“Bayesian Inference and Stochastic Simulation”
An Excursion to 15 Topics

Participants of STA480, Spring 2016,
and Reinhard Furrer, Mattia Molinaro

Version July 14, 2016
Contents

Preface v

Andrea Meier:
1 Bayesian inference 1
  1.1 Introduction .................................................. 1
  1.2 Bayes’ theorem .................................................. 1
  1.3 Posterior distribution ........................................... 2

Kelly Reeve:
2 Choice of priors 7
  2.1 Introduction .................................................. 7
  2.2 Choosing a prior ............................................... 8
  2.3 Comparing MLE and posterior mean ............................ 13

Marco Schleiniger:
3 Loss functions and Bayes estimates 15
  3.1 What we know so far ............................................ 15
  3.2 Loss functions ................................................ 16
  3.3 Bayes estimate ................................................. 16
  3.4 Loss functions for credible regions ........................... 17

Ke Li:
4 Bayesian inference in multiparametric models 21
  4.1 Introduction .................................................. 21
  4.2 Conjugate prior distributions .................................. 21
  4.3 Jeffreys’ and reference prior distributions .................... 23
  4.4 Elimination of nuisance parameters ............................ 25
Marco Kaufmann:
5 Bayesian asymptotics 29
  5.1 Introduction .......................................................... 29
  5.2 Main part .............................................................. 29
  5.3 Exercises .............................................................. 35

Tasnim Hamza:
6 Bayesian prediction 37
  6.1 Introduction ........................................................... 37
  6.2 Prior and posterior predictive distribution ....................... 37
  6.3 Computation of the posterior predictive distribution ........... 41
  6.4 Bayesian model averaging ........................................... 44
  6.5 Supplement: R-Code .................................................. 44

Michael Hediger:
7 Overview AIC, BIC 47
  7.1 Shannon's information theory and the K–L information ........ 47
  7.2 From K–L information to AIC ....................................... 56
  7.3 Bayesian information criterion .................................... 60

Silvia Panunzi:
8 Bayesian model selection 61
  8.1 Introduction ........................................................... 61
  8.2 General framework ................................................... 61
  8.3 Bayes factors .......................................................... 62
  8.4 Marginal likelihood approximation ................................ 64
  8.5 BIC derivation ........................................................ 65
  8.6 Deviance information criterion ..................................... 66
  8.7 Bayesian model averaging ............................................. 67

Rasha Aktaa:
9 Hierarchical models 71
  9.1 Hierarchical models .................................................. 71
  9.2 Dynamic models ....................................................... 73

Amr Elriedy:
10 Monte Carlo methods 77
  10.1 Introduction .......................................................... 77
  10.2 Monte Carlo integration ............................................. 78
10.3 Importance sampling ......................................................... 80
10.4 Rejection sampling ............................................................ 81
10.5 Exercises ........................................................................... 83
10.6 Supplement: R-Code ............................................................. 83

Roman Flury:

11 Gibbs sampling ...................................................................... 85
  11.1 Basic ideas and concepts ..................................................... 85
  11.2 Definition of Gibbs sampling .............................................. 86
  11.3 Applied Gibbs sampling ...................................................... 87
  11.4 Frequentist setting .............................................................. 91
  11.5 Counter example ............................................................... 91

Thimo Schuster:

12 Software for Gibbs sampling .................................................. 93
  12.1 Introduction ...................................................................... 93
  12.2 Software for Gibbs sampling ............................................ 94
  12.3 Exercises ......................................................................... 99

Verena Steffen:

13 Diagnostic tests ...................................................................... 101
  13.1 Principles .......................................................................... 101
  13.2 Exploratory and graphical techniques .................................... 103
  13.3 Gelman and Rubin diagnostic ............................................. 105
  13.4 Geweke diagnostic ............................................................ 107
  13.5 Raftery and Lewis diagnostic ............................................ 108
  13.6 Heidelberger and Welch diagnostic ..................................... 109

Megan Payne:

14 Metropolis–Hastings algorithm ............................................... 111
  14.1 Basic ideas ...................................................................... 111
  14.2 The Metropolis–Hastings algorithm ..................................... 111
  14.3 Example: A mixed normal distribution ................................ 112
  14.4 Tuning parameter ............................................................. 112
  14.5 Acceptance rates ............................................................. 114
  14.6 Burn-in and thinning ......................................................... 114
  14.7 Special cases ................................................................... 115
  14.8 Exercises ......................................................................... 117
Sonja Hartnack:

15 Markov chains and applications 119
  15.1 Introduction to Markov chains ........................................ 119
  15.2 Diagnostic test accuracies .............................................. 121
  15.3 No gold standard models ............................................... 121
  15.4 Veterinary applications ............................................... 124

References 127

Contributor Index 131
Preface

This document has been written as part of the module *STA480 Biostatistical Journal Club* during the spring semester 2016. The module is given in the framework of the Master program in Biostatistics ([www.math.uzh.ch/biostat](http://www.math.uzh.ch/biostat)) and 4 ETCS credits are rewarded upon successful completion.

The following R packages are used: *pscl*, *ggplot2* (Chapter 3); *LearnBayes*, *VGAM* (Chapter 6); *flexmix* (Chapter 7); *rjags*, *ggmcmc*, *mcmcplots* (Chapter 12) and *coda*, *mcmcplots* (Chapter 13). Additionally, the JAGS software is required for Chapter 12.

It is the first time that I have asked the students to use the GitLab environment to compile this document and for many the road was a bit steep and bumpy. I have profited from the process as well and at least in the upcoming semesters, the road will be somewhat smoother. But in my opinion all participating students can be proud of what has been accomplished over the semester.

I thank all participants for their hard work and their contribution.

Reinhard Furrer
July 2016
Chapter 1
Bayesian inference

By:
Andrea Meier

1.1 Introduction
Bayesian Inference is a powerful and important technique in statistics to update probabilities as more information becomes available. In the Bayesian approach we have some basic differences compared to frequentist inference. Where frequent inference treat the data $X$ as random and the unknown parameter $\theta$ as fixed, Bayesian inference treat the unknown parameter $\theta$ as random variable and the data $X$ as observed information. Frequentist inference build point and interval estimates of the parameter $\theta$ as functions of the data $X$. Bayesian inference compute with the prior distribution of $\theta$ and Bayes’ theorem the posterior distribution, which summarizes the information about $\theta$ after observing the data $X = x$. The Bayesian approach does not presume infinite repeatable random experiments.

This chapter follows in structure and content the Chapters 6.1 and 6.2 from Held and Sabanés Bové (2014). If not mentioned otherwise, all examples, formulas and pictures are from these chapters and can be looked up there in a more detailed version.

1.2 Bayes’ theorem
Thomas Bayes (1702-1761) defined his theorem for any two events $A$ and $B$ with $0 < \Pr(A) < 1$ and $\Pr(B) > 0$ as

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$
Bayes’ theorem is a useful and powerful tool in estimating the probability of a hypothesis and making predictions and thus very established in analyzing diagnostic tests.

**Example 1.1.** Suppose a simple diagnostic test for a specific disease with sensitivity \( \Pr(T + |D+) = 0.9 \) and specificity \( \Pr(T - |D-) = 0.9 \). Bayes’ formula can be used to compute the conditional probability of disease given a positive test result, also called *positive predictive value*:

\[
\Pr(D + |T +) = \frac{\Pr(T + |D+) \Pr(D+)}{\Pr(T+)}.
\]

Prevalence \( \Pr(D+) \) of this disease is assumed to be 1%. The unconditional probability of a positive test result \( \Pr(T+) \) can be computed via the law of total probability:

\[
\Pr(T+) = \Pr(T + |D+) \Pr(D+) + \Pr(T + |D-) \Pr(D-),
\]

with \( \Pr(T + |D-) = 1 - \Pr(T - |D-) \).

In the above example, \( \Pr(T+) = 0.9 \cdot 0.01 + 0.1 \cdot 0.99 = 0.108 \), and hence

\[
\Pr(D + |T +) = \frac{0.9 \cdot 0.01}{0.108} \approx 0.083,
\]

meaning if the test was positive, then the disease risk increases from 1.0 % (prior risk) to 8.3 % (posterior risk). This process is also called *Bayesian updating*.

For \( n = 2 \) events we can easily switch to a simpler and more intuitive version of Bayes’ theorem when looking on odds instead of probabilities.

\[
\begin{align*}
\frac{\Pr(D + |T +)}{\Pr(D - |T +)} &= \frac{\Pr(T + |D+)}{\Pr(T + |D-)} \cdot \frac{\Pr(D+)}{\Pr(D-)}.
\end{align*}
\]

A more general formulation of Bayes’ theorem is given by

\[
f(D = d | T = t) = \frac{f(T = t | D = d) f(D = d)}{f(T = t)},
\]

where \( D \) and \( T \) are binary random variables and emphasizing that the theorem relates also to general probability mass functions of the random variables \( D \) and \( T \). This equation also holds if the random variables \( D \) or \( T \) have more than two possible values or are even continuous random variables, in which case \( f(,;) \) would denote the density function.

### 1.3 Posterior distribution

#### 1.3.1 Definition

The posterior distribution contains all the information available about the unknown parameter \( \theta \) after having observed the data \( X = x \) and is thus very essential.

If \( X = x \) denotes the observed realization of a (possibly multivariate) random variable \( X \) with density function \( f(x|\theta) \) and we specified a prior distribution with density function \( f(\theta) \), we can compute the density function \( f(\theta|x) \) of the posterior distribution using Bayes’ theorem

\[
f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}.
\]
For a discrete $\theta$, the integral has to be replaced with a sum. And since $\theta$ is random, we explicitly condition on a specific value $\theta$ and write $f(x|\theta)$, but this is the same as previously $L(\theta) = f(x|\theta)$. The denominator can also be written as

$$\int f(x|\theta)f(\theta)d\theta = \int f(x,\theta)d\theta = f(x),$$

which emphasizes that it does not depend on $\theta$. The quantity $f(x)$ is known as the marginal likelihood and is important for Bayesian model selection. The density of the posterior distribution is therefore proportional to the product of the likelihood and the density of the prior distribution with a subsequent normalization by $1/f(x)$.

### 1.3.2 Statistical inference on $\theta$ based on the posterior distribution only

#### Point estimates

The posterior mean $E(\theta|x)$ is the expectation of the posterior distribution (implicitly often assumed to be finite and thus unique):

$$E(\theta|x) = \int \theta f(\theta|x)d\theta.$$  

The posterior mode $\text{Mod}(\theta|x)$ is the mode of the posterior distribution (not necessarily unique):

$$\text{Mod}(\theta|x) = \arg\max f(\theta|x).$$

When starting with a uniform prior distribution, the posterior mode equals the MLE.

The posterior median $\text{Med}(\theta|x)$ is the median $a$ of the posterior distribution (not necessarily unique):

$$\int_{-\infty}^{a} f(\theta|x)d\theta = 0.5$$  

and  

$$\int_{a}^{\infty} f(\theta|x)d\theta = 0.5.$$

#### Interval estimates

A credible interval with credible level $\gamma$ is defined through two real numbers $t_l$ and $t_u$, that fulfill:

$$\int_{t_l}^{t_u} f(\theta|x)d\theta = \gamma.$$  

The definition implies that the unknown parameter $\theta|x$ lies in the certain credible interval with probability $\gamma$. The easiest way to compute equi-tailed credible intervals is to choose $t_l$ as the $(1-\gamma)/2$ quantile and $t_u$ as the $(1+\gamma)/2$ quantile of the posterior distribution. For discrete parameter spaces $\Theta$, the definition of a credible intervals $I$ has to be changed to

$$\sum_{\theta \in I \cup \Theta} f(\theta|x) \geq \gamma,$$

since it might not be possible to find any interval with exact credible level $\gamma$.

A highest posterior density interval (HPD) describes a $\gamma$ credible interval $I = [t_l, t_u]$ if

$$f(\theta|x) \geq f(\tilde{\theta}|x),$$

for all $\theta \in I$ and all $\tilde{\theta} \notin I$. An HPD interval contains all those parameter values that have higher posterior density than all parameter values not contained in the interval. Although this
property does not always hold, the posterior density ordinates at the limits of an HPD interval are equal under some additional regularity conditions, meaning

\[ f(t_l|x) = f(t_u|x). \]

**Example 1.2.** Doing inference for a proportion where we have \( X = x \) individuals of interests in a sample of size \( n \), we assume that \( X \sim \text{Bin}(n, \pi) \) where \( \pi \) is the unknown probability of this event. For the prior distribution we chose \( \text{Be}(\alpha, \beta) \). The posterior distribution can be approximated with

\[
f(\pi|x) \propto f(x|\pi)f(x) \propto \pi^x(1-\pi)^{n-x}\pi^{\alpha-1}(1-\pi)^{\beta-1} = \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}.
\]

This density describes another Beta-distribution with the parameter \( \text{Be}(\alpha + x, \beta + n - x) \). This allows us to use simple and known formulas to compute point and interval estimates, e.g., \( \alpha/(\alpha + \beta) \) for the mean or the \texttt{qbeta} command for the median and the credible interval in R. This also emphasizes that the posterior distribution is only a weighted version of the prior distribution. To specify this weight we can think that the \( \text{Be}(\alpha, \beta) \) distribution has been arisen from an “improper” \( \text{Be}(0, 0) \) with \( \alpha \) successes in \( \alpha + \beta = N_0 \) trials (\( \text{Be}(0 + \alpha, 0 + \alpha + \beta - \alpha) = \text{Be}(\alpha, \beta) \)). \( N_0 \) denotes the prior sample size and the relative prior sample size \( N_0/N_0 + n \) quantifies the weight of the prior. This weight decreases with increasing \( n \). It is a natural choice to take a uniform Beta-distribution (\( \alpha = \beta = 1 \)) as a prior distribution when not having any prior information. Bayes also used this distribution in his famous essay (Bayes 1763).

**Example 1.3.** Revisiting Example 1.1 under the more realistic assumption, that the prevalence is not known but estimated from a prevalence study, Bayesian updating can be used to asss this uncertainty in the estimates. We assume a realistic prior distribution \( \text{Be}(0.5, 5) \) and update this with the knowledge from the study to \( \text{Be}(0.5 + x, 5 + n - x) \). But how does this posterior distribution of the prevalence enter the estimations of \( \Pr(D+|T+) \)? It is possible to analytically compute the distribution of \( \Pr(D+|T+) \) or to generate a random sample from \( \Pr(D+|T+) \) with samples from the posterior beta distribution of the prevalence. With \( 10^5 \) samples this results in the histogram from Figure 1.1. The histogram overlaps perfectly with the true density function. This forms the basis of Monte Carlo techniques for Bayesian inference.
Figure 1.1: Posterior distribution of the positive predictive value with prevalence estimated by a prevalence study (see also Held and Sabanés Bové, 2014).
Chapter 2

Choice of priors

By: Kelly Reeve

2.1 Introduction

As we have read in the previous chapter, Bayesian analysis relies on the use of Bayes rule in order to make inferences. From Equation (1.1), we see that

\[
Pr(\text{parameters} | \text{data}) = \frac{Pr(\text{parameters} \cap \text{data})}{Pr(\text{data})} = \frac{Pr(\text{data} | \text{parameters}) \times Pr(\text{parameters})}{Pr(\text{data})}.
\]

We define a model to express qualitative knowledge (distribution form, assumptions, etc.) of the research question and this model contains one or more unknown parameters. After evaluating available information, we express our prior beliefs about which values for the parameter are likely through a probability distribution. Upon observing the data, a probability distribution for the parameter given the observed data is computed (Neal, 2004). Probability here describes not only “physical” randomness, such as measurement and labeling errors, but also the uncertainty regarding the true value of the parameter. The prior and posterior probabilities express our degree of belief (certainty) before and after observing the data.

Temporarily ignoring the normalizing constant, we see that

\[
Pr(\text{parameters} | \text{data}) \propto Pr(\text{data} | \text{parameters}) \times Pr(\text{parameters}).
\]

The posterior distribution is proportional to the likelihood multiplied by the chosen prior distribution. We can see that the prior distribution then is a means to computing the posterior distribution, or the distribution of our unknown parameter.
Table 2.1: Conjugate prior distributions for different likelihood functions (Held and Sabanés Bové, 2014).

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate prior distribution</th>
<th>Posterior distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X</td>
<td>\pi \sim \text{Bin}(n,\pi)$</td>
<td>$\pi \sim \text{Be}(\alpha,\beta)$</td>
</tr>
<tr>
<td>$X</td>
<td>\pi \sim \text{Geom}(\pi)$</td>
<td>$\pi \sim \text{Be}(\alpha,\beta)$</td>
</tr>
<tr>
<td>$X</td>
<td>\lambda \sim \text{Po}(e \times \lambda)$</td>
<td>$\lambda \sim \Gamma(\alpha,\beta)$</td>
</tr>
<tr>
<td>$X</td>
<td>\lambda \sim \text{Exp}(\lambda)$</td>
<td>$\lambda \sim \Gamma(\alpha,\beta)$</td>
</tr>
<tr>
<td>$X</td>
<td>\mu \sim \mathcal{N}(\mu, \sigma^2 \text{ known})$</td>
<td>$\mu \sim \mathcal{N}(v, \tau^2)$</td>
</tr>
<tr>
<td>$X</td>
<td>\sigma^2 \sim \mathcal{N}(\mu \text{ known}, \sigma^2)$</td>
<td>$\sigma^2 \sim \Gamma(\alpha,\beta)$</td>
</tr>
</tbody>
</table>

2.2 Choosing a prior

As stated earlier, the prior is chosen based on our knowledge of the problem. It should be neither too restrictive nor too broad. Zero probabilities should not be assigned to realistic possibilities, but spreading probability over all possibilities should also be avoided (Neal, 2004). “Improper” priors may be used, but only when they result in proper posteriors.

Priors are central to Bayesian inference, so determination of priors is a very important step. However, in practice, the available prior information may not be precise enough to lead to an exact determination of a single prior distribution. Several distributions may be consistent with the prior information and there is no unique way of choosing. Often, it is necessary to make a partly arbitrary choice of prior distribution. There may be only vague or unreliable prior information available, or a statistician may wish to provide a prior with as little subjective input as possible, in order to base the inference on the sampling model alone. It is important to note that the choice of prior can have a considerable effect on the subsequent inference (Robert, 2014). There are several common techniques for choosing a prior, but we will only discuss the conjugate prior approach and noninformative approach.

2.2.1 Conjugate Prior Distributions

A classical parametric approach, using limited subjective input, is the conjugate prior approach.

**Definition 2.1.** A family $G$ of probability distributions is conjugate with respect to the likelihood function $f(x|\theta)$ if the posterior distribution $f(\theta|x)$ is in $G$ for every $f(\theta)$ within $G$.

A practical choice of prior is to select a family of distributions so that the posterior distribution is also in this family (Held and Sabanés Bové, 2014). The set of all possible distributions, for example, is trivially conjugate with respect to any likelihood function. However, it is useful to restrict the family of distributions $G$ further based on the specific likelihood function. When $G$ is small and parameterized, going from prior to posterior is simplified to an update of the corresponding parameters, as can be seen in Table 2.1.

**Example 2.1.** Let $X$ be a binomially distributed random variable. The sample size is $n$ and the probability $\pi$ of the event is unknown, but clearly within $[0,1]$. From Table 2.1, we can see that the conjugate for this likelihood is a beta distribution with parameters $\alpha$ and $\beta$. The support of
2.2. CHOOSING A PRIOR

the beta distribution is \([0, 1]\), which is also the parameter space, making beta a sensible choice. The posterior density of \(\pi\) then is

\[
f(\pi|\pi) \propto f(x|\pi) \times f(\pi)
\]

\[
\propto nC_\pi \pi^x(1-\pi)^{n-x} \times \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}
\]

\[
\propto \pi^x(1-\pi)^{n-x} \times \pi^{\alpha-1}(1-\pi)^{\beta-1}
\]

\[
\propto \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}.
\]

As we can see, the prior parameters \(\alpha\) and \(\beta\) have been updated to \(\alpha + x\) and \(\beta + n - x\), where \(x\) is the number of successes and \(n - x\) the number of failures.

**Example 2.2.** Let us also consider a normally distributed random variable with known variance but an unknown mean. The posterior for unknown mean \(\mu\) then is

\[
f(\mu|x) \propto \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \times \exp\left(-\frac{1}{2\tau^2}(\mu - v)^2\right)
\]

\[
\propto \exp\left(-\frac{1}{2}\left[\frac{x}{\sigma^2} + \frac{v}{\tau^2}\right]\right)\]

\[
\propto \exp\left(A\mu^2 + B\mu + C\right),
\]

for some \(A < 0\), \(B\), \(C\) which depend on \(\mu^2\), \(x\), \(\tau^2\). The posterior has the form of a normal distribution, just like the prior. After further manipulation, equation 2.1 becomes

\[
f(\mu|x) \propto \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma^2} + \frac{v}{\tau^2}\right)^2\right],
\]

which yields

\[
\mu|x \sim N\left(\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \times \left(\frac{x}{\sigma^2} + \frac{v}{\tau^2}\right)\right).
\]

This statement can be simplified using inverse variances, also known as precisions. When we let \(\kappa = 1/\sigma^2\), the inverse of the variance in the likelihood, and \(\delta = 1/\tau^2\), the variance of the prior, then the posterior for \(\mu\) simplifies to

\[
\mu|x \sim N\left((\kappa + \delta)^{-1} \times (\kappa x + \delta v), (\kappa + \delta)^{-1}\right).
\]

For simplicity, precision parameters are often used in Bayesian analysis.

It is worth noting that it is sufficient to determine conjugacy for a single member \(X_i\) of a random sample \(X_1:n\) (Held and Sabanés Bové, 2014). If the prior is conjugate to the likelihood function of \(X_i\) then it will result in a posterior distribution of the same family. This posterior will then serve as the prior for the next observation and again result in a posterior of the same family. This process can be continued for all \(X_i\), with change only in the parameters but not the family of the distribution.

The main advantage of using conjugate priors is that the resulting posterior distributions are always computable to a certain extent. They are simple and computationally efficient. The conjugate prior can also be used as the starting point from which to build up prior distributions based on limited prior input. The imprecision can be assessed through additional prior distributions.
It is important to remember that a conjugate prior is not considered noninformative; it influences the resulting inference somewhat. Also, choosing a conjugate prior may force one to ignore part of the prior information if it is not compatible with the prior format. There are other types of priors that allow for limited input but have a more limited influence on the results (Robert, 2014).

2.2.2 Noninformative Prior Distributions

When no prior information is available, conjugate priors are still useful in that they may lead to closed-form expressions for posteriors; however, the choice cannot be justified. In this setting, the only information available is the sample, and so the prior distribution must be derived from the sample distribution. The first uses of the noninformative approach were by Bayes and Laplace (Gelman et al., 2004). When presented problems with no prior information, they decided to use a uniform prior, which would give every value of the parameter the same likelihood.

Other times we wish to minimize the influence of the prior and so choose a “vague” prior with large variance (Held and Sabanés Bové, 2014). Such vague distributions may be “improper.”

**Definition 2.2.** A prior distribution with a density function \( f(\theta) \geq 0 \) is improper when

\[
\int_{\Theta} f(\theta)d\theta = \infty \text{ or } \sum_{\theta \in \Theta} f(\theta) = \infty
\]

for continuous or discrete parameters \( \theta \).

The probability density does not integrate to 1 and is therefore improper. These priors may be used in Bayesian analysis so long as the posterior distribution is proper; however, not all improper priors result in proper posterior distributions.

**Example 2.3.** (Held and Sabanés Bové, 2014) Remember Example 2.1. The binomial model has the conjugate prior \( \pi \sim Be(\alpha, \beta) \), whose density is proper for \( \alpha > 0 \) and \( \beta > 0 \). The limiting case, therefore, is \( \alpha = \beta = 0 \), which is the improper prior distributions known as Haldane’s prior. Using Haldane’s prior as the prior on the binomial’s unknown propability \( \pi \), we see

\[
f(\pi|x) \propto \pi^x (1 - \pi)^{n-x} \times \pi^\alpha (1 - \pi)^{\beta-1} \\
\propto \pi^x (1 - \pi)^{n-x} \times \pi^{-1} (1 - \pi)^{-1} \\
\propto \pi^{x-1} (1 - \pi)^{n-x-1}.
\]

The posterior distribution then is \( Be(x, n-x) \), which is proper for \( x > 0 \) and \( n - x > 0 \).

Some statisticians object to the use of improper priors, however, they can be used as long as they are not interpreted as probability distributions (Robert, 2014). This was a criticism Laplace faced when introducing the use of the uniform prior. A more important criticism of Laplace’s choice deals with invariance under reparameterization. A one-to-one transformation from \( \theta \) to \( \eta = g(\theta) \) seems to lack prior information, but after the transformation we see that the distribution is often no longer constant.
Example 2.4. Let $\eta = g(\theta)$ be a one-to-one differentiable transformation of $\theta$ with a uniform prior density $f_\theta(\theta) \propto 1$. We may obtain the corresponding prior for $\eta$ by employing the change of variables formula, as follows:

$$f_\eta(\eta) = f_\theta[g^{-1}(\eta)] \times \left| \frac{dg^{-1}}{d\eta} \right| \propto (1) \times \left| \frac{dg^{-1}}{d\eta} \right|.$$ 

If $g$ is not linear, $f_\eta(\eta)$ will depend on $\eta$ and will not be uniform. A uniform distribution for one parameterization will only yield another uniform distribution if the transformation is linear. The lack of invariance of the uniform prior can lead to significant variation in the resulting posteriors (Datta and Ghosh, 1996).

Example 2.5. Revisit $X \sim Bin(n, \theta)$. We know $\theta \in [0,1]$. A uniform prior on $\theta$ is then $\pi(\theta) = 1$. Let $\rho$ be a transformation of $\theta$ such that $\rho = \log \theta / (1 - \theta)$, which moves the parameter to the full real line. Knowing $\rho > -1$ and its derivative,

$$\rho^{-1} = \frac{e^\rho}{1 + e^\rho} \cdot \frac{d\rho^{-1}}{d\rho} = \frac{e^\rho}{(1 + e^\rho)^2},$$

we can use the change-of-variables formula to find the prior on $\rho$,

$$\left| \frac{e^\rho}{(1 + e^\rho)^2} \right|.$$ 

While the uniform prior was uninformative for $\theta$, under a simple reparameterization of $\theta$ this is no longer the case. The prior is no longer flat; it has become informative and will influence inference.

2.2.3 Jeffreys’ Prior Distributions

In the previous example, the prior on $\theta$ was appealing because it was flat under that specific parameterization. An improvement, however, would be to create an uninformative prior that also remains uninformative, that is, a prior that is invariant under reparameterization. Such a prior is called Jeffreys’ prior.

Definition 2.3. Also known as Jeffreys’ rule, it is formally defined as

$$f(\theta) \propto \sqrt{J(\theta)},$$

where $J(\theta)$ is the expected Fisher information of the unknown scalar parameter $\theta$.

As a reminder, Fisher information is the negative second derivative of the log-likelihood function, written as

$$I(\theta) = -\frac{d^2 l_\theta}{d\theta^2} = -\frac{dS\theta}{d\theta^2},$$

where $S(\theta)$ is the score function, or first derivative of the log-likelihood function. Here we are interested in the expected value of the Fisher information, which can be seen as a measure of the average information in the data with respect to the unknown parameter (see chapters 2, 4, 6, Held and Sabanés Bové, 2014).
CHAPTER 2. CHOICE OF PRIORS

Figure 2.1: Diagram illustrating the invariance of Jeffreys’ prior (Held and Sabanés Bové, 2014).

Example 2.6 (Invariance of Jeffrey’s prior). The invariance of Jeffreys’ prior under a one-to-one transformation reparamaterization of θ means that if we have \( f_θ(θ) \propto \sqrt{J_θ(θ)} \), then the density function of \( η = g(θ) \) is \( f_η(η) \propto \sqrt{J_η(η)} \). Given \( f_θ(θ) \propto \sqrt{J_θ(θ)} \), we start with the change-of-variables formula

\[
f_η(η) = f_θ(θ) \times \left| \frac{dg^{-1}(η)}{dη} \right| \propto \sqrt{J_θ(θ)} \times \left| \frac{dg^{-1}(η)}{dη} \right| \propto \sqrt{J_η(η)} \times \left| \frac{dg^{-1}(η)}{dη} \right|^2 \propto \sqrt{J_η(η)}
\]

and arrive at \( f_η(η) \propto \sqrt{J_η(η)} \), the result stated earlier. In this example, we first applied the change-of-variables formula and then apply Jeffreys’ rule, but the order of these operations does not matter. Both orders will yield the same result, as illustrated in Figure 2.1.

The expected value of the Fisher information can help us decide on a prior. If this value is dependent on \( θ \), this value will change for different values of \( θ \). The variance-stabilizing transformation \( η = g(θ) \) should then be used to remove this dependence. A locally uniform prior for \( η \) can now be used. However, if the expected value of the Fisher information is independent of \( θ \), a uniform prior may be used directly.

Example 2.7. (Held and Sabanés Bové, 2014) Jeffreys’ prior for the unknown parameter \( θ \) comes from the derivative of the variance-stabilizing transformation \( g(λ) \). Poisson distribution data and the variance-stabilizing transformation \( g(λ) = \sqrt{λ} \) for mean \( λ \) is given. By taking the derivative of \( g(λ) \)

\[
\frac{d\sqrt{λ}}{dλ} = \frac{1}{2} λ^{-1/2} \propto λ^{-1/2}
\]

we obtain \( λ^{-1/2} \) as Jeffreys’ prior for \( f_λ(λ) \). This result, along with other common likelihood-Jeffreys’ prior pairs, is listed in Table 2.2. This resulting prior, like most in Table 2.2 (except for the binomial likelihood), is improper. With a quick comparison of Table 2.1 and Table 2.2, we see that Jeffreys’ priors can be looked at as the limiting cases of the corresponding conjugate priors (Held and Sabanés Bové, 2014).
Table 2.2: Jeffreys’ prior for several likelihood functions (Held and Sabanés Bové, 2014).

