superiority (PPS) can be computed. Remember to compute Monte Carlo standard errors when you use MC techniques (see Held and Sabanés Bové [2020] Section 8.3.1 and 8.3.2).

Monte Carlo sampling can also be used for computation of the area content of an object (for example of a unit circle or any other shape). This demonstrates that MC sampling can be used for integration. Methods such as importance sampling or rejection sampling can be used to explore a distribution. See Section 8.3 of Held and Sabanés Bové [2020] for more

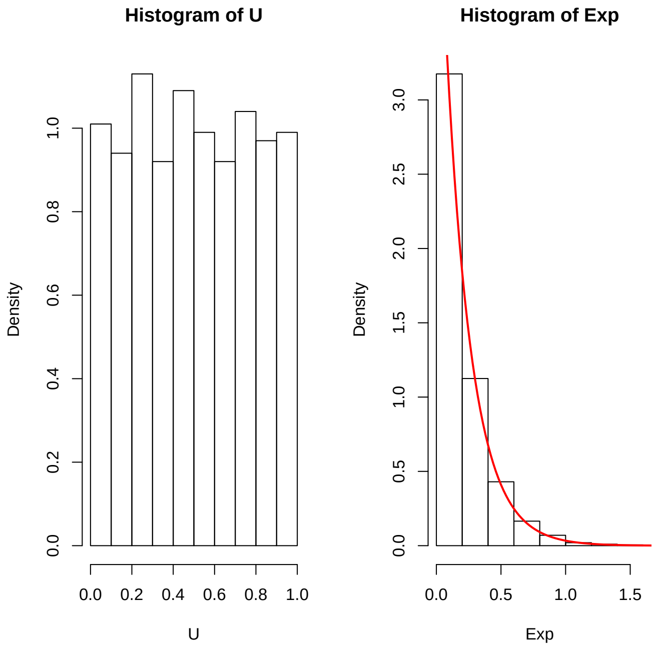


Figure 3.4: Monte Carlo simulation from a uniform distribution on the unit interval $[0, 1]$ (left panel) transformed to a random sample of the Exponential distribution with parameter $\lambda = 5$ by inverse transform sampling (right panel).

details on Monte Carlo methods.

3.5 Worksheet 3

Probability calculus	**Distributions**	**Change of variables formula**
Priors	**MC sampling**	**Asymptotics**

Bayes	**Classical**
Posterior $\propto$ **Likelihood** $\times$ **Prior**	Likelihood

Conjugate Bayes	MCMC sampling	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 3.1: Foundations of Bayesian Methodology: content of the third lecture and material relevant for Worksheet 3.

Secukinumab		**Placebo**
Sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		Data
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 3.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 4

Lecture 4: Markov chain Monte Carlo method

In Chapter 3, we showed that independent Monte Carlo simulations have manifold applications. In particular, they can approximate the area of a complicated object by dropping randomly points on it and computing

$$\widehat{Area} = \frac{\text{Number of points in the object}}{\text{Number of all points}}.$$

This lecture demonstrates that also dependent samples provided by Markov chain Monte Carlo (MCMC) can be useful in this respect. In particular, it overviews the theory behind the Gibbs sampler and the Metropolis-Hastings sampler, two famous techniques for Markov chain Monte Carlo (MCMC) sampling. The `R` code and applications are postponed to Worksheet 4.

Suggested reading: [Held and Sabanés Bové, 2014] Section 8.4, [Ntzoufras, 2009] Chapter 2, [Robert and Casella, 2010] Chapter 7.

4.1 MCMC Sampling in R

Stages of Bayesian analysis (an iterative process):

- Stage 1: model building $(\overbrace{\text{likelihood, parameters,}}^{\text{classical}}\overbrace{\text{priors}}^{\text{Bayesian}})$
- Stage 2: calculation of the posterior distribution (or target distribution)
  - Analytical computation (Conjugate Bayes)
  - Bayesian numerical approximation (INLA, bayesmeta)
  - MCMC sampling: we get a sample either from own samplers or from JAGS, Stan, OpenBUGS, and WinBUGS
- Stage 3: **CODA convergence diagnostics are required for MCMC sampling.** If CODA indicates any problems with MCMC samples, go to Stage 1.

- Stage 4: Analysis of the posterior distribution. Compute descriptive statistics (mean, sd, quantiles, CrI, tail probabilities) of marginal posterior distributions.
- Stage 5: Description of the results for the client.

4.2 Markov chain

Construction of a Markov chain (an iterative procedure, where the value generated in one step depends on the value in the previous step) that eventually "converges" to the target distribution (stationary or equilibrium) $f(\theta \mid \mathbf{y})$. Although MCMC samples are no more independent, we can still approximate areas by MCMC samples.

4.3 The algorithm for MCMC sampling

Markov chain is a stochastic process $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$ such that $f(\theta^{(t+1)} \mid \theta^{(t)}, \ldots, \theta^{(1)}) = f(\theta^{(t+1)} \mid \theta^{(t)})$.

Markov chain must be

- irreducible
- aperiodic
- positive-recurrent

These properties will be discussed together with JAGS and convergence diagnostics (CODA) in Chapter 5 .

If $t \to \infty$ then $\theta^{(t)}$ converges to an equilibrium, which is also called the stationary distribution. This stationary distribution is independent of the initial value $\theta^{(0)}$. Stationarity is needed to guarantee that the sample is identically distributed.

The desirable properties of the algorithm for computation are:

- $f(\theta^{(t+1)} \mid \theta^{(t)})$ is easy to generate
- equilibrium of the Markov chain = $f(\theta \mid \mathbf{y})$ the target posterior distribution. Thus, $\theta^{(t)} \overset{\text{identically distributed}}{\sim} f(\theta \mid \mathbf{y})$, but $\theta^{(t)}$ are no more independent.

4.4 General steps of MCMC algorithm

1. Select an initial value $\theta^{(0)}$.
2. Generate T values until the equilibrium is reached.
3. Monitor convergence diagnostics (if CODA fails, generate more observations, that is, increase T).
4. Cutoff the first B observations (in BUGS it is called burn-in and in Stan warming up).

5. Consider $\theta^{(B+1)}, \theta^{(B+2)}, \ldots, \theta^{(T)}$ as the sample for the posterior analysis (possibly after some tuning).

6. Plot the posterior distribution (usually focus on univariate marginal distributions).

7. Obtain summaries of the posterior distribution (classical: sample mean, median, quantiles, MC-error), effective sample size (ESS) of a MCMC simulation.

Note that an iteration is a cycle of the algorithm that generates a full set of parameter values from the posterior distribution.

4.5 Gibbs sampler

The Gibbs sampler proposed by Geman and Geman [1984] is a special case of the single-component Metropolis-Hastings algorithm, which uses as proposal density $q(\theta' \mid \theta^{(t)})$ the full conditional posterior distribution $f(\theta_j \mid \boldsymbol{\theta}_{-j})$, where $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^T$. In each step of the Gibbs sampler, a candidate value θ_j' of the jth component of the vector of p parameters is proposed by $f(\theta_j' \mid \boldsymbol{\theta}_{-j}, \mathbf{y})$.

Algorithm

1. Set initial values $\theta^{(0)}$.

2. For $t = 1, \ldots, T$ repeat the following steps

   generate $\theta^{(t)}$ (new parameter values) given $\theta^{(t-1)}$ by

   $\theta_1^{(t)}$ from $f(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{y})$

   $\theta_2^{(t)}$ from $f(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{y})$

   $\vdots$

   $\theta_j^{(t)}$ from $f(\theta_j \mid \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{y})$

   $\vdots$

   $\theta_p^{(t)}$ from $f(\theta_p \mid \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{p-1}^{(t)}, \mathbf{y})$

   which are univariate distributions and all other variables except θ_j are held constant at their given values.

3. Save the $\theta^{(t)}$ for the next iterations.

Advantages

- In each step, random values can be generated from univariate distributions for which a wide variety of computational tools exist.

- Frequently, the full conditional posterior distribution has a known form and random values can be easily simulated using standard functions in `R`.

- Gibbs sampler is always moving to a new value, because it accepts the new generated value with acceptance probability 1.

- Gibbs sampler does not require any tuning of proposal distribution.

Drawback Gibbs sampler can be ineffective when the parameter space is complicated or the parameters are highly correlated. It moves slowly.

4.6 Application of the Gibbs sampler

See Exercise 2 of Worksheet 4 (normal example) and the `R` code in the file `04GibbsSampler.R`.

Setting

1. Model: sampling distribution $y_1, \ldots, y_n \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$

2. Data: `set.seed(44566)`, $n = 30$ from $\mathrm{N}(\mu = 4, \sigma^2 = 16)$

3. Priors: independent mean and precision
   $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, $\mu_0 = -3$, $\sigma_0^2 = 4$
   $\frac{1}{\sigma^2} \sim \mathrm{G}(a_0, b_0)$, $a_0 = 1.6$, $b_0 = 0.4$
   Mean $= \frac{a_0}{b_0} = \frac{1.6}{0.4} = 4$ and Var $= \frac{a_0}{b_0^2} = \frac{1.6}{0.4^2} = 10$.

Likelihood

$$\begin{aligned} f(y_1, \ldots, y_n \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \end{aligned} \tag{4.1}$$

Posterior
$f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto f(y_1, \ldots, y_n \mid \mu, \sigma^2) f(\mu, \sigma^2)$
Prior
$f(\mu, \sigma^2) = f(\mu) f(\sigma^2)$ informative and independent
$-\infty < \mu_0 < \infty$, $\sigma_0^2, a_0, b_0 > 0$

Prerequisite $X = \frac{1}{\sigma^2} \sim \mathrm{G}(a_0, b_0)$ [Held and Sabanés Bové, 2014, change of variables formula on page 336].

$f(x) = \frac{b_0^{a_0}}{\Gamma(a_0)} x^{a_0 - 1} \exp(-b_0 x)$
$y = \frac{1}{x} = g(x)$, $x = \frac{1}{y} = g^{-1}(y)$, $\frac{dg^{-1}(y)}{dy} = -\frac{1}{y^2}$

$$\begin{aligned} f(y) &= \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\frac{1}{y}\right)^{a_0 - 1} \exp\left\{-\frac{b_0}{y}\right\} \left|-\frac{1}{y^2}\right| \\ &= \frac{b_0^{a_0}}{\Gamma(a_0)} y^{-(a_0+1)} \exp\left\{-\frac{b_0}{y}\right\} \end{aligned} \tag{4.2}$$

Derivation of full conditional posterior distributions needed for Gibbs sampler.

Posterior

$$\begin{aligned} f(\mu, \sigma^2 \mid y_1, \ldots, y_n) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \\ \times (\sigma_0^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right) \\ \times (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2} \right) \end{aligned} \tag{4.3}$$

The full conditional posterior distribution for μ is obtained by treating the posterior in Equation (4.3) as a function of μ

$$\begin{aligned} f(\mu \mid \sigma^2, y_1, \ldots, y_n) &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \times \exp\left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right) \\ &= \exp\left(-\frac{1}{2} \left[\frac{1}{\sigma^2} \left\{ \sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu \sum_{i=1}^{n} y_i \right\} + \frac{1}{\sigma_0^2} \left\{ \mu^2 + \mu_0^2 - 2\mu_0\mu \right\} \right] \right) \\ &\propto \exp\left(-\frac{1}{2} \left[\mu^2 \left\{ \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right\} - 2\mu \left\{ \frac{\sum_{i=1}^{n} y_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right\} \right] \right) \end{aligned} \tag{4.4}$$

Exercise $p(\theta) \propto \exp\left(-\frac{1}{2} \left(a\theta^2 - 2b\theta \right) \right)$, then $\theta \sim \mathrm{N}\left(\frac{b}{a}, \frac{1}{a} \right)$.

Therefore,

$$\mu^{(t)} \mid (\sigma^2)^{(t-1)}, y_1, \ldots, y_n \sim \mathrm{N}\left(\frac{\frac{\sum_{i=1}^{n} y_i}{(\sigma^2)^{(t-1)}} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}} \right)$$

The full conditional distribution for σ^2 is obtained by treating the posterior in Equation (4.3) as a function of σ^2.

$$\begin{aligned} f(\sigma^2 \mid \mu, y_1, \ldots, y_n) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2} \right) \\ &= (\sigma^2)^{-\frac{n}{2} - (a_0+1)} \exp\left(-\frac{1}{\sigma^2} \left[b_0 + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \right), \end{aligned} \tag{4.5}$$

which is the kernel of an inverse Gamma distribution.

$$(\sigma^2)^{(t)} \mid \mu^{(t)}, y_1, \ldots, y_n \sim \mathrm{InvG}\left(\frac{n}{2} + a_0, b_0 + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu^{(t)})^2 \right).$$

Gibbs algorithm $\boldsymbol{\theta} = (\mu, \sigma^2)$, $p = 2$

- Set initial values
- For t=1, . . . , T

  μ - step: calculate mean $= \frac{\frac{\sum_{i=1}^n y_i}{(\sigma^2)^{(t-1)}} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}$, var $= \frac{1}{\frac{n}{(\sigma^2)^{(t-1)}} + \frac{1}{\sigma_0^2}}$

  generate one random value μ from $\mathrm{N}(\text{mean}, \text{var})$, set $\mu^{(t)} = \mu$

  σ^2 - step: calculate shape $= \frac{n}{2} + a_0$, scale $= b_0 + \frac{1}{2}\sum_{i=1}^n (y_i - \mu^{(t)})^2$,

  generate σ^2 from InvG(shape, scale), set $(\sigma^2)^{(t)} = \sigma^2$.

4.7 Metropolis-Hastings algorithm

The goal of the Metropolis-Hastings algorithm [Metropolis (1953), Hastings (1970)] is to explore target function $f(\theta \mid \mathbf{y})$. See Metropolis [1987] for a historical overview and Hill and Spall [2019] for an easily accessible introduction.

Algorithm

- Set initial values $\theta^{(0)}$.
- For $t = 1, \ldots, T$ repeat the following steps.

  1. Set $\theta = \theta^{(t-1)}$
  2. Generate new candidate parameter values θ' from a proposal distribution $q(\theta^{(t-1)} \mid \theta)$
  3. Calculate the acceptance probability

     $$A = \min\left(1, \frac{f(\theta' \mid \mathbf{y}) q(\theta \mid \theta')}{f(\theta \mid \mathbf{y}) q(\theta' \mid \theta)} = \frac{f(\mathbf{y} \mid \theta') f(\theta') q(\theta \mid \theta')}{f(\mathbf{y} \mid \theta) f(\theta) q(\theta' \mid \theta)}\right)$$

     The chain moves towards direction where $\frac{f(\theta' \mid \mathbf{y})}{q(\theta' \mid \theta)}$ is higher than at the present position (has a higher probability).
  4. Decide randomly whether to accept the proposed value and update $\theta^{(t)} = \theta'$ with probability A otherwise set $\theta^{(t)} = \theta$

     In other words throw a coin, $U \sim \mathrm{Unif}[0, 1]$

     accept if $U < A$

     reject if $U \geq A$

The MH algorithm converges to its equilibrium distribution regardless of whatever proposal distribution q is selected. In practice, the choice of proposal is important. MH can be inefficient if independent proposals are used. To increase the efficiency, use bivariate (multivariate) proposals. Tune also the spread of the proposal to get an acceptance rate of ca. 25%.

Special cases of MH

- random-walk Metropolis
- single component MH
- Gibbs sampler

4.8 Application of the Metropolis-Hastings sampler

See Exercise 3 of Worksheet 4 (mice example with logistic regression) and `R` code in the file `04MHSampler.R`.

1. Data
   - Suppose we have N binomial observations from $\frac{y_i}{n_i}$, $i = 1, \dots, N$
   - Expectation $\mathbb{E}(y_i) = n_i p_i$, where p_i is the corresponding response probability ($\hat{p}_i$ estimated relative frequency of deaths)
2. Logistic model

   Transformation to make linear: $\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$

   $p_i = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$
3. Likelihood

$$
\begin{aligned}
f((\mathbf{y}_i, \mathbf{n}_i, \mathbf{x}_i) \mid \alpha, \beta) &= \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \\
&= \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{n_i - y_i}
\end{aligned}
\qquad (4.6)
$$

4. Priors independent

   $f(\alpha) = \text{N}(0, \sigma^2),\ f(\beta) = \text{N}(0, \sigma^2),\ \sigma^2 = 10^4.$
5. Posterior distribution

$$
f(\alpha, \beta \mid (\mathbf{y}_i, \mathbf{n}_i, \mathbf{x}_i)) \propto \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{y_i} \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^{n_i - y_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\beta^2}{2\sigma^2}}
$$

6. Random walk univariate proposal $\alpha' \sim \text{N}(\alpha, \sigma_\alpha^2)$, $q(\alpha \mid \alpha') = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha - \alpha')^2\right)$,

   $q(\alpha' \mid \alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha - \alpha')^2\right).$

Log-acceptance

$$\begin{aligned}\ln(A^{\alpha}) &= \ln\left(\frac{f(\alpha', \beta \mid \mathbf{y}, \mathbf{n}, \mathbf{x}) q(\alpha \mid \alpha')}{f(\alpha, \beta \mid \mathbf{y}, \mathbf{n}, \mathbf{x}) q(\alpha' \mid \alpha)}\right) \\ &= \ln(f(\alpha', \beta \mid \mathbf{y}, \mathbf{n}, \mathbf{x})) - \ln(f(\alpha, \beta \mid \mathbf{y}, \mathbf{n}, \mathbf{x}))\end{aligned} \tag{4.7}$$

If $\ln(\mathrm{runif}(1)) \leq \ln A^{\alpha}$ then $\alpha \leftarrow \alpha'$, accept α' with probability $\ln(A^{\alpha})$.

7. Random walk univariate proposal. $\beta' \sim \mathrm{N}(\beta, \sigma_{\beta}^2)$

$$\ln(A^{\beta}) = \ln\left(\frac{f(\alpha, \beta' \mid \mathbf{y}, \mathbf{n}, \mathbf{x}) q(\beta \mid \beta')}{f(\alpha, \beta \mid \mathbf{y}, \mathbf{n}, \mathbf{x}) q(\beta' \mid \beta)}\right) \tag{4.8}$$

If $\log(\mathrm{runif}(1)) \leq \log A^{\beta}$ then $\beta \leftarrow \beta'$.

Remarks:

- The user is responsible for tuning of σ_{α}^2 and σ_{β}^2.

- Random walk bivariate proposal. $\binom{\alpha'}{\beta'} \leftarrow \texttt{rmvnorm}\left(\binom{\alpha}{\beta}, \sigma\right)$

$$\ln(A)^{\alpha,\beta} = \ln\left(\frac{f(\alpha', \beta' \mid \mathbf{y}, \mathbf{n}, \mathbf{x}) q(\alpha, \beta \mid \alpha', \beta')}{f(\alpha, \beta \mid \mathbf{y}, \mathbf{n}, \mathbf{x}) q(\alpha', \beta' \mid \alpha, \beta)}\right) \tag{4.9}$$

If $\log(\mathrm{runif}(A)) \leq \log A^{\alpha,\beta}$ then $\binom{\alpha}{\beta} \leftarrow \binom{\alpha'}{\beta'}$. Motivation adjust the shape of the probing proposal density to the shape of the posterior.

In case of a Gibbs sampler, we get $A = 1$. To see this, consider a single component Metropolis algorithm with

$$A = \min\left(1, \frac{f(\theta_j' \mid \boldsymbol{\theta}_{-j}, \mathbf{y}) q(\theta_j \mid \theta_j', \boldsymbol{\theta}_j)}{f(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathbf{y}) q(\theta_j' \mid \theta_j, \boldsymbol{\theta}_{-j})}\right). \tag{4.10}$$

Take $q(\theta_j \mid \theta_j', \boldsymbol{\theta}_{-j}) = f(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathbf{y})$ the full conditional posterior distribution, and $q(\theta_j' \mid \theta_j, \boldsymbol{\theta}_{-j}) = f(\theta_j' \mid \boldsymbol{\theta}_{-j}, \mathbf{y})$. Then,

$$\frac{f(\theta_j' \mid \theta_{-\mathbf{j}}, \mathbf{y}) f(\theta_j \mid \theta_{-\mathbf{j}}, \mathbf{y})}{f(\theta_j \mid \theta_{-\mathbf{j}}, \mathbf{y}) f(\theta_j' \mid \theta_{-\mathbf{j}}, \mathbf{y})} = 1, \tag{4.11}$$

so that the Gibbs proposal will be always accepted.

4.9 Expert system

Figure 4.1 demonstrates an expert system that is used by OpenBUGS for an automatic choice of MCMC samplers.

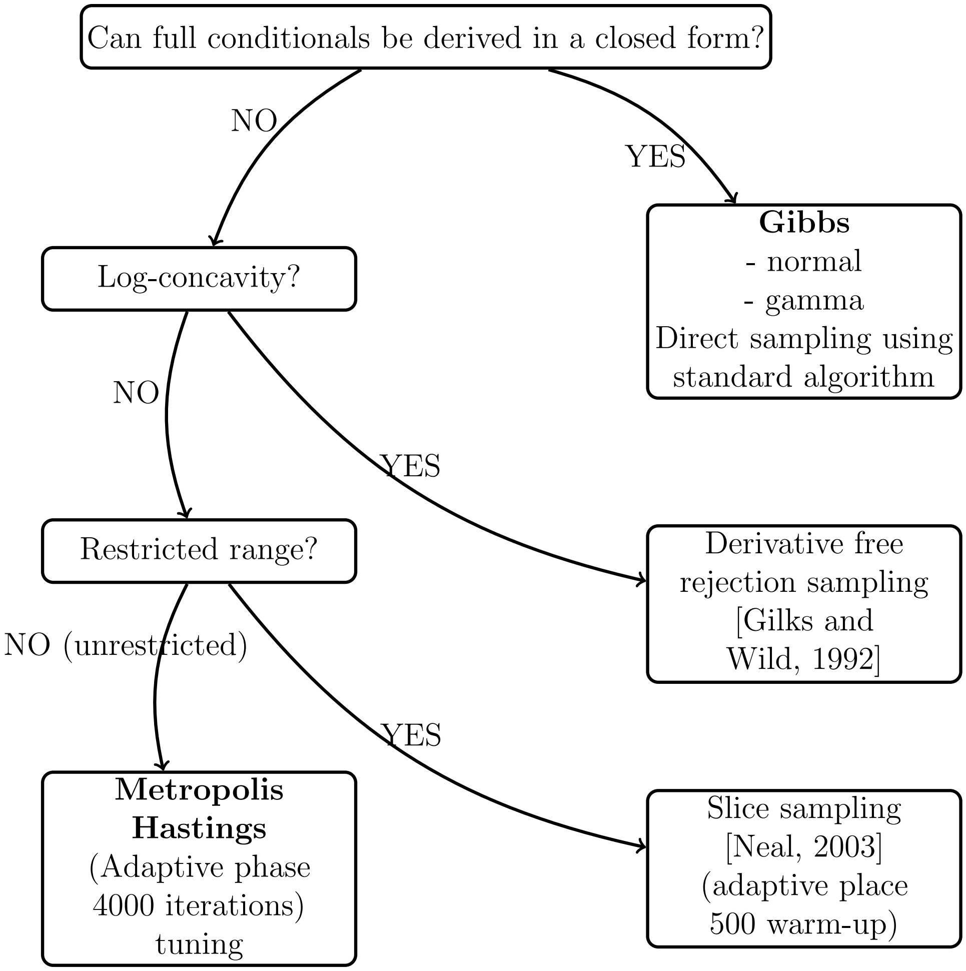

Can full conditionals be derived in a closed form?
NO
YES
Gibbs
- normal
- gamma
Direct sampling using standard algorithm
Log-concavity?
NO
YES
Derivative free rejection sampling [Gilks and Wild, 1992]
Restricted range?
NO (unrestricted)
YES
Metropolis Hastings
(Adaptive phase 4000 iterations) tuning
Slice sampling [Neal, 2003] (adaptive place 500 warm-up)

Figure 4.1: Expert system for the choice of MCMC samplers in OpenBUGS and WinBUGS

4.10 The meaning of priors for the slopes of centered and scaled (standardized) covariates

To stabilize numerical estimation and to express the estimated regression coefficients in common units, covariates in regression models are frequently centered and scaled, i.e. standardized. Centering of a covariate x means that we subtract its mean $\bar{x}$ and use $x - \bar{x}$ for analysis. Scaling of a covariate x means that we divide it by its standard deviation $\hat{\sigma}$ and use $x/\hat{\sigma}$ for analysis. Standardization means that we both center and scale the covariate to get $(x - \bar{x})/\hat{\sigma}$. The mean of both a centered and a standardized covariate is equal to 0. Moreover, the standard deviation of both a scaled and a standardized covariate is equal to 1. Note that if one prefers to scale by $2\hat{\sigma}$, then the standard deviation of the standardized covariate is equal to 0.5 [Gelman, 2008]. Below, we discuss the impact of standardization through centering and scaling on the values of parameter estimates. Moreover, we investigate the meaning of priors in these settings.

Classical paremeter estimates of the linear regression equation expressed in terms of

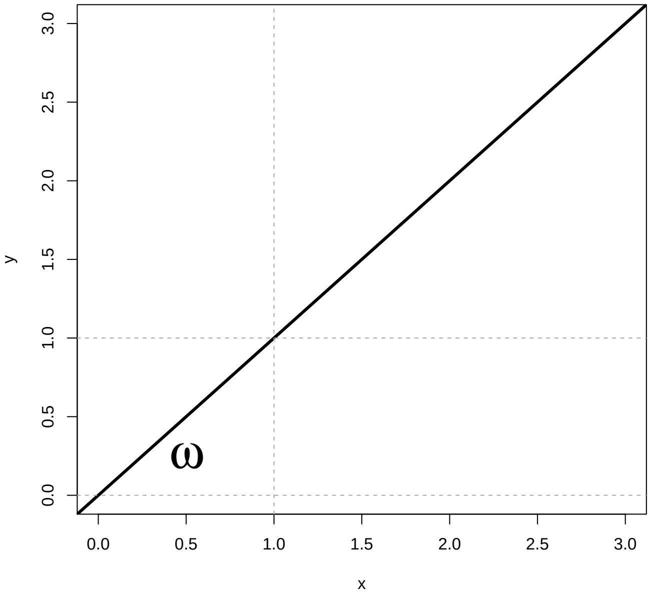


Figure 4.2: Regression line for a non-scaled covariate with the angle $\omega = \pi/4$ (45 degrees) and the slope $\beta = \tan(\pi/4) = 1/1 = 1$.

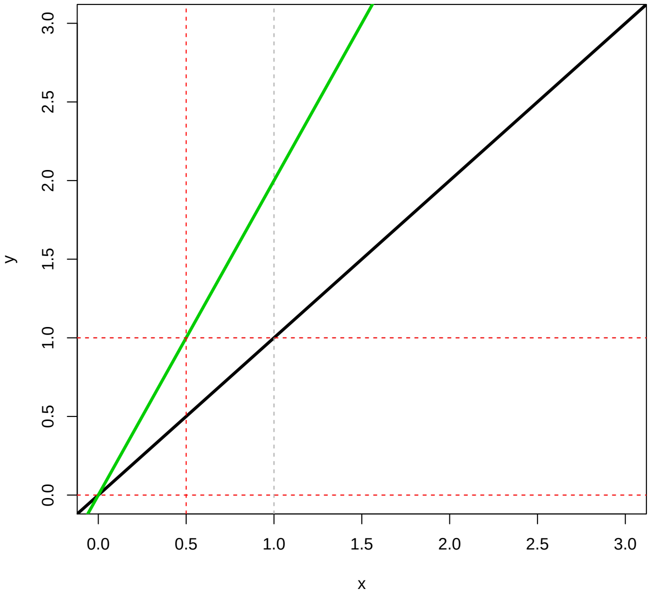


Figure 4.3: Regression lines for non-scaled (black) and scaled with $\hat{\sigma} = 2 > 1$ (green) covariates.

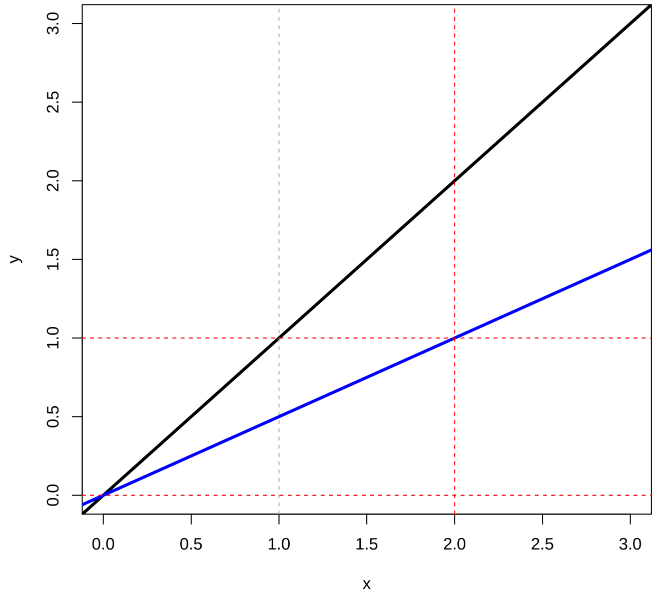


Figure 4.4: Regression lines for non-scaled (black) and scaled with $\hat{\sigma} = 0.5 < 1$ (blue) covariates.

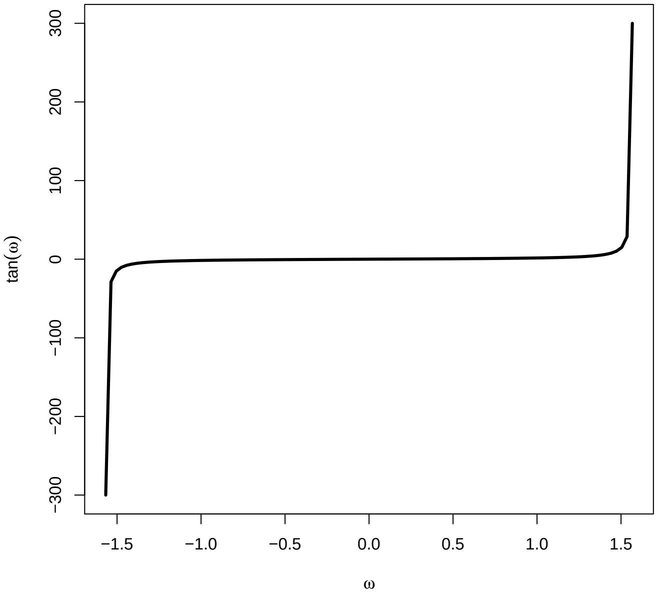


Figure 4.5: Tangens of angles ω ranging between -1.57 and 1.57 radians (89.8 and -89.8 degrees).

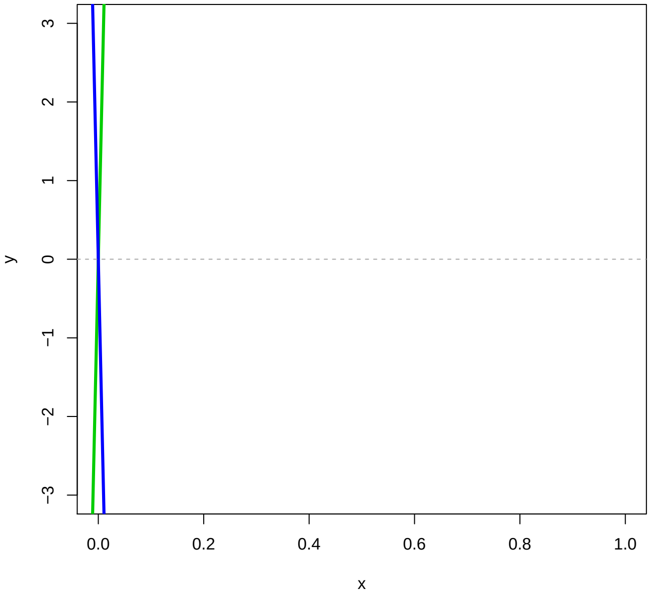


Figure 4.6: Slopes 300 and -300 for angles $\omega = 89.8$ and $\omega = -89.8$ degrees.

original covariates

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$$

are equal to

$$\hat{\alpha} = \bar{y}. - \hat{\beta}_1 \bar{x}_1. - \hat{\beta}_2 \bar{x}_2., \quad \hat{\beta}_1, \quad \text{and} \quad \hat{\beta}_2.$$

In contrast, classical parameters estimates of the linear regression equation expressed in terms of centered covariates

$$y_i = \gamma + \beta_1(x_{1i} - \bar{x}_1.) + \beta_2(x_{2i} - \bar{x}_2.)$$

are equal to

$$\hat{\gamma} = \bar{y}., \quad \hat{\beta}_1, \quad \text{and} \quad \hat{\beta}_2.$$

This demonstrates that centering of covariates does not affect the values of slope estimates.

As we show below, the scaling of covariates x_1 and x_2 by their standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$ impacts the values of slope estimates. Indeed, classical paremeter estimates of the linear regression equation expressed in terms of non-centered but scaled covariates

$$y_i = \alpha + \delta_1 x_{1i}/\hat{\sigma}_1 + \delta_2 x_{2i}/\hat{\sigma}_2$$

are equal to

$$\hat{\alpha} = \bar{y}. - \hat{\delta}_1 \bar{x}_1./\hat{\sigma}_1 - \hat{\delta}_2 \bar{x}_2./\hat{\sigma}_2, \quad \hat{\delta}_1 = \hat{\beta}_1 \hat{\sigma}_1, \quad \text{and} \quad \hat{\delta}_2 = \hat{\beta}_2 \hat{\sigma}_2.$$

Moreover, classical parameters estimates of the linear regression equation expressed in terms of centered and scaled (standardized) covariates

$$y_i = \gamma + \delta_1(x_{1i}/\hat{\sigma}_1 - \bar{x}_{1\cdot}/\hat{\sigma}_1) + \delta_2(x_{2i}/\hat{\sigma}_2 - \bar{x}_{2\cdot}/\hat{\sigma}_2)$$

are equal to

$$\hat{\gamma} = \bar{y}_{\cdot}, \quad \hat{\delta}_1 = \hat{\beta}_1\hat{\sigma}_1, \quad \text{and} \quad \hat{\delta}_2 = \hat{\beta}_2\hat{\sigma}_2.$$

Whereas centering of covariates does not impact the values of slope estimates, scaling greatly affects their values. The estimates of slopes $\hat{\delta}_1$ and $\hat{\delta}_2$ for scaled covariates are either smaller ($\hat{\sigma} < 1$) or larger ($\hat{\sigma} > 1$) than the slope estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained for non-scaled covariates (Figures 4.2–4.4).

To see the impact of scaling on the slope values, we consider the same problem from a different point of view. The slope β in a simple linear regression measures the tangens of the angle ω the line makes with the positive x-axis

$$\beta = \tan(\omega) = \frac{\Delta y}{\Delta x}.$$

If the scale of the covariate x changes, then the slope will be affected by this scaling

$$\delta = \frac{\Delta y}{\Delta x/\hat{\sigma}} = \beta\hat{\sigma}.$$

Again, the estimate of the slope δ for the scaled covariateis either smaller ($\hat{\sigma} < 1$) or larger ($\hat{\sigma} > 1$) than the original coefficient β based on the non-scaled covariate (Figures 4.2–4.4).

Let us focus on the meaning of the prior for the slope in a Bayesian regression model. A prior $\mathrm{N}(0, 100^2)$ assumes that with probability 0.997 the slope can range between $-3 \times 100 = -300$ and $3 \times 100 = 300$. The prior for the slope is actually a prior that we put on the tangens of the angle ω. Slopes 300 and -300 correspond to angles with $\omega = 89.8$ and $\omega = -89.8$ degrees (1.57 and -1.57 radians) (Figures 4.5 and 4.6).

Usually, similar priors are assumed for slopes of Bayesian regression models. Now, we show that the meaning of these priors depends on the meaning of covariates. For example,

$$\beta_1 \sim \mathrm{N}(0, 100^2) \quad \text{and} \quad \beta_2 \sim \mathrm{N}(0, 100^2)$$

priors assumed for slopes based on non-scaled covariates demonstrates that we assume that both slopes have equal properties disregarding the original units of covariates x_1 and x_2. In contrast, priors

$$\delta_1 \sim \mathrm{N}(0, 100^2) \quad \text{and} \quad \delta_2 \sim \mathrm{N}(0, 100^2)$$

assumed for slopes based on scaled covariates correspond to

$$\delta_1 = \beta_1\hat{\sigma}_1 \sim \mathrm{N}(0, 100^2) \quad \text{and} \quad \delta_2 = \beta_2\hat{\sigma}_2 \sim \mathrm{N}(0, 100^2)$$

and to two different priors

$$\beta_1 \sim \mathrm{N}(0, (100/\hat{\sigma}_1)^2) \quad \text{and} \quad \beta_2 \sim \mathrm{N}(0, (100/\hat{\sigma}_2)^2)$$

for non-scaled covariates x_1 and x_2. This demonstrates that the effective meaning of priors always depends on the effective scale of covariates included in the equation of the multiple linear regression.

To standardize or not to standardize, that is the question. There are two different approaches to standardize covariates. First, the function `scale` in `R` can center, scale, and standardize (center and scale) covariates. Second, functions `rescale` and `standardize` in the `arm` package in `R` implement an approach proposed by [Gelman, 2008]. This approach scales continuous covariats by twice standard deviation, so that the standard deviation of the scaled covariate is equal to 0.5. This is an attempt to adjust the meaning of binary and continuous covariates, so that the slope of scaled continuous and binary covariates is defined on an equal footing in terms of a unit change, which is equal to twice the standard deviation of scaled covariates ($2 \times 0.5 = 1$). Moreover, in this setting the same generic default prior can be assumed for all slopes of standardized covariates in a Bayesian regression model for an automatic use [Gelman et al., 2008]. For original covariates, these generic priors for standardized covariates induce covariate-specific priors that highly depend on standard deviations that were used for scaling.

By default, the `scale` function in `R` scales covariates x_1 and x_2 by their corresponding standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$. Note that we can use this function to scale the covariates by a different value. For example, we can modify the call of the `scale` function slightly

```
scale(X, center = FALSE, scale = 2*apply(X, 2, sd, na.rm = TRUE)), 2, summary))
```

to scale covariates x_1 and x_2 stored in a matrix `X` by $2\hat{\sigma}_1$ and $2\hat{\sigma}_2$. Note that the slope γ of a covariate scaled by $2\hat{\sigma}$ is twice as large as the slope δ of a covariate scaled by $\hat{\sigma}$:

$$\gamma = \frac{\Delta y}{\Delta x/(2\hat{\sigma})} = \beta 2\hat{\sigma} = 2\delta, \quad \text{where} \quad \delta = \frac{\Delta y}{\Delta x/\hat{\sigma}} = \beta\hat{\sigma}.$$

For original non-transformed covariates, there are sevaral approaches that use information specific to a particular analysis to elicit covariate-specific priors of slopes [Gelman et al., 2008, Section 1.2]. For example, one can set a prior distribution by eliciting the possible values of outcomes given different combinations of regression imputs. Alternatively, one can elicit a prior distribution by characterizing expecet effects in informativeness ranges.

Although centering of covariates is unnecessary if we are interested in the slope [Gelman, 2008], centering can facilitate interpretation of interactions. Belsley [1991] warns that centering of covariates does not solve the problem of collinearity but rather hides it, so that a potentially harmful impact of collinearity cannot any longer be clearly detected and diagnosed. In any case, we should be aware of the impact of centering and scaling on the values of parameter estimates in a multiple linear regession in both the classical and the Bayesian setting.