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Jeffreys’ prior</th>
<th>Density of Jeffreys’ prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin$(n, \pi)$</td>
<td>$\pi \sim \text{Be}(\frac{1}{2}, \frac{1}{2})$</td>
<td>$f(\pi) \propto [\pi(1-\pi)]^{-\frac{1}{2}}$</td>
</tr>
<tr>
<td>Geom$(\pi)$</td>
<td>$\pi \sim \text{Be}(0, \frac{1}{2})$</td>
<td>$f(\pi) \propto \pi^{-1}(1-\pi)^{-\frac{1}{2}}$</td>
</tr>
<tr>
<td>Po$(\lambda)$</td>
<td>$\lambda \sim \text{Be}(\frac{1}{2}, 1)$</td>
<td>$f(\lambda) \propto \lambda^{-1}$</td>
</tr>
<tr>
<td>Exp$(\lambda)$</td>
<td>$\lambda \sim \Gamma(0, 0)$</td>
<td>$f(\lambda) \propto \lambda$</td>
</tr>
<tr>
<td>$\mathcal{N}(\mu, \sigma^2 \text{ known})$</td>
<td>$\mu \sim \mathcal{N}(0, \infty)$</td>
<td>$f(\mu) \propto 1$</td>
</tr>
<tr>
<td>$\mathcal{N}(\mu \text{ known}, \sigma^2)$</td>
<td>$\sigma^2 \sim \Gamma(0, 0)$</td>
<td>$f(\sigma^2) \propto \sigma^{-2}$</td>
</tr>
</tbody>
</table>

Example 2.8. Let $X_{1:n}$ denote a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution with unknown mean $\mu$ and known variance $\sigma^2$. The expected Fisher information has been calculated to be $J(\mu) = n/\sigma^2$, which does not depend on $\mu$. This means we may use a uniform prior directly (instead of first applying a variance-stabilizing transformation). The Jeffreys’ prior is $f(\mu) \propto \sqrt{n/\sigma^2} \propto 1$, an improper locally uniform prior on the whole real line. Using instead a conjugate prior approach, we see from Table 2.1 that the prior distribution $\mu \sim \mathcal{N}(\nu, \tau^2)$ is appropriate. If we allow $\tau^2$ to approach $\infty$, this prior distribution becomes flat, the Jeffreys’ uniform prior.

Although conjugate priors and uninformative priors have been discussed as separate methods, Jeffreys’ prior can be seen as a link between the two. They do then share common ground.

2.3 Comparing MLE and posterior mean

Although Bayesian and frequentist analysis methods employ different methods, they often obtain similar, if not identical, results. This is especially true for large samples. Table 2.3 compares maximum likelihood estimates to the posterior mean obtained using a Jeffreys’ prior for several different distributions.

Example 2.9. Let us first show how to arrive the MLE of a binomial distribution using the frequentist method and then posterior mean (Table 2.3, first line) using a Bayesian method. Let $X \sim \text{Bin}(n, \pi)$. Using MLE we see

$$L(\pi) = (nC_x)\pi^x(1-\pi)^{n-x}$$
$$l(\pi) = \log(nC_x) + x \log \pi + (n-x) \log(1-\pi)$$
$$S(\pi) = \frac{dl}{d\pi} = \frac{x}{\pi} - \frac{(n-x)}{1-\pi}.$$  

Solving for $S(\pi) = 0$ we obtain $\pi = x/n \propto \bar{x}$. Using the Bayesian method we see

$$f(\pi|x) \propto (nC_x)\pi^x(1-\pi)^{n-x} \times \frac{1}{B(\alpha, \beta)}\pi^{(1/2)-1}(1-\pi)^{(1/2)-1}$$
$$\propto \pi^x(1-\pi)^{n-x} \times \pi^{-1/2}(1-\pi)^{-1/2}$$
$$\propto \pi^{x-1/2}(1-\pi)^{n-x-1/2}.$$. 


Table 2.3: Comparison of MLEs and the posterior means using Jeffreys’ prior (Held and Sabanés Bové, 2014).

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>$\theta_{ML}$</th>
<th>Posterior mean using Jeffreys’ prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin($n, \pi$)</td>
<td>$\bar{x}$</td>
<td>$\frac{n}{n+1} (\bar{x} + \frac{1}{2n})$</td>
</tr>
<tr>
<td>Geom($\pi$)</td>
<td>$\frac{1}{\bar{x}}$</td>
<td>$\frac{1}{(\bar{x} + \frac{1}{2n})}$</td>
</tr>
<tr>
<td>Po($\lambda$)</td>
<td>$\bar{x}$</td>
<td>$\bar{x} + \frac{1}{2\lambda}$</td>
</tr>
<tr>
<td>Exp($\lambda$)</td>
<td>$\frac{1}{\bar{x}}$</td>
<td>$\frac{1}{\bar{x}}$</td>
</tr>
<tr>
<td>$N(\mu, \sigma^2 \text{ known})$</td>
<td>$\bar{x}$</td>
<td>$\bar{x}$</td>
</tr>
<tr>
<td>$N(\mu \text{ known}, \sigma^2)$</td>
<td>$\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$</td>
<td>$\frac{1}{n-2} \sum_{i=1}^{n} (x_i - \mu)^2$</td>
</tr>
</tbody>
</table>

We obtain $\pi|x \sim Be(x + (1/2), n - x + (1/2))$. The expected value of a beta distribution is $\alpha/\alpha + \beta$, so here we have mean $x + (1/2)/n + 1$. Upon substituting $x/n$ for $\bar{x}$, we see these two estimates are, in this case, virtually equivalent.
Chapter 3

Loss functions and Bayes estimates

By:
Marco Schleiniger

3.1 What we know so far

It is now the time that we start to think about how we can make inference in Bayesian statistics. From the previous chapters, it is known that the parameter of interest $\theta$ is a random variable with an appropriate prior density $f(\theta)$, which is the main difference to the frequentist inference in which only the data $X$ is treated as random. Our primary interest, the posterior density $f(\theta|x)$, can be obtained after observing the data $X$. All statistical inference about $\theta$ is based on the posterior distribution. We will have a look at three different point estimates.

The first one is the expectation of the posterior distribution which is also called the posterior mean:

$$E(\theta|x) = \int \theta f(\theta|x)d\theta.$$ 

We can also define the mode of the posterior distribution or posterior mode:

$$\text{Mod}(\theta|x) = \arg \max_{\theta} f(\theta|x).$$

Lastly, also the median of the posterior distribution should be defined. The posterior median is any number $a$ which satisfies

$$\int_{-\infty}^{a} f(\theta|x)d\theta = 0.5 \quad \text{and} \quad \int_{a}^{\infty} f(\theta|x)d\theta = 0.5.$$
3.2 Loss functions

We now know three different possibilities to estimate the unknown parameter $\theta$. For that reason, the question arises which point estimate should be used in a specific application. To answer this question, it would make sense to have some kind of mathematical tool to measure how far the point estimate $\hat{\theta}$ is away from the true parameter value $\theta$. As luck would have it, it turns out that loss functions do exactly that. Three different loss functions will be introduced here. Note that from now on, the parameter estimate $\hat{\theta}$ is going to be denoted as $\hat{a}$ to clarify the difference to the true parameter $\theta$.

By definition, the loss function $l(a, \theta) \in \mathbb{R}$ quantifies the loss encountered when estimating the true parameter $\theta$ by $a$. As it makes the most sense, it is common to set the loss function to zero if $a = \theta$, so $l(\theta, \theta) = 0$.

An often used loss function is the quadratic loss function

$$l(a, \theta) = (a - \theta)^2.$$ 

One can also use a linear loss function. Note that, in this case, the absolute value has to be used, i.e.,

$$l(a, \theta) = |a - \theta|.$$ 

The third possibility for a loss function is the zero–one loss function which is most easily denoted with the indicator function $I$:

$$l(a, \theta) = I(|a - \theta| > \epsilon).$$

Note that in this particular loss function we have an additional parameter $\epsilon > 0$ which has to be chosen carefully.

3.3 Bayes estimate

Now that we know the concept of loss functions, it seems obvious that the loss function of a well estimated parameter $a$ should be as small as possible. The Bayes Estimate does this by minimizing the expectation of the loss function with respect to the posterior distribution. More precisely, the Bayes estimate minimizes:

$$\mathbb{E}[l(a, \theta)|x] = \int_{\Theta} l(a, \theta)f(\theta|x)d\theta.$$ 

Maybe the most interesting result of this topic is that the Bayes estimate with respect to a specific loss function is a unique point estimate that has been covered in Section 3.1. To be more concrete, the Bayes estimate with respect to the quadratic loss is the posterior mean, whereas the Bayes estimate with respect to linear loss is the posterior median. Finally, the posterior mode is the Bayes estimate respecting the zero–one loss function, as $\epsilon \to 0$. Since this result is so crucial, we are going to prove it for the three different point estimates.

Firstly, let us minimize the quadratic loss function

$$\mathbb{E}[l(a, \theta)|x] = \int_{\Theta} (a - \theta)^2f(\theta|x)d\theta.$$
3.4. LOSS FUNCTIONS FOR CREDIBLE REGIONS

Taking the derivative with respect to $a$ and setting it to zero yields

$$2 \times \int (a - \theta) f(\theta | x) d\theta = 0$$

$$\Leftrightarrow \int a f(\theta | x) d\theta = \int \theta f(\theta | x) d\theta$$

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

Since we know that $\int f(\theta | x) d\theta = 1$, we get exactly the posterior mean

$$a = \int \theta f(\theta | x) d\theta = \mathbb{E}[\theta | x].$$

Let us move on to the expected linear loss

$$\mathbb{E}[l(a, \theta)] = \int |a - \theta| f(\theta | x) d\theta$$

$$= \int_{-\infty}^{a} (a - \theta) f(\theta | x) d\theta + \int_{a}^{\infty} (\theta - a) f(\theta | x) d\theta.$$

Using Leibnitz’s integral rule, we get for the derivative with respect to $a$

$$\int_{-\infty}^{a} f(\theta | x) d\theta - (a - (\infty)) f(-\infty | x) \times 0 + (a - a) \times f(a | x) \times 1$$

$$- \int_{a}^{\infty} f(\theta | x) d\theta - (a - a) \times f(a | x) \times 1 + (\infty - a) \times f(\infty | x) \times 0$$

$$= \int_{-\infty}^{a} f(\theta | x) d\theta - \int_{a}^{\infty} f(\theta | x) d\theta.$$

Again, this result has to be set to zero and the fact that $\int_{-\infty}^{\infty} f(\theta | x) d\theta = 1$ leads to

$$\int_{-\infty}^{a} f(\theta | x) d\theta = \int_{a}^{\infty} f(\theta | x) d\theta = 0.5.$$

So by definition, $a$ is the posterior median $\text{Med}(\theta | x)$. What’s left is the expected zero–one loss

$$\mathbb{E}[l(a, \theta)] = \int_{-\infty}^{a-\epsilon} f(\theta | x) d\theta + \int_{a+\epsilon}^{\infty} f(\theta | x) d\theta$$

$$= 1 - \int_{a-\epsilon}^{a+\epsilon} f(\theta | x) d\theta.$$

We minimize the expectation by maximizing $\int_{a-\epsilon}^{a+\epsilon} f(\theta | x) d\theta$ and since $\epsilon \to 0$, the integral can be approximated by $2\epsilon f(a | x)$. We can easily see now that this is maximized through the posterior mode $a = \text{Mod}(\theta | x)$.

3.4 Loss functions for credible regions

It can be expected that there are not only loss functions for point estimates, but also for interval estimates like the credible region which is a generalization of the credible interval. A $\gamma \cdot 100\%$ credible region is defined by a subset $C \subseteq \Theta$ with

$$\int_{C} f(\theta | x) d\theta = \gamma.$$
The only difference to the credible interval is that $C$ does not need to be a real interval in the credible region, so it can also be applied in multivariate cases. Clearly, one can find a lot of subsets $C$ which meet the criterion of a credible region. So what makes a “good” credible region? Firstly, we want $C$ to be small because then it will give us more precise information and secondly, the region should contain the true parameter $\theta$.

From those criteria, it makes sense to construct the following loss function:

$$l(C, \theta) = |C| - I_C(\theta).$$

Note that $C \in A$, whereas $A$ denotes the set of all $\gamma \cdot 100\%$ credible regions for $\theta$. It turns out that $C$ is optimal with respect to this particular loss function $l(C, \theta)$, if for all $\theta_1 \in C$ and $\theta_2 \notin C$,

$$f(\theta_1|x) \geq f(\theta_2|x).$$

One may remember from a previous chapter that in this case, $C$ is exactly a highest posterior density (HPD) region. So all parameter values which are contained in $C$ have a higher posterior density than those which are not contained in $C$.

To illustrate this result, let us consider a Beta Be(10, 5) distribution. Remember that in Bayesian statistics, the Beta distribution is a fairly common prior distribution since it has the parameter space $(0, 1)$ which can, for example, be used for the unknown parameter $\pi$ of a binomial distribution. In that case, also the posterior will be a Beta distribution. We will have a look at three different $95\%$ credible regions. One of them is the equi-tailed region for which we cut off the same amount of probability mass on both ends of the distribution (in our case 2.5%). Another one is a credible region which is shifted to the right compared to the equi-tailed region. Lastly, also the HPD region of that particular distribution has been constructed. One can see the limits of the respective credible regions in Figure 3.1 and the corresponding R-code below.

```r
library(pscl)
library(ggplot2)
x <- seq(0, 1, by = 0.001)
y <- dbeta(x, 10, 5)
## equitailed region
cred1 <- c(qbeta(0.025, 10, 5), qbeta(.975, 10, 5))
## right-shifted region
cred2 <- c(qbeta(0.04, 10, 5), qbeta(.99, 10, 5))
## HPD region
cred3 <- betaHPD(10, 5)
qplot(x, y, geom = "line", xlab = "theta", ylab = "posterior") +
geom_vline(xintercept = cred1, col = "blue") +
geom_vline(xintercept = cred2, col = "darkgreen") +
geom_vline(xintercept = cred3, col = "red") +
geom_hline(yintercept = y[1000*cred3[1]], linetype = "dashed")
```
3.4. LOSS FUNCTIONS FOR CREDIBLE REGIONS

For the sake of simplicity, let us assume that all of those three credible regions do contain the true parameter \( \theta \). So \( I_C(\theta) \) in the loss function (Equation (3.1)) takes the value 1. In order to find the best of the three credible regions respecting the loss function, we now only need to find the region with the smallest size \( |C| \).

\[
\begin{align*}
\text{cred1[2]} - \text{cred1[1]} & \quad \text{## size of equi-tailed region} \\
& \quad \text{## [1] 0.4534368} \\
\text{cred2[2]} - \text{cred2[1]} & \quad \text{## size of right-shifted region} \\
& \quad \text{## [1] 0.4519462} \\
\text{cred3[2]} - \text{cred3[1]} & \quad \text{## size of HPD region} \\
& \quad \text{## [1] 0.4493598}
\end{align*}
\]

As one can see, the HPD credible region has indeed the smallest size. It is thus, the most optimal of the three regions with respect to the loss function, which is in agreement with the result that we have seen in Equation (3.2).

If we look at the point and interval estimate in combination, according to the loss function in Equation (3.1), we want the point estimate to lie within the interval estimate. It can easily be seen that the posterior mode always lies within any HPD interval and the posterior median

![Figure 3.1: Be(5, 10) distribution with the limits of an equi-tailed credible region (blue), an HPD region (red), and a right-shifted region (green). Note that the limits of the HPD region intersect the posterior density at the same height (dashed line).](image)
is within an equi-tailed credible interval under all circumstances. However, for the posterior mean, no such statement can be made, as it does not need to lie within an HPD interval or an equi-tailed credible interval.

Lastly, one can also have a look at how the point and interval estimates behave under a one-to-one transformation $\phi = h(\theta)$ of the parameter $\theta$. It can be stated that only the characteristics which are based on quantiles of the posterior distribution are invariant under such transformations. Therefore, we can assume invariance for the posterior median as well as for the equi-tailed credible intervals. Since estimates like the posterior mode or the posterior mean are not based on quantiles of the posterior distribution, we can generally not assume them to be invariant under one-to-one transformations.
Chapter 4

Bayesian inference in multiparametric models

By: Ke Li

4.1 Introduction

The principal advantages of the Bayesian approach lie in dealing with more than one unknown or unobservable quantity (Gelman et al., 2014). To obtain those unknown information, it should be done first that to obtain the joint posterior distributions of all unknown. The route to achieve this is straightforward: multiplication of the likelihood functions and the prior distributions (Held and Sabanés Bové, 2014). Next, the ultimate aim of bayesian analysis is to derive the marginal posterior distributions of the parameters of interest, namely elimination of nuisance parameters. In many practical problems there is no need to make inferences about many unknown parameters, even though they are necessary for the construction of a model. This kind of parameters are called nuisance parameters (Gelman et al., 2014). We begin this chapter with choice of prior distributions for multiparametric models. In this section, we will exemplify Conjugate prior distributions, Jeffreys’ and Reference prior distributions for different multiparametric models. Then we will introduce the general treatment of nuisance parameters.

4.2 Conjugate prior distributions

Conjugate prior distributions are available for multivariate likelihood functions, as illustrated in the following examples.
Example 4.1 (Normal model). Consider a \( k \)-dimensional multivariate normal distributed data \( y = (y_1, \ldots, y_k)^T \sim \mathcal{N}_k(\mu, \Sigma) \) with a \( k \)-dimensional mean vector \( \mu = (E[y_1], \ldots, E[y_k])^T \), and a \( k \times k \) covariance matrix \( \Sigma = [\text{Cov}(y_i, y_j)], i = 1, 2, \ldots, k; j = 1, 2, \ldots, k \). The density function is

\[
f(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp \left( -\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu) \right),
\]

where \( \mu \) is a \( k \)-dimensional column vector and \( |\Sigma| \) is the determinant of \( \Sigma \) (Gelman et al., 2014). Therefore, the likelihood function for an observation is

\[
f(y|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp \left( -\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu) \right),
\]

and for \( n \) independent observations, \( y_1, \ldots, y_n \), is

\[
f(y_1, \ldots, y_n|\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1}(y_i - \mu) \right) = |\Sigma|^{-n/2} \exp \left( -\frac{1}{2} \text{tr}(\Sigma^{-1} S_0) \right),
\]

where \( S_0 \) is the matrix of “sum of square” relative to \( \mu \),

\[
S_0 = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T.
\]

The conjugate prior distribution for \( \mu \) with known \( \Sigma \) is parameterized as \( \mu \sim \mathcal{N}_k(\mu_0, \Lambda_0) \). Then, the probability density function can be expressed as:

\[
f(\mu|\mu_0, \Lambda_0) = \frac{1}{\sqrt{(2\pi)^k|\Lambda_0|}} \exp \left( -\frac{1}{2}(\mu - \mu_0)^T \Lambda_0^{-1}(\mu - \mu_0) \right).
\]

So posterior distribution for \( \mu \) with known \( \Sigma \) is

\[
f(\mu|y, \Sigma) \propto \exp \left( -\frac{1}{2} (\mu - \mu_0)^T \Lambda_0(\mu - \mu_0) + \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1}(y_i - \mu) \right)
\]

\[
= \exp \left( -\frac{1}{2} (\mu - \mu_n)^T \Lambda_n(\mu - \mu_n) \right),
\]

where

\[
\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}),
\]

\[
\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}.
\]

Therefore, the posterior distribution \( f(\mu|y, \Sigma) \sim \mathcal{N}_k(\mu|\mu_n, \Lambda_n) \).

Example 4.2 (Multinomial model, Held and Sabanés Bové, 2014). We consider a \( k \)-dimensional multinomially distributed random variable \( X \sim \text{M}_k(n, \pi) \) with the unknown probability vector \( \pi = (\pi_1, \ldots, \pi_k)^T \). The density function is

\[
f(x_1, \ldots, x_k; n, \pi_1, \ldots, \pi_k) = Pr(X_1 = x_1, \ldots, X_k = x_k)
\]

\[
= \begin{cases} 
\frac{n!}{x_1! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n, \\
0, & \text{otherwise},
\end{cases}
\]
for non-negative integers \( x_1 \ldots x_k \). Therefore, if \( y \) is the vector of counts of the number of each observation of each outcome, then

\[
f(y|\pi) \propto \prod_{i=1}^{k} \pi_i^{y_i}, \tag{4.1}
\]

where \( \sum_{i=1}^{k} \pi_i = 1 \) and \( \sum_{i=1}^{k} y_i = n \). The conjugate prior distribution for the multinomial distribution is the Dirichlet distribution \( \pi \sim D_k(\alpha_1, \ldots, \alpha_k) \)

\[
f(\pi_1, \ldots, \pi_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_k)}{\prod_{i=1}^{k} \Gamma(\alpha_k)} \prod_{i=1}^{k} \pi_i^{\alpha_k-1},
\]

where \( \pi_k > 0 \) and \( \sum_{k=1}^{k} \pi_k = 1 \), additionally, \( \alpha_k \) represents the actual counts of value \( k \) (Minka, 2003). Posterior distribution for multinomial model: Given the Dirichlet prior, the posterior distribution will be

\[
f(\pi|y) \propto \prod_{i=1}^{k} \Gamma(\alpha_k) \prod_{i=1}^{k} \pi_i^{y_i+\alpha_k-1}.
\]

Therefore, the posterior distribution \( f(\pi|y) \sim D_k(y_i + \alpha_k) \) is also a Dirichlet distribution.

### 4.3 Jeffreys’ and reference prior distributions

Consider there are multiple parameters of interests, say a parameter vector \( \theta = (\theta_1, \ldots, \theta_k)^\top \). Let \( l(\theta) = \log f(x|\theta) \) denote the log-likelihood function, the second order derivative of \( l(\theta) \) with respect to \( \theta \) is a \( k \times k \) matrix (Held and Sabanés Bové, 2014). Therefore, the Jeffreys’ prior in the multi-parametric case is defined as:

\[
f(\theta) \propto \sqrt{|J(\theta)|},
\]

where \( |J(\theta)| \) denotes the determinant of the expected Fisher information matrix. Here, the Fisher information matrix is defined as

\[
|J(\theta)| = -E \left( \frac{\partial^2 l(\theta)}{\partial \theta^2} \right) = -E \left[ \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right]_{i=1,\ldots,k; j=1,\ldots,k}.
\]

**Example 4.3** (Multinomial model, Held and Sabanés Bové, 2014). Consider a multinomial distributed variable \( X \sim M_k(n, \pi) \) with unknown probability vector \( \pi = (\pi_1, \ldots, \pi_k)^\top \). The log-likelihood function in multinomial model is defined as

\[
l(\pi) = \sum_{i=1}^{k} x_i \log(\pi_i).
\]

As known, \( \pi_k = 1 - \sum_{i=1}^{k-1} \pi_i \), thus for the trimmed parameter vector \( \tilde{\pi} \) the log-likelihood function can also be written as the form

\[
l(\tilde{\pi}) = \sum_{i=1}^{k-1} x_i \log(\pi_i) + x_k \log \left( 1 - \sum_{i=1}^{k-1} \pi_i \right).
\]
As a result, the score equation will be
\[
S(\tilde{\pi}) = \frac{\partial}{\partial \tilde{\pi}} l(\tilde{\pi}) = \left( \frac{x_1}{\pi_1} - \frac{x_k}{1 - \sum_{i=1}^{k-1} \pi_i}, \ldots, \frac{x_{k-1}}{\pi_{k-1}} - \frac{x_k}{1 - \sum_{i=1}^{k-1} \pi_i} \right)^T.
\]
As the equation shows, the Fisher information matrix has dimension \((k - 1) \times (k - 1)\)
\[
J(\tilde{\pi}) = -\frac{\partial}{\partial \tilde{\pi}} S(\tilde{\pi}) = \text{diag} \left\{ \frac{x_i}{\pi_i} \right\}_{i=1}^{k-1} + \frac{x_k}{(1 - \sum_{i=1}^{k-1} \pi_i)^2} 11^T.
\]
If we replace \(\tilde{\pi}\) by \(\hat{\pi}_{ML}\), then we obtain the observed Fisher information matrix
\[
J(\hat{\pi}_{ML}) = n \left\{ \text{diag}(\hat{\pi}_{ML})^{-1} + \left(1 - \sum_{i=1}^{k-1} \pi_i \right)^{-1} 11^T \right\}.
\]
Applying the Sherman–Morrison formula (Held and Sabanés Bové, 2014), we can compute the
inverse of the expected Fisher information, with the determinant
\[
|J(\theta)| = \frac{n}{\prod_{i=1}^{k} \pi_i}.
\]
Therefore, the Jeffreys’ prior is
\[
f(\pi) \propto \sqrt{\frac{1}{\prod_{i=1}^{k} \pi_i}} = \prod_{i=1}^{k} \pi_i^{-\frac{1}{2}}.
\]
This is the kernel of a Dirichlet distribution \(\pi \sim D_k(1/2, \ldots, 1/2)\). In binomials, this prior is proper. After multiplying it with the likelihood function expressed as equation (4.1), we get the posterior distribution
\[
f(\pi|y) \sim D_k(1/2 + y_1, \ldots, 1/2 + y_k).
\]
However, application of Jeffreys’ prior in multi-parameter cases is controversial. Even though it
fits the invariance property in this example, it exists one particular problem which needs to be
discussed.

**Example 4.4** (Normal model, Held and Sabanés Bové, 2014). Consider a random sample \(X_{1:n}\) with parameter vector \(\theta = (\mu, \sigma^2)^T\) normally distributed \(X_{1:n} \sim N(\mu, \sigma^2)\).

First, we look at the probability density function of normal distribution
\[
f(x) = (2\pi\sigma^2)^{-1/2} \exp \left( -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right).
\]
Therefore, the log-likelihood function is
\[
l(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.
\]
Based on this equation, we can derive the partial derivative of the log-likelihood are
\[
\frac{\partial l(\theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu),
\]
\[
\frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2.
\]
So the score vector is
\[ S(\theta) = \frac{1}{\sigma^2} \left( \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \right). \]

If we set the score functions to 0, we can compute the MLEs
\[ \hat{\mu}_{ML} = \bar{x}, \]
\[ \hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]

As a consequence, the Fisher information matrix \( J(\theta) \) turns out to be
\[ J(\theta) = -\left( \begin{array}{cc} \frac{\partial^2 l(\theta)}{\partial \mu^2} & \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\theta)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\theta)}{\partial \sigma^4} \end{array} \right) = \frac{1}{\sigma^2} \left( \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \right) \left( \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2\sigma^2} \right). \] (4.2)

According to the equation (4.2), we replace \( \mu \) and \( \sigma^2 \) with \( \hat{\mu}_{ML} \) and \( \hat{\sigma}_{ML}^2 \), then the Fisher information matrix turns out to be
\[ J(\hat{\theta}_{ML}) = \left( \begin{array}{cc} \frac{n}{\hat{\sigma}_{ML}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}_{ML}^4} \end{array} \right). \]

Thus Jeffreys’ prior is:
\[ f(\theta) \propto \sqrt{|J(\theta)|} = \sqrt{\frac{n^2}{2\sigma^6}} \propto \sigma^{-3}. \]

Therefore, this Jeffreys’ prior is proportional to \((\sigma^2)^{-3/2}\) for \(\sigma^2\) and independent with \(\mu\). This contradicts the Jeffreys’ prior if \(\mu\) is known. In the latter case, the Jeffreys’ prior is proportional to \((\sigma^2)^{-1}\).

Alternative approach is reference prior distribution. This minimises the impact of the prior distribution on the posterior distribution. Further speaking, it is different to Jeffreys’ prior for treatment of parameters. In this approach, the parameter vector \(\theta\) is separated into a parameter of interest and a nuisance parameter.

**Example 4.5** (Normal model, Held and Sabanés Bové, 2014). A better approach to the last example is reference prior
\[ f(\theta) \propto \sigma^{-2} \]
for \(\theta = (\mu, \sigma^2)^T\). This can be derived by multiplication of Jeffreys’ prior for \(\mu\) in \(\sigma^2\) known case with Jeffreys’ prior for \(\sigma^2\) in \(\mu\) known case. In this case the reference prior remains identical, no matter which parameter \(\mu\) or \(\sigma^2\) is treated as parameter of interest.

### 4.4 Elimination of nuisance parameters

The route to eliminate the nuisance parameters is straightforward. Suppose a parameter vector \((\theta, \eta)^T\). Also, assume that \(\theta\) is the parameter of interest, and that \(\eta\) is a nuisance parameter. First, we should compute the joint posterior density of \(\theta\) and \(\eta\). Then we should integrate the
CHAPTER 4. BAYESIAN INFERENCE IN MULTIPARAMETRIC MODELS

Figure 4.1: Marginal distributions of a joint normal-gamma distribution with fixed parameters $\nu = 0$, $\lambda = 0.5$, and $\alpha = 2$, $\beta = 1.2$. (Figure taken from Held and Sabanés Bové, 2014).

joint posterior density with respect to $\eta$. This gives us the marginal posterior density of the parameter of interest. Let

$$f(\theta, \eta|x) = \frac{f(x|\theta, \eta) \cdot f(\theta, \eta)}{f(x)}$$

denote the joint posterior density function. The marginal posterior of $\theta$ is then

$$f(\theta|x) = \int f(\theta, \eta|x) d\eta.$$ 

Example 4.6 (Normal-Gamma distribution, Held and Sabanés Bové, 2014). Consider $\theta = (\mu, \kappa)^T \sim \text{NG}(\nu, \lambda, \alpha, \beta)$ follow a normal-gamma distribution. It is defined through factorization

$$f(\theta) = f(\kappa) \cdot f(\mu|\kappa),$$

where $\kappa \sim G(\alpha, \beta)$ and $\mu|\kappa \sim \mathcal{N}(\nu, (\lambda \cdot \kappa)^{-1})$.