4.11 Worksheet 4

Probability calculus	**Distributions**	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ **Likelihood** $\times$ **Prior**		Likelihood
Conjugate Bayes	**MCMC sampling**	Bayesian logistic regression
Predictive distributions	JAGS	Bayesian meta-analysis
Prior elicitation	CODA	Bayesian model selection

Table 4.1: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 4.

Secukinumab		**Placebo**
Bayesian sample size computation		
		Bayesian meta-analysis Prior elicitation
Beta(0.5, 1)		Beta(11, 32)
Data		Data
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 4.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 5

Lecture 5: JAGS and CODA

Recently, Bayesian modeling has become more accessible through general-purpose user-friendly software systems for Bayesian computation. Basically, there are two approaches to arrive at posterior inference: Markov chain Monte Carlo (MCMC) sampling [Gilks et al., 1996] and numerical approximation. There are well known engines for MCMC sampling such as WinBUGS, OpenBUGS, JAGS, Stan or BayesX [Lunn et al., 2009, 2013, Lesaffre and Lawson, 2012, Gelman et al., 2014a]. They represent well the current state-of-the-art of simulation approaches for fitting complex Bayesian models. Unfortunately, Bayesian inference can be very involved and time-consuming. For example, computational MCMC interface may require a week or more to converge. Technical problems related to MCMC algorithms and waiting for the convergence of runs are quite typical [Kuikka et al., 2014].

Just Another Gibbs Sampler (JAGS) is a MCMC sampling engine for Gibbs sampling [Plummer, 2016] which can be conveniently accessed in `R` by `rjags`, `R2jags` or `runjags` interfaces. JAGS can be used for analysis of measurement error models [Muff et al., 2015]. Recently, JAGS received an extension for data smoothing through the well known `mgcv` package [Wood, 2016]. Functionality and output of JAGS is comparable to WinBUGS or OpenBUGS. However, JAGS is more portable between different computer systems.

While MCMC can usually converge to the stationary target distribution as the number of draws approaches infinity, the finite sample behavior of these algorithms can be bad [Vehtari et al., 2021]. Thus, numerical summaries are necessary that can flag potential problems. We must verify empirically that the Markov chain is geometrically ergodic, which is the most important property, if we want a central limit theorem to hold for approximate posterior expectations. Without convergence, the large deviation behavior of estimates cannot be controlled and the obtained MCMC may be useless for practical purposes.

Thus, after each MCMC simulation, a CODA step must be conducted. CODA stands for Convergence Diagnostics and Output Analysis for MCMC samples [Cowles and Carlin, 1996]. It helps when assessing the convergence of MCMC samples to a stationary distribution by providing computational checks. CODA methodology was made available in R by Prof. Martyn Plummer with the `coda` package [Plummer et al., 2006]. Only if MCMC samples estimating a Bayesian model have passed the CODA checks, this model may be released for practical use. See Roy [2020] for a recent overview.

Bayesian models are represented either by a list of model assumptions or graphically by a directed acyclic graph (DAG), which corresponds to these model assumptions. Note that a

DAG rests on expert knowledge about the relations between variables in the model. Different DAGs lead to different models and different posteriors (conclusions). DAGs provide a unified graphical language to lay out and to be clear and explicit about model assumptions. This transparency facilitates communication and discussions.

Recommended reading about JAGS The relevant books for JAGS are [Lunn et al., 2013] and Kruschke [2015]. See also the JAGS manual [Plummer, 2016], Plummer [2003], Wabersich and Vandekerckhove [2014] and optional reading Muff et al. [2015].

`http://jeromyanglim.blogspot.ch/2012/04/getting-started-with-jags-rjags-and.html`

See lectures by Esarey

`http://www.justinesarey.com/teaching`

`https://www.youtube.com/playlist?list=PLAFC5F02F224FA59F`

Recommended reading about CODA Useful literature: Wabersich and Vandekerckhove [2014], CODA on the Framingham example from Muff et al. [2015], Robert and Casella [2004], Cowles and Carlin [1996], Gelman and Rubin [1992], Brooks and Gelman [1998], Plummer et al. [2006], Plummer [2008], Gelman et al. [2014b], Vehtari et al. [2021] Spiegelhalter et al. [2002].

5.1 Introduction to JAGS

BUGS: Bayesian Inference Using Gibbs Sampling
(Since 1980)

- Bayesian Inference
- Graphical modelling
- Simulation-based inference

WinBUGS + OpenBUGS + Examples + Book [Lunn et al., 2009]

JAGS: Just Another Gibbs Sampler
(Since 2003)

- clone of BUGS (Dialect)
- better portable (Computer systems Linux, Windows, MacOSX)
- object-based R inference (rjags, R2jags, runjags)
- connected to CODA
- manual for JAGS and the BUGS book with a manual for JAGS [Lunn et al., 2009]

Notation:

`glm.out <- glm(`$y \sim x$`, family=., data=.)`

$y \sim x$ is the description of a model, queried via extractor functions, `summary(), coef, vcov`

```
library(rjags)
m <- jags.model("code",data,inits,n.chain=2)
list.samplers(m)
# Burn-in
updates(m,n.iter=4000)
```

- `m` is not a fitted model, it is a dynamic object that can be queried to generate from the posterior
- `"code"` name of the file containing a description of the model in BUGS language
- `data` named list of data for observed variables
- `inits` is a list of lists of initial values, one for each chain
- `n.chain` number of chains to run
- `updates` is used for adaptation, burn-in

To generate samples from a given model `m`

`x <- coda.samples(m, variable.names=".", n.iter=1000)`.

- `x` is an object of class mcmc.list `-> coda`
- `coda.samples` is a wrapper function for `"jags.samples"`

Example 1 (all normal)

1. $y_1, \dots, y_n \sim \mathrm{N}(\mu, \sigma^2)$, $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, $\mu_0 = -3$, $\sigma_0^2 = 4$

   precision $\tau = 1/\sigma^2(1/\text{variance}) \sim G(a_0, b_0)$, $a_0 = 1.6$, $b_0 = 0.4$

   Remark: PARAMETRIZATION! Be careful about it!

   R / Stan: $\mathrm{N}(\text{mean}, \text{sd})$

   BUGS / JAGS / INLA: $\mathrm{N}(\mu, \tau^{-1})$, $\tau^{-1} = \sigma^2$, $\tau = \frac{1}{\sigma^2}$

2. Doodle and a graphical model representation (Figure 5.1).

   Remark: For the approximation of WinBUGS doodles in this script, we use ovals for stochastic or logical nodes and boxes for constant nodes. The Appendix A of Ntzoufras [2009] exemplifies the original notation used in WinBUGS doodles.

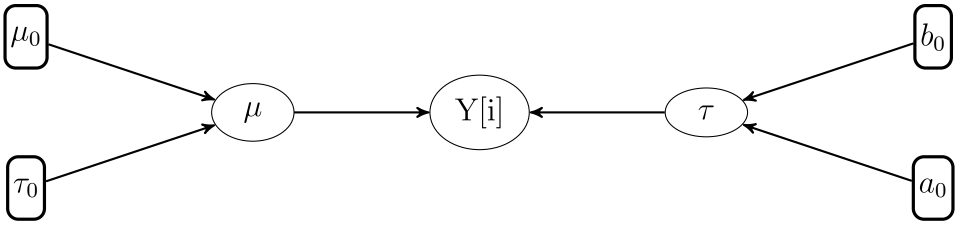


Figure 5.1: Directed Acyclic Graphs (DAG) for the model in Example 1 ($i = 1, \ldots, N$).

3. JAGS CODE

```
model {
# likelihood
for(i in 1:N) {
Y[i] ~ dnorm(mu, tau)
}
# priors
mu ~ dnorm(-3, 0.25) # (mu_0, tau_0)
tau ~ dgamma(1.6, 0.4) # (a_0, b_0)
}
```

- Each variable should appear only once on the left hand side of "∼" relation
- Variable may appear multiple times at the right hand side
- Variables that appear only on the right must be supplied as data

Example 2 (mice logistic regression in JAGS)

1. $y_i \sim \text{Bin}(\text{n[i]}, \text{p[i]})$, $i = 1, \ldots, N$, $\text{logit}(p_i) = a + bx_i$, data are in Collett [2003, p.71].
2. Doodle, DAG in Figure 5.2.
3.

```
model {
# likelihood
for(i in 1:length(Y)) {
Y[i] ~ dbin(p[i], n[i])
logit(p[i]) <- a+b*x[i]
}
# priors
a ~ dnorm(0,1.0E-4)
b ~ dnorm(0,1.0E-4)
}
```

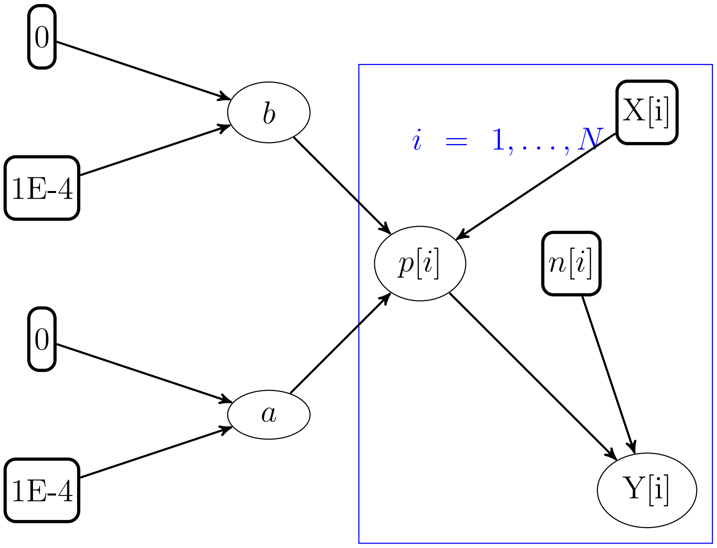


Figure 5.2: Directed Acyclic Graphs (DAG), Mice Logistic Regression example.

Remarks:

- Curly brackets are compulsory for the model.
- "model" describes static relations between variables (declarative according to DAG). It describes the model but not how to draw MCMC samples from it.
- Standard JAGS modules

  basemod - Functions and distributions built into "univariate" samplers

  bugs - Functions and distributions from OpenBUGS Conjugate Samplers

  dic - Deviance statistics

  mix - Distributions for finite mixture models

  glm - Block sampler for generalized (mixed) models (fixed and random effects)

  `logit(p[i]) <- a+b*x[i]` - binary - logit ↔ `p[i] <- ilogit(a+b*x[i])`

  `probit(mu[i]) <- a + b*x[i]` - binary - probit ↔ `mu[i] <- Phi(a + b*x[i])`

  `log(lambda[i]) <- a+b*x[i]` - Poisson - log ↔ `lambda[i] <- exp(a + b*x[i])`

- Extending and Improving JAGS with Modules

  Modules are dynamically loadable extensions to the JAGS library. They provide new

  functions

  distributions

  samplers

  random number generators (RNG)

  monitors

- Reproducible MCMC sampling in JAGS

```
# JAGS: reproducibility of MCMC chains for a model with 3 parameters.
# Provide explicit initial values for alpha, beta1, beta2.
# Use 4 chains with 4 different random number generators provided in
# JAGS and initiate them at different explicit seed values.

inits = list(list(alpha = -0.030, beta1 = -7.474, beta2 = 5.232,
                  .RNG.name = "base::Wichmann-Hill", .RNG.seed = 314159),
             list(alpha = -0.025, beta1 = -6.974, beta2 = 5.732,
                  .RNG.name = "base::Marsaglia-Multicarry", .RNG.seed = 159314),
             list(alpha = -0.035, beta1 = -7.974, beta2 = 4.732,
                  .RNG.name = "base::Super-Duper", .RNG.seed = 413159),
             list(alpha = -0.022, beta1 = -6.674, beta2 = 6.032,
                  .RNG.name = "base::Mersenne-Twister", .RNG.seed = 143915))
```

5.2 BUGS code for a linear regression model

$y_i = \alpha + \beta x_i + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, $y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)$.

Remark: In R: `lm(y ~ x)`.

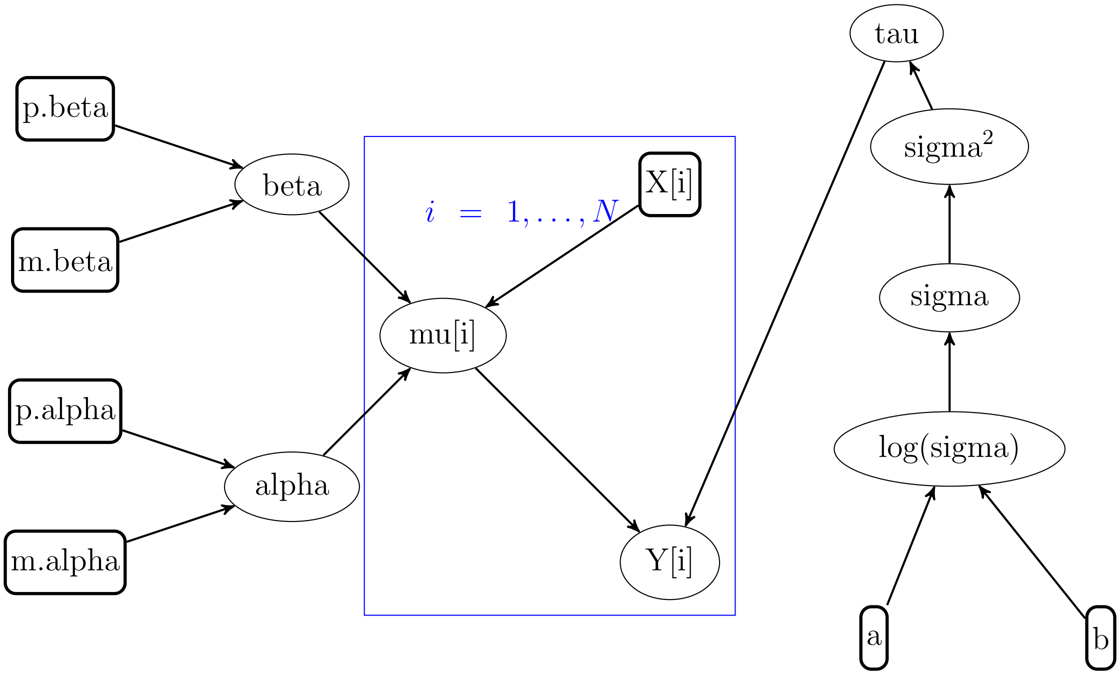


Figure 5.3: Directed Acyclic Graphs (DAG) for a linear regression model.

JAGS CODE

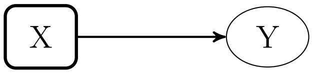
```
model {
# likelihood
for(i in 1:length(Y)) {
Y[i] ~ dnorm(mu[i], tau)
mu[i] <- alpha + beta * x[i]
}
# priors
alpha ~ dnorm(m.alpha, p.alpha)
beta ~ dnorm(m.beta, p.beta)
log.sigma ~ dunif(a,b)
sigma <- exp(log.sigma)
sigma.sq <- pow(sigma, 2)
tau <- 1/sigma.sq
}
```

Remark:

- Need to specify the parameters and data
- parameters need to have explicit prior distributions
- not vectorized "for loops"
- model includes parameter transformations

5.3 What are graphical models

In a graphical model random variables are represented as *nodes*, and the relations between them by *edges*



In this simple model Y is the outcome variable and X is a vector of predictor variables or covariates. Graphical models become interesting when we have multiple variables and the relations between them become more complex.

5.3.1 Why are graphical models useful?

A way of thinking about and representing complex multivariate probability model based on conditional independence relationships.

Use graphs to

- break down complex models into simple components
- communicate the essential structure of the problem

- structural rather than functional relationships
- provide a basis for computations

There are two kinds of graphs

- Directed Acyclic Graph (DAG)
- Conditional Independence Graph (CIG)

Triangulated graphs admit closed form maximum likelihood estimates and considerable computational simplifications. A graph is decomposable, that is, it is triangulated (chordal), if and only if it has no cycle of length ≥ 4 without a chord. DAG is moralized and triangulated to form a chordal graph (cliques exist).

5.3.2 Directed Acyclic Graph (DAG)

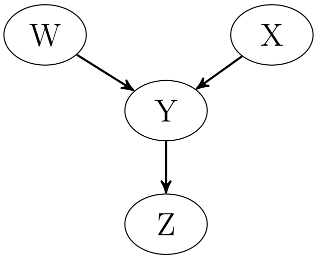


Figure 5.4: A Directed Acyclic Graph (DAG)

Edges in a DAG represent "directed" associations. In Figure 5.4 W and X are called *parents* and Y and Z are called *children*. A DAG describes a decomposition of the joint distribution $p(W, X, Y, Z) = p(W)p(X)p(Y \mid W, X)p(Z \mid Y)$.

5.3.3 Conditional Independence Graphs (CIG)

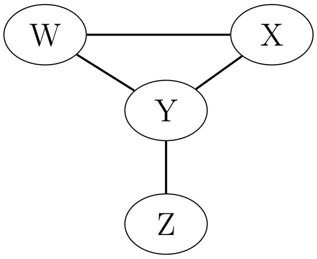
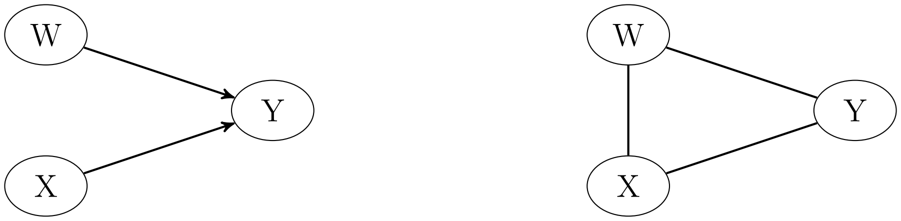


Figure 5.5: A Conditional Independence Graph (CIG)

Edges in a CIG represent symmetric associations. $Z \perp\!\!\!\perp (W, X) \mid Y$ means that Z is conditionally independent of (W, X) given Y.

5.3.4 Moralizing: “Marrying” parent nodes

Conditional independence structure can be deduced from a DAG by moralizing “marrying” parent nodes as in Figure 5.6 (a) and (b).



(a) DAG

(b) CIG

Figure 5.6: Marrying parent nodes to get CIG from a DAG

A CIG defines a multivariate distribution belonging to a very general family of Gibbs distributions. A CIG can be divided into a series of *cliques* (nodes which are all neighbors of each other). For example, there are two cliques in Figure 5.7: (W, X, Y) and (Y, Z).

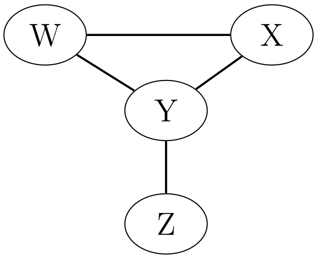


Figure 5.7: There are two cliques in a CIG: (W, X, Y) and (Y, Z)

5.3.5 Gibbs distributions

Joint distribution is

$$p(\omega) = \frac{1}{Z} \exp\left(-\frac{U(\omega)}{T}\right),$$

where $U(\omega) = \sum_{\text{cliques C}} V_C(\omega)$ and Z, T are constants (statistical physics).

- $U \leftarrow$ energy function
- $V_C \leftarrow$ potential that depends only on variables in clique C
- $Z \leftarrow$ is the partition function (normalizing constant)
- $T \leftarrow$ is the temperature.

5.3.6 Gibbs sampling on a graph

- Visit each node in turn sampling from its full conditional distribution given other nodes.
- Full conditionals are needed only up to a multiplicative constant.
- Only neighbors in a CIG make a contribution.
- If CIG is derived from a DAG, then only children, parents, and co-parents contribute.
- Even in a large graph, conditional distributions can be rapidly calculated.

5.3.7 Gibbs sampling from the posterior

If some nodes in the graph are observed then we do not change them but keep them at their observed values.

The stationary distribution of the Markov chain is then the posterior distribution of the unobserved nodes given the observed nodes.

5.3.8 Why does MCMC sampling work?

Stochastic process $(\theta^*_{\mathbf{t}}, \mathbf{t} \in \mathrm{T})$, $\mathrm{T} = 0, 1, 2, \ldots$, discrete time. $P[\theta^*_{\mathbf{t+1}} \in \mathrm{A} \mid \theta^*_{\mathbf{0}}, \ldots, \theta^*_{\mathbf{t-1}}, \theta^*_{\mathbf{t}}] = P[\theta^*_{\mathbf{t+1}} \in \mathrm{A} \mid \theta^*_{\mathbf{t}}]$ takes values in a state space $\mathcal{S}, \mathcal{A} \in \mathcal{S}$. Past $(0, \ldots, t-1)$ and future $(t+1)$ states are independent given the present state t.

Nice Markov chains have the following properties

- Irreducible - you can eventually get from any state to any other state. No matter where it starts, the chain has to be able to reach any other state in a finite number of iterations with positive probability.
- Aperiodic - doesn't "flip - flop" or cycle through states (mixes freely along its possible states)
- Positive recurrent - the expected return time to any state is finite

Aperiodic and Positive recurrent → ergodic,

Irreducible, Aperiodic and Positive recurrent → convergence in distribution possess a unique stationary (equilibrium invariant) distribution.

Convergence settles down eventually to a limiting steady state behavior.

$\forall j \in \mathcal{S}$ stationary distribution $\pi(j) = \sum_{i, i \in \mathcal{S}} \pi(i) P_{ij}(t)$, with transition matrix, kernel $P_{ij} = P[\theta^*_t = j \mid \theta^*_0 = i]$,

$$P(\theta') = \int P(\boldsymbol{\theta}) P_t(\boldsymbol{\theta} \to \theta') d\boldsymbol{\theta}.$$

If $\boldsymbol{\theta}$ is a random sample, from a distribution with density $P(\boldsymbol{\theta})$, so is θ'. Stationary (independent, time homogeneous) θ^*_t and θ^*_{t+k} are sampled from the same distribution.

- Markov chain eventually converges in distribution.

- However, the argument does not tell how quickly.

BUGS Expert system samplers, such as for example Gibbs (full conditional distributions) and Metropolis-Hasting by construction generate such MCMC

1. key property: Reversible transitions

   MH acceptance ratio

   $$P(\theta_i \mid \theta'_{-\mathbf{i}})P_t(\theta_{\mathbf{i}} \to \theta'_{\mathbf{i}}) = P(\theta'_i \mid \theta_{-\mathbf{i}})P_t(\theta'_{\mathbf{i}} \to \theta_{\mathbf{i}})$$

2. key property: MCMC transition kernels can be assembled (multiplied) (for each dimension (i) / clique (c) a suitable sampler can be used)

   $$P_t = P_t^{(C_1)} P_t^{(C_2)} \dots P_t^{(C_k)}$$

   corresponds to one swap of a Conditional Independence Graph (CIG).

   Recall that DAG → CIG → Gibbs distribution:

   $$P(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{1}{Z} \exp\left(-\frac{U(\boldsymbol{\theta} \mid \mathbf{y})}{T}\right),$$

   where Z is a partition function (normalizing constant), T is temperature,

   $$U(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{\text{cliques } c \in C} V_c(\theta_c),$$

   U is energy function, and V_c is the potential depending on variables in a clique c.

   Assume a CIG $= \mathcal{G}(\theta_1, \dots, \theta_d)$ with stochastic nodes $\theta_1, \dots, \theta_d$.

   A Markov blanket on stochastic nodes is a set of nodes $\partial\theta_i$ that separate the node of interest θ_i from others θ in $\mathcal{G} \setminus \{\theta_i, \partial\theta_i\}$.

   $$\theta_i \perp\!\!\!\perp \text{all other nodes in } \mathcal{G} \setminus \{\theta_i, \partial\theta_i\} \mid \partial\theta_i.$$

   Thus, any Markov blanket uses only the local knowledge in a DAG.

5.4 Convergence Diagnostics and Output Analysis (CODA)

Situation: We have an unknown target posterior distribution with parameter $\theta_1, \dots, \theta_d$.

Our goal is to learn about

- the marginal posterior density in one direction
- $\mathbb{E}(\theta_i \mid \mathbf{y})$
- $\sqrt{\mathrm{Var}(\theta_i \mid \mathbf{y})}$

- percentiles

Vats and Knudson [2021] stress that there are two main types of convergence that are relevant for MCMC sampling:

1. the convergence of the t-step Markov transition to the stationary distribution;

2. the convergence of the sample statistics to the truth.

The first type of convergence is known as the burn-in problem. The initial chunk of samples should be discarded, because the Markov chain is not close enough to the stationary target distribution. The second type of convergence is addressed either by assessing the convergence of empirical distribution functions [Boone et al., 2014] or by assessing the convergence of moments of functions of interest. Note that moment-based convergence diagnostics estimate the mean, variance or quantiles and the MCMC is said to have converged when the sample statistics are close enough to the truth. Most of the moment-based convergence diagnostics require the central limit theorem for a Markov chain to be valid. See, for example, Hobert and Geyer [1998], Jones [2004], and Jones and Hobert [2004] for conditions under which a CLT for a Markov chain is possible.

5.4.1 (Markov chain) Monte Carlo Data Set

Assume that after a burn-in of b samples stationarity has been achieved for $i = 1, \ldots, d$ parameters (stochastic nodes).

b burn-in		stochastic nodes	
		θ_i^*	θ_j^*
$t = 1$	$b + 1$	θ_{1i}	θ_{1j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t = M$	$b + M$	θ_{Mi}	θ_{Mj}

Then columns of the MCMC data set form a stationary time series (stationary distribution). It doesn't depend on time $p(\theta^*_{\mathbf{t}} \mid \mathbf{y}) = p(\theta^*_{\mathbf{t+k}} \mid \mathbf{y})$.

- marginal posteriors form a histogram
- sample mean

  $$\bar{\theta}^*_{.i} \mid \mathbf{y} = \frac{1}{M} \sum_{t=1}^{M} \theta^*_{ti}$$

  is an approximation of $\mathbb{E}(\theta^*_i \mid \mathbf{y}) = \int_{-\infty}^{\infty} u f_{\theta^*_i \mid \mathbf{y}}(u) du$
- sample standard deviation

  $$SD(\theta^*_i \mid \mathbf{y}) = \sqrt{\frac{1}{M-1} \sum_{t=1}^{M} (\theta^*_{ti} - \bar{\theta}^*_{.i})^2}$$

- sample percentiles
$$\hat{F}_{\theta_i^*|\mathbf{y}}(q) = \frac{1}{M}\sum_{t=1}^{M} I[\theta_{ti}^* \leq q] = 0.025,$$
solve for q

- these estimates are consistent, they can be made arbitrary close to the truth as $M \to \infty$ (convergence in distribution)

Actually, MCMC samples contain more information:

- correlations (θ_i^*, θ_j^*)
- functions of parameters $\theta_{d+1}^* = \theta_i^* + \theta_j^*$
- $p[\theta_i^* \leq \text{cutoff}]$
- posterior predictive distributions $p(y^* \mid y) = \int p(y^* \mid \theta)p(\theta \mid y)d\theta$

5.4.2 CODA and its scope of application

WARNING: BUGS homepage, "MCMC sampling can be dangerous!"
Steps (JAGS follows it)

- initialization
- burn-in $1, \ldots, b$
- monitoring (sampling) $b+1, \ldots, b+M$ (assumption stationary)

We will try to answer the following questions

Q1 What should I use for starting values?

Q2 How long should burn-in period be? (How do you know when the Markov chain reaches equilibrium (stationary distribution)?)

Q3 How long do you need to monitor the chain to get results of sufficient MC accuracy? (< 0.001)

Questions Q1, Q2 are important and related to each other.

- two (or four) different starting values
- Idea: Start/initialize the chain somewhere near a measure of center of the relevant posterior distribution (mean, mode) (use maximum likelihood justified by weakly informative priors).

  Remark: This can be problematic if the posterior is multimodal. For example, in a bimodal normal mixture $Y = \alpha X_1 + (1-\alpha)X_2$.

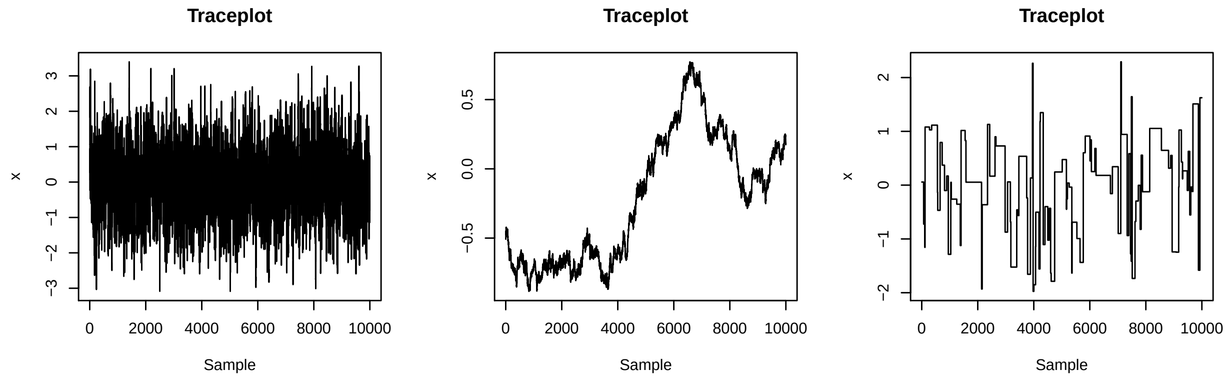


Figure 5.8: Examples of good mixing(left), bad mixing (middle, right). Different types of mixing obtained for one chain for different standard deviation values of the proposal. Left panel: good mixing obtained for a suitable standard deviation. Middle panel: bad mixing obtained for a too small standard deviation. Right panel: bad mixing obtained for a too large standard deviation.

5.4.3 Good and bad mixing

Different mixing types

Solutions to poor mixing/autocorrelation

- Adaptation - better tuning of samplers is necessary (MH a proposal tuning)
- Sampling method choice
- Thinning: Keep only every nth iteration and throw the rest away.

Example: Thinning: "solving bad mixing" by thinning

Choose for example every 20th observation. Thinning a chain by factor 20 means that you have to run it 20 times longer. A run which took 1 hour, now takes nearly 1 day.

5.4.4 Traceplots vs rank plots

Vehtari et al. [2021] propose using rank plots for each chain instead of trace plots. Rank plots are histograms of ranked posterior draws (ranked over all chains) plotted separately for each chain. If all chains explore the same posterior (mix well, have converged), the ranks in each chain should be uniform. If some of the chains have a different location or spread, this will result in non-uniform ranks. Thus, good mixing of the chains leads to similar rank plots for all chains.

Rank plots for two chains in Figure 5.9 are provided in Figures 5.10–5.12. Whereas rank plots in Figure 5.10 are almost uniform, rank plots in Figures for 5.11 and 5.12 are not uniform. This demonstrates that rank plots can indicate cases when chains have not converged and do not mix well. See also `mcmc_rank_hist` function provided in the package `bayesplot` [Gabry et al., 2019].

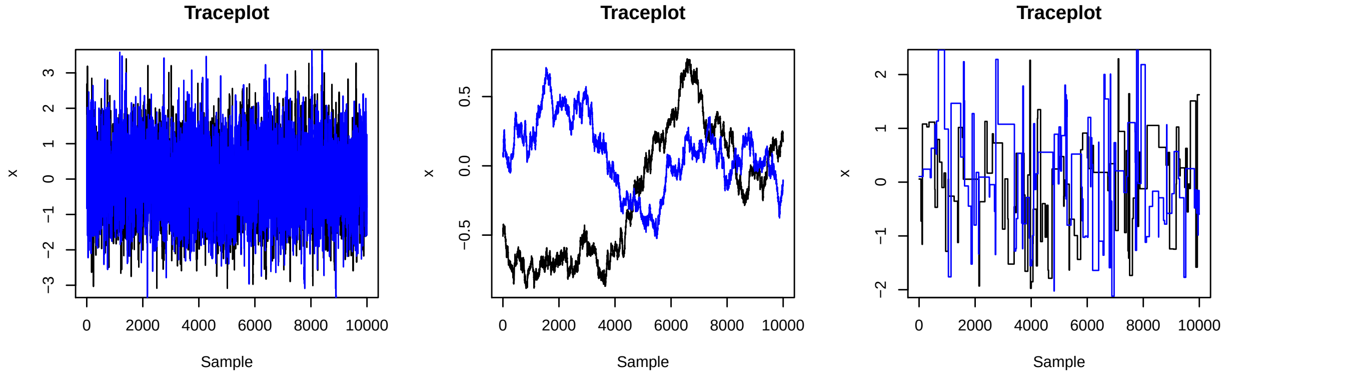


Figure 5.9: Different types of mixing for two chains for different standard deviation values of the proposal. Left panel: good mixing obtained for a suitable standard deviation. Middle panel: bad mixing obtained for a too small standard deviation. Right panel: bad mixing obtained for a too large standard deviation.

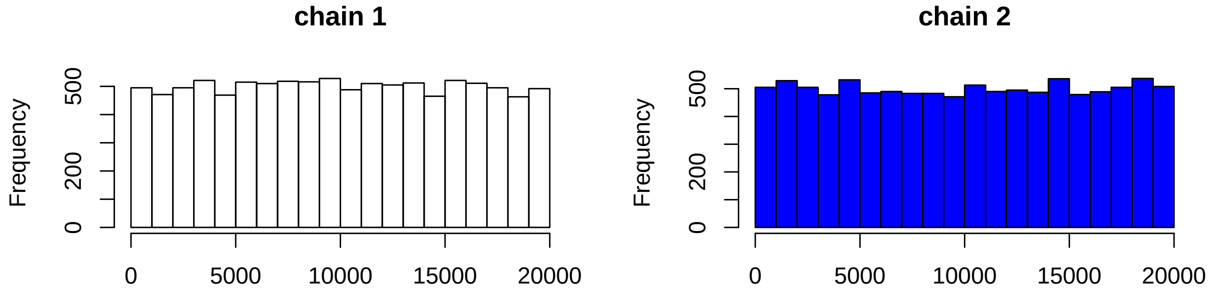


Figure 5.10: Rank plots of two chains for good mixing obtained for a proposal with a suitably chosen standard deviation.

5.4.5 Autocorrelation and Effective Sample Size

-See 05ess.R [Geyer, 1992]

-coda effective sample size

- We know that MCMC converges in distribution to the stationary distribution.
- We do not know how quickly.
- Autocorrelation in the draws affects only efficiency but not sampling validity.
- If autocorrelation is large, then we are learning about $P(\boldsymbol{\theta} \mid \mathbf{y})$ at a slower rate than iid.

Autocorrelation function Autocorrelation at lag k

$$\text{ACF}(k) = \text{corr}(\theta_t^*, \theta_{t+k}^*)$$

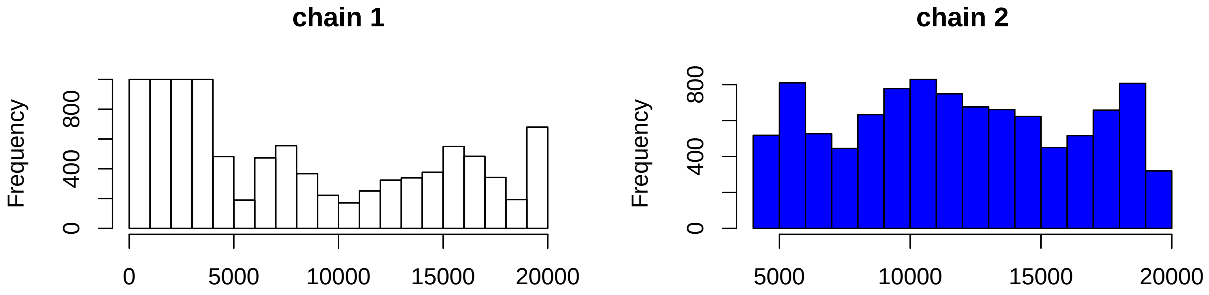


Figure 5.11: Rank plots of two chains for bad mixing obtained for a proposal with a too small standard deviation.

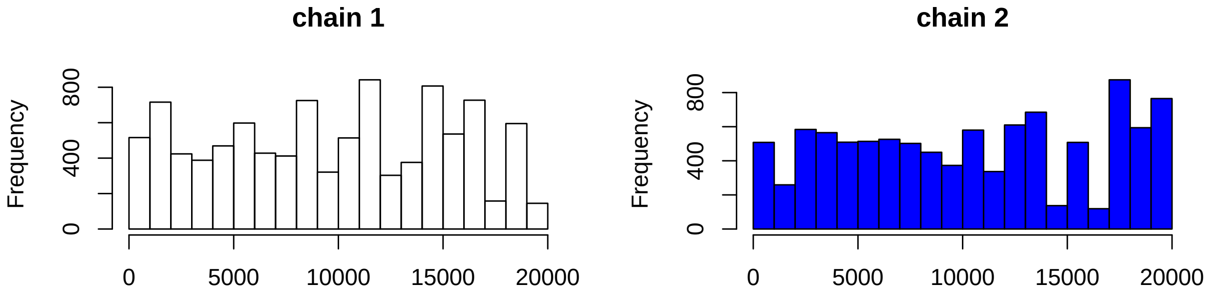


Figure 5.12: Rank plots of two chains for bad mixing obtained for a proposal with a too large standard deviation.

quantifies the amount of mixing of a Markov chain.

ACF estimation methods

- Batch means

- Spectral methods

- Autoregressive process

- Truncation

The Effective Sample Size (Kruschke [2015]) addresses the question "Is the effective sample size large enough to get a stable estimate of uncertainty?" and measures the amount of information in M samples from a Markov chain

$$\mathrm{ESS} = N_{\mathrm{eff}} = \frac{M}{1 + 2\sum_{k\geq 1} \mathrm{ACF}(k)}.$$

Truncation $k = 1, \ldots, w$, $\mathrm{ACF}(w) < 0.1$ (or < 0.05) and $\mathrm{ACF}(w) > \mathrm{ACF}(w+1)$. ESS shows how many independent draws contain the same amount of information as the dependent sample obtained by the MCMC algorithm. The higher ESS the better. ESS must be large enough to get stable inferences for quantities of interest.

Vehtari et al. [2021] recommend computing ESS on the rank-normalized sample. This quantity is well defined even if the chains do not have finite mean or variance. The rank-normalized ESS computes the ESS of a sample from a rank-normalized version of the quantity of interest. Rank-normalization uses the rank transformation followed by the inverse normal transformation. First, the ith value $\theta^{(i)}$ in the MCMC chain is replaced by its rank $r^{(i)}$ within the pooled draws from all chains. Second, these ranks are transformed to normal scores $z^{(i)}$ using the inverse normal transformation $z^{(i)} = \Phi^{-1}\left(\frac{r^{(i)}-3/8}{M-1/4}\right)$. If normalized ranks (normal scores) are used for continuous variables, then normality assumptions are fulfilled. Vehtari et al. [2021] suggest that if the rank-normalized ESS is less than 400, the computed expectations are unlikely to be good approximations to the actual target expectations.

Vehtari et al. [2021] warn that convergence of Markov chains expressed by the bulk-ESS, which assesses how well the location (mean and median) of the distribution is explored and reported, can differ from the tail-ESS for extreme quantiles. Reliable estimation of quantile estimates has a high practical relevance, because decisions are made based on whether or not a specific quantile is above or below a fixed value (for example, if the posterior credible interval contains zero).