The marginal density of $\mu$ is obtained by integration and turns out to be a $t$ distribution. Then, we have that

$$\mu \sim t\left(\nu, \frac{\beta}{\alpha \cdot \lambda}, \frac{2\alpha}{\beta}\right).$$

For illustration, Figure 4.1 presents the marginal densities of $\mu$ and $\kappa$, and Figure 4.2 displays the corresponding joint normal-gamma density.
4.4. ELIMINATION OF NUISANCE PARAMETERS

Figure 4.2: Density function $f(\theta)$ of a normal-gamma distribution for $\theta = (\mu, \kappa)^T$. The parameters are fixed with $\nu = 0$, $\lambda = 0.5$, and $\alpha = 2$, $\beta = 1.2$. In this figure, the mode is at $\mu = 0$, $\kappa = 1.25$. (Figure taken from Held and Sabanés Bové, 2014).
Chapter 5

Bayesian asymptotics

By: Marco Kaufmann

5.1 Introduction

In the previous chapters, basic properties of Bayesian statistics and the corresponding characteristics were discussed. To complete these aspects, the asymptotic behaviour of the posterior distribution and an application of it are covered here.

As many estimators are often assumed as correct, even though this is just approximately the case, it is of high relevance to show and prove that the assumed properties or behaviours hold. In other words, to show the corresponding consistency is true (Walker, 2004). Here, the aspect of asymptotic behaviour of the posterior distribution will be first derived in the discrete and then in the continuous case and both illustrated with an example. Consequently, the importance of the previously stated will be shown in context of the Laplace approximation.

For the entire chapter, it is taken for granted that the prior is not degenerate and the Fisher regularity conditions apply as assumed by Held and Sabanés Bové (2014). Furthermore, the theoretical background as well as the exercises were taken from Held and Sabanés Bové (2014).

5.2 Main part

5.2.1 Discrete asymptotics

It is obvious that in the discrete case, the posterior distribution, in which we are interested, will never be a probability density function but a probability mass function. So unlike for example with the central limit theorem, the asymptotic posterior probability mass function will not ever
be normally distributed around the true parameter value. However, as will be proved in in a bit, the posterior probability mass function will become more and more centred around the true parameter value with increasing sample size. Eventually, with infinite sample size, it will be equal to the true value. This property is commonly referred to as **posterior consistency**.

**Proof.** Let \( \theta \) denote a discrete parameter with countable parameter space \( \Theta \) and true parameter value \( \theta_0 \). \( p_i = \Pr(\theta = \theta_i) \) is the prior probability mass function for \( i = 0, 1, 2, \ldots \) and \( X_{1:n} \) represents a random sample from a distribution with density function \( f(x|\theta) \). We are interested in the posterior probability mass function \( f(\theta_i|x_{1:n}) \).

\[
f(\theta_i|x_{1:n}) = \frac{f(x_{1:n}|\theta_i)f(\theta_i)}{f(x_{1:n})} = \frac{p_i f(x_{1:n}|\theta_i)}{\sum_{j=1}^{n} p_j f(x_{1:n}|\theta_j)} = \frac{p_i \prod_{k=1}^{n} f(x_k|\theta_i)}{\sum_{j=1}^{n} p_j \prod_{k=1}^{n} f(x_k|\theta_j)} = \frac{\exp(p_i + S_j^{(n)})}{\sum_{j=1}^{n} \exp(p_j + S_j^{(n)})},
\]

with \( S_j^{(n)} = \sum_{k=1}^{n} \log(f(x_k|\theta_j)/f(x_k|\theta_0)) \) for \( j = 1, \ldots, n \). Consequently, we look at the Kullback–Leibler discrepancy which is defined as follows

\[
D(\theta_0||\theta_i) = \int f(x|\theta_0) \log\left(\frac{f(x|\theta_0)}{f(x|\theta_i)}\right)dx = E_{X|\theta_0}[\log\left(\frac{f(X|\theta_0)}{f(X|\theta_i)}\right)].
\]

With the law of large numbers and the property that the Kullback–Leibler discrepancy is just equal to zero if \( j \) of \( \theta_j \) is equal to 0, otherwise bigger, we end up at the following formula

\[
\lim_{n \to \infty} \frac{1}{n} S_j^{(n)} = E_{X|\theta_0}[\log\left(\frac{f(X|\theta_j)}{f(X|\theta_0)}\right)] = -D(\theta_0||\theta_i),
\]

which suggests that \( \lim_{n \to \infty} 1/n S_j^{(n)} \) is equal to zero for \( j = 0 \) and otherwise smaller. As a consequence, it turns out that \( \lim_{n \to \infty} S_j^{(n)} \) is equal to zero for \( j = 0 \) and otherwise \( -\infty \). If we plug this result into formula (5.1), we get the following

\[
\lim_{n \to \infty} f(\theta_i|x_{1:n}) = \begin{cases} 
1, & \text{for } i = 0, \\
0, & \text{otherwise},
\end{cases}
\]

which is exactly what we wanted to show. For \( n \to \infty \), the posterior probability for \( \theta_0 \) converges to 1 and the **posterior consistency** is proven. \( \square \)

Another interesting aspect should be mentioned here. If the true parameter value \( \theta_0 \) is not included in the parameter space \( \Theta \), the posterior probability gets centred more and more around the parameter contained in \( \Theta \) with smallest Kullback–Leibler divergence. Figure 5.1 in the example section shows partially this phenomenon.

**Example 5.1.** Held and Sabanés Bové (2014) simulated data from a binomial distribution with parameters \( n = 10, 100, 1000 \) and \( \theta \) using a discrete uniform distribution on \( \Theta = 0.05, 0.10, \ldots, 0.95 \). Consequently inference for the given proportion a) \( \theta_0 = 0.25 \) and b) \( \theta_0 = 0.33 \) was estimated. The corresponding results can be seen in Figure 5.1. The top part represents the posterior distribution, the lower one the Kullback–Leibler discrepancy between \( \theta \) and \( \theta_0 \).
5.2. MAIN PART

Figure 5.1: Inference for proportion from simulation of binomial distribution. \( n = 10, 100, 1000 \), \( \theta_0 = 0.25 \) in a), \( \theta_0 = 0.33 \) in b). Posterior distribution on top, Kullback–Leibler divergence at the bottom. (Figure taken from Held and Sabanés Bové, 2014).

5.2.2 Continuous asymptotics

Not very surprisingly, the asymptotic posterior distribution behaves differently in the continuous case compared to the already covered discrete case. Given the previously mentioned Fisher regularity conditions apply, every continuous posterior distribution is asymptotically normal, centred around the MLE with covariance equal to the inverse observed Fisher information.

\[
f(\theta|x_{1:n}) \sim \mathcal{N}(\hat{\theta}_n, I(\hat{\theta}_n)^{-1}).
\]

As we are talking about asymptotics, the obtained statement can be altered, as for example in the likelihood case, because several aspects are asymptotically identical. Consequently, three other formulations of the formula exist. Using the expected instead of the observed Fisher information

\[
f(\theta|x_{1:n}) \sim \mathcal{N}(\hat{\theta}_n, J(\hat{\theta}_n)^{-1}),
\]

using the posterior mode \( \text{Mod}(\theta|x_{1:n}) \) as well as the negative curvature of the log posterior density at the mode \( C_n \), if available

\[
f(\theta|x_{1:n}) \sim \mathcal{N}(\text{Mod}(\theta|x_{1:n}), (C_n)^{-1}),
\]

or using the posterior mean \( \text{E}(\theta|x_{1:n}) \) and the posterior covariance \( \text{Cov}(\theta|x_{1:n}) \) given this is possible

\[
f(\theta|x_{1:n}) \sim \mathcal{N}(\text{E}(\theta|x_{1:n}), \text{Cov}(\theta|x_{1:n}))).
\]

The proof for the first relationship, from which the others are derived, goes as follows.

\textbf{Proof.} Assume \( X_{1:n} \) is a random sample from \( f(x|\theta) \) and we have an unknown parameter vector \( \theta \). The posterior density is consequently given as

\[
f(\theta|x_{1:n}) \propto f(\theta)f(x_{1:n}|\theta) = \exp[\log(f(\theta)) + \log(f(x_{1:n}|\theta))]. \tag{5.2}\]
Taylor expansion of the two log-terms around the prior mode \( \mathbf{m}_0 \) (log(\( f(\theta) \))) and the MLE \( \hat{\theta}_{ML} \) (log(\( f(x_{1:n} | \theta) \))) leads to

\[
\log(f(\theta)) \approx \log(f(\mathbf{m}_0)) - \frac{1}{2}(\theta - \mathbf{m}_0)^T \mathbf{I}_0(\theta - \mathbf{m}_0)
\]

and

\[
\log(f(x_{1:n} | \theta)) \approx \log(f(x_{1:n} | \hat{\theta}_n)) - \frac{1}{2}(\theta - \hat{\theta}_n)^T \mathbf{I}(\hat{\theta}_n)(\theta - \hat{\theta}_n).
\]

The negative curvature of \( \log(f(\theta)) \) at the mode \( \mathbf{m}_0 \) is \( \mathbf{I}_0 \), \( \mathbf{I}(\hat{\theta}_n) \) the observed Fisher information matrix. In the given setting, the posterior density is therefore proportional to (plug obtained expression into Equation (5.2))

\[
f(\theta | x_{1:n}) \propto \exp\left(-\frac{1}{2}(\theta - \mathbf{m}_n)^T \mathbf{I}_n(\theta - \mathbf{m}_n)\right), \tag{5.3}
\]

with

\[
\mathbf{I}_n = \mathbf{I}_0 + \mathbf{I}(\hat{\theta}_n)
\]

and

\[
\mathbf{m}_n = \mathbf{I}_n^{-1}(\mathbf{I}_0 \mathbf{m}_0 + \mathbf{I}(\hat{\theta}_n)\hat{\theta}_n).
\]

For large \( n \), \( \mathbf{I}_0 \) will be negligibly small and we get the posterior distribution of \( \theta \)

\[
f(\theta | x_{1:n}) \stackrel{a}{\sim} \mathcal{N}(\hat{\theta}_n, \mathbf{I}(\hat{\theta}_n)^{-1}),
\]

with mean equal to the MLE and covariance equal to the inverse observed Fisher information. (Held and Sabanés Bové, 2014)

Due to the above property, we can get a Bayesian interpretation of the MLE as asymptotic posterior mode as well as the relationship that the borders of a Wald confidence-interval for \( \theta \) are asymptotically identical with the ones obtained from a credible interval, given confidence and credible levels are the same (Held and Sabanés Bové, 2014).

**Example 5.2.** Data was simulated from a Bin\( (n, \pi = 0.1) \) distribution with prior \( \text{Be}(1/2, 1/2) \) for \( \pi \). The normal approximation based on this simulation given increasing \( n \) was compared to the “true” posterior density. The three approximations mentioned above (in this case observed Fisher information is equal to expected Fisher information) were used (Held and Sabanés Bové, 2014). The results can be seen in Figure 5.2.

It is obvious from the figure that the approximations get better for larger numbers of \( n \) and more and more represent the exact posterior density. This is consistent with the theory and expectations from the theoretical background discussed above.

### 5.2.3 Laplace approximation

Following Held and Sabanés Bové (2014), the Laplace approximation is used to approximate integrals that look as follows

\[
I_n = \int_{-\infty}^{+\infty} \exp(-nk(u))du,
\]

given the properties that \( k(u) \) is a twice differentiable, convex function with minimum at \( u = \hat{u} \) are true. Without going into too much detail, we will directly look at the finished formulas (for
5.2. MAIN PART

Figure 5.2: Approximation accuracy of three different approximations compared to true posterior density with increasing number of repetitions. (Figure taken from Held and Sabanés Bové, 2014.)

more information, Appendix C.2.2 Held and Sabanés Bové (2014)). In the univariate case we obtain

\[ I_n \approx \exp(-nk(\tilde{u})) \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}nk(u - \tilde{u})^2\right) du = \exp(-nk(\tilde{u}))\sqrt{\frac{2\pi}{nk}}, \]

with \( \kappa = d^2k(\tilde{u})/du^2 > 0 \). The multivariate case yields

\[ I_n \approx \left(\frac{2\pi}{n}\right)^{\frac{p}{2}} |K|^{-\frac{1}{2}} \exp(-nk(\tilde{u})), \]

with \( K \) denoting the \( p \times p \) Hessian of \( k(u) \) at \( \tilde{u} \).

So what does Laplace approximation have to do with Bayesian asymptotics? The Laplace approximation can be used to calculate characteristics of the posterior distribution of an unknown parameter \( \theta \), for example the posterior expectation. To do this, several approximations
CHAPTER 5. BAYESIAN ASYMPTOTICS

are made and exactly there Bayesian asymptotics or more their properties come into play. More about this after outlining the approach. Clearly, the procedure can be applied in uni- as well as in multi parameter settings, due to simplicity reasons, just the univariate one will be shown here though. Assume the same conditions and setting as in the asymptotics case apply and we want to estimate the posterior expectation of a given positive function \( g(\theta) \)

\[
E(g(\theta)|x_{1:n}) = \int g(\theta)f(\theta|x_{1:n})d\theta. \tag{5.4}
\]

We recall that the posterior distribution is defined as

\[
f(\theta|x_{1:n}) = \frac{f(x_{1:n}|\theta)f(\theta)}{f(x_{1:n})} = \frac{f(x_{1:n}|\theta)f(\theta)}{\int f(x_{1:n}|\theta)f(\theta)d\theta},
\]

which, plugged into Equation (5.4), leads to

\[
E(g(\theta)|x_{1:n}) = \int \frac{g(\theta)f(x_{1:n}|\theta)f(\theta)d\theta}{\int f(x_{1:n}|\theta)f(\theta)d\theta}.
\]

Rewriting this gives

\[
E(g(\theta)|x_{1:n}) = \int \frac{\exp(-nk(\theta))d\theta}{\exp(-nk(\theta))d\theta},
\]

with

\[-nk(\theta) = \log(f(x_{1:n}|\theta)) + \log(f(\theta))\]

and

\[-nk_g(\theta) = \log(g(\theta)) + \log(f(x_{1:n}|\theta)) + \log(f(\theta)).\]

Here the Laplace approximation comes into play. Separately applied to the denominator as well as the numerator yields following approximation for the posterior expectation

\[
E(g(\theta)|x_{1:n}) \approx \sqrt{\frac{\hat{\kappa}}{\hat{\kappa}_g}} \exp[-n(k_g(\hat{\theta}_g) - k(\hat{\theta}))],
\]

where \( \hat{\theta} \) and \( \hat{\theta}_g \) are the minima of \( k(\theta) \) and \( k_g(\theta) \), \( \hat{\kappa} = d^2k(\hat{\theta})/d\theta^2 \) and \( \hat{\kappa}_g = d^2k_g(\hat{\theta}_g)/d\theta^2 \) the curvature of \( k(\theta) \) and \( k_g(\theta) \) at the given minima. Attention, unfortunately, the Laplace approximation does not stay the same if \( \theta \) gets parametrised differently, so in other words is not invariant to changes. This can on the other hand also be used as an advantage if applied appropriately. For example if one chooses the function \( g(\theta) \) as the variance stabilizing transformation, the approximation can get even better. Here we return to the Bayesian asymptotic normality (in the continuous case). The Laplace approximation assumes a normal distribution of the posterior density as given for the approximation. With increasing \( n \), the approximation gets better and consequently is asymptotically exact (Held and Sabanés Bové, 2014).

**Example 5.3.** Suppose we have a prior for the success probability of \( \text{Be}(0.5, 0.5) \) and a random sample which comes from a \( \text{Bin}(n, \pi) \) distribution. Due to this, we know that the posterior density is \( \text{Be}(x + 0.5, n - x + 0.5) \) and the posterior expectation equals to \( x + 0.5/n + 1 \). As the posterior expectation is known, we can check how accurate our approximations with the Laplace transformation are. For this purpose, one approximation with \( g(\theta) = \theta \), so the simple posterior
5.3 Exercises

Further examples and exercises can be found in Chapters “6.8 Exercises (Bayesian Inference)” and “8.6 Exercises (Numerical Methods for Bayesian Inference)” in the book of Held and Sabanés Bové (2014).

Table 5.1: Laplace approximations compared to real posterior mean. Relative error in brackets. (Table taken from Held and Sabanés Bové, 2014.)

| Observation | Posterior mean | $\hat{E}(\pi | x)$ | $\hat{E}_1(\pi | x)$ | $\hat{E}_2(\pi | x)$ |
|-------------|----------------|---------------------|---------------------|---------------------|
| $x/n$       | $n$            | $E(\pi | x)$         | $0.5630 (-0.0349)$  | $0.5797 (-0.0062)$  |
| 0.6         | 5              | 0.5833              | 0.5940 (-0.0021)    | 0.5950 (-0.0005)    |
| 0.6         | 20             | 0.6123              | 0.6190 (-0.0018)    | 0.6200 (-0.0012)    |
| 0.6         | 100            | 0.6190              | 0.6260 (-0.0021)    | 0.6260 (-0.0012)    |

mean, another with $g(\theta) = \arcsin(\sqrt{\theta})$, the variance stabilizing transformation, are calculated and applied for varying values of $x$ and $n$ (Held and Sabanés Bové, 2014). See Table 5.1.

From the table one can conclude that with increasing $n$, both approximations get better and are already with a sample size of 100 very precise. The variance stabilizing estimate appears to be even more accurate than the standard one, exactly as we expected. We looked at these errors because we used optimisation (Laplace approximation) rather than integration, which is just asymptotically exact.
Chapter 6

Bayesian prediction

By:
Tasnim Hamza

6.1 Introduction

Prediction techniques have been used for a variety of purposes. Beside estimating unobserved future values, we also can use the prediction to for model verification. For example by fitting a model for a part and predicting the remaining parts. In addition we can estimate some special values such as the missing values in historical records and the outliers in some cases.

All Bayesian inference is based on posterior distributions, including predictions of hypothetical data. So how one would predict a new observation based on a model and existing data.

Let’s clarify the idea of prior and posterior predictive distribution by a simple example. Suppose we are tossing a coin for 100 times, then the unbiased result that we expect to get is 50 heads, so probably our prior prediction distribution will be centered on 50 (before we actually do the experiment). Now, let’s do the experiment and see the result. If we actually get 80 heads, we can update our prior and compute the posterior prediction which is somehow lies between the prior prediction and the likelihood function, see Figure 6.1.

In this chapter we will introduce the prior and posterior predictive concepts, and how we can compute their distributions, and illustrate this in multiple examples. Finally, we introduce the model averaging prediction.

6.2 Prior and posterior predictive distribution

Suppose we have a random sample $X_{1:n}$ with corresponding parameter $\theta$, so we want to predict a new value $Y$, then we have two approaches according to Bayesian. The first approach is the
prior prediction which gives us an initial estimation of \( Y \) based on the prior knowledge of our parameter \( \theta \). After we see the data, we update the prior distribution and compute the posterior distribution of \( \theta \), then we can improve our prediction and compute the posterior predictive distribution (Held and Sabanés Bové, 2014). More specifically, we have the following informal definitions

- **Prior prediction** is the weighted average of the likelihood and the prior distribution, and is computed by the following formula:

\[
f(y) = \int_{\theta \in \Theta} f(y, \theta) \, d\theta \quad \text{by Bayes’ rule}
\]

\[
= \int_{\theta \in \Theta} f(y|\theta) \times f(\theta) \, d\theta.
\]

- **Posterior prediction** is the marginal probability of \( Y \) given that we observed \( X \), which is computed by taking the integral of the joint probability density function \( f(y, \theta|x) \) across the range of \( \theta \). It is computed by the following formula:

\[
f(y|x_{1:n}) = \int_{\theta \in \Theta} f(y, \theta|x_{1:n}) \, d\theta \quad \text{by Bayes’ rule}
\]

\[
= \int_{\theta \in \Theta} f(y|\theta, x_{1:n}) \times f(\theta|x_{1:n}) \, d\theta
\]

\[
= \int_{\theta \in \Theta} f(y|\theta) \times f(\theta|x_{1:n}) \, d\theta.
\]

The last line holds because we assume that \( Y \) independent of \( X_{1:n} \) in presence of the parameter \( \theta \); in other words, \( \theta \) gives us all what we need to know about the sample \( X_{1:n} \). Notice firstly that

\[\text{Figure 6.1:} \quad \text{We see how our posterior prediction for the tossing coin experiment influenced by our prior and the likelihood.}\]
as the sample size increases the posterior approaches to the likelihood, which means that it is more influenced by the likelihood than the posterior distribution. The second note is that the uncertainty in Bayesian prediction comes from the deviation of an observation from its expected value and the uncertainty about the parameter, which makes the Bayesian prediction more precise.

**Plug-in prediction.** The predictive density distribution of \( y \) in the plug-in method is

\[
f(y) = f(y; \hat{\theta}_{MLE}),
\]

where \( \hat{\theta}_{MLE} \) is the maximum likelihood estimate based on the random sample \( X_{1:n} \). This is easily computed and a simple approach but as we see it ignores the uncertainty in the MLE.

However, as we mentioned before, the Bayesian prediction incorporates the uncertainty in MLE and integrates over all choices of the the parameter \( \theta \), which makes it more powerful than the plug in method (Held and Sabanés Bové, 2014).

On the other way, the Bayesian prediction converges to plug in prediction, if the sample size \( n \) tends to infinity

\[
f(y|x_{1:n}) \xrightarrow{n \to \infty} f(y; \hat{\theta}_{MLE}).
\]

This holds because the posterior distribution is asymptotically normal with mean equal MLE and variance approaches zero (Held and Sabanés Bové, 2014). Now, lets illustrate the computation of the predictive distribution using different choice of prior distribution.

**Example 6.1** (Normal Distribution, Held and Sabanés Bové, 2014). Suppose the random sample \( X_{1:n} \sim N(\mu, \sigma^2) \), and suppose we choose a Jeffrey’s prior for \( \mu; f(\mu) \propto 1 \), so the posterior distribution is

\[
\mu|x_{1:n} \sim N\left( \bar{x}, \frac{\sigma^2}{n} \right).
\]

Now, we want to compute the posterior predictive distribution

\[
f(y|x_{1:n}) = \int f(y|\mu)f(\mu|x_{1:n}) \, d\mu
\]

\[
\propto \int \exp \left[ -\frac{1}{2} \left( \frac{y - \mu)^2}{\sigma^2} + \frac{n(\mu - \bar{x})^2}{\sigma^2} \right) \right] \, d\mu
\]

\[
\propto \int \exp \left\{ -\frac{C}{2} (\mu - c)^2 - \frac{1}{2} \frac{(y - \bar{x})^2}{2(1 + \frac{1}{n})\sigma^2} \right\} \, d\mu
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{(y - \bar{x})^2}{2(1 + \frac{1}{n})\sigma^2} \right\} \int \exp \left\{ -\frac{C}{2} (\mu - c)^2 \right\} \, d\mu
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{(y - \bar{x})^2}{2(1 + \frac{1}{n})\sigma^2} \right\} \sqrt{2\pi \frac{\sigma}{\sqrt{n+1}}},
\]

yielding

\[
Y|x_{1:n} \sim N\left( \bar{x}, \frac{\sigma^2}{1 + \frac{1}{n}} \right).
\]

Notice that, when \( n \) tends to infinity \( Y|x_{1:n} \sim N(\bar{x}, \sigma^2) \) where \( \bar{x} \) is the MLE of the population mean, so this is an evidence that the Bayesian converges to plug in prediction for large \( n \).
Example 6.2 (Binomial distribution). If $x \sim \text{Bin}(n, \theta)$ and the prior distribution of $\theta$ is $\text{Be}(\alpha, \beta)$, so the posterior distribution of $\theta|x$ is $\text{Be}(\alpha + x, \beta + n - x)$. Now we want to determine the distribution of the prior predictive is as follows

$$f(x) = \int_0^1 f(x|\theta) \times f(\theta) \, d\theta$$

$$= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \, d\theta$$

$$= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n+\beta-x)}{\Gamma(x+1)\Gamma(n-x+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}.$$

The posterior prediction is computed in the same way but we use the posterior distribution which is also beta distributed:

$$f(y|x) = \int_0^1 f(y|\theta) \times f(\theta|x) \, d\theta.$$

Hence, $Y|x \sim \text{Be-Bin}(m, \alpha + x, \beta + n - x)$.

Now, lets illustrate the idea numerically in a disease case. Suppose we have a population and we are interested in the diseased people. Let $\theta$ be the proportion of the diseased people in the population. Notice that $0 < \theta < 1$ so may be the beta distribution is a good choice for $\theta$ i.e $\theta \sim \text{Be}(a, b)$. Now, we want to take a sample to estimate our parameter $\theta$. So, let $X$ be the number of diseased people in a sample of size $n$ and suppose $X \sim \text{Bin}(n, \theta)$. Now we can use our sample to estimate the likelihood of $\theta$.

Suppose we take a new sample of size $m$, so what is our prior and posterior prediction of $Y$ which is the number of the diseased people in this new sample. We saw before how we can compute them by the formulas. Now lets illustrate our idea by multiple plots.

In Figure 6.2, we choose a uniform distribution $\text{Be}(1, 1)$, which means that we do not have a previous knowledge about the distribution of $\theta$, we see that the prior prediction is a mirror of the prior distribution. Also, we notice that the posterior distribution influenced by the likelihood and the prior, also the posterior predictive is between the likelihood and the posterior distribution.

Next, in Figure 6.3 we specified our prior knowledge and chose $\text{Be}(3, 2)$ as prior distribution and then the posterior distribution is influenced more by the prior distribution, and also the prior and posterior predictive distributions reflect the changes in the prior distribution.

Lets see the effect when we increased the sample size for the original and new samples and fix the other variables. As we see in Figure 6.4, the posterior is more influenced by the likelihood than the prior. Although, the prior predictive distribution still reflects what occurs in the prior distribution, while the posterior predictive distribution approaches the likelihood function more.
6.3. **COMPUTATION OF THE POSTERIOR PREDICTIVE DISTRIBUTION**

![Prior distribution](image1)

![Posterior distribution](image2)

![Likelihood function](image3)

![Prior prediction](image4)

![Posterior Prediction](image5)

**Figure 6.2:** We choose a Be(1, 1) distribution as a prior distribution and our data assumed be Bin(10, 1), and then plot our prediction for our new sample of size 10 and with disease proportion 0.1.

**6.3 Computation of the posterior predictive distribution**

In general, we can compute the prior prediction and the posterior prediction by using the following formulas, respectively Held and Sabanés Bové (2014)

\[
f(y) = \frac{f(y|\theta)f(\theta)}{f(\theta|y)} \tag{6.1}
\]

\[
f(y|x) = \frac{f(y|\theta, x)f(\theta|x)}{f(\theta|x, y)} \tag{6.2}
\]

So if we can compute the densities in the above formulas it is easier to use them, and this occurs in the case if \(f(\theta)\) is conjugate with respect to the likelihood function \(f(x|\theta)\), then \(\theta|x\) and \(\theta|x, y\)
belong to the same family of distributions as \( \theta \). So we can compute the prior and posterior directly by plugging-in the densities and the likelihood without explicit integration. We will illustrate this by the following example.

Example 6.3 (Poisson Distribution). Let \( X \sim \text{Po}(e_x \lambda) \), \( Y \sim \text{Po}(e_y \lambda) \), and \( \lambda \sim \Gamma(\alpha, \beta) \) be the priori. Then the posterior distributions as follow

\[
\lambda | x \sim G(\tilde{\alpha}, \tilde{\beta})
\]

\[
\lambda | x, y \sim G(\tilde{\alpha} + y, \tilde{\beta} + e_y).
\]
6.3. COMPUTATION OF THE POSTERIOR PREDICTIVE DISTRIBUTION

We increased the sample size for the original data \( n = 100 \) and the new sample \( m = 100 \) and maintain the proportion of the diseased people. Now, the prior and posterior prediction distributions follow simply by Equations (6.1) and (6.2)

\[
f(y|x) = \frac{(e_x\lambda)^y}{y!} \exp(-e_y\lambda) \frac{\tilde{\beta}^\alpha}{\Gamma(\tilde{\alpha})} \lambda^{\tilde{\alpha}-1} \exp(-\tilde{\beta}\lambda) \frac{\tilde{\beta}^\alpha}{(\tilde{\beta}+e_y)^{\tilde{\alpha}+y}} \chi^{\tilde{\alpha}+y-1} e^{-(\tilde{\beta}+e_y)\lambda} = \frac{\tilde{\beta}^\alpha}{\Gamma(\tilde{\alpha})} \frac{e_y^y \Gamma(\tilde{\alpha} + y)}{y!} (\tilde{\beta} + e_y)^{-(\tilde{\alpha}+y)}.
\]

So \( Y|X \sim \text{Po-}\Gamma(\tilde{\alpha} = \alpha + x; \tilde{\beta} = \beta + e_x, e_y) \), and we can compute the prior prediction in the same way.
6.4 Bayesian model averaging

For different models, say $M_1, \ldots, M_K$ with corresponding posterior predictive density $f(y|x, M_k)$ and posterior probabilities $\Pr(M_k|x)$, we can compute the model-averaged posterior predictive density as follows (Held and Sabanés Bové, 2014):

$$f(y|x) = \sum_{k=1}^{K} f(y|x, M_k) \times \Pr(M_k|x).$$

Model averaging is a very important technique for improving Bayesian prediction. This is true because when we choose our true model according to the posterior probabilities of the models, this means that we have a zero-one loss function for deciding the optimal model, which might be inappropriate if the models are close to each other in their description of the data generation process (Held and Sabanés Bové, 2014). On the other hand, when we use a model averaging approach, we are sure that the true model is included, and further considering the uncertainty of all models.

The idea of this approach is simple and based on all the good models, where their contributions to the model depends on the posterior probability $\Pr(M_k|x)$. The model-averaged posterior prediction expectation and variance are

$$E(Y|x) = E \left\{ E(Y|x, M) \right\} = \sum_{k=1}^{K} E(Y|x, M_k) \times \Pr(M_k|x)$$

and

$$\text{Var}(Y|x) = E \left\{ \text{Var}(Y|x, M) \right\} + \text{Var} \left\{ E(Y|x, M) \right\},$$

respectively. We come back to Bayesian model averaging in Section 8.7.

6.5 Supplement: R-Code

This section contains the R-Code used to construct Figure 6.2, and with slightly modified parameters Figures 6.3 and 6.4.

```r
library(LearnBayes)
library(VGAM)

##### prior parameters
a <- 1
b <- 1

### Data
n <- 10
x <- 1

##### Prior Distribution
theta <- seq(0, 1, length = 100)
hx <- dbeta(theta, a,b)
layout(matrix(c(1,2,3,3,4,5), 3, 2, byrow = TRUE))
plot(theta, hx, type = "l", xlab = "Theta", ylab = "pdf",
```
### Posterior distribution

\[
p_{\text{post}}(\theta) = \text{dbeta}(\theta, a+x, b+n-x)
\]

```r
post <- dbeta(theta, a+x, b+n-x)
plot(theta, post, type = "l", xlab="Theta", ylab="pdf",
     main="Posterior distribution")
```

### Likelihood Function

\[
\text{likelihood} = \binom{n}{x} \theta^x (1-\theta)^{n-x}
\]

```r
likelihood <- choose(n,x) * theta^x * (1-theta)^(n-x)
plot(theta, likelihood, type="l", xlab="Theta", ylab="likelihood",
     main="Likelihood function")
```

### Prior Predictive

\[
p_{\text{prior}}(\theta | y) = \text{dbetabinom.ab}(y, m, a, b)
\]

```r
prior.pred <- 1
m <- 10
y <- 0:m
for (i in 1:(m+1)) {
    prior.pred[i] = dbetabinom.ab(y[i], m, a, b)
}
plot(y, prior.pred, xlab="Number of diseased people", ylab="pmf",
     main="Prior prediction")
```

### Posterior Predictive

\[
p_{\text{post}}(\theta | y) = \text{dbetabinom.ab}(y, m, x+a, b+n-x)
\]

```r
pos.pred <- 1
y <- 0:m
m <- 10
for (i in 1:(m+1)) {
    pos.pred[i] = dbetabinom.ab(y[i], m, x+a, b+n-x)
}
plot(y, pos.pred, xlab="Number of diseased people", ylab="pmf",
     main="Posterior Prediction")
```
Chapter 7

Overview AIC, BIC

By: Michael Hediger

The main goal of the following chapter is to give a comprehensive and detailed description of Akaike’s Information Criterion (AIC; Akaike, 1973, Akaike, 1974, Akaike, 1977). This widely used likelihood based model selection criterion has its roots in a discipline called information theory, which was developed by Claude Shannon in the mid-1940s. Therefore, we will use Shannon’s essay on communication theory (Shannon, 1948) to give a motivation for the Kullback–Leibler (K–L) “information” or “distance” between two probability distributions (Kullback and Leibler, 1951). The construction and understanding of the Kullback–Leibler information will serve as the unique anchor point in the establishment of AIC. The heart of this chapter will focus on the path from K–L information to AIC. This path will be presented in the form of a mathematical derivation, which will allow us to feel comfortable with the final structure of AIC. As an outlook to the next chapter we will introduce the closely related Bayesian information criterion (BIC) or Schwarz criterion (Schwarz, 1978), which considers model selection from a Bayesian point of view. Motivation and derivation of the BIC will be reserved for the next chapter.

7.1 Shannon’s information theory and the Kullback–Leibler information

The core of Shannon’s information theory starts with the realization that every kind of information extracted from a message or event can be digitized in terms of the pair \{0, 1\}. This pair is said to be one bit of information. Here it is worth considering that the term “message” can be seen as a substitute for the familiar notation of an “event” in a probabilistic framework (Example 7.2). As a first intuitive result, one would argue that a message with high information
content will be digitized with a high amount of bits. As a consequence a message with high information content requires a high amount of storage capacity. Shannon discovered a brilliant and simple relation between the number of bits needed to digitize such an event and the probability of its occurrence. We will discover this simple relation with a first example.

**Example 7.1.** (Information and probability) If one is interested in the message “The Mountain”, the information content we could extract out of this particular message is strongly dependent on the context. Imagine we would give you this message in Switzerland, where the probability of “The mountain” to occur is rather high, the amount of information we could extract out of this message would be rather poor. Now consider the same message in Holland, where the probability of the message “The mountain” is low — we would definitely gain much more information. In other words, the message “The mountain” would need to be digitized with more bits in Holland compared to Switzerland.

This simple example reveals a beautiful relation which can be phrased as follows: If $\pi$ is the probability of the message, its information is related to $1/\pi$. Shannon used this relation to record his formal definition of information,

$$I(\pi) = \log_2 \left( \frac{1}{\pi} \right).$$

(7.1)

Equation (7.1) quantifies information as the number of bits needed to digitize a message which has occurrence probability $\pi$. It will soon become clear why he actually used the logarithm to quantify information. First we will use the given definition to quantify the information of the event “head” in a single coin toss.

**Example 7.2.** (Information of the event “head” in a coin toss) Assume a random variable $X \sim \text{Bin}(n = 1, \pi)$; $\pi = \frac{1}{2}$, and we are only interested in the event “head”, which is one of the two possible outcomes of the binary random variable $X$. Following the definition of information, one would need $\log_2(1/\pi) = 1$ bit to quantify the information of the event “head”. This result can be interpreted in a probabilistic framework and is in perfect accordance with Example 7.1. If the event “head” would be associated with a smaller probability $\pi < 1/2$ (consider “The mountain” in Holland), we would get $I(\pi) > 1$. On the other hand, if the event “head” gets closer to a purely deterministic event $\pi > 1/2$, the number of bits required to digitize “head” would be $I(\pi) < 1$. The lower limit of information would be an event which occurs with probability equal to one — In the coin toss example this would be represented with a coin equipped with two equal sides.

The previous example showed that we are able to calculate the information of one single event. In a next step, Shannon extended its definition of information of one single event to a concept which allows to calculate the average amount of information over the whole sample space. As a final goal, one could speak of the average amount of information in context of a given probability distribution.

This concept is known as Shannon’s entropy. To get a closer look, we will first calculate the average amount of information given one has a sample of $n$ events.
Example 7.3. (Entropy of a coin tossing experiment) Assume a sample of \( N \) Bernoulli random variables
\[
X_i \text{iid} \sim \text{Bin}(n = 1, \pi); \quad \pi = \frac{1}{2}, \quad i \in \{1, 2, \ldots, N\}.
\]
The two possible events are donated as \( \omega_1 = \text{head} \) and \( \omega_2 = \text{tail} \). We will denote this by \( \omega_j \), where \( j \in \{1, 2\} \). If the term \( p_j \) donates the relative amount of \( \omega_j \) in \( N \) trials we can quantify the total amount of observed \( \omega_j \)'s as \( k_j = Np_j \). The next step would be to quantify the total amount of information of a sample of size \( N \), given every single event contributes the information
\[
I(p_j) = \log_2(1/p_j).
\]
This total information would be given as
\[
2 \sum_{j=1}^{2} k_j I(p_j) = 2 \sum_{j=1}^{2} Np_j \log_2 \left( \frac{1}{p_j} \right) = N \sum_{j=1}^{2} p_j \log_2 \left( \frac{1}{p_j} \right).
\]
Therefore, the average amount of information extracted out of that sample is given by
\[
2 \sum_{j=1}^{2} p_j \log_2 \left( \frac{1}{p_j} \right).
\]

The given example can be seen as a motivation for the formal definition of Shannon’s entropy. We will give this definition of entropy in the context of a given discrete probability distribution \( P \).

Assume a discrete random variable \( X \) with possible events \( \{\omega_1, \ldots, \omega_n\} \), and let its probability mass function (pmf) be given as \( P_m \). Further assume that each possible event is observed with probability \( P_m(X(\omega_i)) \) where \( i \in \{1, 2, \ldots, n\} \). Shannon’s entropy, defined as the average number of bits provided by the probability distribution \( P \) is given as
\[
H(P) = \sum_{i=1}^{n} P_m(X(\omega_i)) \log_2 \left( \frac{1}{P_m(X(\omega_i))} \right) = \mathbb{E}_{P_m} \left[ \log_2 \left( \frac{1}{P_m(X(\omega_i))} \right) \right]. \quad (7.2)
\]
From now, and throughout this chapter we will consider a slight abuse of notation, therefore we will use \( P \) to describe the probability distribution in general, whereas \( P_m(x) \) will refer to the according probability mass function. As a matter of fact, we will not differentiate between \( P \) and \( P_m \) in terms of notation. We will adapt this simplification in both, the continuous and discrete case.

Shannon’s equation has its name because of the close relation to a famous quantity in statistical thermodynamics called Gibbs entropy (Gibbs, 1878), named after Josiah Willard Gibbs. Gibbs entropy is given by
\[
S_G = -k_b \sum_{i} \pi_i \log_{10}(\pi_i), \quad (7.3)
\]
where \( k_b \) is known as the Boltzmann constant and \( \pi_i \) is said to be the probability of a micro-state which is part in describing a macro-state. The term entropy is closely tied to disorder and if one has a large number of low probability micro-states the entropy of a macro-state is said to be high. This kind of interpretation can be well adapted to Shannon’s entropy, where the macro-state reflects the probability distribution \( P \) and each single event can be seen as a micro-state. Even further back in time, Ludwig Eduard Boltzmann provided the famous equation (Boltzmann, 1872)
\[
S_B = k_b \log_{10}(W), \quad (7.4)
\]
where \( W \) is the total amount of equiprobable micro-states. Therefore Equation (7.4) and (7.3) are perfectly equivalent if one assumes \( \pi_i = 1/W \) for all micro-states.

Up to now we motivated Shannon’s entropy for a discrete random variable \( X \). If we assume that the random variable \( X \) lives in a continuous setting we have to replace the sum with the integral and write

\[
H(f) = \int f(x)I(x) \, dx = -\int f(x) \log_2(f(x)) \, dx = -E_f[\log_2(f(x))].
\] (7.5)

Equation (7.5) describes the average amount of bits that are necessary to describe a continuous probability distribution \( f \) defined on a sample space \( \Omega \). Therefore if one is interested in the entropy of a probability distribution associated to one single realization of the according random variable, Equation (7.5) or 7.2 are the way to go.

The following example gives further insights into Shannon’s concept of entropy and describes the use of the logarithm in the definition of information.

**Example 7.4. (Entropy of the joint distribution of two Bernoullies)** Assume a couple of independent random variables given as

\[
X \sim \text{Bin}(n = 1, \pi_1) \quad \text{and} \quad Y \sim \text{Bin}(n = 1, \pi_2); \quad \pi_1 = \pi_2 = \frac{1}{2}.
\]

To keep it general we will denote the two possible events belonging to \( X \) as \( \omega_i \), where \( i \in \{1, 2\} \) and the events for \( Y \) are given by \( \omega_j \), and \( j \in \{1, 2\} \) accordingly. In a next step we will define the random variable \( X \cap Y \) as the joint distribution of \( X \) and \( Y \). The probability space \( \{\omega_1, \omega_2\} \times \{\omega_1, \omega_2\} \) can be visualized in Table 7.1.

**Table 7.1:** Probability space of a bivariate Bernoulli.

<table>
<thead>
<tr>
<th></th>
<th>( \omega_1 )</th>
<th>( \omega_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega_1 )</td>
<td>( \omega_1 \cap \omega_1 )</td>
<td>( \omega_1 \cap \omega_2 )</td>
</tr>
<tr>
<td>( \omega_2 )</td>
<td>( \omega_2 \cap \omega_1 )</td>
<td>( \omega_2 \cap \omega_2 )</td>
</tr>
</tbody>
</table>

In terms of Shannon’s bits one would realize that the distribution \( X \cap Y \) needs the double amount of information compared to \( X \). This statement becomes clear if one describes the probability space of the joint in terms of its realizations \( \{0, 1\} \times \{0, 1\} \) and realizes that one needs a second couple \( \{0, 1\} \) to describe \( X \cap Y \). We will now use Shannon’s entropy to verify the above statement. Given the statement of independence one gets for the event \( \{X(\omega_1) \cap Y(\omega_1)\} \), say both Bernoullies give “head”, the according probability as \( P(\{X(\omega_1) \cap Y(\omega_1)\}) = P(\{X(\omega_1)\})P(\{Y(\omega_1)\}) \). In a final step we can write Shannon’s entropy for the random variable \( X \cap Y \) as

\[
H( X \cap Y ) = \pi_1 \pi_1 \log_2 \left( \frac{1}{\pi_1 \pi_1} \right) + 2 \left( \pi_1 \pi_2 \log_2 \left( \frac{1}{\pi_1 \pi_2} \right) \right) + \pi_2 \pi_2 \log_2 \left( \frac{1}{\pi_2 \pi_2} \right)
\]

\[
= \frac{1}{4} \log_2(4) + \frac{1}{2} \log_2(4) + \frac{1}{4} \log_2(4)
\]

\[
= 2.
\]
The above calculations confirmed the intuitive guess that a couple of one bits probability distributions is worth two bits. This would not be the case when Shannon would not have used the logarithm for his definition of information. We will see later that if one is interested in comparing information between probability distributions, one is not forced to use \( \log_2(.) \). This is due to the strict monotone behavior of the logarithm. For now we will use \( \log(.) \), to state the fact that any form of useful base is appropriate.

In the last part of this sub-chapter we will use Shannon’s entropy to motivate the K–L information of two probability distributions.

Equation (7.2) describes the average amount of information associated to the discrete probability distribution \( P \). For now assume that \( P \) is the truth. Therefore our data, if any is there, will be a realization of the truth distribution \( P \).

In a next thought, assume a second discrete probability distribution \( Q \), defined on the same sample space belonging to the random variable \( X \). One can think of it as the fitted probability model \( Q \) on the data \( x \) and \( Q \) should resemble the truth.

This allows us to define the average amount of information necessary to describe the distribution of \( Q \), given one has the particular events belonging to \( X \). This term is known as cross entropy and is defined as

\[
H(P, Q) = \sum_x P(x) \log \left( \frac{1}{Q(x)} \right) = E_P \left[ \log \left( \frac{1}{Q(x)} \right) \right].
\]  

(7.6)

Now the following argument is useful. If the truth is \( P \) and \( P \) describes a macro-process, where each realization \( x \) belongs to this macro-process \( P \), then disorder is assumed to be minimal if we use \( P \) to describe each of the events belonging to \( X \). This can be translated in terms of Shannon’s bits: the average amount of bits, needed to describe a probability distribution \( Q \), is minimized when \( Q \) is the truth.

This fundamental statement is given by the Gibbs’ inequality, named after Josiah Willard Gibbs. It reads as follows:

\[
H(P, Q) \geq H(P)
\]

\[
\sum_x P(x) \log \left( \frac{1}{Q(x)} \right) \geq \sum_x P(x) \log \left( \frac{1}{P(x)} \right).
\]  

(7.7)

**Proof.** To proof the inequality (7.7) we first note that the natural logarithm satisfies

\[
x - 1 \geq \ln(x), \quad \forall x \geq 0.
\]

Although we will use the natural logarithm for the proof of (7.7) the inequality holds for every given base, since

\[
\log_b(x) = \frac{\ln(x)}{\ln(b)}.
\]

Given these two prior statements we can use the two probability distributions \( P \) and \( Q \) to write

\[
- \sum_x P(x) \ln \left( \frac{Q(x)}{P(x)} \right) \geq - \sum_x P(x) \left( \frac{Q(x)}{P(x)} - 1 \right) = - \sum_x Q(x) + \sum_x P(x) = - \sum_x Q(x) + 1 = 0.
\]
Therefore we have

\[ - \sum_x P(x) \ln(Q(x)) \geq - \sum_x P(x) \ln(P(x)) \]

\[ H(P, Q) \geq H(P), \]

and equality if and only if \( Q = P \).

It is Shannon’s entropy and Gibbs inequality which allows a formal definition of the K–L information between two discrete probability distributions \( P \) and \( Q \).

Assume a discrete random variable \( X \) with possible realizations \( \{x_1, \ldots, x_i, \ldots, x_n\} \), where the probability of each event is given by \( P(X = x_i) = p_i \) and \( P \) is said to be the truth. Further we have the approximating distribution given as \( Q \), with \( Q(X = x_i) = q_i \) as the according probability of the \( i \)th event when \( Q \) is used instead of \( P \). Now we can define the K–L information between \( P \) and \( Q \) as:

\[ I(P, Q) = \sum_{i=1}^n p_i \log \left( \frac{p_i}{q_i} \right) \]

(7.8)

\[ = \sum_{i=1}^n p_i \log(p_i) - \sum_{i=1}^n p_i \log(q_i) \]

\[ = H(P, Q) - H(P). \]

The previous effort allows us to make a simple interpretation of the K–L information in terms of Shannon’s information criterion. First of all, Gibbs inequality gives right away that \( I(P, Q) \geq 0 \) and equality if and only if \( P = Q \). The term \( H(P, Q) \) describes the amount of information needed when \( Q \) is used instead of \( P \) and therefore \( H(P, Q) - H(P) \) gives the extra information necessary, when we approximate \( P \) with \( Q \). From this point of view it becomes clear that we are talking about a directed distance between \( P \) and \( Q \) and \( I(P, Q) \neq I(Q, P) \) — The roles of \( P \) as the truth and \( Q \) as the approximating distribution are not exchangeable.

The definition of the K–L information for a discrete random variable allows us to adapt the form of (7.8) for a continuous setting. Assume now \( X \) is a continuous random variable with proper density \( f \). In the following \( f \) will represent the truth and we use \( g \) as the approximating density. Therefore we can write the K–L information for \( f \) and \( g \) as

\[ I(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x|\theta)} \right) dx. \]

(7.9)

In Equation (7.9) we explicitly wrote the parameter \( \theta \) to highlight the fact that full reality \( f \) is fixed and only \( g \) varies over a space of models indexed by \( \theta \). The above results clearly highlight the following model selection criteria:

If one is interested in the best model, one has to minimize the extra information needed, when we use \( g \) to approximate the truth \( f \). As mentioned above, \( I(f, g) \) is minimized when we have \( g = f \). This minimum also involves the parameter \( \theta \). If \( f \) is the known truth it has one specific non-stochastic parameter value fixed at \( \theta_0 \). In conclusion \( I(f, g) \) approaches a minimal value if \( g \) is the truth and \( \theta = \theta_0 \). This first model selection criteria is purely theoretical, since we will never get the true parameter \( \theta_0 \). One would argue that the true parameter \( \theta_0 \) is not
the only problem since we will clearly never discover \( f \). We will see below that \( f \) will drop as a constant and we are left with a somehow familiar problem, finding a parameter value \( \theta \) which best approximates the truth \( \theta_0 \) given the data. Nevertheless we will use the next example as a first demonstration of model selection based on K–L information when we are aware of \( \theta_0 \).

**Example 7.5.** (K–L information and model selection) For this example assume the following true model for the random variable \( X \):

\[
X \sim \mathcal{N}(\mu, \sigma^2); \quad \mu = 0, \quad \sigma^2 = 1.
\]

In a next step consider 3 approximating models \( g_i \), where \( g_i \neq f, \quad \forall g_i \).

We will use the following code to compute the K–L information for each \( g_i \) and the truth \( f \).

```r
require("flexmix")

# Building models:
x <- seq(-5, 12, length=250)
y <- cbind(n1=dnorm(x, mean=0, sd=1), n2=dnorm(x, mean=0.1, sd=1),
          n3 = dnorm(x, mean=1, sd=1), n4 = dnorm(x, mean=2, sd=1))

# KLdiv(y, eps = 1e-16)
```

The four densities can be be seen in Figure 7.1 and the results of the R function `KLdiv(.)` are summarized in Table 7.2.

<table>
<thead>
<tr>
<th>( g_i )</th>
<th>Approximating model</th>
<th>( I(f, g_i) )</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_1 )</td>
<td>( \mathcal{N}(\mu = 0.1, \sigma^2 = 1) )</td>
<td>0.004999977</td>
<td>1</td>
</tr>
<tr>
<td>( g_2 )</td>
<td>( \mathcal{N}(\mu = 1, \sigma^2 = 1) )</td>
<td>0.499999767</td>
<td>2</td>
</tr>
<tr>
<td>( g_3 )</td>
<td>( \mathcal{N}(\mu = 2, \sigma^2 = 1) )</td>
<td>1.999999761</td>
<td>3</td>
</tr>
</tbody>
</table>

The given example already gives some useful ideas about the properties of K–L based model selection. First of all, we did compare three theoretical densities to a given true density \( f \). In any of the models \( g_i \), we correctly assumed a Gaussian distribution to approximate \( f \). Clearly we could have constructed the theoretical minimum K–L value when using \( g_i \) as \( \mathcal{N}(\mu = 0, \sigma^2 = 1) \). Models \( g_2 \) and \( g_3 \) have fixed parameters \( \mu = 1 \) and \( \mu = 2 \) respectively. These parameters are assumed to be known and with \( I(f, g_2) \) and \( I(f, g_3) \) we get the true K–L information, a oriented distance between two known normal distributions with different parameters \( \mu \). The same applies for \( I(f, g_1) \) but with model \( g_1 \) we are pretty close to the true parameter value \( \mu_0 = 0 \) and therefore this model is doing the best job in approximating the truth \( f \).

As mentioned earlier, we will drop the truth \( f \) as a constant in the concept of real world problems. This fact will be motivated in the following few lines.

The K–L information between \( f \) and \( g \) can be written as

\[
I(f, g) = \int f(x) \log(f(x)) \, dx - \int f(x) \log(g(x|\theta)) \, dx
= E_f[\log(f(x))] - E_f[\log(g(x|\theta))].
\]
Therefore $I(f, g)$ can be seen as the difference between two statistical expectations. This expectation is taken with respect to the truth $f$. As a matter of fact, the previous two lines uncover an important conclusion: Whenever we will use $I(f, g)$ as a measure of distance between the truth $f$ and some approximating models $g_i$, the first expectation on the right will remain as a constant — a reference point so to say. This result motivates $I(f, g)$ as a measure of relative directed distance,

$$I(f, g) = C - \mathbb{E}_f[\log(g(x|\theta))],$$

or equivalently

$$I(f, g) - C = -\mathbb{E}_f[\log(g(x|\theta))].$$

Instead of taking $I(f, g)$ as a absolute measure, we have the opportunity to go with the relative measure $I(f, g) - C$ and therefor the term $\mathbb{E}_f[\log(g(x|\theta))]$ becomes the measurement of interest.

Assume two models $g_1$ and $g_2$, and $I(f, g_1) < I(f, g_2)$, then $I(f, g_1) - C < I(f, g_2) - C$, and thus $-\mathbb{E}_f[\log(g_1(x|\theta))] < -\mathbb{E}_f[\log(g_2(x|\theta))]$.

The previous lines put the relative K–L information, $-\mathbb{E}_f[\log(g(x|\theta))]$, into the center of attraction. Nevertheless in real world situations we would not simply come up with some parameters $\theta$ - we would rather want to come up with a good estimation of $\theta_0$ right away. Therefore we want to infer $\theta_0$ according to the given data at hand. However, we have not assumed any given data yet. In Example 7.5 we have just calculated the distance between theoretical densities.

The last part of this section will deal with the upcoming consequences if one takes parameter estimation into account. The observed data, given as $x_{1:n}$, has to be seen as a set of $n$ independent realizations of the true random variable $X$. In Example 7.5 we would write the maximum likelihood estimate (MLE) of given data as $\hat{\mu}$. Now, it is worth considering that $\hat{\mu}$ will be different for every sample of size $n$. If we would have the capacity to sample the whole population, in the

---

**Figure 7.1:** The probability distributions of four normal models
sense that \( n \to \infty \), one would assume that also \( \hat{\theta} \) reaches the true \( \mu_0 \). This property in likelihood theory is known as consistency of the maximum likelihood estimate. However, we will never have access to the whole sample population and therefore we will never see \( g(x|\theta_0) \). What we will see, is a estimate of the form \( g(x|\hat{\theta}) \). This estimated model will lead to an estimated relative K–L information \(-E_f[\log(g(x|\hat{\theta}))]\). Here we have to notice a subtly but important detail: If one would just use \(-E_f[\log(g(x|\theta_0))]\) as a measure of relative K–L information, one would not account for the fact that the maximum likelihood estimate \( \hat{\theta} \) is actually equipped with additional randomness. Given that context we would have estimated \( \hat{\theta} \) based on a realization of the random sample \( X_{1:n} \) and according probability density function given as

\[
g(x_{1:n}; \theta) = \prod_{i=1}^{n} g(x_i|\theta) \tag{7.10}
\]

Following this line of reasoning one would write the estimated relative K–L information based on a realization of \( X_{1:n} \) as

\[
[H(f,g)]_{1:n} = -E_x[\log(\prod_{i=1}^{n} g(x_i|\hat{\theta}))] = -\int f(x)\left(\sum_{i=1}^{n} \log(g(x_i|\hat{\theta}))\right) dx = -\left(\int f(x_1) \log(g(x_1|\hat{\theta})) dx + \cdots + \int f(x_n) \log(g(x_n|\hat{\theta})) dx\right). \tag{7.11}
\]

The random variables \( X_i \) are also said to be identically distributed and therefore we have \( x_i = x, \forall i \in \{1, \ldots, n\} \) upon taking the expectation with respect to \( f \). As a result, The estimated relative K–L information of a sample of size \( n \) can be written as:

\[
[H(f,g)]_{1:n} = n[H(f,g)].
\]

We will write this estimated relative K–L information of a random sample as

\[
[H(f,g)]_{1:n} = -E_{x_{1:n}}[\log(g(x_{1:n}|\hat{\theta}))].
\]

So once I have assumed a structure \( g \), the given data of size \( n \) does only affect \([H(f,g)]_{1:n}\) through the random component \( \hat{\theta}(Y) \). In other terms, \( x_{1:n} \) become the identical integrating variables referring to the output domain of the random variable \( X \) and we introduced a second random variable \( Y \), which takes over the role of the data. This fact allows us to differentiate between the integrating variables \( x_{1:n} \) and the actual data, donated as \( y_{1:n} \). So the estimated relative K–L information can be seen as a random variable given as \(-E_{x_{1:n}}[\log(g(x_{1:n}|\hat{\theta}(Y)))]\), where the random sample \( Y_{1:n} \) does clearly not affect the integrating variables \( x_{1:n} \). One could also argue that \( x_{1:n} \), in the estimated relative K–L information, will be integrated out anyway no matter what kind of realization of \( Y_{1:n} \) we get. In a next step it is natural to take

\[
T_0 = -E_{\hat{\theta}}[E_{x_{1:n}}[\log(g(x_{1:n}|\hat{\theta}(y)))] \tag{7.12}
\]

as a selection target for applied model selection approaches. In Equation (7.12) the MLE is with respect to \( y_{1:n} \) and the two expectations are for the random samples \( X_{1:n} \) and \( Y_{1:n} \) (therefore \( \hat{\theta} \),
both taken with respect to the truth \( f \). Further we have \( \hat{\theta}(Y_{1:n}) \) is independent on \( X_{1:n} \) in the context of that double expectation. Thus, if one assumes that \( g \) is in fact the truth (Example 7.5) and forced to use the MLE as an estimator for \( \theta_0 \), one rather has to minimize the expected estimated K–L information \( (T_0) \). In the derivation of AIC we will see, where the justification of the assumption \( f = g \) comes to play.

If one considers the developed theory regarding information theory, this expected value can be interpreted as the expected estimated cross entropy between the truth \( f \) and the approximating model \( g \). From this point of view, it gets clear right away, that the model \( g_i \) with the smaller expected estimated cross entropy will also have the smallest absolute \( I(f, g_i) \). It was Hirotsugu Akaike who proposed to estimate \(-E_f [E_f \log(g(x_{1:n} | \hat{\theta}(y)))]\) to get the applied K–L based model selection criteria. We will see in the next section how we can actually estimate the expected estimated cross entropy. A careful and comprehensive treatment of AIC was given by Burnham and Anderson (2002).