Vats and Knudson [2021] in package `stableGR` provide an alternative function for computation of the ESS of a MCMC sample. Their approach is connected to the Gelman-Rubin $\widehat{R}$ estimate.

5.4.6 CODA (Overview)

Convergence diagnostics and output analysis (coda provides an object-based infrastructure).

		Convergence to stationarity	Convergence to ergodic average
number of chains 1	graphical	traceplots	NA
	numerical	• Heidelberger & Welch (run length control based on mean) • Geweke (lack of convergence)	• ESS (Effective Sample Size) • Raftery & Lewis (run length control based on quantiles)
number of chains > 1	graphical	rank plots	NA
	numerical	NA	• BGR (Gelman, Rubin, Brooks), $\widehat{R}$, (Lack of convergence using multiple parallel chains)

Frequentistic Hypothesis testing Q2: How do we know if Markov chain has converged?

WE DON'T KNOW

- We can only look for evidence that it has not converged
- This is the usual situation in hypothesis testing
- But the consequences of the type II error are severe: Invalid INFERENCE

		Diagnostic test	
		non-significant	significant
Truth, convergence to stationarity	yes	$1-\alpha$	α
	no	β	$1-\beta$

Diagnostic tests are frequentist tests and there is a danger of false negative results, type II error β. Therefore, MCMC sampling can be dangerous.

5.4.7 Gelman / Rubin / Brooks BGR (Convergence to ergodic average)

See Robert and Casella [2004, p.497], Cowles and Carlin [1996, Section 2.1], Gelman and Rubin [1992], Brooks and Gelman [1998] and `gelman.diag` / `gelman.plot` in R.

Q2 : Idea: Run multiple chains (hope to detect multimodality) from widely dispersed starting values and preform an Analysis of Variance to see if the between-chain variability (B) is large in relation to the average variability within ($W+B$) the (pooled) chain (if so this would indicate more than one mode). The idea is that if separate chains have not mixed well, the variance of all chains ($W+B$) taken together (pooled) should be higher than the variance of individual chains (W).

Run many (at least 2 short chains and compare late parts (second half) of the chains).

$$\sqrt{\widehat{R}} \approx \sqrt{\frac{W+B}{W}} \rightarrow 1, \quad \text{if} \quad B \rightarrow 0.$$

W is within-chains variability, B is between-chains variability. Chains mix well and converge, if $\widehat{R} \rightarrow 1$.

- Needs widely differing starting points
- Use latter half of each chain
- If chains have not converged, they will be overdispersed
- No testing
- “Shrink factor” $\widehat{R}$ and Upper Confidence Interval
- Uses normal approximation to derive $\widehat{R}$.

$\widehat{R}$ is also called psrf the "potential scale reduction factor".

Gelman et al. [2014a] proposed split-$\widehat{R}$, which in addition to $\widehat{R}$ compares the first half of each chain to the second half, to detect the lack of convergence within each chain. Recently, Vehtari et al. [2021] showed that the traditional $\widehat{R}$ can fail to correctly diagnose convergence failures when the chain has a heavy tail or when the variance varies across the chains and proposed an alternative rank-based diagnostic. They recommend that at least four chains should be run by default and the threshold applied to $\widehat{R}$ should be 1.01($\widehat{R} < 1.01$). They only recommend relying on the $\widehat{R}$ estimate to make decisions about the quality of the chains if the ESS is large enough.

Vats and Knudson [2021] argue that a cutoff of $\widehat{R} \leq 1.1$ is too high to yield reasonable estimates of target quantities. They show that there is a one-to-one correspondence between $\widehat{R}$ value and the effective sample size (ESS) for estimating the mean of the posterior:

$$\widehat{R} \approx \sqrt{1 + \frac{\texttt{n.chain}}{\text{ESS}}},$$

where `n.chain` is the number of chains. They show that $\widehat{R} = 1.1$ corresponds to ESS $=$ $5 \times$ `n.chain` (5 independent samples per chain), which is too low to estimate the mean with any reasonable certainty. In contrast, $\widehat{R} = 1.01$ corresponds to ESS $= 50 \times$ `n.chain` (50 independent samples per chain), which is more appropriate for estimation of the mean. The diagnostic proposed by Vats and Knudson [2021] is available for public use in the `R` package `stableGR`:

5.4.8 Raftery & Lewis (Convergence to ergodic average)

See Robert and Casella [2004, p.500], Cowles and Carlin [1996, Section 2.2] and `raftery.diag` in R (1 chain).

Idea: Run diagnostics based on a criterion of accuracy of estimation of the quantile q.

It is a non-parametric approach.

- Two state Markov-Chain theory

  $$Z^{(t)} = I_{\theta^t \leq \theta_q}$$

  discretization of a continuous chain to get a binary control

- Pseudo-transition Matrix

  $$P = \left[\begin{smallmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{smallmatrix}\right]$$

- Needs a pre-run to set up P preliminary chain to estimate α, β

  Q2 (Burn-in) (M) needed

  Q3 Total N sample size N needed

  N_{min} minimal sample needed (independent),

  $I = \frac{M+N}{N_{\text{min}}}$.

  The (in)dependence factor I, indicates to which extend autocorrelation inflates the required sample size.

- $I > 5$ strong autocorrelation. It is a crude estimate of the **thinning** interval.

Quantiles which are closer to median need more N than quantiles further away.

5.4.9 Geweke (Convergence to stationarity)

See Robert and Casella [2004, p.508], Cowles and Carlin [1996, Section 2.3] and `geweke.diag` in R.

Q2 : Detects cases when equilibrium has not been reached.
Idea:

- consider a single long run
- test for equality of means between early and late sections of the chain
- essentially, two independent sample t-tests

Example
Assume
A corresponds to early 10 % (frac1) of the chain
B corresponds to next 50 % (frac2) of the chain.

$$Z = \frac{\bar{\theta}^{*A} - \bar{\theta}^{*B}}{\sqrt{\text{Var}(\theta^{*A}) + \text{Var}(\theta^{*B})}},$$

Does $\mid Z \mid > 2$?
`geweke.plot`: 2.5 % of Z-scores may be below -2 and 2.5 % above 2.
Properties: Univariate
It is unclear how to choose the width of the early and the late window?

5.4.10 Heidelberger & Welch (Convergence to stationarity)

See Robert and Casella [2004, p.509], Cowles and Carlin [1996, Section 2.10] and `Heidel.diag` (1 chain) in R.

1. Stationarity test Q2

   Basically a Kolmogorov-Smirnov Test

   At each iteration 10% are removed of the first half of the chain.

   Information: if passed, iteration and p-value are provided.

   If all iterations have been used and did not pass, then failed.

2. When all passed then a Halfwidth test Q3 is applied on the retained chain.

   95 % CI (mean) with portion of the chain which has passed stationarity test

   Coefficient of variation $= \frac{S}{\mu} <$ eps $= 0.1$ desired accuracy,

   where S is half width of 95 % CI and μ is mean.

If > eps then fails.

If the sample is not sufficient to estimate the mean with sufficient accuracy then increase it by 1.5.

5.5 Worksheet 5

Probability calculus	**Distributions**	Change of variables formula
Priors	MC sampling	Asymptotics

Bayes	**Classical**
Posterior ∝ Likelihood × Prior	Likelihood

Conjugate Bayes	**MCMC sampling**	Bayesian logistic regression
Predictive distributions	**JAGS**	Bayesian meta-analysis
Prior elicitation	**CODA**	Bayesian model selection

Table 5.1: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 5.

Secukinumab		**Placebo**
Bayesian sample size computation		
		Bayesian meta-analysis **Prior elicitation**
Beta(0.5, 1)		**Beta(11, 32)**
Data		Data
	Classical analysis	
Posterior (S)		Posterior (P)
	Posterior probability of superiority	

Table 5.2: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 6

Lecture 6: Bayesian meta-analysis and empirical Bayes

In conjugate Bayes analysis, we considered models of the type $f(\theta \mid y) \propto f(y \mid \theta)f(\theta)$. In hierarchical models, the prior for θ depends on another parameter ψ, called hyperparameter, which is given again a prior distribution. This leads to $f(\theta \mid y) \propto \int f(y \mid \theta)f(\theta \mid \psi)f(\psi)d\psi$. Another expression useful for this computation is $f(\theta \mid y) = \int f(\theta \mid y, \psi)f(\psi \mid y)d\psi \propto \int f(\theta \mid y, \psi)f(y \mid \psi)f(\psi)d\psi$.

In this section, we take a closer look at Bayesian meta-analysis expressed by a normal-normal hierarchical model (NNHM). The Bayesian NNHM is useful for synthesis of evidence from several already published studies. This model is the simplest Bayesian hierarchical model, but presents challenges typical in more complex Bayesian hierarchical models. We show that the Bayesian NNHM encompasses a wide range of assumptions about the similarity of these studies, which is expressed by the between-study standard deviation τ. We also show that two limiting models emerge for $\tau = 0$ (pooling) and $\tau = \infty$ (independent and unrelated studies). Finally, we demonstrate that similar inference for random effects can be obtained from both the full Bayesian meta-analysis expressed by NNHM and by empirical Bayes estimates.

6.1 Bayesian meta-analysis

Let us consider historical data of responders in placebo and treatment groups provided in Table 6.1. These data were used for prior elicitation by Baeten et al. [2013]. Note that

$$y = \log(\mathrm{OR}) = \log(p_{\mathrm{P}}/(1-p_{\mathrm{P}})) - \log(p_{\mathrm{T}}/(1-p_{\mathrm{T}})) = \log(x_{\mathrm{P}}/(n_{\mathrm{P}}-x_{\mathrm{P}})) - \log(x_{\mathrm{T}}/(n_{\mathrm{T}}-x_{\mathrm{T}}))$$

and

$$\sigma = \mathrm{SE}(\log(\mathrm{OR})) = \sqrt{1/x_{\mathrm{P}} + 1/(n_{\mathrm{P}} - x_{\mathrm{P}}) + 1/x_{\mathrm{T}} + 1/(n_{\mathrm{T}} - x_{\mathrm{T}})}$$

[Held and Sabanés Bové, 2020, p. 137–138].

One possible analysis of data in Table 6.1 would be to disregard separate studies and to compute an overall estimate based on 518 responders out of 834 patients in the treatment group and on 127 responders out of 513 patients in the placebo group. This leads to OR =

	treatment		placebo		y	σ
labels	responders	total	responders	total	log(OR)	SE(log(OR))
1	120	208	23	107	-1.605	0.274
2	18	38	12	44	-0.875	0.469
3	107	150	19	51	-1.433	0.341
4	26	45	9	39	-1.518	0.485
5	82	138	39	139	-1.323	0.256
6	16	20	6	20	-2.234	0.742
7	126	201	9	78	-2.556	0.383
8	23	34	10	35	-1.654	0.524

Table 6.1: Historical data of responders in placebo and treatment groups used for prior elicitation by Baeten et al. [2013] with Adalimumab (Studies 1,2), Etanercept (Studies 3-6), and Infliximab (Studies 7,8).

0.201 with 95%CI(OR) (0.157, 0.257) and log(OR) $= -1.606$ with 95%CI(log(OR)) (-1.851, -1.361). However, this approach is justified if the following assumption is true: all studies are independent and identical realizations of the same underlying process and there is no additional heterogeneity between studies at all. For data in Table 6.1 we doubt that this assumption is true. Therefore, we need a method that is capable of synthesizing evidence from several different studies.

A fruitful assumption is the assumption of exchangeability of random variables $Y_1, \ldots, Y_k$ generating observations $y_1, \ldots, y_k$. Random variables $Y_1, \ldots, Y_k$ are exchangeable when their joint distribution does not change when these random variables are permuted [Spiegelhalter et al., 2004, Section 3.4]. This means that random variables $Y_1, \ldots, Y_k$ are similar but not identical. In particular, exchangeable random variables are not independent. By the theorem by de Finetti (1930), exchangeable random variables can be represented by *i.i.d.* variables drawn from a distribution, which depends on an unknown parameter θ, and the unknown parameter θ has itself a prior:

$$f(y_1, \ldots, y_k) = \int \prod_{i=1}^{k} f(y_i \mid \theta) f(\theta) d\theta.$$

We consider the full Bayesian meta-analysis expressed by the Bayesian normal-normal hierarchical model (NNHM) with three levels of hierarchy: the sampling model expressed by the likelihood, the random effects model, and priors.
Likelihood:

$$y_i \sim \mathrm{N}(\theta_i, \sigma_i^2), \tag{6.1}$$

for $i = 1, \ldots, k$.
Random effects:

$$\theta_i \sim \mathrm{N}(\mu, \tau^2) \tag{6.2}$$

Priors: $\mu \sim \mathrm{N}(\nu, \gamma^2)$ with $\nu = 0$, $\gamma = 4$, and $\tau \sim |\mathrm{N}(0, A^2)| = \mathrm{HN}(A)$ with $A = 0.5$.

We denote the within-study standard deviation of the i-th study by σ_i. This value is assumed to be known (fixed). The between-study standard deviation is denoted by τ. This value governs the heterogeneity of random effects.

For the Bayesian NNHM, the Bayes theorem is expressed by

$$f(\mu, \tau, \boldsymbol{\theta} \mid (y_1, \sigma_1), \ldots, (y_k, \sigma_k)) = f((y_1, \sigma_1), \ldots, (y_k, \sigma_k) \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mu, \tau) f(\mu) f(\tau) C^{-1}, \quad (6.3)$$

where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$, and $C = f((y_1, \sigma_1), \ldots, (y_k, \sigma_k))$ is a normalizing constant obtained by integrating out parameters μ, τ and $\boldsymbol{\theta}$ in the numerator of Equation (6.3). By taking the logarithm of both sides in Equation (6.3), the following formula for the log-posterior is obtained:

$$\log(\text{Posterior}) = \log(\text{Likelihood}) + \log(\text{Random-effects model}) + \log(\text{Prior}) - \log(C). \quad (6.4)$$

Equation (6.4) leads directly to the approximation

$$\log(\text{Posterior}) \approx \log(\text{Likelihood}) + \log(\text{Random-effects model}) + \log(\text{Prior}), \quad (6.5)$$

where

$$\begin{array}{rcl}
\log(\text{Posterior}) & = & \log(f(\mu, \tau, \boldsymbol{\theta} \mid (y_1, \sigma_1), \ldots, (y_k, \sigma_k))), \\
\log(\text{Likelihood}) & = & \log(f((y_1, \sigma_1), \ldots, (y_k, \sigma_k) \mid \boldsymbol{\theta})) \\
 & = & \displaystyle\sum_{i=1}^{k} \left\{-0.5 \log(2\pi\sigma_i^2) - (y_i - \theta_i)^2 (2\sigma_i^2)^{-1}\right\}, \\
\log(\text{Random-effects model}) & = & \log(f(\boldsymbol{\theta} \mid \mu, \tau)) \\
 & = & \displaystyle\sum_{i=1}^{k} \left\{-0.5 \log(2\pi\tau^2) - (\theta_i - \mu)^2 (2\tau^2)^{-1}\right\}, \\
\log(\text{Prior}) & = & \log(f(\mu)) + \log(f(\tau)) \\
 & = & -0.5 \log(2\pi\gamma^2) - \mu^2 (2\gamma^2)^{-1} \\
 & + & 0.5 \log(2) - 0.5 \log(\pi A^2) - \tau^2 / 2A^2.
\end{array}$$

We apply NNHM defined in Equations (6.1) and (6.2) to historical data of responders in placebo and treatment groups provided in Table 6.1. Note that NNHM is only an approximation for data based on log(OR) estimates and their standard errors. Note also that a Generalized Evidence Synthesis approach would be better suited for these data [Prevost et al., 2000, Sutton and Abrams, 2001, Verde and Ohmann, 2015, Verde, 2021], because three different treatments (Adalimumab in Studies 1, 2, Etanercept in Studies 3–6, and Infliximab in Studies 7,8) contribute to these historical data.

6.2 Introduction to `bayesmeta`

The Bayesian meta-analysis expressed by the full Bayesian NNHM in Equations (6.1) and (6.2) can be conveniently fitted with `bayesmeta` [Röver, 2020]. This package fits the Bayesian

NNHM numerically [Röver and Friede, 2017]. Thus, `bayesmeta` dispenses with MCMC sampling and posterior convergence diagnostics. See [Röver, 2020] for an accessible description of the functionality of this package and the justification of priors for μ and the between-study standard deviation τ.

```
res1 <- bayesmeta(y = df[ , "y"],
                  sigma = df[ , "sigma"],
                  labels = df[ , "labels"],
                  mu.prior.mean = 0, mu.prior.sd = 4,
                  tau.prior = function(t){dhalfnormal(t, scale = 0.5)},
                  interval.type = "central")
```

A summary of the result provides numbers supporting the forest plot in Figure 6.1. Figure 6.2 shows marginal posterior distributions of all model parameters.

```
summary(res1)

##  'bayesmeta' object.
## data (8 estimates):
##            y     sigma
## 1 -1.6054775 0.2740073
## 2 -0.8754687 0.4691896
## 3 -1.4329256 0.3412963
## 4 -1.5176304 0.4853221
## 5 -1.3229761 0.2563070
## 6 -2.2335922 0.7420210
## 7 -2.5556757 0.3832411
## 8 -1.6538897 0.5238200
##
## tau prior (proper):
## function(t){dhalfnormal(t, scale=tau_scale)}
## <bytecode: 0x000000001ad55d08>
##
## mu prior (proper):
## normal(mean=0, sd=4)
##
## ML and MAP estimates:
##                    tau        mu
## ML joint     0.2094171 -1.592280
## ML marginal  0.2852879 -1.590235
## MAP joint    0.1614761 -1.585174
## MAP marginal 0.2334117 -1.587618
##
## marginal posterior summary:
##                 tau         mu      theta
```

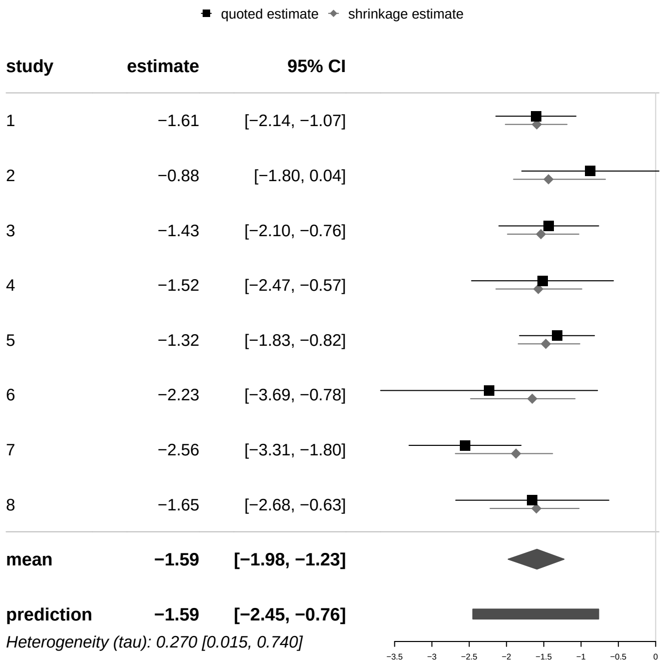


Figure 6.1: Forest plot provided by `bayesmeta` with data from Table 6.1.

```
## mode      0.2334117 -1.5876182 -1.5805059
## median    0.2702386 -1.5919563 -1.5884808
## mean      0.2949284 -1.5946544 -1.5946544
## sd        0.1941244  0.1879906  0.4002409
## 95% lower 0.0153332 -1.9777397 -2.4509871
## 95% upper 0.7397310 -1.2281569 -0.7646060
##
## (quoted intervals are central, equal-tailed credible intervals.)
##
## Bayes factors:
##             tau=0        mu=0
## actual  1.0209152 3.11865e-05
## minimum 0.7030068 1.23805e-06
##
## relative heterogeneity I^2 (posterior median): 0.3343206
```

Note that the full Bayesian meta-analysis expressed by NNHM provides inference on random effects $\theta_1, \dots, \theta_k$ that lies inbetween inference provided by two models: a pooled one (Figure 6.3) and a model which assumes that all studies are independent (Figure 6.4).