### 7.2 From K–L information to AIC

The main purpose of this part is to get a useful expression of the expected estimated cross entropy for each approximating model \( g_i \), given as

\[
T = E_{\hat{\theta}(Y)} \left[ E_X[\log(g(\theta | \hat{\theta}(y)))] \right] = \int_{D_2} f(\hat{\theta}) \left[ \int_{D_1} f(x) \log(g(x | \hat{\theta})) \, dx \right] d\hat{\theta}.
\] (7.13)

The above selection target \( T \) is a generalized version of the target given in Equation (7.12). For now we will omit the minus term and we will add it at the very end. To catch up with the explanations given above, we still recognize two independent random samples of the form \( X \) and \( Y \). Where \( Y \) can be seen as the \( n \)-dimensional random vector given as

\[
Y = (Y_1, Y_2, \ldots, Y_n)^T
\]

and with \( \hat{\theta} \) we have the \( K \)-dimensional parameter vector given as

\[
\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)^T.
\]

The estimated cross entropy will vary with different realizations of the random vector \( Y \) and we will again highlight this fact with

\[
\hat{\theta}(Y) = (\hat{\theta}_1(Y), \hat{\theta}_2(Y), \ldots, \hat{\theta}_k(Y))^T.
\]

The random vector \( X \) can be seen as the \( n \)-dimensional integrand in the \( n \)-dimensional integral over the inner domain \( D_1 \). Where the inner integration domain covers the whole sample space of \( X_i \), where \( i \in (1, \ldots, n) \). Upon taking the second expectation, integration will cover the whole parameter space \( \Theta \) through the outer domain \( D_2 \) in any of the directions \( \hat{\theta}_j, j \in (1, \ldots, K) \). From a conceptual point of view, we would have to integrate out \( X \) as a first step, which results in a function only dependent on \( \hat{\theta} \). The next step would again involve integration over \( D_2 \) to get rid of \( \hat{\theta} \) (hence \( y \)) — The result will be the real valued \( T \). The main problem is now to come up with a useful estimator of the selection target \( T \).
The first step involves a second order Taylor series expansion for the multivariable function \( \log(g(x|\hat{\theta})) \) around \( \theta_0 \) near \( \hat{\theta} \):

\[
\log(g(x|\hat{\theta})) \approx \log(g(x|\theta_0)) + \left[ \frac{\partial \log(g(x|\theta_0))}{\partial \theta} \right] \mathbf{T} \left[ \hat{\theta} - \theta_0 \right] + \frac{1}{2} \left[ \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2} \right] \left[ \hat{\theta} - \theta_0 \right].
\]

(7.14)

In this first approximation, the term

\[
\left[ \frac{\partial \log(g(x|\theta_0))}{\partial \theta} \right]
\]

represents the \( K \times 1 \) column vector of the first partial derivatives of \( \log(g(x|\hat{\theta})) \) with respect to \( \theta_1, \ldots, \theta_K \) evaluated at \( \theta = \theta_0 \):

\[
\left[ \frac{\partial \log(g(x|\theta_0))}{\partial \theta} \right] = \left( \frac{\partial \log(g(x|\theta_0))}{\partial \theta_1}, \ldots, \frac{\partial \log(g(x|\theta_0))}{\partial \theta_K} \right)^T.
\]

Further we have involved the notation

\[
\left[ \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2} \right] = \left[ \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta_i \partial \theta_j} \right]; \quad i \in \{1, \ldots, K\}, \quad j \in \{1, \ldots, K\},
\]

which denotes the \( K \times K \) matrix of second mixed partial derivatives of \( \log(g(x|\hat{\theta})) \) with respect to \( \theta_1, \ldots, \theta_K \), evaluated at \( \theta = \theta_0 \). The derived selection target \( T \) motivates us to take the expectation of the second order approximation with respect to \( X \):

\[
E_x \left[ \log(g(x|\hat{\theta})) \right] \approx E_x \left[ \log(g(x|\theta_0)) \right] + E_x \left[ \left[ \frac{\partial \log(g(x|\theta_0))}{\partial \theta} \right] \mathbf{T} \left[ \hat{\theta} - \theta_0 \right] \right] + \frac{1}{2} \left[ \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2} \right] \left[ \hat{\theta} - \theta_0 \right].
\]

(7.15)

In Equation (7.15) we basically used a second order extension to approximate the estimated relative K–L information. Now, if one assumes that \( g \) resembles the truth, in the sense that \( g = f \), we can use the following lines to simplify Equation (7.15):

\[
\frac{\partial}{\partial \theta} \int f(x) \log \left( \frac{f(x)}{g(x|\theta_0)} \right) \, dx = \frac{\partial}{\partial \theta} \int f(x) \log(f(x)) \, dx - \frac{\partial}{\partial \theta} \int f(x) \log(g(x|\theta_0)) \, dx = 0.
\]

(7.16)

The term \( \int f(x) \log(f(x)) \, dx \) in (7.16) was said to be a constant and therefore its partial derivative is zero. In conclusion we have:

\[
E_x \left[ \left[ \frac{\partial}{\partial \theta} \log(g(x|\theta_0)) \right] \right] = 0.
\]

Using this, together with the assumed independence of \( \hat{\theta} \) and \( X \), we can write (7.15) as:

\[
E_x \left[ \log(g(x|\hat{\theta})) \right] \approx E_x \left[ \log(g(x|\theta_0)) \right] - \frac{1}{2} \left[ \hat{\theta} - \theta_0 \right]^T J(\theta_0) \left[ \hat{\theta} - \theta_0 \right].
\]

(7.17)

In Equation (7.17) we made use of the fact that the expected Fisher information matrix \( J(\theta_0) \) evaluated at the true parameter \( \theta_0 \) is given by:

\[
J(\theta_0) = E_x \left[ -\left[ \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2} \right] \right] = E_x \left[ I(\theta_0) \right] = I(\theta_0).
\]

(7.18)
This expectation is also taken with respect to the truth $f$ and the above statement is valid under the assumption $f = g$. $I(\theta)$ denotes the Fisher information matrix of $\theta$ belonging to the true density $f$.

In a next step we can take the expectation of (7.17) with respect to $\hat{\theta}$ and we observe

$$E_{\hat{\theta}}E_x \left[ \log(g(x|\hat{\theta})) \right] \approx E_x \left[ \log(g(x|\theta_0)) \right] - E_{\hat{\theta}} \left[ \frac{1}{2} (\hat{\theta} - \theta_0)^T J(\theta_0) (\hat{\theta} - \theta_0) \right]$$

$$= E_x \left[ \log(g(x|\theta_0)) \right] - \frac{1}{2} \text{tr} \left[ J(\theta_0) E_{\hat{\theta}} \left[ (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \right] \right]$$

$$= E_x \left[ \log(g(x|\theta_0)) \right] - \frac{1}{2} \text{tr} \left[ J(\theta_0) \Sigma \right].$$

In the given equation the occurrence of the trace function can be explained as follows: Consider a column vector $v \in \mathbb{R}^K$ where a quadratic form fulfills

$$vQv^T = \text{tr} \left[ Qv v^T \right]; \quad Q \in \mathbb{R}^{K \times K}.$$

Taking expectation with respect to $v$ we get

$$E_v \left[ vQv^T \right] = \text{tr} \left[ E_v \left[ Qv v^T \right] \right].$$

Where the last equality holds because the trace is linear.

We can summarize the result of this first part as:

$$T \approx E_x \left[ \log(g(x|\theta_0)) \right] - \frac{1}{2} \text{tr} \left[ J(\theta_0) \Sigma \right]. \quad (7.19)$$

There is still one problem in Equation (7.19). We have approximated the expected estimated relative K–L information in terms of $\theta_0$, the true unknown parameter vector. In a next step we have to somehow relate the given target $T$ to the actual log-likelihood evaluated at $\hat{\theta}$. This approach makes sense because large sample theory will ensure that $\hat{\theta} = \theta_0$ for $n \to \infty$. Therefore we will do a second expansion, but this time of $\log(g(x|\theta_0))$ around $\hat{\theta}(X)$. In a sense this step refers to the usual likelihood based approach: Getting the MLE based on a sample $x_{1:n}$ through maximizing the log-likelihood. Note, that the log-likelihood of $\theta$ given a sample $x_{1:n}$ can be written in the multivariate framework:

$$l(\theta; x) = \log(f(x; \theta)).$$

Here we can again treat the vector $x$ as our data, hence $x$ is still a realization of the random vector $X$ and since we are after a expectation with respect to $x$ we can easily switch between the notation $y$ and $x$. We obtain the second expansion as:

$$\log(g(x|\theta_0)) \approx \log(g(x|\hat{\theta})) + \left[ \frac{\partial \log(g(x|\hat{\theta}))}{\partial \theta} \right]^T [\theta_0 - \hat{\theta}]$$

$$+ \frac{1}{2} [\theta_0 - \hat{\theta}]^T \left[ \frac{\partial^2 \log(g(x|\hat{\theta}))}{\partial \theta^2} \right] [\theta_0 - \hat{\theta}]. \quad (7.20)$$

The MLE $\hat{\theta}$ satisfies the score equations

$$\frac{\partial \log(g(x|\hat{\theta}))}{\partial \theta} = 0.$$
Upon taking the necessary expectation, we get
\[
E_x \left[ \log(g(x|\theta_0)) \right] \approx E_x \left[ \log(g(x|\tilde{\theta})) \right] - \frac{1}{2} \text{tr} \left[ E_x \left[ I(\tilde{\theta})|\theta_0 - \tilde{\theta}|[\theta_0 - \tilde{\theta}]^T \right] \right].
\]

Now consistency of the ML estimator ensures
\[
I(\tilde{\theta}) = I(\theta_0),
\]
for \( n \to \infty \)— the observed Fisher information \( I(\tilde{\theta}) \) evaluated at the MLE approaches the Fisher information at the true parameter vector. As a consequence, we can write (7.19) as:
\[
E_x \left[ \log(g(x|\theta_0)) \right] \approx E_x \left[ \log(g(x|\tilde{\theta})) \right] - \frac{1}{2} \text{tr} \left[ E_x \left[ I(\theta_0)|\theta_0 - \tilde{\theta}|[\theta_0 - \tilde{\theta}]^T \right] \right].
\]

Recall that the first expansion (7.19) was of the form
\[
T \approx E_x \left[ \log(g(x|\theta_0)) \right] - \frac{1}{2} \text{tr} \left[ J(\theta_0)\Sigma \right].
\]

If one substitutes (7.19) into (7.21) we approach a promising result:
\[
T \approx E_x \left[ \log(g(x|\tilde{\theta})) \right] - \text{tr} \left[ J(\theta_0)\Sigma \right].
\]

Note, in this equation \( x \) refers to the observed data and we again consider one single random sample \( X \) involved. The final goal is to come up with a proper, therefore unbiased estimate of \( T \). The right term is simply the expected value of the log-likelihood evaluated at the MLE and therefore we can use \( \log(g(x|\tilde{\theta})) \) as an unbiased estimate for \( E_x \left[ \log(g(x|\tilde{\theta})) \right] \). This brings us one step further and we write:
\[
\hat{T} \approx \log(g(x|\tilde{\theta})) - \hat{\text{tr}}[J(\theta_0)\Sigma].
\]

Now in a last step we recall that the true covariance matrix is given by
\[
\Sigma = I(\theta_0)^{-1}.
\]

Taking all together we can write a unbiased estimate of the expected estimated cross entropy as
\[
\hat{E}_x \left[ [H(f;g)]_{1:n} \right] \approx - \log(g(x|\tilde{\theta})) + \hat{\text{tr}}[I_n] = - \log(g(x|\tilde{\theta})) + K,
\]
where \( I_n \) represents the \( K \times K \) identity matrix and we recovered the minus sign. This selection criteria was developed under the assumption \( f = g \) and therefore justification of the term \( K \) as an estimator of \( \hat{\text{tr}}[J(\theta_0)\Sigma] \) becomes valid by considering a model structure \( g \), which comes
close to the true model $f$. Observing the structure of $\hat{T}$ reveals another useful property: If one just adds more parameters to the approximating model, the first term of (7.25) tends to decrease slow compared to the increase of the second term on the right. Therefore the given model selection criterion is equipped with an internal balance between under and overfitting and we often talk about a penalty term $K$.

If one multiplies the derived model selection criteria by a factor of two one gets

$$AIC = -2 \log(g(x|\hat{\theta})) + 2K. \quad (7.26)$$

The factor two was added for “historical reasons”, to somehow resemble the likelihood ratio test statistic which has a proper distribution under $H_0$.

We will now briefly introduce a related model selection criterion which has its justification in the Bayesian context.

### 7.3 Bayesian information criterion

Very similar in structure, we define the Bayesian information criterion

$$BIC = -2 \log(g(x|\hat{\theta})) + \log(n)K. \quad (7.27)$$

Here $n$ is present as the sample size and somehow $n$ affects the penalty term on the right. Comparing $AIC$ and BIC we realize that for $n \geq 8$ the factor two in the penalty of $AIC$ gets overtaken. This is an advantage for BIC compared to AIC if one is interested to penalize to complex models, complex in a sense that there are too many parameters involved in the model structure. This specific character of BIC comes to play, when we intend to compare several models which are close to the truth $f$. In that situation we might be not interested in the subtle details concerning the differences in the log-likelihoods. Therefore these models are anyway somehow sufficient and we have reached a level where we just want to stick to a simple model. $AIC$ will penalize exactly these rather complex models with its penalty term given as $2K$. BIC will do the same job as AIC considering the first term on the right of (7.27) but it will magnifies the difference between a overfitted model and a simpler version of it given a sample size $n$. This is a nice quality of the Bayesian counterpart in model section and will lead to a clear decision against the overfitted model.
Chapter 8
Bayesian model selection

By: Silvia Panunzi

8.1 Introduction

In Chapter 7 we studied Model Selection from a Non-Bayesian point of view. The Bayesian approach here presented takes into account two additional components: the prior parameters distributions in the different models and the prior probabilities assigned at each model that is going to be evaluated. In this chapter we first provide the definition of the Bayes factor (Kass and Raftery, 1995), the Bayesian information criterion (BIC; Schwarz, 1978), the Deviance information criterion (DIC; Spiegelhalter et al., 2002) and we finally discuss the mechanism of Bayesian Model Averaging to treat the model uncertainty (BMA; Hoeting et al., 1999). We mainly follow the structure elaboration used in Ando (2010).

8.2 General framework

Suppose that we have a set of models $M_1, \ldots, M_r$ and we want to select the best one from them. Each model $M_k$ is characterized by a probability density $f_k(X_n|\theta_k)$, where $\theta_k$ is a $p_k$ dimensional vector of unknown parameters with prior probability $\pi_k(\theta_k)$.

The idea of the Bayesian approach is to choose the model with the highest posterior probability among the candidates. We define the posterior probability of the model $M_k$ for a specific dataset $X_n = \{x_1, \ldots, x_n\}$ given by

$$
\Pr(M_k|X_n) = \frac{\Pr(M_k) \int f_k(X_n|\theta_k)\pi_k(\theta_k)d\theta_k}{\sum_{j=1}^r \Pr(M_j) \int f_j(X_n|\theta_j)\pi_j(\theta_j)d\theta_j}.
$$
where \( f_k(X_n|\theta_k) \) and \( \Pr(M_k) \) represent the likelihood function and the prior probability for model \( M_k \), respectively; the latter together with the prior distribution for the unknown parameters \( \pi_k(\theta_k) \) specify the initial model uncertainty. Choosing a model that maximizes \( \Pr(M_k|X_n) \) is equal to choose the one with maximum

\[
\Pr(M_k) \int f_k(X_n|\theta_k)\pi_k(\theta_k)d\theta_k.
\]

The quantity

\[
\Pr(X_n|M_k) = \int f_k(X_n|\theta_k)\pi_k(\theta_k)d\theta_k,
\]

is the marginal probability of the data under model \( M_k \), it is also called integrated likelihood because it is obtained by integrating (not maximizing) the joint density of \((X_n, \theta_k)\) over \( \theta_k \). It is also the predictive probability of seeing the data that actually were observed, calculated before any data became available. Another relevant aspect to deal with, is the choice of the prior model probabilities \( \Pr(M_1), \ldots, \Pr(M_r) \), the simplest and popular way is to choose uniform priors

\[
\Pr(M_k) = \frac{1}{r}.
\]

This is a non-informative prior because it treats all models equally. However, in general, we might consider some characteristics of the models that in this case are ignored, for example the number of parameters. In linear regression models we place higher probability on more parsimonious models and we specify a Poisson prior distribution

\[
\Pr(M_k) \propto \lambda^{p_k}\exp(-\lambda),
\]

where \( p_k \) is the number of predictors in the current model and the parameter \( \lambda \) adjust the expectation of \( p_k \). Naturally the property \( \Pr(M_1) + \cdots + \Pr(M_r) = 1 \) has always to be satisfied.

### 8.3 Bayes factors

#### 8.3.1 Definition

The Bayes factor is one of the fundamental quantity for comparing models and for testing hypothesis in Bayesian analysis. It is defined as the ratio of marginal probabilities of two models, in this sense it allows us to make pairwise comparison of models based on the posterior probabilities. Suppose we want to measure the evidence for a model \( M_k \) versus \( M_j \) given the data, we can compute the Bayes factor:

\[
BF_{kj} = \frac{Pr(X_n|M_k)}{Pr(X_n|M_j)}.
\]

We can easily note that

\[
\frac{Pr(M_k|X_n)}{Pr(M_j|X_n)} = \frac{Pr(X_n|M_k)}{Pr(X_n|M_j)} \times \frac{Pr(M_k)}{Pr(M_j)}
\]

therefore the Bayes factor is also given as the ratio of posterior odds and prior odds

\[
BF_{kj} = \frac{Pr(M_k|M_n)}{Pr(M_j|M_n)} \div \frac{Pr(M_k)}{Pr(M_j)} = \frac{\text{Posterior Odds}(M_k|M_j)}{\text{Prior Odds}(M_k|M_j)}.
\]
8.3. BAYES FACTORS

Table 8.1: Kass and Raftery’s scale of evidence for Bayes Factor($M_k, M_j$).

<table>
<thead>
<tr>
<th>$2\log_e(BF_{kj})$</th>
<th>$(BF_{kj})$</th>
<th>Evidence against $M_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 2</td>
<td>1 – 3</td>
<td>Barely worth mentioning evidence</td>
</tr>
<tr>
<td>2 – 6</td>
<td>3 – 20</td>
<td>Positive</td>
</tr>
<tr>
<td>6 – 10</td>
<td>20 – 150</td>
<td>Strong</td>
</tr>
<tr>
<td>$&gt; 10$</td>
<td>$&gt; 150$</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

When $M_k$ and $M_j$ are equally probable a priori so that $Pr(M_k) = Pr(M_j)$, the Bayes factor reduces to the the posterior odds in favor of model $M_k$. When the marginal probabilities of the model compared are single distributions with no free parameters, the Bayes factor is the classical Likelihood Ratio. Indicating as $\theta_{MLEk}$ and $\theta_{MLEj}$ the maximum likelihood estimates for the two models and supposing that the prior densities $\pi_k(\theta_k)$ and $\pi_j(\theta_j)$ consist of a point mass at those values, we reduces to

$$BF_{kj} = \frac{f_k(X_n|\theta_{MLEk})}{f_j(X_n|\theta_{MLEj})} = LR.$$  

However in the case of the likelihood ratio statistic we know that the parameters are eliminated by maximization rather than by integration and we also know the models have to be nested, here this condition is not required.

8.3.2 Interpretation

Kass and Raftery suggested interpreting twice the natural logarithm of the Bayes factor to have it in the same scale as the familiar deviance and likelihood ratio test statistics. Table 8.1 gives the categories that seem to be appropriate guidelines and a rough descriptive statement about standards of evidence in scientific investigation (Kass and Raftery, 1995).

Here they speak in term of evidence against model $M_j$, but we can equally speak in term of evidence in favor of model $M_k$. It is also interesting to view the logarithm of the marginal probability of the data as a predictive score and interpret the log Bayes factor as the difference in predictive scores; such approach allows us to regardless on viewing one of the models as the “true” and design the Bayes factor to choose the model that will, on average, give better out of sample predictions. To deep in the concept of the log-BF as “weight of evidence” we remand on Good (1991).

8.3.3 Calculating the Bayes factor

In most of the cases the integral (8.1) is intractable analytically and must be computed by numerical methods, this can be avoided in conjugate families. We rewrite it leaving out the conditioning on $M_k$ for readability as $Pr(X_n) = \int f(X_n|\theta)\pi(\theta)d\theta$ and noting that the marginal likelihood is the denominator of the posterior distribution, we have

$$Pr(X_n) = \frac{f(X_n|\theta)\pi(\theta)d\theta}{\pi(\theta|X_n)}.$$  

As we have seen in Chapter 2 in conjugate cases the proportionality constants $f(X_n|\theta)$, $\pi(\theta)$ and $\pi(\theta|X_n)$ are known (see Table 2.1) and $Pr(X_n)$ can be computed.
8.3.4 Bayes factor and the choice of priors

As we have previously seen, in order to calculate the Bayes factor the prior distributions $\pi_k(\theta_k)$ on the parameters of each model must to be specified; in contrast with Bayesian point estimates, the Bayes factors tend to be sensitive to that choice. In Bayesian estimation priors are often picked for convenience because in large samples their effect is asymptotically negligible, in Bayesian model selection this is not so.

From the definition of the marginal probability we already know that improper priors cannot be applied otherwise $\Pr(X_n|M_k)$ would be indeterminate. Another issue occurs using a prior with very large spread, also called vague prior, in the effort to make it non-informative. Comparing two nested models it can be shown that this choice will force the Bayes factor to favor the simplest one because the posterior probability of this will always converge to 1. This is the so called Lindley’s paradox and describes a situation in which Bayesian and frequentist approaches to a hypothesis testing problem give different results (Kass and Raftery, 1995).

As Kass and Raftery (1995) specify, when an estimate is found to be distant from the null hypothetical value the data seem to disagree with the null hypothesis, but it can happen that BF turns out to be in favor of this hypothesis and this means the data are more unlikely under the alternative one.

To deal with the prior problem several approaches has been developed. Some authors propose Pseudo-Bayes factors that can be adapted in different contests but here we are going to deep in as alternative the use of the Bayesian Information Criteria introduced in Chapter 7, under the assumption that $\log \pi(\theta) = O_p(1)$. We will show that under a large sample situation this prior specification has no effect on the approximation error of the marginal likelihood by this criterion.

8.4 Marginal likelihood approximation

When the integral (8.1) cannot be evaluated analytically, we can use the Laplace’s Method explained in Chapter 5 to derive an approximation of the marginal likelihood. What we have to assume is that the posterior density $\pi(\theta|X_n)$ (proportional to the quantity under the integral) is highly peaked about its maximum, which is the posterior mode $\tilde{\theta}$; this will usually be the case if the likelihood function is highly peaked near its maximum (case of large samples). Recall that theta is defined as a vector of parameters. We also define $s(\theta) = \log[f(X_n|\theta)\pi(\theta)]$, with $s(\tilde{\theta})' = 0$.

We derive the Taylor expansion of $s(\theta)$ about $\tilde{\theta}$ as

$$s(\theta) = s(\tilde{\theta}) - \frac{n}{2}(\theta - \tilde{\theta})^T s(\tilde{\theta})(\theta - \tilde{\theta}),$$

and the exponential approximation

$$\exp\left\{ s(\theta) \right\} \approx \exp\left\{ s(\tilde{\theta}) \right\} \times \exp\left\{ \frac{n}{2}(\theta - \tilde{\theta})^T s(\tilde{\theta})(\theta - \tilde{\theta}) \right\}.$$

In the right part of the equation we can recognize the multivariate normal density with mean $\tilde{\theta}$ and covariance matrix $n^{-1}s(\tilde{\theta})^{-1}$, where $s(\tilde{\theta})$ is minus the Hessian of $n^{-1}\log[f(X_n|\theta)\pi(\theta)]$.
evaluated at the posterior mode. The integration of this approximation yields
\[
\Pr(X_n|M) = \int \exp \left\{ s(\theta) \right\} \approx f(X_n|\hat{\theta})\pi(\hat{\theta}) \times \frac{(2\pi)^{p/2}}{n^{p/2}|S(\hat{\theta})|^{1/2}},
\]
(8.2)
where \( p \) is the dimension of \( \theta \). The Bayes factor can be now approximated as
\[
\text{BF}_{kj} = \frac{P(X_n|M_k)}{P(X_n|M_j)} \approx \frac{f_k(X_n|\hat{\theta}_k)\pi_k(\hat{\theta}_k)}{f_j(X_n|\hat{\theta}_j)\pi_j(\hat{\theta}_j)} \times \frac{|S_j(\hat{\theta}_j)|^{1/2}}{|S_k(\hat{\theta}_k)|^{1/2}} \times \frac{2\pi}{n} \frac{p_k-p_j}{2},
\]
therefore it depends on the ratio of the likelihood function evaluated at the posterior modes, the ratio of the prior evaluated at the posterior modes, on the difference between the dimensions of the models and on the Hessian terms (Ando, 2010).

### 8.5 BIC derivation

Considering the case of \( \log \pi(\theta) = O_p(1) \), that means the prior information can be ignored for a sufficiently large sample size, and assuming that we are in presence of a set of \( n \) independent observations, the formula in (8.2) can be transformed replacing \( \hat{\theta} \) with the MLE \( \hat{\theta} \)
\[
\Pr(X_n|M) \approx f(X_n|\hat{\theta})\pi(\hat{\theta}) \times \frac{(2\pi)^{p/2}}{n^{p/2}|J(\hat{\theta})|^{1/2}},
\]
where \( J(\hat{\theta}) \) is the observed information matrix, the negative Hessian of \( n^{-1} \log f(X_n|\theta\pi(\theta)) \) evaluated at the maximum likelihood estimator.

Taking minus twice the logarithm of the previous formulation we obtain
\[
-2\Pr(X_n|M) \approx -2\log f(X_n|\hat{\theta}) - 2 \log \pi(\hat{\theta}) + p \log(n) + \log |J(\hat{\theta})| - p \log(2\pi),
\]
then ignoring the term of order \( O_p(1) \) and higher order terms we find the Schwarz’s Bayesian Criterion:
\[
-2 \log[\Pr(X_n|M)] \approx -2 \log f(X_n|\hat{\theta}) + p \log(n) = \text{BIC}.
\]
(8.3)
From this derivation we can also have a rough approximation of the logarithm of the Bayes factor
\[
\log(BF_{kj}) = \log[\Pr(X_n|M_k)] - \log[\Pr(X_n|M_j)] \approx (\text{BIC}_j - \text{BIC}_k)/2.
\]
The error of the approximation by BIC (8.3) remain constant for increasing sample size, while the approximation error relative to the true Bayes factor tends to 0:
\[
\frac{2\log[\Pr(X_n|M)] - \text{BIC}}{2\log[\Pr(X_n|M)]} \to 0,
\]
because both the log marginal likelihood and the BIC are increasing when \( n \to \infty \).

We can conclude that selecting the model with smallest BIC is asymptotically equivalent in selecting the model with the largest posterior probability.

For a particular choice of prior distribution, it is demonstrated that the error is of a much smaller order of magnitude than \( O_p(1) \). The prior here discussed is called Unit Information Prior \( \theta \sim N_p(\theta_0, J_1(\theta_0)^{-1}) \) and it is designed to contain the same amount of information as one observation, this quantity of information is expressed by the expected Fisher Information \( J_1(\theta_0) \) at the prior mean. In this case the error of the approximation \(-2 \log[\Pr(X_n|M)] - \text{BIC} = O_p(n^{-1/2})\).
8.6 Deviance information criterion

Another criterion quite often used in the Bayesian framework is the Deviance Information Criterion (DIC), this is considered a particular useful tool for model selection when the posterior distributions of the models have been obtained by Markov Chain Monte Carlo (MCMC) simulations (see Chapter 10).

In 2002 Spiegelhalter et al. (2002) pointed out that the posterior mean of the expected log likelihood can be taken as a Bayesian measure of fit. He defined a Bayesian measure of model complexity $pD$ the posterior mean of the deviance minus the deviance evaluated at the posterior mean $\hat{\theta} = E[\theta|y]$:

$$pD = D(\theta) - D(\hat{\theta}) = E_{\theta|X_n} - 2\log f(X_n|\theta) + 2\log f(X_n|\hat{\theta}),$$

that can be rewritten as

$$pD = 2\log f(X_n|\hat{\theta}) - 2 \int \log f(X_n|\theta)\pi(\theta|X_n) d\theta.$$

The Deviance Information Criterion is expressed as:

$$DIC = D(\hat{\theta}) + 2pD = D(\theta) + pD,$$

with the first term that is considered a Bayesian measure of ‘adequacy’ because it includes itself a measure of fit with a penalization $D(\hat{\theta}) = D(\hat{\theta}) + pD$. As for AIC and BIC models with lower DIC value have to be preferred over those with higher value.

Expanding the log likelihood around the posterior mean yields

$$\log f(X_n|\theta) \approx \log f(X_n|\bar{\theta}) + (\theta - \bar{\theta})^T L_\theta'(\theta - \bar{\theta}) + 1/2(\theta - \bar{\theta})^T L_\theta''(\theta - \bar{\theta}),$$

with $L_\theta'$ and $L_\theta''$ first and second derivatives with respect to $\theta$.

Taking the expectation with respect to the posterior distribution

$$\int \log f(X_n|\theta)\pi(\theta|X_n)d(\theta) \approx \log f(X_n|\bar{\theta}) + \frac{1}{2} tr(L_\theta'' V(\theta)),$$

with $V(\theta)$ that correspond to the posterior covariance matrix of theta.

For a large sample size and with prior information negligible, we know that we can approximate the posterior distribution $\pi(\theta|X_n)$ of the model parameters by a multivariate normal distribution with mean the posterior mode $\tilde{\theta}$ and covariance matrix the observed Fisher information $J(\tilde{\theta})$ matrix.

We can also replace the posterior mode with the posterior mean because for $n \to \infty$ they converge to the same value, in this way we have

$$pD \approx tr(J(\tilde{\theta})V(\theta)) \approx tr(I) = p,$$

i.e $pD$ will be approximate by the true number of parameters. Therefore DIC $\approx$ AIC for large sample sizes.

The advantage of this information criterion is that while AIC and BIC require the calculation of the likelihood at its maximum estimate, DIC simply compute the $\hat{\theta}$ as the average of the
samples $\theta^{(1)}, \ldots, \theta^{(B)}$ from the posterior. Then $\bar{\theta} \approx B^{-1}$ is approximated by the average of the samples. As well as the AIC and BIC, it is an asymptotically approximation that become more accurately increasing the sample size.

We have to specify that this approach assumes that the specified parametric family of probability distributions that generate future observations contains the true model. Furthermore, the data are used twice in the construction of $p_D$ therefore DIC tends to select overfitted models.

8.7 Bayesian model averaging

In the previous sections we have seen various types of Bayesian model selection criteria to select the best model among a set of candidates. However any approach that select a single model and then makes inference conditionally on it ignores the uncertainty in model selection. To incorporate model uncertainty into the decisions, a basic idea is to average the competing models, this method is called Bayesian Model Averaging (BMA) and it was here introduced at the end of Chapter 6.

Always suppose that $M_1, \ldots, M_r$ are being considered, and a quantity of interest $\Delta$ is defined for every model; its posterior distribution given model $M_k$ is

$$
Pr(\Delta|X_n, M_k) = \sum_{k=1}^{r} Pr(\Delta|X_n, \theta_k, M_k) Pr(\theta_k|X_n, M_k).
$$

This formula can be used to make inference about $\Delta$ conditionally on model $M_k$, but without conditioning on a determinate model we can obtain the posterior probability as

$$
Pr(\Delta|X_n) = \sum_{k=1}^{r} Pr(M_k|X_n) Pr(\Delta|X_n, M_k).
$$

(8.4)

In this way we account for the uncertainty about model form by weighting the conditional posterior densities according to the posterior probabilities of each model (Kass and Raftery, 1995).

Similarly we can obtain the model averaged estimates of quantities of interest, e.g the mean and variance:

$$
E(\Delta|X_n) = \sum_{k=1}^{r} Pr(M_k|X_n) E[\Delta|X_n, M_k] = \sum_{k=1}^{r} Pr(M_k|X_n) \hat{\Delta}_k,
$$

and

$$
Var(\Delta|X_n) = \sum_{k=1}^{r} Pr(M_k|X_n) \left[Var[\Delta|X_n, M_k] + E[\Delta|X_n, M_k]^2\right] - E[\Delta|X_n]^2
$$

$$
= \sum_{k=1}^{r} Pr(M_k|X_n) \left[Var[\Delta|X_n, M_k] + \hat{\Delta}_k^2\right] - E[\Delta|X_n]^2.
$$
8.7.1 BMA difficulties

Even if BMA seems to be an attractive solution to model uncertainty, it presents several difficulties, for example (Hoeting et al., 1999):

- the numbers in terms (8.4) can be enormous,
- the integral implicit in (8.4) can be hard to compute. In this case the MCMC method for simulations can be applied,
- specification of $\Pr(M_k)$ must be done carefully,
- choosing the class of models over which to average becomes a fundamental issue to deal with.

Under a situation where the number of averaging models $r$ involved is very large, one practical solution to reduce the computational amount is the Occam’s Window method. This method averages over a reduced set of parsimonious models excluding models that are much less likely than the most likely model and models containing effects for which there is no evidence, i.e., models that have more likely submodels nested within them. As Held and Sabanés Bové (2014) report, BMA arises naturally from integrating out the model from the posterior distribution, but there are also frequentist model averaging in which the weights are derived from information criteria, e.g., BIC. We know that BIC can be seen as an approximation of twice the log marginal likelihood of a model hence $\Pr(X_n|M_k) = \exp \Pr(-\text{BIC}_k/2)$, assuming $\Pr(M_k) = 1/r$ (giving equal probability to all candidate models) the weight $w_k$ for $M_k$ is its resulting approximate posterior probability

$$w_k = \frac{\exp \Pr(-\text{BIC}_k/2)}{\sum_{j=1}^{r} \exp \Pr(-\text{BIC}_j/2)}.$$

8.7.2 BMA examples

Example 8.1 (Raftery, 1995). Consider the situation where the quantity of interest is one of the regression parameters $\beta_1$, some of the models typically specify $\beta_1 = 0$ and so the posterior probability $\Pr(\beta_1 = 0|D) \neq 0$ (D here are our data). We are particularly interested in measuring

$$\Pr(\beta_1 \neq 0|D) = \sum_{k=1}^{r} \Pr(M_k|D),$$

where $M_1, \ldots, M_r$ is the set of models that include $\beta_1$.

Of interest it is also the size of the effect, given that it is nonzero. The posterior distribution of this is

$$\Pr(\beta_1|D, \beta_1 \neq 0) = \sum_{k=1}^{r} \Pr'(M_k|D) \Pr(\beta_1|D, M_k),$$

with $\Pr'(M_k|D) = \Pr(M_k|D)/\Pr(\beta_1 \neq 0|D)$. 
This can be summarized by its posterior means and standard deviation:

$$E(\beta_1|D, \beta_1 \neq 0) \approx \sum_{k=1}^{r} \Pr'(M_k|D)\hat{\beta}_1(k)$$

$$SD^2(\beta_1|D, \beta_1 \neq 0) \approx \sum_{k=1}^{r} \Pr'(M_k|D)[se^2_1(k) + \hat{\beta}_1(k)] - E[\beta_1|D, \beta_1 \neq 0]^2,$$

where $\hat{\beta}_1(k)$ and $se^2_1(k)$ are the MLE and standard error of $\beta_1$ under model $M_k$.

**Example 8.2** (Wasserman, 2000). We observe $n$ independent flips of a coin. Denote the outcomes by $Y_n = \{y_1, \ldots, y_n\}$ where each $Y_i$ is either 0 or 1 (corresponding to tails and heads on the coin). Let $P(Y_i = 1)$ be the unknown probability of observing $Y_i = 1$. We have two theories: theory one says that the coin is fair, i.e., $\theta = 1/2$; theory two says that the coin is not fair, i.e., $\theta \neq 1/2$. The probability function for a single toss is $Pr(y|\theta) = \theta^y(1 - \theta)^{1-y}$ where $y \in [0, 1]$. The two theories correspond to two sets of probability distributions:

$$M_1: \theta = 1/2,$$

$$M_2: \theta \in [0, 1], \theta \neq 1/2$$

First we want a measure of evidence in favor of a theory over the other:

$$BF_{21} = \frac{Pr(M_2|Y_n)}{Pr(M_1|Y_n)} \div \frac{Pr(M_2)}{Pr(M_1)},$$

choosing $Pr(M_2) = Pr(M_1) = 0.5$ the Bayes factor reduces to the posterior odds

$$BF_{21} = \frac{Pr(M_2|Y_n)}{Pr(M_1|Y_n)} = \frac{\int_0^1 L(\theta)\pi(\theta)}{L(1/2)},$$

where $L(\theta) = \prod_{i=1}^{n} Pr(y_i|\theta)$. We also know that for a particular choice of prior:

$$\log(BF) \approx \log L(\hat{\theta}) - \log L(1/2) - 1/2 \log(n) = b,$$

with $\hat{\theta} = \frac{\Sigma y_i}{n}$ that is the maximum likelihood estimate under $M_2$.

Therefore we have $Pr(M_1|Y_n) \approx 1/(1 + \exp Pr(b))$ and $Pr(M_2|Y_n) \approx \exp Pr(b)/(1 + \exp Pr(b))$ because $BF \approx \exp Pr(b)$. Now we want to make prediction on a future outcome:

$$Pr(Y_{n+1} = 1|Y_n) = \sum_{k=1}^{2} Pr(M_k|Y_n) Pr(Y_{n+1} = 1|Y_n, M_k)$$

$$= Pr(M_1|Y_n) Pr(Y_{n+1} = 1|Y_n, M_1) + Pr(M_2|Y_n) Pr(Y_{n+1} = 1|Y_n, M_2).$$

In the equation above we know that $Pr(Y_{n+1} = 1|Y_n, M_1) = 0.5$ because $M_1$ says that the coin is fair; $Pr(Y_{n+1} = 1|Y_n, M_2) = \int_0^1 \theta \pi(\theta|y_n)$ is the mean of $\theta$ under $M_2$ and $\pi(\theta|y_n) = L(\theta)\pi(\theta) \int_0^1 L(\theta)\pi(\theta)$ is the posterior density of $\theta$ under $M_2$.

For $n$ large: $Pr(Y_{n+1} = 1|Y_n, M_2) \approx \theta$ and we can write

$$\log(BF) \approx \log L(\hat{\theta}) - \log L(1/2) - 1/2 \log(n),$$
with $\hat{\theta} = \frac{\sum y_i}{n}$ that is the maximum likelihood estimate under $M_2$. Therefore we have $\Pr(M_1|Y_n) \approx 1/(1 + \exp Pr(b))$ and $\Pr(M_2|Y_n) \approx \exp Pr(b)/(1 + \exp Pr(b))$ because $BF \approx \exp Pr(b)$. Now we want to make prediction on a future outcome:

$$
\Pr(Y_{n+1} = 1|Y_n) = \sum_{k=1}^{2} \Pr(M_k|Y_n) \Pr(Y_{n+1} = 1|Y_n, M_k)
$$

$$
= \Pr(M_1|Y_n) \Pr(Y_{n+1} = 1|Y_n, M_1) + \Pr(M_2|Y_n) \Pr(Y_{n+1} = 1|Y_n, M_2).
$$

In the equation above we know that $\Pr(Y_{n+1} = 1|Y_n, M_1) = 0.5$ because $M_1$ says that the coin is fair; $\Pr(Y_{n+1} = 1|Y_n, M_2) = \int_0^1 \theta \pi(\theta|y_n)$ is the mean of $\theta$ under $M_2$ and $\pi(\theta|y_n) = L(\theta)\pi(\theta)/\int_0^1 L(\theta)\pi(\theta)$ is the posterior density of $\theta$ under $M_2$.

For $n$ large: $\Pr(Y_{n+1} = 1|Y_n, M_2) \approx \hat{\theta}$ and we obtain:

$$
\Pr(Y_{n+1} = 1|Y_n) \approx \frac{1}{2} \times \frac{1}{1 + \exp Pr(b)} + \hat{\theta} \times \frac{\exp Pr(b)}{1 + \exp Pr(b)}.
$$
Chapter 9
Hierarchical models

By:
Rasha Aktaa

9.1 Hierarchical models

Hierarchical models are mathematical descriptions of complex data structures involving multiple parameters such that the values of some parameters meaningfully depend on the values of other parameters. A hierarchical model allows us to entertain a much richer class of models that can better capture our statistical understanding of the problem than simpler approaches. We introduce hierarchical models using the following example:

Example 9.1. Let’s consider students (data) grouped into classes (2nd level), which are grouped into schools (1st level). If we study the students’ scores (which have the parameters $\beta$, $\sigma^2$) in math, for example, they will depend on the teaching style (which has parameters $\mu$, $\tau^2$), which is dependent on the annual income of the school, for independent probability distributions $F_\sigma$ and $F_\tau$. The joint density for the model parameters $(\beta, \mu, \sigma^2, \tau^2)$ is

$$p(\beta, \mu, \sigma^2, \tau^2) = p(\beta \mid \mu, \tau^2)p(\mu)p(\sigma^2)p(\tau^2).$$

We consider that:

1st level : $\beta \mid \mu, \tau^2 \sim N(\mu, \tau^2)$,
2nd level : $\mu \sim N(b_0, B_0)$,
$\sigma^2 \sim F_\sigma$ and $\tau^2 \sim F_\tau$.

Example 9.2. Another example that was discussed in Gamerman and Lopes (2006), on page: 60, 2.4, showed normal or classical case, they considered observations $y_{ij} \sim N(\beta_i, \sigma^2 I), j =$
1, \ldots, n_i, i = 1, \ldots, d$, collected from $d$ groups with different means $\beta_i$ but the same dispersion. A generalization towards a normal regression model is given by:

\[
y \mid \beta_1, \phi \sim \mathcal{N}(X_1 \beta_1, \phi^{-1} I_n),
\]

\[
\beta_1 \mid \beta_2 \sim \mathcal{N}(X_2 \beta_2, C),
\]

\[
\beta_2 \sim \mathcal{N}(b, B),
\]

\[
\phi \sim \Gamma(n_0/2, n_0\sigma^2_0/2),
\]

for appropriate design matrices $X_1$ and $X_2$, covariance matrices $C$ and $B$, vector $b$ and scalars $n_0$ and $\sigma^2_0$. The joint distribution of all variables in the problem, $(y, \beta_1, \beta_2, \phi)$ is given by:

\[
p(y, \beta_1, \beta_2, \phi) = p(y \mid \beta_1, \phi)p(\beta_1 \mid \beta_2)p(\beta_2)p(\phi). 
\] (9.1)

The hierarchical character of the model becomes clear through the successive conditional specification of the joint distribution in (9.1). Now we can discuss: when $\beta_2$ is known, the prior does not depend on the probabilistic specification of $\beta_2$, replacement of $b_1$ by $X_2 \beta_2$ in the regression model is all that needs to be done. Hence, the full conditional posterior for $\beta_1$ is $\mathcal{N}(b_\phi, B_\phi)$ and for $\phi$ it is $\Gamma(n_1/2, n_1S\beta/2)$ where $b_\phi = B_\phi(C^{-1}X_2 \beta_2 + \phi X_1^T)$, $B_\phi^{-1} = C^{-1} + \phi X_1^T X_1$, $n_1 = n + n_0$ and $n_1S\beta = n_0\sigma^2_0 + (y - X_1 \beta_1)^T(y - X_1 \beta_1)$.

When $\beta_2$ is unknown, it is not possible to obtain its marginal distribution in closed form, but its full conditional posterior distribution is

\[
p(\beta_2 \mid \beta_1, \phi) \propto p(\beta_1 \mid \beta_2)p(\beta_2)
\]

\[
\propto \exp \left\{ -\frac{1}{2}(\beta_1 - X_2 \beta_2)^T C^{-1}(\beta_1 - X_2 \beta_2) \right\} \times \exp \left\{ -\frac{1}{2}(\beta_2 - b)^T B^{-1}(\beta_2 - b) \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2}(\beta_2 - b^*)^T B^{*-1}(\beta_2 - b^*) \right\},
\]

where $b^* = B^*(B^{-1}b + X_2^T \beta_1)$ and $B^* = (B^{-1} + X_2^T C^{-1} X_2)^{-1}$.

And we can give the following example for more understanding.

**Example 9.3.** Consider several coins minted from the same factory, a head-biased factory will tend to produce head-biased coins. If we denote the bias of the factory by $\omega$, and we denote the biases of the coins as $\theta_s$ (where the subscript indexes different coins), then the credible values of $\theta_s$ depend on the values of $\omega$. The estimate of the bias of any one coin depends on the estimate of factory bias, which is influenced by the data from all the coins.

Indeed, in this sense all Bayesian models are hierarchical, in that a prior for $\theta$ sits above the model for $y$, the latter indexed by the parameter $\theta$. As a conclusion, we can say, that:

- **Simple, hierarchical model:**
  \[
p(y, \beta_1, \beta_2) \propto p(y \mid \beta_1, \beta_2) 
  \text{likelihood},
  \]
  \[
p(\beta_1 \mid \beta_2)p(\beta_2) 
  \text{prior}.
  \]

- **Non-hierarchical model:**
  \[
p(y, \beta_1, \beta_2) \propto p(y \mid \beta_1, \beta_2) 
  \text{likelihood (data model)},
  \]
  \[
p(\beta_1)p(\beta_2) 
  \text{prior (parameter model)}.
  \]
9.2 Dynamic models

A large class of models with time-varying parameters, adequate to the modelling of time series and regression, was presented by Harrison and Stevens (1976). Dynamic Bayesian models are developed for application in nonlinear, non-normal time series and regression problems, providing dynamic extensions of standard generalized linear models. The structure of the models depends on the time evolution of underlying state variables, and the feedback of observational information to these variables is achieved using linear Bayesian prediction methods.

**Definition 9.1.** Dynamic linear models are defined by a pair of equations, called the observation equation and the evolution or system equation, respectively given by:

\[
\begin{align*}
\text{observation equation} & \quad y_t = F_t^T \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2), \\
\text{system equation} & \quad \beta_t = G_t \beta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W_t), \\
\text{prior} & \quad \beta_1 \sim N(a, R),
\end{align*}
\]

- where \( y_t \) is a sequence of observations through time, conditionally independent given \( \beta_t \) and \( \sigma_t^2 \),
- \( F_t \) is a vector of explanatory variables,
- \( \beta_t \) is a \( d \)-dimensional vector of regression coefficients or state parameters at time \( t \),
- \( G_t \) is a \( d \times d \) matrix describing the parametric evolution,
- The errors \( \epsilon_t \) and \( \omega_t \) are mutually independent,
- \( \sigma_t^2 \) and \( W_t \) are the error variances respectively associated to the univariate observation and the \( d \)-dimensional vector of parameters, and
- \( a \) and \( R \) are a \( d \)-dimensional vector and a \( d \times d \) covariance matrix of the prior.

And also we have to know that, the choice of \( F_t \) and \( G_t \) depends on the desired model and the nature of the series one wishes to describe. A complete specification of the model requires full description of the variances \( \sigma_t^2 \) and \( W_t \). In general they are assumed to be constant in time with \( \sigma_t^2 \) typically larger than the entries of \( W_t \) in applications.

9.2.1 Sequential inference

Sequential inference methods have played a crucial role in many of the technological marvels that we use today, from GPS and navigation systems to machine learning. So the main aspect of a dynamic model shows that at any time \( t \), inference can be based on the updated distribution of \( \beta_t \mid y^t \), where \( y^t = \{y_t, y^{t-1}\} \) with \( y^0 \) describing the initial information available, including the values of \( F_t \) and \( G_t \), \( \forall t \). Here we have three basic operations:

- evolution
- prediction
- updating
Consider that at time $t - 1$, the updated distribution is $\beta_{t-1} \mid y^{t-1} \sim \mathcal{N}(m_{t-1}, C_{t-1})$. The system equation can be written as $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(G_t \beta_{t-1}, W_t)$ hence we can conclude

$$\beta_t \mid y^{t-1} \sim \mathcal{N}(a_t, R_t)$$

with $a_t = G_t m_{t-1}$, and $R_t = G_t C_{t-1} G_t^T + W_t$. One-step ahead prediction can be made by noting that $p(y_t, \beta_t \mid y^{t-1}) = p(y_t \mid \beta_t)p(\beta_t \mid y^{t-1})$. We have the marginal distribution for the joint distribution of $y_t, \beta_t \mid y^{t-1}$

$$y_t \mid y^{t-1} \sim \mathcal{N}(f_t, Q_t)$$

with $f_t = F_t^T a_t$, and $Q_t = F_t^T R_t F_t + \sigma_t^2$. And the updated posterior distribution is obtained by $p(\beta_t \mid y^t) = p(\beta_t \mid y_t, y^{t-1}) \propto p(y_t \mid \beta_t)p(\beta_t \mid y^{t-1})$. The resulting posterior is

$$\beta_t \mid y^t \sim \mathcal{N}(m_t, C_t)$$

with $m_t = a_t + A_t e_t$ and $C_t = R_t - A_t A_t^T Q_t$, where $A_t = R_t F_t / Q_t$, $e_t = y_t - f_t$, and from the identity $C_t^{-1} = R_t^{-1} + F_t^T F_t \sigma_t^{-2}$. Applications of sequential analysis: We can see it in the clinical trials, in a randomized trial with two treatment groups, group sequential testing may for example be conducted in the following manner: After $n$ subjects in each group, i.e., a total of $2n$ subjects, are available, an interim analysis is conducted. That means, a statistical test is performed to compare the two groups, if the null hypothesis is rejected, the trial is terminated. Otherwise, the trial continues. Another $n$ subjects per group are recruited. The statistical test is performed again, now including all $4n$ subjects. If the null is rejected, the trial is terminated. Otherwise, it continues with periodic evaluations until a maximum number of interim analyses have been performed. At this point, the last statistical test is conducted, and the trial is discontinued.

### 9.2.2 Smoothing

One of the attractive features of dynamical models is that estimation and forecasting can be developed sequentially, as new data become available. However, in time series analysis one often has observations on $y_t$, $t$ for a certain period, $t = 1, \ldots, T$, and wants to retrospectively reconstruct the behavior of the system, for studying the socioeconomic construct or physical phenomenon underlying the observations. The joint distribution of $y^n = \{y_1, \ldots, y_n\}$, and $\beta = (\beta_1^T, \ldots, \beta_n^T)^T$ is

$$p(y^n, \beta) = \prod_{t=1}^n p(y_t \mid \beta_t) \prod_{t=2}^n p(\beta_t \mid \beta_{t-1})p(\beta_1).$$

Therefore, the full conditional density of $\beta_t$ is

$$p(\beta_t) \propto p(y_t \mid \beta_t)p(\beta_{t+1} \mid \beta_t)p(\beta_t \mid \beta_{t-1})$$

$$\propto p_N(y_t; F_t^T \beta_t, \sigma^2)p_N(\beta_{t+1}; G_{t+1} \beta_t, W_{t+1})p_N(\beta_t; G_t \beta_{t-1}, W_t)$$

$$= p_N(\beta_t; b_t, B_t),$$

where $b_t = B(t \sigma_t^{-2} F_t y_t + G_{t+1} W_{t+1}^{-1} \beta_{t+1} + W_{t+1}^{-1} G_{t+1} \beta_{t-1})$, and $B_t = (\sigma_t^{-2} F_t F_t^T + G_{t+1} W_{t+1}^{-1} G_{t+1} W_{t+1}^{-1})$, for $t = 2, 3, \ldots, n - 1$. The distribution of parameters at time $t$ can be revised after data at times subsequent to $t$ becomes available. And for a set of distributions $p(\beta_t \mid y^{t+k})$, for $k$ integer, it can be considered that
9.2. DYNAMIC MODELS

- When $k > 0$, then we can get $p(\beta_t \mid y^{t+k})$; $k = 1, 2, \ldots$. They are called smoothed or filtered distribution of the parameters.

- When $k = 0$, we get $p(\beta_t \mid y^t)$, then it is the updated distribution.

- When $k < 0$, here we get $p(\beta_t \mid y^{t+k})$; $k = -1, -2, \ldots$. They are prior distribution.

We have to know that, in dynamic models, the smoothed distribution $p(\beta \mid y^n)$ is more commonly used, and it has density:

$$p(\beta \mid y^n) = p(\beta_n \mid y^n) \prod_{t=1}^{n-1} p(\beta_t \mid \beta_{t+1}, \ldots, \beta_n, y^n)$$

$$= p(\beta_n \mid y^n) \prod_{t=1}^{n-1} p(\beta_t \mid \beta_{t+1}, y^t).$$

This equation follows from the fact given $\beta_{t+1}, \beta_t$ is independent of all quantities indexed by times larger than $t$. Now, integrating the last equation with respect to $\{\beta_1, \ldots, \beta_{t-1}\}$ gives

$$p(\beta_t, \ldots, \beta_n \mid y^n) = p(\beta_n \mid y^n) \prod_{k=t}^{n-1} p(\beta_k \mid \beta_{k+1}, y^t),$$

for $t = 1, \ldots, n-1$, and

$$p(\beta_t, \beta_{t+1} \mid y^n) = p(\beta_{t+1} \mid y^n)p(\beta_t \mid \beta_{t+1}, y^t),$$

for $t = 1, \ldots, n-1$. Finally, this equation provides a simple and recursive form to obtain the marginal posterior distribution of $\beta_t \mid y^n$. 

Chapter 10
Monte Carlo methods

By:
Amr Elriedy

10.1 Introduction

The posterior distribution is one of the most important quantities in Bayesian inference. It is an updated version of the prior distribution of a parameter $\theta$ given the data. It can be determined using

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

provided that the likelihood $f(x|\theta)$ and the prior distribution $f(\theta)$ are known.

Any Bayesian statistical inference about $\theta$ exists in the posterior distribution $f(\theta|x)$. The posterior distribution can be used to compute some point estimates for $\theta$, such as the posterior mean,

$$
E(\theta|x) = \int \theta f(\theta|x) d\theta,
$$

or posterior median, mode, variance, etc. See Chapter 3 for a detailed exposition.

Thus, integrals are essential in Bayesian inference, and, sometimes analytical computation of the integrals falls short (e.g., multi-dimensional $\theta$). Hence, other methods are needed to compute multi-dimensional integrals. One of those methods is Monte Carlo integration. This chapter provides a general explanation of the Monte Carlo integration method, and describes two other approaches that can be used to improve the accuracy of Monte Carlo integration, i.e., importance sampling and rejecting sampling.
10.2 Monte Carlo integration

Definition 10.1 (Monte Carlo integration). The term *Monte Carlo integration* indicates a special use of a Monte Carlo method, where we randomly sample uniformly over some domain $V \in \mathbb{R}^s$ and use the produced sample $\{X_1, \ldots, X_n\}$ to construct an estimator for the integral

$$\int_V f(x) \, dx,$$

where $f$ is a real-valued function defined over $V$ (Lemieux, 2009).

To explain Monte Carlo integration, let us first look at an univariate function. We will start with one dimensional setting to ease the explanation. Suppose we have a function $f(x)$ defined over an interval $A \in \mathbb{R}$. The aim is to calculate

$$I(f) = \int_A f(x) \, dx.$$

If $f(x)$ is “simple”, we can use basic calculus to compute a closed form solution for $I(f)$. But, in some cases a closed solution is not possible. For example, if $f(x)$ is the probability density function (pdf) of a standard normal random variable and $A = [0, c]$ for some real constant $c > 0$. This can be written as:

$$I(f) = \int_0^c f(x) \, dx,$$

Where,

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

So, $I(f) = \Phi(c) - \Phi(0) = \phi(c) - \frac{1}{2}$, where $\phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal random variable. Since, there is no closed form solution for $\Phi(c)$, this means that $I(f)$ must be approximated.

Numerical integration methods can be applied to approximate $I(f)$. But, they often fail to spot the region of importance for the function, and they can not compute efficiently highly (or even moderately) multidimensional integrals (Robert and Casella, 2009).

Therefore, Monte Carlo integration can be used to approximate equations like (10.1). This equation can be rewritten as a function of $g$. For example, with $g(\theta) = \theta^2$, we compute an estimate of $\mathbb{E}(\theta^2|x)$, which is used to estimate the posterior variance: $\hat{\text{Var}}(\theta|x) = \hat{\mathbb{E}}(\theta^2|x) - \hat{\mathbb{E}}(\theta|x)^2$. Thus,

$$\mathbb{E}(g(\theta)|x) = \int g(\theta) f(\theta|x) \, d\theta,$$  \hspace{1cm} (10.2)

can be approximated by

$$\hat{\mathbb{E}}(g(\theta)|x) = \frac{1}{M} \sum_{m=1}^M g(\theta^{(m)}),$$  \hspace{1cm} (10.3)

for suitable values $\{\theta^{(m)}\}$.

The main advantages of Monte Carlo integration is that: first, it depends on the Strong Law of Large Numbers (SLLN), which states, that if $\theta^{(1)}, \ldots, \theta^{(M)}$ is a realisation of iid random variables with density function $f(\theta)$, then equation (10.3) will converge to equation (10.2) as $M$
approaches \( \infty \). Second, it depends on Monte Carlo sampling techniques, which is used to generate uniformly distributed random iid from a probability density function.

We now considering the associated estimator of \((10.3)\), e.g., \(\{\theta^{(m)}\}\) are random variables. Then this estimator has the following variance and standard error:

\[
\text{Var} \left[ \hat{E}(g(\theta)|x) \right] = \frac{1}{M} \int \left[ g(\theta) - E(g(\theta)|x) \right]^2 f(\theta|x)d\theta,
\]

\[
\text{se}\left[ \hat{E}(g(\theta)|x) \right] = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \left[ g(\theta^{(m)}) - \hat{E}(g(\theta)|x) \right]^2 / (M - 1).
\]

According to the Central Limit Theorem (CLT), for large \(m\), the asymptotic variance of the estimate becomes \(\hat{E}(g(\theta)|x) - E(g(\theta)|x)\)/\(\text{se}[\hat{E}(g(\theta)|x)]\). Then, the speed of convergence of \(\hat{E}(g(\theta)|x)\) can be assessed, since the convergence takes place at a speed \(O(\sqrt{m})\) \cite{Robert2009}.

**Example 10.1 (Binomial model, Held and Sabanés Bové, 2014).** Consider a posterior distribution with \(\text{Be}(4.5, 1.5)\) for a success rate of \(\pi\) in a binomial model. We are interested in computing the posterior expectation, the posterior probability of \(\Pr(\pi < 0.5|x)\), and the quantiles 2.5\%, 50\%, and 97.5\% using Monte Carlo integration method. Quantities can be written in the general form \(E(g(\pi)|x)\) with \(g(\pi) = \pi \) and \(g(\pi) = I_{0,0.5}(\pi)\).

So using Monte Carlo integration method and taking \(M = 10000\) (number of random samples), the estimated \(E(g(\pi)|x)\) and \(\Pr(\pi < 0.5|x)\) were \(0.751\), and \(0.083\) respectively with quantiles values of \(0.375\), \(0.779\), \(0.978\) for 2.5\%, 50\%, and 97.5\% quantiles.

The true values for \(E(g(\pi)|x)\) and \(\Pr(\pi < 0.5|x)\) are \(0.75\), and \(0.088\) respectively, and \(0.371\), \(0.779\), and \(0.977\) for 2.5\%, 50\%, and 97.5\% quantiles. Therefore, the Monte Carlo estimations are nearly equal to the true values, with very small standard errors of 0.0016, and 0.0028 for \(E(g(\pi)|x)\), and \(\Pr(\pi < 0.5|x)\) respectively.

As mentioned earlier, Monte Carlo integration can be used to estimate several posterior characteristics, such as variance, quantiles, and highest posterior density (HPD) interval.

**Example 10.2 (HPD estimation, Held and Sabanés Bové, 2014).** In the previous binomial model example, the posterior expectations of the proportion \(\pi\) was determined using Monte Carlo Estimation. In this example, Monte Carlo methodology will be used to estimate the 95 \% of HPD.

To do so, we use the property of HPD intervals to have minimal length among all credible intervals at a certain fixed level. We hence calculate all possible credible intervals (based on the ordered samples) with fixed coverage of the whole sample and choose the one with minimal length as a Monte Carlo estimate of the true HPD interval.

For example, if \(M = 100\) and we are interested to estimate the 95\% HPD interval. We first order the sample and obtain the ordered sample; \(\{\theta^1, \ldots, \theta^m\}\). Then the empirical credible intervals of 95 \% empirical coverage are \(\left[\theta^1, \theta^{0.05}\right], \left[\theta^2, \theta^{0.05}\right], \left[\theta^3, \theta^{0.05}\right], \left[\theta^4, \theta^{0.05}\right], \left[\theta^5, \theta^{0.05}\right], \left[\theta^6, \theta^{100}\right]\), and finally the one with the smallest length is picked as Monte Carlo estimate of the 95\% HPD interval. So the estimated 95\% HPD interval using Monte Carlo is \(0.441, 0.998\).
10.3 Importance sampling

Importance sampling is a modification to the Monte Carlo integration method for several reasons, most importantly, is to decrease variance. For example, many density functions of interest have significant weight in only a few regions, the simple Gaussian integral is mostly affected by the area located near the central peak. In a Monte Carlo integration scheme, points are sampled uniformly wasting considerable effort sampling the tails of the Gaussian.

Also, sometimes it is more complex to sample from the target distribution. Therefore, importance sampling introduce a sampling density, or an alternative density to sample from. So, the posterior expectation in of equation (10.1), and its Monte Carlo estimation can be rewritten as such:

\[
\hat{E}[g(\theta|x)] = \frac{1}{\sum_{m=1}^{M} W_m} \sum_{m=1}^{M} W_m g(\theta^m),
\]

(10.4)

where \(h(\theta)\) is the alternative probability density function for sampling and \(W_m = f(\theta^m|x)/h(\theta^m)\), which represents the weights of the random samples drawn from \(h(\theta)\) density. The weights \(W_m\) are known as the importance weights, and it is clear, that they depend on the ratio of the two densities (not only the alternative one). In expectation (perfect situation), the weights \(W_m\) are equal to one since,

\[
\int \frac{f(\theta|x)}{h(\theta)} h(\theta) d\theta = \int f(\theta|x) d\theta = 1.
\]

Note that, if \(W_m\) is never larger than one whenever \(f(\theta^m|x)\) is nonzero, then we are guaranteed that the importance sampling estimator will have a smaller variance than the naive Monte Carlo estimator (Lemieux, 2009).