The pooled model is based on the assumption that the true location θ of individual studies is equal for all studies and $\theta_1 = \ldots = \theta_k = \theta$. This corresponds to assuming that the between-study variability $\tau = 0$, so that the index i of each individual study may be ignored. This assumption lets the model for random effects collapse to θ, a normal distribution with

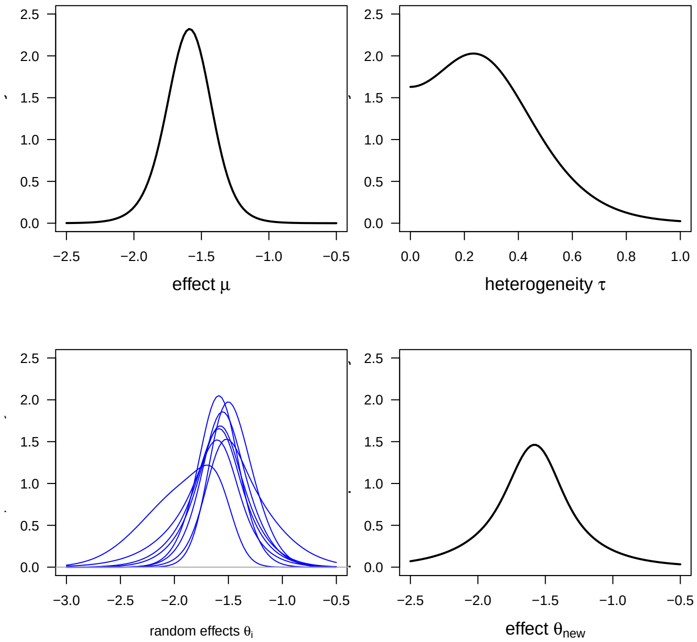


Figure 6.2: Marginal posterior distributions for all parameters in the model provided by `bayesmeta` with data from Table 6.1.

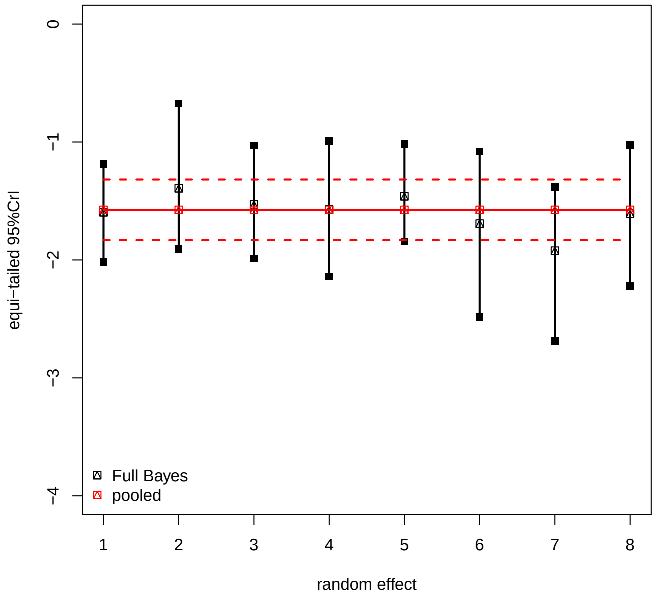


Figure 6.3: Posterior equi-tailed 95% credible intervals of random effect θ_i parameters provided by the Bayesian NNHM (black) and by the pooled model (red) for data from Table 6.1.

location θ and no variability $\tau^2 = 0$. This assumption changes the NNHM model defined in Equations (6.1) and (6.2) to
Likelihood:

$$y_i \sim \mathrm{N}(\theta, \sigma_i^2), \tag{6.6}$$

for $i = 1, \ldots, k$
Prior:

$$\theta \sim \mathrm{N}(\nu, \gamma^2) \tag{6.7}$$

A sequential application of the Bayes theorem leads to the posterior

$$\theta \mid y_1, \ldots, y_k \sim \mathrm{N}\left(\frac{\sum_{i=1}^{k} \frac{y_i}{\sigma_i^2} + \frac{\nu}{\gamma^2}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2} + \frac{1}{\gamma^2}}, \Big(\sum_{i=1}^{k} \frac{1}{\sigma_i^2} + \frac{1}{\gamma^2}\Big)^{-1}\right). \tag{6.8}$$

This result extends the results obtained in Equation (3.9).

Figure 6.3 shows the result for data from Table 6.1, when the between-study heterogeneity $\tau = 0$, $\nu = 0$, and $\gamma = 4$. The pooled overall posterior mean is equal -1.575 and the pooled posterior standard deviation is equal 0.131.

Note that if the prior is very flat ($\gamma \to \infty$) we get:

$$\theta \mid y_1, \ldots, y_k \sim \mathrm{N}\left(\frac{\sum_{i=1}^{k} \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}, \Big(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\Big)^{-1}\right).$$

The location of this distribution is the classical overall mean estimate: an average of individual estimates each weighted by its precision. This fixed-effect model should be rather called a common-effect model [Borenstein et al., 2017, Röver, 2020, Röver et al., 2021].

The assumption of independence of studies leads to estimates of random effects shown in Figure 6.4. The term "independence" in this case means that random effects $\theta_1, \ldots, \theta_k$ are not related to each other. This means that knowledge gained from one study is irrelevant for other studies. Studies under consideration provide different and not comparable pieces of information. In a matter of fact, it is not reasonable to pool the evidence from such disparate studies at all. This situation corresponds to the assumption that the between-study standard deviation $\tau = \infty$ [Spiegelhalter et al., 2004, Section 3.17].

For one i-th study $y_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$. If we assume that the prior for θ_i is uniform, we get that $f(\theta_i \mid y_i) \propto f(y_i \mid \theta_i)$. Thus, $\theta_i \mid y_i \sim \mathrm{N}(y_i, \sigma_i^2)$. Under these assumptions Figure 6.4 shows estimates obtained for data from Table 6.1. These estimates are more heterogeneous than those obtained by the full Bayes approach.

Figure 6.5 compares the estimates discussed above and provides additional estimates (green) obtained by an empirical Bayes argument. Empirical Bayes estimates clearly differ from the estimates obtained by pooled (red) and independent (blue) models but they are close to marginal posterior estimates of random effects provided by the full Bayesian meta-analysis (black).

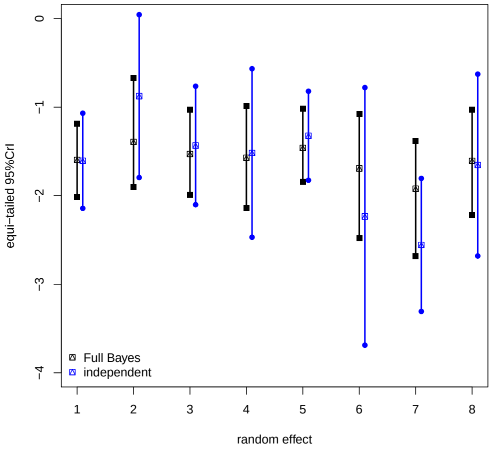


Figure 6.4: Posterior equi-tailed 95% credible intervals of random effect θ_i parameters provided by the Bayesian NNHM (black) and by the independent model (blue) for data from Table 6.1.

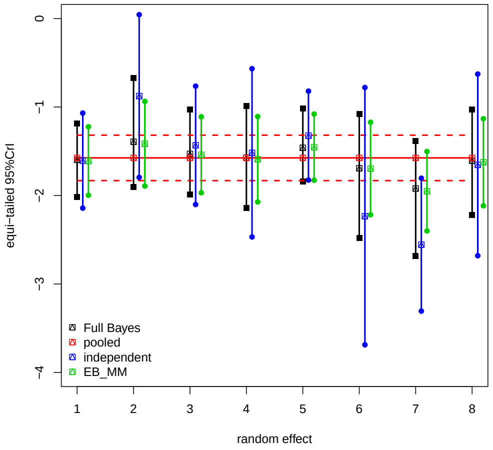


Figure 6.5: Posterior equi-tailed 95% credible intervals of random effect θ_i parameters provided by the Bayesian NNHM (black), by the pooled model (red), by the independent model (blue), and by the empirical Bayes estimates with method of moments (green) for data from Table 6.1.

6.3 Empirical Bayes

This section addresses the empirical Bayes estimates (green) shown in Figure 6.6 and explains why these estimates are close to estimates obtained by the full Bayesian approach. We explain also how these results are obtained without application of the full Bayes method.

Empirical Bayes lies inbetween full Bayesian and classical methods. It uses Bayesian methods together with observations in the data set to obtain empirical estimates of hyperparameters of a prior. Indeed, these hyperparameters are not fixed based on any prior knowledge (full Bayes), but are estimated empirically from data. Although empirical Bayes methods are not full Bayesian methods, they provide reasonable results and are useful in applications. Moreover, they provide a way to stop an infinite hierarchy that arises when we acknowledge additional levels of uncertainty in a Bayesian hierarchical model.

A Bayesian hierarchical model can be computed approximately to get

$$f(\theta \mid y) = \int f(\theta \mid y, \psi) f(\psi \mid y) d\psi \propto \int f(\theta \mid y, \psi) f(y \mid \psi) f(\psi) d\psi.$$

The empirical Bayes method uses

$$f(\theta \mid y) \approx f(\theta \mid y, \hat{\psi})$$

instead, where $\hat{\psi}$ is the marginal maximum likelihood estimator of the hyperparameter. This method avoids both the computation of the integral and the choice of a hyperprior $f(\psi)$. Empirical Bayes has also good frequentist properties justified by the famous result of Charles Stein.

Note that the empirical Bayes approach uses the data y twice. First, to select the hyperparameters of the prior $\hat{\psi}$. Second, to compute the posterior according to the Bayes formula $f(\theta \mid y, \hat{\psi})$.

The empirical Bayes approach provides some shrinkage to the prior mean. Thus, it is a compromise between total independence of random effects and the assumption of complete pooling. However, the true between-study uncertainty can be underestimated if we choose a particular value $\hat{\psi}$ of hyperparameters instead of averaging over different plausible values as it is done by the full Bayes approach.

Now, we describe the method used to compute empirical Bayes estimates shown in Figure 6.6. Similarly to the full Bayes approach, the empirical Bayes argument uses the likelihood and the model for random effects but does not use any priors for hyperparameters. Instead, values of hyperparameters are derived either numerically by maximizing a profile likelihood or by a method of moments [Spiegelhalter et al., 2004, Sections 3.17 and 3.18.2].
Likelihood:

$$y_i \sim \mathrm{N}(\theta_i, \sigma_i^2), \tag{6.9}$$

for $i = 1, \ldots, k$.
Random effects:

$$\theta_i \sim \mathrm{N}(\mu, \tau^2) \tag{6.10}$$

The posterior distribution for the i-th random effect can be derived similarly to Equation (3.9):

$$\theta_i \mid y_i \sim \mathrm{N}\left(\frac{\frac{y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}, \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)^{-1} \right). \tag{6.11}$$

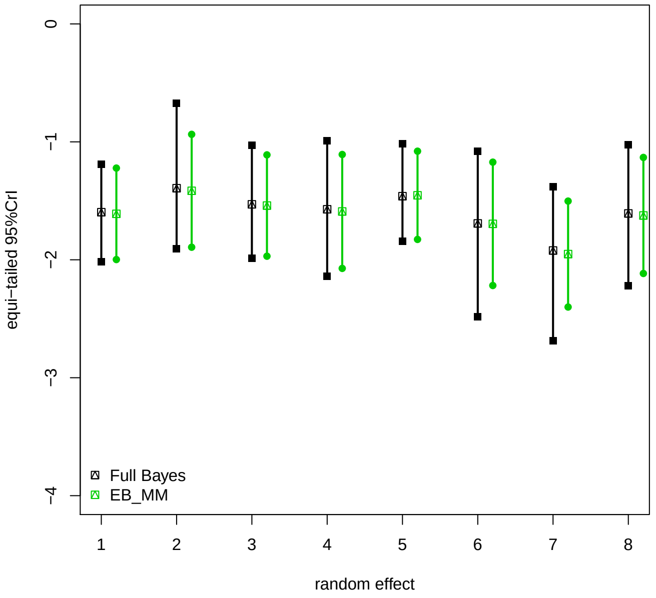


Figure 6.6: Posterior equi-tailed 95% credible intervals of random effect θ_i parameters provided by the Bayesian NNHM (black) and by the empirical Bayes estimates with method of moments (green) for data from Table 6.1.

This distribution is used for empirical Bayes inference for the i-th random effect in Figure 6.6. Thus, the empirical Bayes approach leads to inference for each random effect having narrower equi-tailed 95% intervals (precision: $1/\sigma_i^2 + 1/\tau^2$) than if they were assumed independent (precision: $1/\sigma_i^2$). Moreover, the location of each random effect is shrunk towards the mean:

$$\frac{\frac{y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} = \frac{\tau^2}{\sigma_i^2 + \tau^2} y_i + (1 - \frac{\tau^2}{\sigma_i^2 + \tau^2})\mu = \frac{\tau^2}{\sigma_i^2 + \tau^2} y_i + \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}\mu.$$

Thus, the degree of pooling or shrinkage towards the mean μ is determined by the factor $\sigma_i^2/(\sigma_i^2 + \tau^2)$. This factor depends on both the between-study variability τ^2 and the within-study variability σ_i^2.

The estimates $\hat{\mu}$ and $\hat{\tau}$ of unknown parameters μ and τ in Equation 6.11 are derived from data. The main argument for computation of estimates of hyperparameters μ and τ needed in Equation 6.11 is based on the prior predictive evaluated at observed data, the so called marginal likelihood. Therefore, we can use results obtained for the prior predictive distribution of model $Y \mid m \sim \mathrm{N}(m, \sigma^2)$ with $m \sim \mathrm{N}(\mu, \tau^2)$ provided in Section 3.2.2. Following Equation (3.11), we get:

$$y_i \sim \mathrm{N}\left(\mu, \sigma_i^2 + \tau^2\right).$$

We denote the precision of the prior predictive distribution by $w_i = 1/(\sigma_i^2 + \tau^2)$. Note that w_i is a function of τ. The independence of observations $y_1, \ldots, y_k$ leads to a marginal log-likelihood for the whole sample

$$l(\mu, \tau) = -\frac{1}{2}\Big(\sum_{i=1}^{k} w_i (y_i - \mu)^2 - \sum_{i=1}^{k} \log(w_i)\Big). \tag{6.12}$$

Differentiation of $l(\mu, \tau)$ with respect to μ and setting the first derivative equal 0 leads to a data-based estimate of μ:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}. \tag{6.13}$$

If we get an estimate for $\hat{\tau}_{\mathrm{MM}}$ by a classical meta-analysis with random effects (method-of-moments approach), we can plug in $\hat{\tau}_{\mathrm{MM}}$ in Equation (6.13) to get an explicit value of $\hat{\mu}_{\mathrm{MM}}$ [Spiegelhalter et al., 2004, Sections 3.17 and 3.18.2]. For data from Table 6.1 we obtained $\hat{\tau}_{\mathrm{MM}} = 0.286$ and $\hat{\mu}_{\mathrm{MM}} = -1.615$. These estimates were inserted in Equation 6.11 to get the empirical Bayes inference shown in Figure 6.6.

Alternatively, we can plug in the formula for $\hat{\mu}$ from Equation (6.13) in Equation (6.12) to get a profile log-likelihood

$$l(\tau) = -\frac{1}{2}\Big(\sum_{i=1}^{k} w_i \big(y_i - \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}\big)^2 - \sum_{i=1}^{k} \log(w_i)\Big) \tag{6.14}$$

with $w_i = 1/(\sigma_i^2 + \tau^2)$ to get a function of only one parameter τ. This function $l(\tau)$ can be maximized numerically to get a $\hat{\tau}_{\mathrm{ML}}$ estimate. Then, we can plug in $\hat{\tau}_{\mathrm{ML}}$ in Equation (6.13) to get an explicit value of $\hat{\mu}_{\mathrm{ML}}$. Estimates $\hat{\mu}_{\mathrm{ML}}$ and $\hat{\tau}_{\mathrm{ML}}$ can be inserted into Equation 6.11 to get the empirical Bayes inference. For data from Table 6.1 we obtained $\hat{\tau}_{\mathrm{ML}} = 4.102259\mathrm{e}{-}05$ and $\hat{\mu}_{\mathrm{ML}} = -1.609$ estimates.

6.4 Worksheet 6

Probability calculus	**Distributions**	Change of variables formula
Priors	MC sampling	Asymptotics
Bayes		**Classical**
Posterior $\propto$ **Likelihood** $\times$ **Prior**		Likelihood
Conjugate Bayes	**MCMC sampling**	Bayesian logistic regression
Predictive distributions	**JAGS**	**Bayesian meta-analysis**
Prior elicitation	**CODA**	Bayesian model selection

Table 6.2: Foundations of Bayesian Methodology: content of the lecture relevant for Worksheet 6.

Secukinumab		**Placebo**
Bayesian sample size computation		
		Bayesian meta-analysis **Prior elicitation**
Beta(0.5, 1)		**Beta(11, 32)**
Data		**Data**
	Classical analysis	
Posterior (S)		**Posterior (P)**
	Posterior probability of superiority	

Table 6.3: Individual project: A sketch of analysis steps leading to the results provided in Table 2 of Baeten et al. [2013]. For your individual project you are asked to conduct this analysis in several small steps and provide a report of your findings.

Chapter 7

Lecture 7: Priors and Bayesian model selection

The Bayes methodology is the only known scientific method that conditions on the observed data. It follows probabilistic rules and is, therefore, coherent. The Bayes methodology is flexible and allows the user to implement complicated models with little effort. Moreover, it quantifies the probability of evidence in favor of a hypothesis and provides an opportunity to monitor evidence as it accumulates.

The subjective part of the Bayes methodology is the prior. Although the prior is subjective, it is always well specified and is always open for inspection and discussion. When different reasonable priors yield substantially different answers, there may be an inherent scientific uncertainty, with conclusions depending on prior beliefs.

7.1 Priors

We agree with Lambert et al. [2008] that all priors contribute some information to posterior. Therefore, truly non-informative priors do not exist.

7.1.1 Suggested reading about priors

- Lesaffre and Lawson [2012] book chapter 5
- Garthwaite et al. [2005] Prior elicitation (psychological aspects)

  experts tend to be overconfident (spread is too narrow)

  expert opinions differ (ask several experts)

  answer / opinion can depend on how the questions have been asked
- Johnson et al. [2010]
- Consonni et al. [2018]
- Morris et al. [2014], Morris “MATCH”

`http://optics.eee.nottingham.ac.uk/match/uncertainty.php`

openly discuss and elicit priors online

- Beta Buster (Beta priors given prevalence, sensitivity and specificity of a test)
- JASP "A Fresh Way to do Statistics" `https://jasp-stats.org/how-to-use-jasp/`

7.1.2 General Suggestions

1. Think about the range (spread) and the location of the prior
2. How much mass is assigned to which values
3. Prior Effective Sample size (PESS)

   PESS ="How many observations is a prior worth"? (much literature)

   In short

   PESS $\approx$ approximately proportional to the prior's precision (inverse of its variance) .

   For example, Beta distribution

   $x \sim \text{Beta}(a, b)$

   Posterior $f(\theta \mid \mathbf{y}) \sim \text{Beta}(a + k, b + (n - k))$

   $\mathbb{E}(X) = \frac{a}{a+b}$

   $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$

   $\text{PESS} = a + b = \frac{\mathbb{E}(X)(1-\mathbb{E}X)}{\text{Var}(X)} - 1$

   Suggestion: try not to assign a stronger prior than your data. However, in certain cases the prior must be stronger than data.