An alternative way to write (10.4) is

\[
\hat{E}(g(\theta)|x) = \sum_{m=1}^{M} \hat{W}_m g(\theta^m),
\]

where \(\hat{W}_m\) are normalized versions of the importance weights:

\[
\hat{W}_m = \frac{W_m}{\sum_{n=1}^{M} W_n}.
\]

The importance weights are a main component in standard error estimation, as we can see in the following equation:

\[
\text{SE}[\hat{E}(g(\theta)|x)] = \frac{1}{\sum_{m=1}^{M} w_m} \left[ \sum_{m=1}^{M} w_m^2 [g(\theta^m) - \hat{E}(g(\theta)|x)]^2 \right].
\]

Therefore, choosing an alternative density that minimize the sum of the importance weights is essential if variance reduction is the aim.

In Robert and Casella (2009) some conditions were mentioned for choosing a good alternative density:
1. $h(\theta)$ must be $> 0$ whenever $f(\theta|x) \neq 0$.

2. $h(\theta)$ should be proportional to $f(\theta|x)$ and with thicker tails ($h(\theta)$ is *absolutely continuous* with respect to $f(\theta|x)$).

3. It should be easy to simulate values from $h(\theta)$.

4. It should be easy to compute the density of $h(\theta)$ for any value $\theta$ you might realize.

**Example 10.3** (Importance sampling see Held and Sabanés Bové, 2014). Here, Example 10.1 was used to estimate posterior mean $E(\pi|x)$, and the posterior probability $\text{Pr}(\pi < 0.5|x)$. However, the importance sampling estimation was used instead of normal Monte Carlo integration and the alternative density follow an uniform distribution.

So using Importance sampling method and taking sampling from an uniform distribution the values of estimated posterior mean $E(\pi|x)$, and the posterior probability $\text{Pr}(\pi < 0.5|x)$ are 0.753 and 0.0863 respectively. Their standard errors are 0.0018 and 0.0021. Hence, the estimated values using importance sampling are nearly similar to the true values 0.75, 0.0877, and to the values obtained from the original Monte Carlo method.

Also, the standard error of the estimated posterior probability $\text{Pr}(\pi < 0.5|x)$ is 0.0021, which is smaller than the standard error we got using the original Monte Carlo integration method. This proves that importance sampling can reduce variance.

Importance sampling can be an effective method to decrease variance and increase the accuracy of Monte Carlo integration. If the alternative density is chosen correctly. But, in fact the task of identifying a good new density remains an important research problem. In Lemieux (2009) some methods were suggested that can help in this choice; e.g., *exponential twisting/tilting*, *adaptive importance sampling*, and *sampling importance resampling*.

Importance sampling suffers also from the fact that samples are drawn from an alternative density, so some parameters that belongs to the original density can not be correctly estimated such as Quantiles and HPD intervals. Therefore, in the next section rejection sampling method is introduced as an alternative to importance sampling.

### 10.4 Rejection sampling

The idea of rejection sampling is instead of sampling from the target posterior distribution with density $f_X(x)$, simulation is done from another two distributions, the first simulation is done from an arbitrarily distribution with density $f_Z(z)$ under the assumption that there exists an $a \geq 1$ then $f_X(z) \leq a \cdot f_Z(z)$. The second is from a probability density function with uniform distribution $[0,1]$.

Then, the rejection sampling algorithm is simply consists of the following steps:

1. Draw independent random variable $Z$ from $f_Z(z)$.

2. Calculate an acceptance probability $\alpha$ for $Z$: $\alpha = \frac{f_X(z)}{a \cdot f_Z(z)}$.

3. Draw a value $U$ from the Uniform $[0,1]$ distribution.
4. Accept \( Z \) as a draw from \( f_Z(z) \) if \( \alpha \geq U \). Otherwise, reject \( Z \) and go back to step 1.

**Example 10.4** (Rejection sampling, Held and Sabanés Bové, 2014). Similar to Example 10.3, rejection sampling was used to compute posterior mean \( E(\theta|x) \) and the posterior probability \( \text{Pr}(\theta < 0.5|x) \) based on a uniform proposed distribution with \( f_Z(\theta) = 1 \). Here the value of \( \alpha \) is the mode of the target density \( f(\theta|x) \).

The results of using rejection sampling to estimate posterior mean \( E(\theta|x) \) and the posterior probability \( \text{Pr}(\theta < 0.5|x) \) are 0.749 and 0.0883, respectively. The standard errors corresponding to these values are 0.0016 and 0.0028. The true values as previously mentioned are 0.75 and 0.0877.

As it can be seen, the results are very similar, and that is because the accepted samples belong to the target distribution. So technically sampling is done from the true distribution. So, To understand why this works, we can observe the following equations seen in Held and Sabanés Bové (2014):

\[
\text{Pr}[Z \leq x | a \cdot U \cdot f_Z(Z) \leq f_X(z)] = \frac{\text{Pr}[Z \leq x, a \cdot U \cdot f_Z(Z) \leq f_X(Z)]}{\text{Pr}[a \cdot U \cdot f_Z(Z) \leq f_X(Z)]} \\
= \frac{\int_{-\infty}^{x} \text{Pr}[a \cdot U \cdot f_Z(Z) \leq f_X(Z)|Z = z] f_Z(z) \, dz}{\int_{-\infty}^{\infty} \text{Pr}[a \cdot U \cdot f_Z(Z) \leq f_X(Z)|Z = z] f_Z(z) \, dz}. \tag{10.5}
\]

Since,

\[
\text{Pr}[a \cdot U \cdot f_Z(Z) \leq f_X(Z)|Z = z] = P[U \leq a \cdot f_Z(z)] = \frac{f_X(z)}{a \cdot f_Z(z)}.
\]

Thus, \( f_X(z)/a \cdot f_Z(z) \leq 1 \) and \( U \sim (0,1) \) by assumption. Therefore, Equation (10.5) can be written as:

\[
\frac{\int_{-\infty}^{x} a \cdot f_Z(z) f_Z(z) \, dz}{\int_{-\infty}^{\infty} a \cdot f_Z(z) f_Z(z) \, dz} = \frac{\int_{-\infty}^{x} f_X(z) \, dz}{\int_{-\infty}^{\infty} f_X(z) \, dz} = \int_{-\infty}^{x} f_X(z) \, dz = F_X(x).
\]

Conditional on the event \( E = [a \cdot U \cdot f_Z(Z) \leq f_X(Z)] \), the distribution function of the random variable \( Z \) is \( F_X \) with density \( f_X \). The probability to fulfill the event \( E \) is

\[
\text{Pr}[a \cdot U \cdot f_Z(Z) \leq f_X(Z)] = \int_{-\infty}^{\infty} \frac{f_X(z)}{a \cdot f_Z(z)} f_Z(z) \, dz = \int_{-\infty}^{\infty} \frac{f_X(z)}{a} \, dz = a^{-1}.
\]

Hence, the efficiency of rejection sampling depends on \( a \). Trials to sample a random variable from the proposed density then, either accept or reject it are independent. Therefore, the trials are geometrically distributed with parameter \( 1/a \). So the number of trials expected till a random variable is accepted, is equal to \( (a) \). Hence, if \( (a) \) is large, rejection sampling will not be very practical in variance reduction. So, \( (a) \) must be chosen as small as possible while satisfying the condition

\[
f_X(z) \leq a \cdot f_Z(z).
\]

Finally, rejection sampling is considered a reliable, and easy to implement method for variance reduction in Monte Carlo integration. However, it is still suffers from the difficulty to choose a good probability density function.
10.5  Exercises

For additional examples and exercises, the reader is referred to “3.4 Exercises (Monte Carlo integration)” found in Robert and Casella, 2009.

10.6  Supplement: R-Code

This section contains the R-Code used to calculate the relevant quantities of this chapter.

Code for Example 10.1 in Section 10.2.

```r
M <- 10000
theta <- rbeta(M, 4.5, 1.5)
Etheta <- mean(theta)
se.Etheta <- sqrt(var(theta)/M)
Ptheta <- mean(theta < 0.5)
se.Ptheta <- sqrt(var(theta < 0.5)/M)
quant.A <- quantile(theta, probs = c(2.5, 50, 97.5)/100)
quant.1 <- quant.A[[1]]
quant.2 <- quant.A[[2]]
quant.3 <- quant.A[[3]]
## True Values ##
true.1.1 <- 4.5/(1.5 + 4.5)
true.1.2 <- pbeta(0.5, 4.5, 1.5)
```

Code for Example 10.2 in Section 10.2.

```r
thetaorder <- theta[order(theta)]
M <- 10000
level <- 0.95
n.cis <- round(M * (1-level)) + 1
size <- numeric(n.cis)
for(i in seq_len(n.cis)){
  lower <- thetaorder[i]
  upper <- thetaorder[M - n.cis + i]
  size[i] <- upper - lower
}
size.min <- which.min(size)
HPD.lower <- thetaorder[size.min]
HPD.upper <- thetaorder[M - n.cis + size.min]
HPD.1 <- c(HPD.lower, HPD.upper)
```

Code for Example 10.3 in Section 10.3.
\begin{verbatim}
M <- 10000
u <- runif(M)
w <- dbeta(u, 4.5, 1.5)
sum.w <- sum(w)
Etheta.u <- sum(u * w) / sum(w)
se.Etheta.u <- sqrt(sum((u - Etheta.u)^2 * w^2)) / sum(w)
Ptheta.u <- sum((u < 0.5) * w) / sum(w)
se.Ptheta.u <- sqrt(sum(((u < 0.5) - Ptheta.u)^2 * w^2)) / sum(w)

Code for Example 10.4 in Section 10.4:

## posterior parameters and mode:
alpha <- 4.5
beta <- 1.5
mode <- (alpha - 1) / (alpha + beta - 2)
a2 <- dbeta(mode, alpha, beta)
## number of samples to be produced:
M <- 10000
## vector where the samples will be stored:
theta <- numeric(M)
trials <- numeric(M)
## for each sample:
for(m in seq_along(theta)) {
  k2 <- 0
  while(TRUE) {
    k2 <- k2 + 1
    ## sample random variables
    u2 <- runif(1)
z2 <- runif(1)
    ## check for acceptance, then exit the loop
    if(u2 <= dbeta(z2, alpha, beta) / a2)
      break
  }
  theta[m] <- z2
  trials[m] <- k2
}
mean.trials <- mean(trials)
## estimate posterior mean of theta:
Etheta.rs <- mean(theta)
se.Etheta.rs <- sqrt(var(theta) / M)
Ptheta.rs <- mean(theta < 0.5)
se.Ptheta.rs <- sqrt(var(theta < 0.5) / M)
\end{verbatim}
Chapter 11

Gibbs sampling

By:
Roman Flury

11.1 Basic ideas and concepts

11.1.1 Introduction

The purpose of this chapter is to introduce the basic concepts of the Gibbs sampling algorithm and to stress the requirements for its applicability. Due to comprehensibility Gibbs sampling is elaborated with the help of simple examples which also can be solved analytically. In prospect of Chapter 13 “Diagnostic tests” a rigorous mathematical proof for the convergence of Gibbs sampling is disclaimed, but to make the reader aware of possible cases where Gibbs sampling does not converge — or less strictly spoken: fails — a “counter example” is given.

11.1.2 Motivation of sampling

Assume one knows the joint probability density \( f(x_1, \ldots, x_d) \) for a \( d \in \mathbb{N} \) and one is interested in the expectation of the marginal density:

\[
f(x) = \int \ldots \int f(x, y_1, \ldots, y_d) \, dy_1 \ldots dy_d,
\]

which is per definition:

\[
E[X] = \int x f(x) \, dx.
\]

To calculate the expectation analytically or numerically can get extremely difficult or even impossible, if \( d \) is getting larger, due to the curse of dimensionality, see Bellman (1957). But
imagine if one could sample from the marginal distribution, then the law of large numbers states for the expectation:

\[ E[X] \approx \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{as } n \text{ is getting large enough.} \]

The same holds for the variance and other features of the marginal distributions.

### 11.2 Definition of Gibbs sampling

The content of this section is gathered from Gamerman and Lopes (2006), nevertheless the notation was changed for the sake of consistency.

**Definition 11.1 (Full conditional distributions).** Assume \( f(\theta_1, \ldots, \theta_d) \) is a probability density function where \( d \in \mathbb{N} \) and \( \forall i \in \{1, \ldots, d\} \), \( \theta_i \in \Theta_i \). Then the full conditional distribution are defined as:

\[ f_i(\theta_i) = f(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d) = f(\theta_i|\theta_{-i}). \]

**Definition 11.2 (Gibbs sampling algorithm).** The distribution of interest is \( f(\theta) \) s.t. \( \theta = (\theta_1, \ldots, \theta_d)^T \), where \( \theta_i \in \mathbb{R}^{n \times n}, \forall i \in \{1, \ldots, d\} \) and \( n \in \mathbb{N} \); i.e., every parameter can be a scalar, a vector or a matrix.

**Assumption** The full conditional distributions \( f_i(\theta_i) = f(\theta_i|\theta_{-i}), \forall i \in \{1, \ldots, d\} \) are available. In other words they are completely known and of a closed form.

**Initialization**

\[ \theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})^T \quad \text{and} \quad j := 1. \]

**jth step** Obtain a new value

\[ \theta_1^{(j)} \sim f(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}) \]
\[ \theta_2^{(j)} \sim f(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) \]
\[ \vdots \]
\[ \theta_d^{(j)} \sim f(\theta_d|\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{d-1}^{(j)}). \]

**Change counter** Set \( j \) to \( j + 1 \) and return to the previous step until convergence is reached. As soon as convergence is reached, the sampled value \( \theta^{(j)} \) is a draw from the density distribution \( f(\theta) \).

**Remark 11.1 (Convergence).** As the number of iterations increases, the sampling process approaches it’s equilibrium, i.e., the distribution of every \( \theta^{(j)} \) converges to \( f(\theta) \), its true marginal distribution. Formal convergence conditions for the Gibbs sampler were established by Roberts and Smith (1994).
11.3. APPLIED GIBBS SAMPLING

Remark 11.2 (MCMC). The next step depends only on the current position and on no previous positions, i.e., Gibbs sampling is a Markov chain Monte Carlo (MCMC) process. For a proper introduction into MCMC methods the reader is recommended to Chapter 4 “Markov Chains” of Gamerman and Lopes (2006).

Remark 11.3 (Cycled sampling). The parameters are cycled through in their natural order. The reason that parameters are cycled rather than selected randomly is that for complex models with many parameters, it would take too many steps to visit every parameter by random chance alone, even though they would be visited about equally often in the long run.

Definition 11.3 (Burn-in). The initial steps which the sampling process needs to reach its equilibrium is called the burn-in period. The deletion of the corresponding samples shall help to improves the accuracy of the resulting sample. Of course through this deletion the sample size is reduced and therefore its variance enlarged (Gamerman and Lopes, 2006).

Definition 11.4 (Thinning). Each sample is probably correlated with the nearby samples due to a dependency of the full conditional distributions. Therefore some values can be over-represented and others under-represented. To mitigate this correlation one can thin the sample out, i.e., instead of using every sample from every step one uses only every \( m \)th step (usually \( m \in \{10, 20, 50, 100, \ldots \} \)).

11.3  Applied Gibbs sampling

Example 11.1 (Coin flips). Assume one is interested in the proportions of two different coins which are flipped several times. Inference for the two binomial proportions \( \theta_1 \) and \( \theta_2 \) shall be made. Therefore denote \( N_j \) as the number of flips, \( z_j \) as the number of observed heads \( (D = \{N_1, z_1, N_2, z_2\}) \) and \( y_{ij} \) the \( i \)th individual flip where \( j \in \{1, 2\} \). Furthermore it is assumed that the performance of one coin has no influence on the performance of the other; i.e., the coins are independent.

Posterior via exact formal analysis From previous chapters it is known that the beta distribution is conjugate to the Bernoulli likelihood for single proportions. Recall the beta probability density function:

\[
 f(\theta | a, b) = \theta^{(a-1)} (1-\theta)^{(b-1)} / B(a, b) \]

where \( B(a, b) = \int_0^1 \theta^{(a-1)} (1-\theta)^{(b-1)} d\theta \) is the beta normalizing function. The independence of the coins implies for the likelihood function:

\[
 f(D | \theta_1, \theta_2) = \prod_{y_{1i} \in D_1} f(y_{1i} | \theta_1, \theta_2) \prod_{y_{1i} \in D_2} f(y_{2i} | \theta_1, \theta_2)
 = \prod_{y_{1i} \in D_1} \theta_1^{y_{1i}} (1-\theta_1)^{(1-\theta_1)} \prod_{y_{2i} \in D_2} \theta_2^{y_{2i}} (1-\theta_2)^{(1-\theta_2)}
 = \theta_1^{\sum_{i=1}^{y_{1i}} (N_1-z_1)} \theta_2^{\sum_{i=1}^{y_{2i}} (N_2-z_2)}.
\]

The prior distribution is:

\[
 f(\theta_1, \theta_2) = f(\theta_1) f(\theta_2)
 = \theta_1^{(a_1-1)} (1-\theta_1)^{(b_1-1)} \theta_2^{(a_2-1)} (1-\theta_2)^{(b_2-1)}.
\]
Then the resulting posterior function $f(\theta_1, \theta_2|D) \propto \text{likelihood} \times \text{prior}$ is:

$$f(\theta_1, \theta_2|D) = \frac{\theta_1^{z_1+a_1-1}(1-\theta_1)^{(N_1-z_1+b_1)-1}}{\text{Be}(\ldots)} \times \frac{\theta_2^{z_2+a_2-1}(1-\theta_2)^{(N_2-z_2+b_2)-1}}{\text{Be}(\ldots)}$$

$$\Rightarrow \theta_1, \theta_2|D \sim \text{Be}(z_1+a_1, N_1-z_1+b_1) \times \text{Be}(z_2+a_2, N_2-z_2+b_2).$$

**Draw from Posterior via Gibbs sampling** Following the Definition 11.2 derive the full conditionals first:

$$f(\theta_1|\theta_2, D) \propto \theta_1^{z_1}(1-\theta_1)^{(N_1-z_1)-1}\theta_1^{a_1-1}(1-\theta_1)^{(b_1)-1} = \frac{\theta_1^{z_1+a_1-1}(1-\theta_1)^{(N_1-z_1+b_1)-1}}{\text{Be}(\ldots)}$$

$$\Rightarrow \theta_1|\theta_2, D \sim \text{Be}(z_1+a_1, N_1-z_1+b_1)$$

analogue $\Rightarrow \theta_2|\theta_1, D \sim \text{Be}(z_2+a_2, N_2-z_2+b_2)$.

The vigilant observer recognizes that the full conditionals do not depend on each other. Therefore, if Gibbs sampling is applied, this two distributions do not change during the sampling process, as visualized in Figure 11.1. But from this very simple example one can derive an intuition of how the algorithm works. Note, the rectangular pattern is due to Remark 11.3.

An animated version of the Gibbs sampling process of this example is available as Shiny application under roflur.shinyapps.io/GibbsSampler-CoinFlips.

---

**Figure 11.1:** Gibbs sampling

**Figure 11.2:** True contours
11.3. APPLIED GIBBS SAMPLING

11.3.1 Classical Bayesian setting

Example 11.2 (Normal Gamma). Consider the typical Bayesian model with:

$$\text{likelihood: } Z | \beta, \sigma^2 \sim \mathcal{N}(\beta, \sigma^2), \quad \text{priors: } \beta \sim \mathcal{N}(m, v), \quad \frac{1}{\sigma^2} \sim \Gamma(a, b).$$

Posterior distribution of $\beta, \sigma^2|Z = z_1$ derived analytically

Likelihood: $f(z_1|\beta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left( -\frac{1}{2} \frac{(z_1 - \beta)^2}{\sigma^2} \right)$

Prior: $f(\beta, \sigma^2) = f_X(m, v) \times f_\Gamma(a, b) \propto \exp \left( -\frac{1}{2} \frac{(\beta - m)^2}{v} \right) \times \sigma^{-2(a-1)} \exp \left( -b \frac{1}{\sigma^2} \right)$.

pdf of $X \sim \Gamma(a, b)$: $f(x) = \frac{\beta^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ where $\Gamma(x)$ is the gamma function.

$$f(\beta, \sigma^2|z_1) = f(z_1|\beta, \sigma^2) f(\beta) f(\sigma^2)$$

$$= \sigma^{-2} \exp \left( -\frac{1}{2} \frac{(z_1 - \beta)^2}{\sigma^2} + \frac{(\beta - m)^2}{v} \right)$$

where $(*) = -\frac{1}{2} \beta^2 \left( \frac{1}{\sigma^2} + \frac{1}{v} \right) + \beta \left( \frac{z_1}{\sigma^2} - \frac{m}{v} \right) - \frac{z_1^2}{2\sigma^2} - \frac{m^2}{2v} - b \frac{1}{\sigma^2}$

$$\propto \sigma^{-2(a-1)-1} \exp \left( -\frac{1}{2} \beta^2 \left( \frac{1}{\sigma^2} + \frac{1}{v} \right) + \beta \left( \frac{z_1}{\sigma^2} + \frac{m}{v} \right) - \frac{z_1^2}{2\sigma^2} - b \frac{1}{\sigma^2} \right).$$

Derivation of the full conditionals $\beta|\sigma^2, Z$ and $\sigma^2|\beta, Z$

$$f(\beta|Z = z_1, \sigma^2) = \exp \left( -\frac{1}{2\sigma^2} (z_1 - \beta)^2 \right) \exp \left( \frac{1}{2v} (\beta - m)^2 \right)$$

$$\propto \exp \left( -\frac{1}{2} \beta^2 \left( \frac{1}{\sigma^2} + \frac{1}{v} \right) + \beta \left( \frac{z_1}{\sigma^2} + \frac{m}{v} \right) \right)_{\beta = a^{-1}} \exp \left( -\frac{1}{2} \beta^2 \left( \frac{1}{\sigma^2} + \frac{1}{v} \right) + \beta \left( \frac{z_1}{\sigma^2} + \frac{m}{v} \right) \right)_{\beta = b}$$

$$\Leftrightarrow \exp \left( -\frac{1}{2a} \left( \beta + \frac{b}{2a} \right)^2 \right); \quad \text{recognize the variance: } a^{-1}, \text{ and the mean: } \frac{b}{2a}$$

$$\Leftrightarrow \beta|Z, \sigma^2 \sim \mathcal{N} \left( \frac{z_1 v + m \sigma^2}{v + \sigma^2}, \frac{v \sigma^2}{v + \sigma^2} \right)$$

and

$$f(\sigma^2|Z = z_1, \beta) \propto \frac{1}{\sigma^2} \exp \left( -\frac{1}{2\sigma^2} (z_1 - x\beta)^2 \right) \exp \left( -b \frac{1}{\sigma^2} \right)$$

$$= \sigma^{-2(a-1)-1} \exp \left( -\frac{1}{2} \left( \frac{z_1 + x \beta}{\sigma^2} + b \right) \right)$$

$$\Leftrightarrow \sigma^2|Z, \beta \sim \Gamma \left( -(2a - 1), \frac{(z_1 - \beta)^2}{2} + b \right).$$
One can see that in this example the two full conditionals do depend on each other. Therefore they are updated in every sampling step with the sample from the previous step, see Figure 11.3 for a visualization. Moreover this figure illustrates the Burn–In phase of the sampling process, since the full conditional distributions did not reach yet an equilibrium. An animation of this example is provided on roflur.shinyapps.io/GibbsSampler-NormalGamma/.
11.4 Frequentist setting

Remark 11.4 (Gibbs sampling in frequentist setting). Gibbs sampling is not restricted to the Bayesian setting only. If one considers the assumptions of its Definition 11.2 one can see that those are applicable to arbitrary multivariate distributions also.

Example 11.3 (Bivariate Normal). Suppose

\[ Z \sim \mathcal{N}(\mu, \Sigma), \quad \text{where} \quad Z = (X, Y)^\top, \quad \mu = (\mu_X, \mu_Y)^\top, \quad \Sigma = \begin{bmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix} \]

\( \rho \in [-1, 1] \) is the correlation coefficient between \( X \) and \( Y \).

Then the full conditionals are (Mardia et al., 1979):

\[
X|Y = y \sim \mathcal{N} \left( \mu_X + \frac{\sigma_X^2}{\sigma_Y^2} \rho(y - \mu_Y), (1 - \rho^2) \sigma_Y^2 \right)
\]

\[
Y|X = x \sim \mathcal{N} \left( \mu_Y + \frac{\sigma_Y^2}{\sigma_X^2} \rho(x - \mu_X), (1 - \rho^2) \sigma_X^2 \right).
\]

Similar to the examples above one can now apply Gibbs sampling. In Figure 11.4 the mean vector is set to zero \( \mu = 0 \) and \( \rho = .9 \). Moreover to improve the samples the burn-in period of 1000 samples is removed and thinned out to every twenties sample. Figure 11.5 shows the resulting sample with the corresponding true contour lines.

![Figure 11.4: Sampling steps](image1)

![Figure 11.5: True bivariate normal](image2)

11.5 Counter example

Example 11.4 (Artificial Gamerman and Lopes, 2006). Consider \( \theta = \theta_1 \) and \( \theta_2 \) which are uniformly distributed over two disjoint regions \( A = A_1 \times A_2 \) and \( B = B_1 \times B_2 \) in the same
plane. With the corresponding probabilities $p_A$ and $p_B$ adding up to 1. Assume also disjoint projections of $A_1$ and $B_1$ on the $\theta_1$ axis and of $A_2$ and $B_2$ on the $\theta_2$ axis. This implies that the full conditionals are also uniform distributed, but over the region dependent on the starting point. A sampling process that starts in region $A$ leads to sampling $\theta_1$ over $A_1$ which leads to sampling $\theta_2$ over $A_2$. The process continues again in $A_1$ and will never reach the region $B$. If one chooses a starting value from region $B$ the sampling process behaves analogue.

**Remark 11.5.** This example shows that it is possible that convergence fails. Moreover it seems that to satisfy convergence, one needs at least some assumptions on the domain, such as a connected projections of the domain. This should motivate the reader to justify sampled observations with suitable tests, as introduced in Chapter 13 “Diagnostic tests”.
Chapter 12

Software for Gibbs sampling

By:

Thimo Schuster

12.1 Introduction

The previous chapter showed the idea and theory of the Gibbs sampler. This chapter does focus on the software implementation of Gibbs sampling and shows an example of it.

By today, there are different softwares that can be used for Bayesian inference. Four of them are:

- **BUGS**: Is an acronym for “Bayesian inference Using Gibbs Sampling”. It was initialized in 1989 at the Biostatistics Unit in Cambridge. There are two versions: OpenBUGS and WinBUGS.

- **JAGS**: This stands for “Just another Gibbs sampler”. It was initially released in 2007 by Martyn Plummer.

- **Stan**: Is named after Stanislaw Ulam, who was a pioneer of Monte Carlo methods. The first version was released in 2012.


In this chapter, information and an example is provided for JAGS. For BUGS, Stan and INLA, a summary about the important properties is given.

All of the four mentioned software packages can be accessed via R. This is done via freely available R-packages. An overview is given in Table 12.1. For the example of JAGS, the R-package *rjags* has been used.
**Table 12.1:** Gibbs sampling software and the corresponding R-packages

<table>
<thead>
<tr>
<th>Software</th>
<th>R-package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUGS</td>
<td>BRugs / R2openBUGS</td>
<td>R-package for OpenBUGS</td>
</tr>
<tr>
<td></td>
<td>R2WinBUGS</td>
<td>R-package for WinBUGS</td>
</tr>
<tr>
<td>JAGS</td>
<td>rjags</td>
<td>Standard R-package for JAGS</td>
</tr>
<tr>
<td></td>
<td>runjags</td>
<td>Facilitates the use of parallel processing for multiple chains</td>
</tr>
<tr>
<td>Stan</td>
<td>rstan</td>
<td>Standard R-package for Stan</td>
</tr>
<tr>
<td>INLA</td>
<td>INLA</td>
<td>Standard R-package for INLA</td>
</tr>
</tbody>
</table>

### 12.2 Software for Gibbs sampling

#### 12.2.1 BUGS

The BUGS project started in 1989 at the Biostatistics Unit in Cambridge. It is the oldest of the four discussed software packages and grew out of work in artificial intelligence. Resulting was a windows version of the BUGS software, today known as WinBUGS, which was released in 1997. WinBUGS is written in Component Pascal, which makes it difficult to run on other operating systems than Windows. There exist some workarounds using Wine, but it was initially designed for Windows only.

BUGS stands for “Bayesian Inference using Gibbs Sampling”. Of course this also describes its possible applications: Using Markov Chain Monte Carlo (MCMC) methods, BUGS conducts Bayesian inference. The user basically only needs to specify a statistical model, and, depending on the application, provide some data. BUGS will then figure out an appropriate MCMC method on its own, depending on the desired output (Kruschke, 2010).

In 2005, an open source version of BUGS was released: OpenBUGS. This resulted in the development of WinBUGS being stopped. The last update was released in 2007, but WinBUGS is still available for download. Nevertheless, the development focus now is on OpenBUGS. The major change between OpenBUGS and WinBUGS is, that OpenBUGS also runs on Linux, not only on Windows. Furthermore, it features several new functions and distributions. All these reasons make it easy to say that by today, OpenBUGS should be preferred to WinBUGS.

WinBUGS can be accessed via R with R2WinBUGS, which calls WinBUGS in the background where the analysis is performed. The R-packages for OpenBUGS are called BRugs nad R2openBUGS. R2openBUGS is the adaption of R2WinBUGS to OpenBUGS, whereas BRugs runs OpenBUGS analyses fully interactively from within R.