4. Prior-data conflict (disagreement)

   Stephen Senn:

   "Bayesian - one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule".

5. "Predictions of today are priors of tomorrow." Learning (unrestricted Bayesian) corresponds to direct pooling. It can be done only when very strong assumptions are fulfilled (independence and exchangeability).

   Not recommended: $\text{prior}_1 \xrightarrow{\text{data}_1} \text{posterior}_1 = \text{prior}_2 \xrightarrow{\text{data}_2} \text{posterior}_2$.

   Are conditions for data_1 and data_2 identical? This procedure converges to a point (Dirac) distribution when $n \to \infty$.

   Think about

   $\text{prior}_1 \xrightarrow{\text{data}_1} \text{posterior}_1 \to \text{prediction}_1 = \text{prior}_2 \xrightarrow{\text{data}_2} \text{posterior}_2$

   How much spread is OK?

Quantifies expected differences in condition between the studies.

$f(\mathbf{y}^* \mid \mathbf{y}) = \int f(\mathbf{y}^* \mid \theta) f(\theta \mid \mathbf{y}) d\theta$

6. Sensitivity of posterior inference to prior assumptions.

7.1.3 Classification of prior distributions

(a) weakly informative	(b) informative
diffuse pessimistic vague	non-diffuse optimistic
(PESS small) priors precision small variance large	(PESS large) priors precision large variance small
little is known about the quantity of interest a priori	we are convinced where the truth is
"Let data speak for themselves" do not contaminate the likelihood much results resemble frequentist (classic) analysis	substantial prior knowledge is available it can be extracted from historical or contextual data that are relevant for the current data set

An alternative classification of prior distributions

A Priors in a conjugate analysis are related to the likelihood.
Assume

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \theta_i + \epsilon_i, \quad \theta_i \sim \mathrm{N}(0, \tau^2), \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

B For fixed effects (regression coefficients), use informative or weakly informative priors:

$$\alpha, \beta_1, \beta_2 \sim \mathrm{N}(\mu, \gamma^2), \quad \mu = 0, \quad \gamma^2 = 10^4.$$

C For random effects, think very carefully about their precision, variance, and standard deviation. Use priors on $\mathbb{R}^+$ such as for example half-normal, half-Cauchy, gamma, and log-normal. Remember that

$$\text{standard deviation} = \tau = 0 \rightarrow \text{variance} = \tau^2 = 0 \rightarrow \text{precision} = \frac{1}{\tau^2} = \infty,$$

$$\text{precision} = \frac{1}{\tau^2} = 0 \rightarrow \text{variance} = \tau^2 = \infty \rightarrow \text{standard deviation} = \tau = \infty.$$

Think what is the NULL model when $\tau = 0$.

7.2 Some computational tricks connected to priors

Be aware of challenges and possible solutions, ideas.

Ex1 Translation, Uniform for sigma → Gamma for precision.
Derivation of parameters
- using sampling MC,"Unif $\to \sigma$","$\tau \leftarrow 1/\sigma^2$"
- matching the moments (mean of the τ sample, variance of the τ sample)

$$\mathbb{E}(G(a,b)) = \frac{a}{b}$$

$$\text{Var}(G(a,b)) = \frac{a}{b^2}$$

$a = \frac{\text{mean}^2}{\text{Var}}$, $b = \frac{\text{mean}}{\text{Var}}$

Ex2 Gamma prior for precision vs priors for SD

Ex3 Tail adjusting
- PC (penalized complexity) priors [Simpson et al., 2017]
- tail-adjustment [Spiegelhalter et al., 2004, Neuenschwander et al., 2010]

Exponential distribution
$f(x) = \theta e^{-\theta x}$
$P[X > U] = \alpha \to U = -\frac{\ln(\alpha)}{\theta} \to \theta = -\frac{\ln(\alpha)}{U}$,
U is the quantile.

- U context, $\alpha = 0.05, 0.01, 0.001$
- Decide how much weight you allow to put at extreme values.

Ex4 Gamma prior (conjugate) Leukemia data / example
See [Rosner et al., 2021, Section 3.2.4]
Likelihood $\sim$ Exp(θ), ($f(t) = \theta e^{-\theta t}$)
Data $y_1, \ldots, y_n$
prior $\theta \sim \text{G}(a,b)$
posterior $f(\theta \mid y_1, \ldots, y_n) \sim G(a+n, b+\sum_{i=1}^n y_i)$
We are given an informal statement:
"The median survival time is around 20 weeks and we believe that the median is greater than 5 weeks with 95 % certainty".

1. Median $= -\frac{\ln(0.5)}{\theta} = \frac{\ln(2)}{\theta} = 20$. Then $\theta = \frac{\ln(2)}{20} = 0.0347$ (property of exponential distribution)

   Center the Gamma prior around that value by setting it to the mode $\left(\frac{a-1}{b}\right)$ (or mean $\frac{a}{b}$) of the Gamma prior.

   $\theta = \frac{a-1}{b} = 0.0347$, $a = 1 + 0.04347 \cdot b$

2. We want

   $0.95 = P[\text{Median} > 5] = P[\frac{\ln(2)}{\theta} > 5] = P[\theta < \frac{\ln(2)}{5}]$, $\theta \sim G(a,b)$, $\frac{\ln(2)}{5} = 0.1386$

Look for a minimum (root) of a function.

`pgamma(0.1386, shape = 1+0.0347 * x, rate = x ) - 0.95`, where `shape = a, rate = b`

result is b (optimal), knowing b, compute a

Ex5 Prior for the Weibull shape parameter (p)

Density $f(t) = cpt^{p-1}\exp(-ct^p)$ and hazard function $h(t) = cpt^{p-1}$ with $c = \exp\{\beta_0^{\text{PH}} + \beta_1^{\text{PH}} x\}$

$p < 1$ deceleration

$p = 1$ constant hazard

$p > 1$ acceleration

Ex6 Inference of the priori on smoothing in `rw1` vs gam-smoothing (degrees of freedom).

- many degrees of freedom cause oversmoothing	- low number of degrees of freedom (rigid linear fit)
- weakly informative prior	- very informative prior
- precision small	- precision large
- PESS is small	- PESS is large
- variance large	- variance small
- we do not impose much knowledge	- we impose very much knowledge
- there is much freedom	- there is no freedom

$$\log f(\theta \mid \mathbf{y}) \approx \underbrace{\log f(\mathbf{y} \mid \theta)}_{\text{likelihood}} \quad + \quad \underbrace{\log f(\theta)}_{\text{weight of the prior}}$$

The weight of the prior is like a penalty (smoothing).

7.3 Bayesian model selection

All inference is based on assumptions, which define a model that describes stochastic properties of the underlying true mechanism that generates data. Usually, there is a set of several candidate models that describe the true mechanism. All of them are approximations of the underlying true mechanism. The goal of model selection is to select one model from the set of candidate models that is as close as possible to the underlying true mechanism.

Usually, the modelling process passes a sequence of stages:

1. Model criticism:

   Given a current model, we evaluate its adequacy to represent data. For example, we use predictive distributions.

2. Model extension:

Any shortcomings exposed at the first stage stimulate extensions of the current model. Go to the first stage to critically assess these extensions.

3. Model comparison and selection:

   Compare candidate models that were generated at stages 1–2. The goal is to choose one final model.

4. Model averaging:

   Combine candidate models that were generated at stages 1–3. Compute a weighted average of the estimate of interest.

7.3.1 Bayes factors revisited

We have already considered Bayesian model selection in Section 1.7.1, which addressed Bayes factors and p-values. The methodology leading to Bayes factors can be generalized to comparison of two different models $\mathcal{M}_1$ and $\mathcal{M}_2$, where $P[\mathcal{M}_1] + P[\mathcal{M}_2] = 1$. The model is treated as an additional parameter lying in the set of the model space. In this case, the likelihood can be reformulated as $f(y \mid \theta_i, \mathcal{M}_i)$, the prior as $f(\theta_i \mid \mathcal{M}_i)$, and the marginal likelihood defined as

$$f(y \mid \mathcal{M}_i) = \int f(y \mid \theta_i, \mathcal{M}_i) f(\theta_i \mid \mathcal{M}_i) d\theta_i, \tag{7.1}$$

for $i = 1, 2$. The posterior probability for model $\mathcal{M}_i$ can be obtained by the Bayes theorem:

$$P[\mathcal{M}_i \mid y] = \frac{f(y \mid \mathcal{M}_i) P[\mathcal{M}_i]}{f(y \mid \mathcal{M}_1) P[\mathcal{M}_1] + f(y \mid \mathcal{M}_2) P[\mathcal{M}_2]}.$$

The posterior chance (odds) $P[\mathcal{M}_1 \mid y]/P[\mathcal{M}_2 \mid y]$ can be rewritten as the product of the Bayes factor $BF_{12}(y)$ and the prior chance (odds) $P[\mathcal{M}_1]/P[\mathcal{M}_2]$:

$$\frac{P[\mathcal{M}_1 \mid y]}{P[\mathcal{M}_2 \mid y]} = BF_{12}(y) \frac{P[\mathcal{M}_1]}{P[\mathcal{M}_2]}, \tag{7.2}$$

where

$$BF_{12}(y) = \frac{f(y \mid \mathcal{M}_1)}{f(y \mid \mathcal{M}_2)}. \tag{7.3}$$

$BF_{12}(y)$ summarizes the evidence provided by the data in favor of one model as opposed to another. The Bayes factor can be also rewritten as the ratio of posterior and priors odds for two models:

$$BF_{12}(y) = \frac{P[\mathcal{M}_1 \mid y]}{P[\mathcal{M}_2 \mid y]} \Big(\frac{P[\mathcal{M}_1]}{P[\mathcal{M}_2]} \Big)^{-1}.$$

Therefore, $BF_{12}(y)$ can be interpreted as the ratio of the posterior chance of model $\mathcal{M}_1$ and the prior chance of $\mathcal{M}_1$. The value $BF_{12}(y) > 1$ indicates that the data y increased the probability of model $\mathcal{M}_1$ and the value $BF_{12}(y) < 1$ indicates that the data y decreased the probability of model $\mathcal{M}_1$. The Bayes factor $BF_{12}(y)$ is identical to the likelihood ratio if both models $\mathcal{M}_1$ and $\mathcal{M}_2$ do not contain any unknown parameters and are completely specified. Otherwise, the marginal likelihood $f(y \mid \mathcal{M}_i)$ of the data y in Equation (7.1) must be evaluated.

Note that the marginal likelihood has an automatic build-in penalty for model complexity. To see this, consider the Bayes formula:

$$f(\theta \mid y, \mathcal{M}) = \frac{f(y \mid \theta, \mathcal{M}) f(\theta \mid \mathcal{M})}{f(y \mid \mathcal{M})}.$$

We solve this equation for the marginal likelihood $f(y \mid \mathcal{M})$ and take the logarithm to get:

$$\log(f(y \mid \mathcal{M})) = \underbrace{\log(f(y \mid \theta, \mathcal{M}))}_{\text{log-likelihood}} + \underbrace{\log(f(\theta \mid \mathcal{M})) - \log(f(\theta \mid y, \mathcal{M}))}_{\text{penalty}} \tag{7.4}$$

The log-likelihood in Equation (7.4) increases with increasing number of parameters (model complexity). The penalty in Equation (7.4) works in cases when the posterior $f(\theta \mid y, \mathcal{M})$ sharpens up with respect to the prior $f(\theta \mid \mathcal{M})$. This produces a negative penalty, because $\log(f(\theta \mid \mathcal{M})) < \log(f(\theta \mid y, \mathcal{M}))$. Thus, procedures based on the marginal likelihood seek sparsity and are consistent (select the right model if it is there). As a side remark: the classical Bayesian Information Criterion (BIC) is an approximation of the marginal likelihood.

Bayes factors $BF_{12}(y)$ are based on marginal likelihoods and have several advantages. They are easy to interpret, they work well for non-nested models, they are symmetric, they allow multiple comparisons ($BF_{13} = BF_{12}BF_{23}$), have an automatic penalty for model complexity, and are consistent. On the downside, Bayes factors cannot be used with improper and vague priors. For computation of Bayes factors of conjugate Bayes models all proportionality constants are important. For complex non-conjugate, hierarchical Bayesian models, the computation of the marginal likelihood $f(y \mid \mathcal{M})$ can be very difficult. There are different approaches to compute marginal likelihoods such as Monte Carlo and numerical approximations. See [Held and Sabanés Bové, 2020, Chapter 7] for more details.

For model selection, the prior is not asymptotically negligible. Thus, Bayesian model selection strongly depends on the prior distribution of model parameters. For more details on Bayesian model selection in regression, see Feng et al. [2008].

7.3.2 Deviance Information Criterion (DIC)

Model selection measures such as AIC and BIC are not convenient for models fitted by MCMC sampling, because the posterior mode or maximum needs to be known. Therefore, Spiegelhalter et al. [2002] proposed the Deviance Information Criterion (DIC) as a Bayesian measure of model complexity and fit:

$$\text{DIC} = \overline{D(\theta)} + p_{\text{D}}, \tag{7.5}$$

where

$$D(\theta) = -2\log(f(y \mid \theta)) + C$$

is the deviance with a constant C that cancels out in all calculations that compare different models,

$$\overline{D(\theta)} = \mathbb{E}_{\theta|y} D(\theta)$$

is the measure of the goodness of fit (the larger the value, the worse the fit), and

$$p_{\mathrm{D}} = \overline{D(\theta)} - D(\bar{\theta})$$

is the measure of model complexity. The smaller DIC value in Equation (7.5), the better is the model. See [Held and Sabanés Bové, 2020, Chapter 7.2.3] for more details.

The approximation used to derive DIC holds asymptotically when the model parameters have a normal posterior distribution and the effective number of parameters p_{D} is much smaller than the sample size. DIC tends to underpenalize complex models when applied to complex Bayesian hierarchical models. Therefore, Plummer [2008] proposed improved versions of DIC such as for example the penalized expected deviance, which penalizes complex models more severely. The function `dic.samples` in package `rjags` is useful for computation of improved versions of DIC in practice.

As an alternative to DIC, a widely applicable Bayesian information criterion (WAIC) has been proposed [Watanabe, 2010, 2013]. `INLA` computes WAIC in addition to DIC to support model comparisons and choice. See Vehtari et al. [2017] for additional details.

7.3.3 Bayesian model averaging

If the goal is to choose one single model, we can select the maximum-a-posteriori (MAP) model with the highest probability $P[\mathcal{M}_i \mid y]$. However, model selection can be of less interest than model averaging. For example, if the goal is to specify the posterior distribution of a parameter θ in view of I different model candidates $\mathcal{M}_1, \ldots, \mathcal{M}_I$, we can calculate the marginal posterior distribution of θ:

$$f(\theta \mid y) = \sum_{i=1}^{I} f(\theta \mid y, \mathcal{M}_i) P[\mathcal{M}_i \mid y].$$

This approach accounts for model uncertainty in the estimation of the parameter θ. See [Held and Sabanés Bové, 2020, Chapter 7.2.4] for more details. A similar model averaging approach could be applied to get predictions in view of I different model candidates $\mathcal{M}_1, \ldots, \mathcal{M}_I$.

Chapter 8

Outlook: INLA and Stan

This chapter provides some extra material for fitting Bayesian hierarchical models with INLA and Stan.

8.1 Integrated Nested Laplace Approximation (INLA)

LAPLACE Approximation $\neq$ Integrated Nested Laplace Approximation

INLA - apply Laplace approximation several times

INLA reading: Tierney and Kadane [1986] (Section 4), Rue et al. [2009], Rue and Held [2005], Blangiardo and Cameletti [2015] (Sections 4.7.1, 4.7.2 and 4.9), Robert and Casella [2004] (pages 107-110 and 120-122), Opper and Archambeau [2009], Goutis and Casella [1999], Rue et al. [2009], Zuur et al. [2017], Wang et al. [2018], Martins et al. [2013], Rue et al. [2017], Gerber and Furrer [2015], Muff et al. [2015], Gómez-Rubio [2020], and Martino and Riebler [2020].

Additional INLA reading:

1. Cameletti et al. [2012] about spatio-temporal models and SPDE in INLA;
2. Sørbye and Rue [2013] on the `scale.model` option in INLA;
3. Bivand et al. [2015] about spatial models, INLA and INLABMA;
4. Ferkingstad and Rue [2015] about a copula correction for logistic and Poisson models;
5. Simpson et al. [2017] about penalized complexity (PC) priors.

INFO:

- Blangiardo and Cameletti [2015] introduce their own notation
- R-INLA package (`https://www.r-inla.org/`) and help with notation of Rue et al. [2009]

INFO: Replicability of results computed with INLA:

To speed up computations, INLA conducts calculations in parallel. This produces very similar but still different results, as we cannot control how the run-time system does the parallelization (which will be different each time). To get identical numerical results with INLA, use:

```
library(INLA)
inla.setOption(smtp='taucs', num.threads="1:1")
```

8.1.1 INLA Models

y_i is assumed to belong to an exponential family, where the mean μ_i is linked to a structured additive predictor η_i through a link-function $g(.)$, so that $g(\mu_i) = \eta_i$. The structured predictor η_i accounts for effects of various covariates in an additive way:

First stage: likelihood $\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i,$$

where $f^{(j)}(u_{ji})$ are unknown smooth functions of the covariates $\mathbf{u}$, β_k are the linear slopes coefficients of covariates $\mathbf{z}$, ϵ_i is unstructured random term, $(\mathbf{y}, \mathbf{z})$ are the data, linear regression.

Latent Gaussian Markov random field. Assign Gaussian prior to each

$$\mathbf{x} = \left\{\alpha, f^{(j)}(.), \beta_k, \epsilon_i\right\},$$

$\mathbf{x}$ the vector of all latent Gaussian variables (random field).

Second stage: latent field $\pi(\mathbf{x} \mid \boldsymbol{\theta})$ forms a Gaussian Linear Markov Random Field.

Third stage: priors for hyperparameters

Let $\boldsymbol{\theta}$ be a vector of hyperparameters for the latent vector.

- Assume priors for the hyperparameters. They do not need to be Gaussian, $\pi(\boldsymbol{\theta})$.
- $\mid \boldsymbol{\theta} \mid \leq 6$!

Posterior

$$\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x} \mid \boldsymbol{\theta}) \prod_{i \in T} \pi(y_i \mid \mathbf{x}, \boldsymbol{\theta}).$$

Approximated by the INLA approach.

- $y_i, \ldots, y_n \sim \mathrm{N}(\mu, \sigma^2)$, $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, $1/\sigma^2 \sim \mathrm{G}(\alpha_0, \beta_0)$

  Normal example → INLA is exact.

Remark

Do an Exploratory Data Analysis (EDA) prior to INLA analysis. INLA reacts to outliers etc.

Remarks (Restrictions)

* Restriction to 3-stage models.

* Not all OpenBUGS examples from Volumes I, II can be translated to INLA. Nevertheless, there is still a large number of useful models that can be fitted by INLA [Wang et al., 2018].

* $\mid \boldsymbol{\theta} \mid \leq 6$

* Only marginal posterior densities on output ⇒ Linear Combinations

Remarks (Advantages)

* Approximation (very fast approximation) can be computed within seconds

* Modular structure, modules can be easily combined with each other

* No CODA necessary!!!

Differences of MCMC algorithms and INLA approximation

MCMC sampling JAGS, STAN, OpenBUGS Gibbs, MH Joint posterior distribution sum $= \alpha + \beta$ can be easily computed	INLA approximates the joint posterior distribution and shows only marginal posteriors No direct samples Linear combination $\alpha + \beta$ must be computed separately

8.1.2 Modular Structure of INLA

1. Likelihoods $\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$ ($\mathbf{x}$ is the latent field, not data)

   Poisson

   Gaussian

   Student-t

   Negative Binomial

   Zero Inflated

   Exponential, Weibull, Cox

2. Latent models $\pi(\mathbf{x} \mid \boldsymbol{\theta})$

   iid

   rw1 (Random walk), rw2 (gam smoothing, `mgcv`)

   besag

   BYM (Besag, York, Mollié)

   mec, meb (combine modules)

3. Priors $\pi(\boldsymbol{\theta})$

   β normal

   τ loggamma

   log t normal, expression / table

Remark: Internal computations are conducted on the log-scale.

8.2 Stan

Stan -"Sampling through Adaptive Neighborhoods"

8.3 Reading about Stan

The relevant books for STAN are Gelman et al. [2014a], Kruschke [2015], Korner-Nievergelt et al. [2015], and Rosner et al. [2021].

Relevant and useful books for physics and Hamiltonian are Greiner [2008], Beiser [2003], Fliessbach [1995a], Fliessbach [1995b], Grosche et al. [2003].

See interfaces to Stan in R: `rstan` and the R Interface to CmdStan `CmdStanR`.

Learn more about STAN:

1. Part 1 (Lectures 1 and 2): `https://www.youtube.com/watch?v=pHsuIaPbNbY`;

2. Part 2 (Lecture 3): `https://www.youtube.com/watch?v=xWQpEAyI5s8`;

3. HMC: Neal [2011];

4. HMC: Betancourt and Girolami [2013];

5. NUTS: Hoffman and Gelman [2014];

6. NUTS: `http://de.slideshare.net/xianblog/hmc-bi-p`;

7. Kucukelbir et al. [2015] on ADVI (automatic differentiation variational inference) in Stan

8.3.1 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a Markov chain Monte Carlo (MCMC) method that uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior (see, e.g., Betancourt and Girolami [2013] for more details). It uses an approximate Hamiltonian dynamics simulation based on numerical integration which is then corrected by performing a Metropolis acceptance step.

The algorithm: The HMC starts at a specified initial set of parameters (it can be user-specified or generated randomly). For a given number of iterations, a new momentum vector is sampled and the current value of parameters is updated using the leapfrog integrator with discretization time stepsize and number of steps n_leapfrog. Then a Metropolis acceptance step is applied, to decide whether to update the new state or to keep the existing state.

No-U-Turn Sampler (NUTS) automatically selects an appropriate n_leapfrog in each iteration in order to allow the proposals to traverse the posterior without doing unnecessary work. The motivation is to maximize the expected squared jump distance at each step and avoid random-walk behavior that arises in random-walk Metropolis or Gibbs samplers when there is correlation in the posterior. For more details on NUTS algorithm see Hoffman and Gelman [2014].

The algorithm: NUTS generates a proposal by starting at an initial position determined by the parameters drawn in the last iterations. Next, it generates an independent unit-normal random momentum vector. Thereafter, it evolves the initial system both forwards and backwards in time to form a balanced binary tree. At each iteration of the NUTS algorithm the treedepth is increased by one, doubling n_leapfrog and effectively doubling the computation time. The algorithm terminates, if the NUTS criterion is satisfied or the depth of the completed tree hits the maximum depth allowed. Instead of standard Metropolis step, the final parameter value is selected with slice sampling (See MacKay Lecture 13, 1:07).

Configuring the no-U-turn sampler involves putting a cap on the treedepth that it evaluates during each iteration, which is controlled by maximum depth parameter.

$$\text{The number of leapfrog steps} \leq 2^{\text{maxDepth}-1}.$$

accept-stat used by NUTS for slice and Metropolis rejection is acceptance probability averaged over samples in the slice. For HMC without NUTS accept-stat is the standard Metropolis acceptance probability.

n_divergent is the number of leapfrog transitions with diverging error. Because NUTS terminates at the first divergence this will be either 0 or 1 for each iteration. The average value of n_divergent over all iterations is therefore the proportion of iterations with diverging error. Stan uses a symplectic integrator to approximate the exact solution of the Hamiltonian dynamics and when the stepsize is too large relative to the curvature of the log posterior this approximation can diverge.

If there are any post-warmup iterations for which n_divergent = 1 then the results may be biased and should not be used. One should try to rerun the model with a higher target acceptance probability until n_divergent = 0 for all post-warmup iterations.

step-size is the integrator step size used in the Hamiltonian simulation. If step-size is too large, the leapfrog integrator will be inaccurate and too many proposals will be rejected. The too small step-size leads to long simulation times per interval.

n_leapfrog is the number of leapfrog steps (calculations) taken during the Hamiltonian simulation. If n_leapfrog is too small, the trajectory traced out in each iteration will be too short. If n_leapfrog is too large, the algorithm will do too much work on each iteration.

treedepth is the depth of tree used by NUTS. Tree depth is an important diagnostic tool for NUTS. For example, treedepth = 0 occurs when the first lepfrog step is immediately rejected and the initial state returned, indicating extreme curvature and poorly chosen stepsize. On the other hand, treedepth = max_treedepth indicates that NUTS is taking many leapfrog steps and being terminated prematurely to avoid excessively long execution time.

8.3.2 JAGS (BUGS) vs STAN

Remarks

1. BUGS visits only one node at a time

   Stan moves in the entire space of parameters according to Newtons laws

2. adaptation and burn-in → warm-up for NUTS (specify tuning parameters and find the stationary distribution)

   convergence → finding and exploring the typical set

   Chain diagnostics (CODA) and Sampler Diagnostics (Stan)

3. BUGS expert system to choose the sampler

   STAN directly executable code

4. BUGS order of commands in the model does not matter

   Stan important block-order

5. BUGS no gradients necessary

   Stan requires computation of gradients

6. BUGS semi-automatic tune/adapt

   Stan monitor $\widehat{R}$ potential scale reduction, ESS

Language differences

- ;

- vectorization

- BUGS N(μ, precision), `dnorm`

  Stan N(μ, sd), `normal`

- Blocks / declarations / boundedness

  data

  transformed data

  parameters

  transformed parameters

  model (required)

  generated quantities

- real theta;

  Stan can produce improper models easily if the prior is not specified

- reparametrizations (BUGS and Stan)
- extractor for postprocessing
- Stan transformation to unconstrained space
- Stan HMC bias difficulties with short- and long-tailed distributions
- Stan only continuous parameters
- Stan

  no missing values in data

  but a missing // commented out likelihood is possible

Acknowledgement:

We thank Sona Hunanyan for typing an earlier version of this script for the Bayesian Data Analysis lecture in the spring term 2018.

Bibliography

D. Baeten, X. Baraliakos, J. Braun, J. Sieper, P. Emery, D. van der Heijde, I. McInnes, J. van Laar, R. Landewé, P. Wordsworth, J. Wollenhaupt, H. Kellner, J. Paramarta, J. Wei, A. Brachat, S. Bek, D. Laurent, Y. Li, Y.A. Wang, A.P. Bertolino, S. Gsteiger, A.M. Wright, and W. Hueber. Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *The Lancet*, 382:1705–1713, 2013.

M.J. Bayarri and J.O. Berger. An interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

A. Beiser. *Concepts of Modern Physics. 6th Edition.* McGraw-Hill, 2003.

D.A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* John Wiley & Sons Inc, 1991.

M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. *arXiv:1312.0906v1*, 2013.

R.S. Bivand, V. Gómez-Rubio, and H. Rue. Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software*, 63(20):1–30, 2015.

M. Blangiardo and M. Cameletti. *Spatial and Spatio-temporal Bayesian Models with R-INLA.* Wiley, 2015.

E.L. Boone, J.R.W. Merrick, and M.J. Krachey. A Hellinger distance approach to MCMC diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849, 2014. URL `https://doi.org/10.1080/00949655.2012.729588`.

M. Borenstein, J.P.T. Higgins, L.V. Hedges, and H.R. Rothstein. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1):5–18, 2017. URL `https://doi.org/10.1002/jrsm.1230`.

S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

M. Cameletti, F. Lindgren, D. Simpson, and H. Rue. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, pages 1–23, 2012. URL `http://dx.doi.org/10.1007/s10182-012-0196-3`.

D. Collett. *Modelling Binary Data. Second Edition.* Chapman & Hall/CRC, 2003.

G. Consonni, D. Fouskakis, B. Liseo, and I. Ntzoufras. Prior distributions for Objective Bayesian analysis. *Bayesian Analysis*, 13(2):627–679, 2018. doi: https://doi.org/10.1214/18-BA1103. URL https://projecteuclid.org/euclid.ba/1523671250.

M.K. Cowles and B.P. Carlin. Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

L. Feng, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481): 410–423, 2008. doi: 10.1198/016214507000001337. URL https://doi.org/10.1198/016214507000001337.

E. Ferkingstad and H. Rue. Improving the inla approach for approximate bayesian inference for latent gaussian models. *(arXiv:1503.07307v2)*, 2015.

T. Fliessbach. *Quantenmechanik. Lehrbuch zur Theoretischen Physik III. 2. Auflage.* Spektrum Akademischer Verlag, 1995a.

T. Fliessbach. *Statistische Physik. Lehrbuch zur Theoretischen Physik IV. 2. Auflage.* Spektrum Akademischer Verlag, 1995b.

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society, Series A*, 182(2):389–402, 2019. doi: https://doi.org/10.1111/rssa.12378. URL https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12378.

P.H. Garthwaite, J.B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–700, 2005.

A. Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, 2008. URL https://doi.org/10.1002/sim.3107.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7(4):457–511, 1992.

A. Gelman, A. Jakulin, M.G. Pittau, and Y-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2 (4):1360–1383, 2008.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis. 3rd Edition.* Chapman & Hall/CRC Press, Boca Raton, 2014a.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014b.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. doi: 10.1109/TPAMI.1984.4767596.

F. Gerber and R. Furrer. Pitfalls in the implementation of Bayesian hierarchical modeling of areal count data: An illustration using BYM and Leroux models. *Journal of Statistical Software*, 63(Code Snippet 1):1–32, 2015. URL http://www.jstatsoft.org/.

C.J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992. URL https://www.jstor.org/stable/2246094.

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo.* Chapman & Hall, 1996.

W.R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992.

V. Gómez-Rubio. *Bayesian Inference with INLA*. Chapman & Hall/CRC, 2020. URL https://becarioprecario.bitbucket.io/inla-gitbook/.

S.N. Goodman. Toward evidence-based medical statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, 130(12):1005–1013, 1999a. URL https://doi.org/10.7326/0003-4819-130-12-199906150-00019.

S.N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999b. URL https://doi.org/10.7326/0003-4819-130-12-199906150-00008.

C. Goutis and G. Casella. Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224, 1999. doi: 10.1080/00031305.1999.10474463. URL https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1999.10474463.

W. Greiner. *Klassische Mechanik II. 8. Auflage.* Verlag Harri Deutsch, 2008.

G. Grosche, V. Ziegler, D. Ziegler, and E. Zeidler, editors. *Teubner-Taschenbuch der Mathematik, Teil II. 8. Auflage.* Teubner, 2003.

S. Hartnack and M. Roos. Teaching: confidence, prediction and tolerance intervals in scientific practice: a tutorial on binary variables. *Emerging Themes in Epidemiology*, 18(17), 2021. URL https://doi.org/10.1186/s12982-021-00108-1.

L Held and M. Ott. On p-values and Bayes Factors. *Annual Review of Statistics and Its Application*, 5:393–419, 2018. URL https://doi.org/10.1146/annurev-statistics-031017-100307.

L. Held and D. Sabanés Bové. *Applied Statistical Inference. Likelihood and Bayes.* Springer, 2014.

L. Held and D. Sabanés Bové. *Likelihood and Bayesian Inference: With Applications in Biology and Medicine.* Springer, 2020. URL `https://www.springer.com/gp/book/9783662607916`.

S.D. Hill and J.C. Spall. Stationarity and convergence of the Metropolis-Hastings algorithm: Insights into theoretical aspects. *IEEE Control Systems Magazine*, 39(1):56–67, 2019. doi: 10.1109/MCS.2018.2876959. URL `https://ieeexplore.ieee.org/document/8616920`.

J.P. Hobert and C.J. Geyer. Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, 67:414–430, 1998.

M.D. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

A.A. Johnson, M.Q. Ott, and M. Dogucu. *Bayes Rules! An Introduction to Applied Bayesian Modeling.* Chapman and Hall/CRC, 2022. URL `https://www.bayesrulesbook.com`.

S.R. Johnson, G.A. Tomlinson, G.A. Hawker, J.T. Granton, and B.M. Feldman. Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of Clinical Epidemiology*, 63:355e–369, 2010.

G.L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004. URL `https://doi.org/10.1214/154957804100000051`.

G.L. Jones and J.P. Hobert. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32(2):784–817, 2004. URL `https://doi.org/10.1214/009053604000000184`.

F. Korner-Nievergelt, T. Roth, S. von Felten, J. Guélat, B. Almasi, and P. Korner-Nievergelt. *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN.* Elsevier, 2015.

J.K. Kruschke. *Doing Bayesian Data Analysis. A Tutorial with R, JAGS and Stan. 2nd Edition.* Elsevier, 2015.

A. Kucukelbir, R. Ranganath, A. Gelman, and D.M. Blei. Automatic variational inference in Stan. *Neural Information Processing Systems*, 2015.

S. Kuikka, J. Vanhatalo, H. Pulkkinen, S. Mäntyniemi, and J. Corander. Experiences in Bayesian inference in Baltic salmon management. *Statistical Science*, 29(1):42–49, 2014.

P.C. Lambert, A.J. Sutton, P.R. Burton, K.R. Abrams, and D.R. Jones. Comments on "Trying to be precise about vagueness" by Stephen Senn, Statistics in Medicine 2007; 26:1417-1430. *Statistics in Medicine*, 27(4):619–622, 2008. URL `https://doi.org/10.1002/sim.3043`.

L.M. Leemis and J.T. McQueston. Univariate Distribution Relationships. *The American Statistician*, 62(1):45–53, 2008. doi: 10.1198/000313008X270448. URL `https://doi.org/10.1198/000313008X270448`.

E. Lesaffre and A.B. Lawson. *Bayesian Biostatistics*. John Wiley & Sons, 2012.

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009.

D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS Book: A Practical Introduction to Bayesian Analysis.* Chapman & Hall/CRC, 2013.

G.M. Martin, D.T. Frazier, and C.P. Robert. Computing Bayes: Bayesian computation from 1763 to the 21st century. *arXiv:2004.06425*, 2020. URL `https://arxiv.org/abs/2004.06425`.

S. Martino and A. Riebler. Integrated Nested Laplace Approximations (INLA). In *Wiley StatsRef: Statistics Reference Online*, pages 1–19, 2020. URL `https://doi.org/10.1002/9781118445112.stat08212`.

T.G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.

N. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science*, 15:125–130, 1987.

D.E. Morris, J.E. Oakley, and J.A. Crowe. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling* & *Software*, 52:1e–4, 2014.

S. Muff, A. Riebler, L. Held, H. Rue, and Ph. Saner. Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 64(2):231–252, 2015.

R.M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

R.M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC Press, 2011.

B. Neuenschwander, G. Capkun-Niggli, M. Branson, and D.J. Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010. URL `https://doi.org/10.1177/1740774509356002`.

I. Ntzoufras. *Bayesian Modeling Using WinBUGS.* John Wiley & Sons, Inc., 2009.

M. Opper and C. Archambeau. The variational gaussian approximation revisited. *Neural Computation*, 21:786–792, 2009.

M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22.* Vienna, Austria, 2003.

M. Plummer. Penalized loss function for Bayesian model comparison. *Biostatistics*, 9(3): 523–539, 2008.

M. Plummer. *JAGS: Just Another Gibbs Sampler*, 2016. URL `http://mcmc-jags.sourceforge.net`. version 4.2.0.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11, 2006.

T.C. Prevost, K.R. Abrams, and D.R. Jones. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine*, 19(24):3359–3376, 2000. URL `https://doi.org/10.1002/1097-0258(20001230)19:24<3359::AID-SIM710>3.0.CO;2-N`.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods, Second Edition.* Springer, 2004.

C.P. Robert and G. Casella. *Introducing Monte Carlo Methods with R.* Springer, 2010.

G.L. Rosner, P.W. Laud, and W.O. Johnson. *Bayesian Thinking in Biostatistics*. Chapman & Hall/CRC, 2021. URL `https://github.com/BTB-RLJ`.

J.N. Rouder and R.D. Morey. Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73(2):186–190, 2019. URL `https://doi.org/10.1080/00031305.2017.1341334`.

C. Röver. Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software*, 93(6):1–51, 2020. doi: 10.18637/jss.v093.i06. URL `https://www.jstatsoft.org/article/view/v093i06`.

C. Röver and T. Friede. Discrete approximation of a mixture distribution via restricted divergence. *Journal of Computational and Graphical Statistics*, 26(1):217–222, 2017.

C. Röver, R. Bender, S. Dias, Schmidli H. Schmid, C.H., S. Sturtz, S. Weber, and T. Friede. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, page Early View, 2021. doi: https://doi.org/10.1002/jrsm.1475. URL `https://onlinelibrary.wiley.com/doi/full/10.1002/jrsm.1475`.

V. Roy. Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020. URL `https://doi.org/10.1146/annurev-statistics-031219-041300`.

H. Rue and L. Held. *Gaussian Markov Random Fields. Theory and Applications*. Chapman & Hall/CRC, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B.*, 71(2):319–392, 2009. URL `https://doi.org/10.1111/j.1467-9868.2008.00700.x`.

H. Rue, A. Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson, and F.K. Lindgren. Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Applications*, 4 (March):395–421, 2017.

D. Simpson, H. Rue, A. Riebler, T.G. Martins, and S.H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017.

S.H. Sørbye and H. Rue. Tutorial: Scaling IGMRF-models in R-INLA. 2013. URL `www.r-inla.org/examples/tutorials/tutorial-on-option-scale-model`.

D.J. Spiegelhalter, L.S. Freedman, and K.B. Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3): 357–416, 1994. URL `http://www.jstor.com/stable/2983527`.

D.J. Spiegelhalter, K.R. Abrams, and J.P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley & Sons, Chichester, 2004.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B.*, 64(4):583–616, 2002. ISSN 1369-7412. URL `https://doi.org/10.1111/1467-9868.00353`.

A.J. Sutton and K.R. Abrams. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4):277–303, 2001. URL `https://doi.org/10.1177/096228020101000404`.

L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

D. Vats and C. Knudson. Revisiting the Gelman-Rubin diagnostic. *Statistical Science*, 36(4): 518–529, 2021. doi: doi.org/10.1214/20-STS812. URL `https://arxiv.org/abs/1812.09384`.

A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27:1413–1432, 2017.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P. Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi: 10.1214/20-BA1221. URL `https://doi.org/10.1214/20-BA1221`.

P.E. Verde. A bias-corrected meta-analysis model for combining, studies of different types and quality. *Biometrical Journal*, 63(2):406–422, 2021. URL `https://doi.org/10.1002/bimj.201900376`.

P.E. Verde and C. Ohmann. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods*, 6(1):45–62, 2015. doi: 10.1002/jrsm.1122. URL `https://doi.org/10.1002/jrsm.1122`.

D. Wabersich and J. Vandekerckhove. Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, 46:15–28, 2014.

X. Wang, Y.R. Yue, and J.F. Faraway. *Bayesian Regression Modeling with INLA.* Chapman & Hall/ CRC Press, 2018.

S. Watanabe. Asymptotic equivalence if Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11: 3571–3594, 2010.

S. Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.

S.N. Wood. Just Another Gibbs Additive Modeller: Interfacing JAGS and mgcv. *arXiv:1602.02539v1*, 2016.

A. Zuur, E.N. Ieno, and A.A. Savaliev. *Beginner's Guide to Spatial, Temporal, and Spatial-Temporal Ecological Data Analysis with R-INLA. Volume I: Using GLM and GLMM.* Highland Statistics Ltd., 2017.