#### 12.2.2 JAGS

JAGS was build after the model of BUGS in 2007. It retained many of its design features including a very similar syntax, but uses different samplers and offers a better usability. One other advantage over BUGS is that JAGS is written in C++, which allows portability to all available systems. In the statistical community, there is an ongoing debate about which software
should be used. At the moment, JAGS seems to be the software of choice - it includes all the main advantages of OpenBUGS and fixed some of its shortcomings.

JAGS can be used from R with the package \texttt{rjags}. Its functionality and syntax is very similar to \texttt{BRugs}, which makes it easy to translate programs.

There is another R-package for JAGS: \texttt{runjags}. With this package, it is possible to run different chains on different cores in the computer simultaneously. With \texttt{rjags}, different chains will be computed sequentially on a single core. This is not a problem for simple models but can be slow for larger ones with much data. With \texttt{runjags}, a slightly different syntax and different functions have to be used (see Kruschke, 2015).

\subsection*{12.2.3 Stan}

Stan is the most recently released of the four discussed softwares. There are two main differences to BUGS and JAGS. First, Stan is based on a programming language which is more flexible and especially useful for complex models. This means that Stan is able to provide solutions where JAGS or BUGS fail. Second, Stan uses Hamiltonian Monte Carlo (HMC) instead of the Gibbs sampler to compute MCMC chains. This method can be more effective than the samplers from BUGS and JAGS.

But there are also some downsides: Stan is not able to perform inference for discrete unknowns, whereas JAGS and BUGS are. Furthermore, there are some models which cannot yet be directly expressed in Stan. This might change in the future, as Stan is still undergoing development (compare Kruschke, 2015 and Gelman \textit{et al.}, 2015).

Stan can be accessed from R via the package \texttt{rstan}. The estimation of a posterior distribution works in a similar way as for JAGS: First, the data must be prepared and loaded. Second, the model needs to be specified. Finally, inference can be conducted and the outcome can be analysed. However, the syntax is different and Stan requires a little bit more programming expense.

\subsection*{12.2.4 INLA}

INLA uses a different approach for conducting Bayesian Inference than Stan, BUGS and JAGS. It is an acronym for “integrated nested Laplace approximation”, which is an alternative to MCMC. INLA is designed for latent Gaussian models (LGM), a specific model characterisation. There is still a wide range of possible models - but this design makes it a little less flexible than the MCMC approach, which is not designed for a specific characterisation.

Nevertheless, INLA is a very effective tool for estimating LGM’s as it performs direct numerical calculation of the posteriors. It does not need the time-intensive MCMC-part, which makes it faster than MCMC-based software tools (see Lindgren and Rue, 2015).

INLA can be accessed via R with the package \texttt{INLA}.

\subsection*{12.2.5 Example with JAGS: coin bias}

As it is easiest to understand the usage of a Gibbs sampling software via an example, the Bayesian estimation of the bias of a coin is considered (see Kruschke, 2015).
Consider a coin which is tossed 50 times. If the coin is unbiased, it should come up heads in about 50% of all cases. If this is not the case, for example if the coin comes up heads in 75% of all cases, the coin is biased. This example was already introduced in chapter 11. With JAGS, the posterior distribution of this example is approximated via an MCMC sample. For doing this, JAGS needs the data and an appropriate model.

**Load data into R** The first step of approximating the posterior distribution via JAGS is to load the data into R. In order to use the data with JAGS, a list containing the data and the total number of observations needs to be created (this corresponds to `dataList` in Code Chunk 12.2.1).

**Code Chunk 12.2.1.**

```r
y <- c(0,1,0,0,0,0,0,0,1,0,0,0,1,1,1,0,0,1,0,0,0,0,1,1,0,0,0,1,0,0,0,1,0,0,1,1,0,1,0,0,0,0)
# This is the outcome of 50 coin flips.
Ntotal <- length(y)  # Compute the total number of flips.
dataList <- list(  # Put the information into a list.
  y = y,
  Ntotal = Ntotal
)
```

**Model definition** Once the data is available, the model for JAGS can be constructed. It consists of two parts: The likelihood function and the prior distribution. The likelihood function links the data to the parameters, whereas the prior distribution specifies the probability of possible parameters without taking into account the data.

As the outcome of a single coin flip can be denoted as 0 (tails) or 1 (heads), the Bernoulli distribution is used as a likelihood function. As prior distribution, the beta density is considered. For the model specification, this information has to be stored as a string. The model is defined in a language JAGS understands - see for example the definition of the likelihood function for each data point in Code Chunk 12.2.2. The single values are assigned to their Bernoulli density by ∼; in R, one would either use `=` or `<-`. For the beta distribution, starting values of \((\alpha, \beta) = (2, 2)\) were chosen.
12.2. SOFTWARE FOR GIBBS SAMPLING

Code Chunk 12.2.2.

```r
modelString <- "model {
    for (i in 1:Ntotal) {
        y[i] ~ dbern(theta)   # Likelihood function
    }
    theta ~ dbeta(2, 2)    # Prior distribution
}
"
writeLines(modelString, con="TEMPmodel.txt")
# Save model as temporary text file
```

Initial values for MCMC chains In order to run a JAGS-model, it is not necessary to specify initial values for the parameter(s). If nothing is specified, JAGS uses default values as initial values. However, convergence can sometimes be reached faster with self-chosen starting values. The maximum likelihood estimate (MLE) of the parameters would be one possibility. The posterior distribution is supposed to be quiet close to the likelihood function, therefore the MLE should also be quiet close to the parameters of the posterior distribution.

For the Bernoulli likelihood function the MLE is \( \theta = \frac{N_{\text{event}}}{N} \), which is equivalent to \( \frac{\text{sum}(y)}{\text{length}(y)} \) in the R-Code (compare Code Chunk 12.2.3). The result of this is again stored in a list, which makes it possible to send it to JAGS.

Code Chunk 12.2.3.

```r
thetaInit <- sum(y) / length(y) # Compute the MLE for a Bernoulli distribution

initsList <- list(list(theta = thetaInit,
                   .RNG.seed = 45611,
                   .RNG.name = "base::Super-Duper"),
                   list(theta = thetaInit,
                   .RNG.seed = 54318,
                   .RNG.name = "base::Wichmann-Hill"),
                   list(theta = thetaInit,
                   .RNG.seed = 22368,
                   .RNG.name = "base::Marsaglia-Multicarry"))
# Save the initial values in a list of lists, each with a seed (for reproducibility) and the MLE
```

Generate MCMC chains As the data is ready, the model is defined and initial values are set, JAGS can finally start generating the MCMC sample from the posterior distribution.

In a first step, the R-function `jags.model` (from the package `rjags`) is called, using the data, model and intial values specified. This function constructs a `jagsModel`-object, which basically
includes appropriate MCMC samplers that were figured out by JAGS. As one can see in Code Chunk 12.2.4, also the number of chains (\(n.chains\)) and the number of steps for adapting the samplers (\(n.adapt\)) are handed over to the function. These values can be changed for tuning purposes.

Code Chunk 12.2.4.

```r
jagsModel <- jags.model(file = "TEMPmodel.txt", # Model
data = dataList, # Data
inits = initsList, # Initial values
n.chains = 3, # Chains
n.adapt = 1000) # Adaption steps
```

In a next step, the chains are runned for some number of steps to accomplish burn-in. Burn-in iterations are the first steps in a MCMC chain which are discarded once the chain was constructed. The reason is that the chain needs to “burn-in”: Initial samples tend to be not completely valid because the Markov chain has not yet stabilized. This can be a consequence of the chosen starting values.

Be aware of the difference between burn-in and the adaption steps from Code Chunk 12.2.4. While burn-in makes sure that the first steps of the Markov chain are discarded, JAGS uses the adaption steps to find an optimal set of parameters for the samplers. This means: Once a good sample generating algorithm was found (after some adaption steps), a sequence of samples can be generated. The first \(n\) of those samples will be discarded due to the definition of a burn-in period.

The burn-in can be done via `update()` (compare Code Chunk 12.2.5). This function changes the internal state of the `jagsModel` by \(n.iter\) iterations.

Code Chunk 12.2.5.

```r
update(jagsModel, # jagsModel created
       n.iter=500) # Burn-in period
```

Now that the burn-in period has been specified, JAGS can create MCMC samples that will be used for the estimation of the posterior distribution. This is done via `coda.samples()` from `rjags`. The name refers to its output which will be stored as a `coda`-object. Code chunk 12.2.6 also shows that one needs to specify which variables should be monitored by JAGS. In models with many parameters, it would be computationally expensive to monitor all of them by default.

Code Chunk 12.2.6.

```r
codaSamples <- coda.samples(
    jagsModel, # jagsModel including burn-in
    variable.names = c("theta"), # Variables monitored by JAGS
    n.iter = 10000) # Number of iterations per chain
```

The resulting `coda`-object `codaSamples` now contains three MCMC chains with 10'000 steps each, estimated after a burn-in period of 500 steps.
**Resulting posterior distribution**  A plot of the density of the three generated MCMC chains shows, that all of them are very similar to each other (compare Figure 12.1). Another peculiarity is that $\theta$ has its maximum at approximately 0.3 for all three chains, which is basically the MLE that was defined as an initial value.

This can also be seen from the `summary()` output (compare Code Chunk 12.2.7). 3 chains have been calculated with a chain length of 10’500 steps each. The iterations 501 - 10’500 were used for inference. This corresponds to the defined burn-in period of 500 steps and results in a sample size of 10’000 steps per chain. The mean for $\theta$ is 0.315, with 95% confidence interval from 0.200 to 0.444. This coincides with the observed value from the density plot.

```r
## 1. Empirical mean and standard deviation for each variable, 
##    plus standard error of the mean: 
##
##    Mean     SD    Naive SE  Time-series SE 
## 0.3144821 0.0624731 0.0003607   0.0003607
##
## 2. Quantiles for each variable: 
##
## 2.5%  25%  50%  75%  97.5%  
## 0.2002 0.2706 0.3122 0.3563 0.4425
```

A next step would be to further examine the MCMC samples to check if the chains are suitable representative of the posterior distribution. This topic will be discussed in Chapter 13.

### 12.3 Exercises

Exercises for JAGS and STAN can be found in Kruschke (2015). For BUGS-exercises, Kruschke (2010) can be considered. Kruschke also provides further information and additional applications of the different software packages.
Figure 12.1: Plot of the density of the three generated MCMC chains for the coin flip example.
Chapter 13

Diagnostic tests

By:
Verena Steffen

From the theory of Markov chains, one expects the chains to converge to a stationary distribution. However, there is no guarantee that the chains converged after, say, \( M \) draws. Although convergence cannot be proven, there exist both visual, and statistical tests to conclude convergence. Convergence diagnostic thus aims to answer the simple question (White, 2010): has the sampler been given a sufficient adaptation (“burn-in”) period to justify the claim that the samples drawn come from the approximate posterior distribution?

There exist theoretical attempts to study convergence, however these are not commonly used because —according to Cowles and Carlin (1996)— they rely on sophisticated mathematics and need a separate solution for each problem. Usually the theoretical approaches are based on distance measures and establishing bounds on distribution functions generated from a chain. Also Gamerman and Lopes (2006) recommend the use of statistical tests, which lead to equivalent results if used properly.

This chapter aims to give an overview about diagnostic tests for convergence, and can be used as a practical guide to analyze convergence of MCMC samples using the R packages \texttt{coda} (Plummer et al., 2006) and \texttt{mcmcplots} (Curtis, 2015). The focus will be on exploratory and graphical techniques, as well as more formal methods like the “Gelman and Rubin Diagnostic”, the “Geweke Diagnostic”, and the “Raftery and Lewis Diagnostic”.

13.1 Principles

Gamerman and Lopes (2006) categorizes the statistical approaches based on their methodology. Table 13.1 gives an overview about the methods described in this chapter and their categorization.
Table 13.1: Categorization of the presented methods according to their methodology.

<table>
<thead>
<tr>
<th>Informal convergence monitors</th>
<th>Exploratory and Graphical Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergence prescription</td>
<td>Raftery and Lewis</td>
</tr>
<tr>
<td>Formal convergence methods:</td>
<td></td>
</tr>
<tr>
<td>• Time series analysis</td>
<td>Geweke</td>
</tr>
<tr>
<td>• Multiple chains</td>
<td>Heidelberger and Welch</td>
</tr>
<tr>
<td>• Methods based on conditional distributions</td>
<td>Gelman and Rubin</td>
</tr>
</tbody>
</table>

Example 13.1. In an example adapted from White (2010), the principles of convergence diagnostics shall be explained. It will also be used to demonstrate the usage of the coda package in R. The following code simulates data from a linear model.

```r
N <- 1000
x <- 1:N
epsilon <- rnorm(N, 0, 1)
y <- x + epsilon
write.table(data.frame(X = x, Y = y, Epsilon = epsilon),
            file = "chapter13.data",
            row.names = FALSE,
            col.names = TRUE)
```

The Bayesian model for the example is saved in chapter13.bug:

```r
model {
  for (i in 1:N){
    y[i] ~ dnorm(y.hat[i], tau)
    y.hat[i] <- a + b * x[i]
  }
  a ~ dnorm(0, .0001)
  b ~ dnorm(0, .0001)
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 100)
}
```

Here, non-informative normal priors for $a$ and $b$, and a non-informative uniform prior for $\sigma$ is used. The data comes from a normal distribution with mean $a + b \times x[i]$ and precision $\tau$, which is just $1/\sigma$.

The model can be run with rjags:

```r
jags <- jags.model("chapter13.bug",
                   data = list("x" = x,
                                "y" = y,
                               })
```
After it has run over a fair amount of samples, we can draw inferences for a and b, which should be close to the values of 0 and 1, respectively. In general, one should also check tau, which is ignored here for the sake of space.

### 13.2 Exploratory and graphical techniques

For a first overview about the MCMC output, exploratory techniques can be used. This section describes some of the most commonly used tools from the coda and mcmcplots packages in R.

#### 13.2.1 Summary statistics

Most R users are familiar with the `summary()` function for different classes. In coda it is used for objects of the class mcmc to generate summary statistics for the posterior. Using it produces an output for each variable (in this example a and b) that gives the mean, standard deviation, naive standard error of the mean *ignoring* autocorrelation, and a time-series based standard error based on an estimate of the spectral density at 0. We also get the 2.5, 25, 50, 75, and 97.5% quantiles of the sample.

```r
summary(samples)
```

```r
##
## Iterations = 1101:2100
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable, plus standard error of the mean:
##
## Mean     SD  Naive SE Time-series SE
## a -0.0623 0.0595 9.410e-04 2.508e-03
## b  1.0001 0.0001 1.624e-06 4.196e-06
##
## 2. Quantiles for each variable:
##
## 2.5%   25%   50%   75%   97.5%
```
13.2.2 Trace and density plots

In a trace plot the value of the parameter is plotted against the iteration number. Those who are familiar with time-series analysis, may recognize that this is simply a time-series plot. It is useful to spot if the chain gets stuck in certain areas of the parameter space, which indicates bad mixing. The density plots visualize the density of each parameter. In coda, both plots can simply be produced using the plot() function.

In this example there is no indication for bad mixing. In addition, both parameters look more or less normally distributed, see Figure 13.1.

13.2.3 Running mean plots

Another graphical technique is the so-called running mean plot. Therefore, the parameter mean up to the current iteration is plotted against the iteration number. A convenient implementation thereof in R can be found in the package mcmcplots.

Figure 13.2 shows the output from running the rmeanplot() command. As indicated before, the used example converges. If we, however, had bad mixing, this could be spotted in a running mean plot by change of mean.
13.2.4 Autocorrelation

High autocorrelations within chains indicate slow mixing and, usually, slow convergence. The autocorrelations can be calculated with `autocorr()`, or with `autocorr.diag()` for different lags. A graphical method for assessing the autocorrelation is implemented with `autocorr.plot()`.

Figure 13.3 shows the output from the plotting function. After about 10 lags, the autocorrelation is nearly disappeared.

13.3 Gelman and Rubin diagnostic

The method was originally proposed by Gelman and Rubin (1992). It is based upon classical ANOVA. First, \( m \) different starting values that are overdispersed with regard to the target distribution are chosen. Then \( m \) chains are run parallel for each \( 2n \) iterations, of which \( n \) are
discarded to avoid the burn-in period.

Then, for all scalar parameters, assume a random variable $X$ that has mean $\mu$ and variance $\sigma^2$ under the target distribution. Let $X_{tj}$ denote the $t^{th}$ of the $n$ iterations of $X$ in chain $j$, we estimate the between chain variance $B/n$, and the within chain variance $W$, defined by

$$B/n = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{X}_j - \bar{X})^2$$

and

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{t=1}^{n} (X_{jt} - \bar{X}_j)^2.$$ 

Having observed these estimates, we can estimate $\sigma^2$ by a weighted average of $B$ and $W$, 

$$\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{B}{n},$$

which would be unbiased if the starting points were drawn from the target distribution, but overestimates $\sigma^2$ if the starting distribution is appropriately overdispersed. The estimates $\hat{\mu}$ and $\hat{\sigma}^2$ could be used if $X$ followed an approximate normal distribution. This approximation can be improved by accounting for the sampling variability of $\hat{\mu}$ and $\hat{\sigma}^2$ by using an approximate $t$-distribution with location $\hat{\mu}$, variance $\hat{V} = \hat{\sigma}^2 + B/mn$, and degrees of freedom $d$, which can be estimated via the method of moments,

$$d \approx \frac{2\hat{V}}{\text{Var}(\hat{V})}.$$ 

This approximate posterior accounts both for the estimated variance and for the uncertainty in its estimation.

Now, convergence can be monitored by estimating the factor by which the estimated scale of the posterior distribution for $X$ will shrink as $n \to \infty$: \n
$$R = \frac{\hat{V}}{\sigma^2}.$$ 

We call $R$ the scale reduction factor, or SRF. To account for the fact that the degrees of freedom are estimated, a correction factor is included for $R$. This correction factor was incorrectly defined as $d/(d - 2)$ in the original paper by Gelman and Rubin (1992). In a follow up paper Brooks and Gelman (1998) introduce the factor $(d + 3)/(d + 1)$.

Finally, since $\sigma^2$ is unknown, $R$ will be overestimated by

$$\hat{R}_c = \frac{d + 3}{d + 1} \frac{\hat{V}}{W}.$$ 

The value of $\hat{R}_c$, which is called potential scale reduction factor, or PSRF, can be interpreted as a convergence diagnostics as follows. If $\hat{R}_c$ is large, this suggests either that the estimate of the variance $\hat{\sigma}$, can be further decreased by more simulations or, that further simulation will decrease $W$, since the simulated sequences have not yet made a full tour of the target distribution. If $\hat{R}_c$ is close to 1, we can conclude that each of the $m$ sets of $n$ simulated observations is close to the target distribution.

In short, the steps involved for each parameter, are
13.4 Geweke diagnostic

Geweke (1992) proposes a convergence diagnostic based on standard time-series methods. It is based on a single chain and is appropriate when convergence of the mean (of some function) of the sampled variable is of interest. The chain is divided into 2 “windows” containing the first 10% and the last 50% of the iterates. If the whole chain is stationary, the means of the values early and late in the sequence should be similar. The convergence diagnostic \( Z \) is the difference between the two means divided by the asymptotic standard error of their difference. As \( n \) goes to infinity the sampling distribution of \( Z \) goes to \( N(0,1) \) if the chain has converged.

Figure 13.4: Gelman plot of the simulated linear regression example with the median shrink factor (solid line) and the 97.5% quantile (dashed line).

1. run \( m \geq 2 \) chains of length \( 2n \) from overdispersed starting values,
2. discard the first \( n \) draws in each chain,
3. calculate the within-chain and between-chain variance,
4. calculate the estimated variance of the parameter as a weighted sum of the within-chain and between-chain variance, and
5. calculate the potential scale reduction factor \( \hat{R}_c \).

The PSRF (or \( \hat{R}_c \)) can be plotted conveniently with the coda function \texttt{gelman.plot}. In Figure 13.4, we can see such a plot for the linear regression example introduced earlier. The PSRF \( \hat{R}_c \) gets closer to 1 as the number of iterations increases, and therefore convergence can be concluded.
Hence values of $Z$ which fall in the extreme tails of $\mathcal{N}(0,1)$ indicate that the chain has not yet converged. In coda a plot function of the $Z$-scores is implemented with `geweke.plot()`.

Simply said, it is a comparison of means of the first 10% and the last 50% of a chain. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke’s statistic has an asymptotically standard normal distribution. The test statistic is a standard $Z$-score: the difference between the two sample means divided by its estimated standard error (taking into account autocorrelation). The Geweke plot (see Figure 13.5) shows $Z$-scores with different numbers of iterations discarded from the chain. The first one is calculated with all samples in the chain, and the last one with only the second half of the chain.

### 13.5 Raftery and Lewis diagnostic

Since the question of main interest before running a MCMC sampler is how many iterations are needed for the distribution to be close enough to the posterior, Raftery and Lewis (1992) presented a method that determines the number of iterations, and burn-ins before the sampling takes place. It can be applied to many different kinds of posterior distributions, however, it fails when there are high posterior correlations between the parameters, and it is not ideal for hierarchical models.

When the interest is in calculating particular quantiles of the posterior distribution of a function $U$ of the parameter $\theta$. Suppose that one is interested in estimating $\Pr(U \leq u \mid y)$ to be within $\pm r$ with probability $s$, where $U$ is a function of $\theta$. The method then finds the approximate number of iterations required to do this when the correct answer is $q$ (the quantile). For example, if $q = 0.5$, with $r = 0.005$ and $s = 0.95$, then we require that the cumulative distribution function of the 0.025%-quantile will be estimated within $\pm 0.005$ with probability 0.95.

The conceptional details of this methods can be found in Raftery and Lewis (1992), and a summary in Cowles and Carlin (1996).

In R the Raftery and Lewis diagnostic is implemented as `raftery.diag()`.
13.6 Heidelberger and Welch diagnostic

An approach that was originally published in the framework of operations research by Heidelberger and Welch (1983) —in short— calculates a test statistic to reject or fail the null hypothesis that the Markov chain is from a stationary distribution. The method provides a way to construct a confidence interval of prespecified width for the mean when there is an “initial transient” —that is, when the simulation does not start off in its stationary distribution. Convergence can then be diagnosed with a hypothesis test with the $H_0$: the sequence of iterates is from a stationary process. In Markov chains, this is equivalent to uniform ergodicity. The test statistic then is distributed as a Brownian bridge. For the application of the Heidelberger and Welch statistic the user has to specify the maximum number of possible iterations that can be run, and $\epsilon$, the desired half-width for confidence intervals. First, 10% of the specified maximum number of iterations are run and the stationary test is applied to the first 10% of these iterations. If the null hypothesis is rejected, then the first 10% of the iterations are discarded and the stationary test is repeated without them. This is done until $H_0$ is not rejected any more. If the half-width of the confidence interval is less than $\epsilon$ times the sample mean of the retained iterates, then the process stops and sample mean and confidence intervals are reported. More details about this method can be found in the original paper, or again in Cowles and Carlin (1996). In R the Heidelberger and Welch diagnostic is implemented as `heidel.diag()`.

heidel.diag(samples[[1]])

```r
heidel.diag(samples[[1]])
```

## Stationarity start p-value
## test iteration
## a passed 1 0.988
## b passed 1 0.980
##
## Halfwidth Mean Halfwidth
## test
## a failed -0.0615 9.69e-03
## b passed 1.0001 1.65e-05

In our example 3746 iterations are needed that the cumulative distribution function of the 0.025%-quantile of $a$ and $b$ will be estimated within $\pm0.005$ with probability 0.95.
Chapter 14

Metropolis–Hastings algorithm

By:
Megan Payne

14.1 Basic ideas

Suppose we have a posterior distribution, with density \( f(\theta) \), that we want to study in some detail. As we learned earlier with the Gibbs sampler, we can sample from the distribution if we know the full conditional distribution of every parameter, \( \theta_k \). However if the full conditional distribution is not known, a different method must be used. So instead, we will simulate realizations from it and study those. To do this, we will simulate a Markov chain whose stationary distribution is our posterior using the Metropolis-Hastings algorithm.

14.2 The Metropolis–Hastings algorithm

Definition 14.1 (Stationary distribution). A distribution with density function \( f \) is said to be a stationary distribution of a chain if there are transition probabilities \( P(x, y) \) such that

\[
\int_{x \in S} f(x) P(x, y) = f(y),
\]

\( \forall y \in S \) where \( S \) is the state space for continuous distribution \( f \).

To sample from \( f \), construct a transition kernel \( p(\theta, \phi) \) using the reversibility condition of a chain such that

\[
f(\theta)p(\theta, \phi) = f(\phi)p(\phi, \theta).
\]

If this equation is satisfied, it is a sufficient condition for \( f \) to be the equilibrium or stationary distribution of the chain. The requirement on \( p \) is somewhat restrictive so \( p \) is further broken
down into an arbitrary transition kernel \( q(\theta, \phi) \) and a probability \( \alpha(\theta, \phi) \) so that \( \forall (\theta, \phi) \\
p(\theta, \phi) = q(\theta, \phi) \alpha(\theta, \phi) \).

The basic Metropolis-Hastings algorithm from Gamerman and Lopes (2006) goes as follows,

1. Initialize a counter, \( j = 1 \) and set an initial value for \( \theta^{(0)} \).
2. Generate a proposed new value \( \phi \) from the density \( q(\theta^{(j-1)} \ldots) \).
3. Evaluate the acceptance probability of the move \( \alpha(\theta^{(j-1)}, \phi) \), which will be shortly defined. If the move is accepted, then assign \( \phi \) to \( \theta^{(j)} \). If the move is not accepted then \( \theta^{(j)} = \theta^{(j-1)} \) and the chain remains for this turn.
4. Iterate the counter from \( j \) to \( j + 1 \) and start again at step 2 until convergence to the stationary distribution is reached.

The natural question from here is how to determine the acceptance or rejection of the move. In step 3, the acceptance probability is defined as \( \alpha_j = \alpha(\theta^{(j-1)}, \phi) \). To evaluate this probability, compare the value of \( \alpha_j \) with a randomly generated value from the uniform distribution. To do this draw \( u \) from \( U(0, 1) \) and if \( u \leq \alpha_j \), the move is accepted, otherwise the move is rejected.

Hastings (1970) suggested the following test ratio
\[
\alpha(\theta^{(j-1)}, \phi) = \min \left\{ 1, \frac{f(\phi) q(\phi, \theta)}{f(\theta) q(\theta, \phi)} \right\},
\]
as an acceptance probability where \( \theta \) is the current position of the chain and \( \phi \) is the proposed move for the chain.

### 14.3 Example: A mixed normal distribution

Suppose the distribution \( f(\theta) \) to be sampled from is a mixed normal distribution. In this case, consider
\[
f(\theta) = 0.3 f_N(\theta; \mu_1, \sigma_1^2) + 0.7 f_N(\theta; \mu_2, \sigma_2^2) \quad (14.1)
\]
where \( f_N \) is a normal density and \( \mu_1 = 0.5, \sigma_1 = 1, \mu_2 = 4, \) and \( \sigma_2 = 0.5 \), respectively. Setting the number of samples to 500, 5000, 50,000 and 500,000, it is clear to see in Figure 14.1 the histogram of values approximate the normal distribution defined in Equation (14.1). This example will be expanded upon throughout this chapter.

### 14.4 Tuning parameter

The tuning parameter can also be seen as the step size between \( \theta^{(j)} \) and \( \phi \). Set the tuning parameter to be small and the chain has a difficult time traversing the whole parameter space, an indication of bad mixing. Another side effect of having too small of a tuning parameter is the acceptance rate for the proposed next step is quite large. If the tuning parameter is too big, the acceptance rate is small and the chain takes longer to converge. This also results in
14.4. TUNING PARAMETER

Figure 14.1: Histograms of random walk chain with variance of $\epsilon = 3$, initial value $\theta^{(0)} = 10$. Successive draws of 500, 5000, 50000, and 500000 samples.

the algorithm spending too much time in the tails of the distribution. For this reason, it is important to find a tuning parameter that is a good size.

To illustrate the concept of tuning parameters, the mixed normal example will be displayed graphically using trace plots. It is worth noting that trace plots should only be used as a way to eyeball convergence to the stationary distribution. Convergence should be checked analytically using methods described in Chapter 13. More information about traceplots can be found in 13.2.2.

14.4.1 Example: Mixed normal with variation of tuning parameters

Figure 14.2 illustrates how different tuning parameters affect the mixing and acceptance of proposed values in the chain. The first trace plot has tuning parameter 0.5, which is quite small. Notice the large acceptance rate indicated by the few number of flat spots where the position did not move. Another aspect to note of this example is how the algorithm stays on the second half of the mixed normal where the mean $\mu_2 = 4$.

In contrast, the second plot has tuning parameter 3. The acceptance rate is 33%, which falls within the guidelines from Gamerman and Lopes (2006) of 20% to 50% for acceptance rates. Visually, the trace plot appears to mix well but convergence should be evaluated using techniques described in Chapter 13.

The last trace plot has tuning parameter equal to 10. This larger tuning parameter allows for quite a large distance between the current chain value and the proposed chain value. With
an acceptance rate of only 15%, convergence to the stationary distribution might take a long time.

\begin{figure}
\centering
\begin{subfigure}[b]{0.4\textwidth}
\includegraphics[width=\textwidth]{figure1}
\caption{Iteration, tuning = 0.5}
\end{subfigure}
\begin{subfigure}[b]{0.4\textwidth}
\includegraphics[width=\textwidth]{figure2}
\caption{Iteration, tuning = 3}
\end{subfigure}
\begin{subfigure}[b]{0.4\textwidth}
\includegraphics[width=\textwidth]{figure3}
\caption{Iteration, tuning = 10}
\end{subfigure}
\end{figure}

\textbf{Figure 14.2:} Trace plot of random walk chain with three different variances for $\epsilon$, 0.5, 3, and 10, initial value $\theta^{(0)} = 10$.

\subsection{14.5 Acceptance rates}

When programming the Metropolis-Hastings algorithm, it is important to keep track of the acceptance rate of the proposed step in the Markov chain. Ideal acceptance rates have been proposed Gamerman and Lopes (2006) and range from 20% to 50%.

\subsection{14.6 Burn-in and thinning}

To account for the initial journey from the starting value, $\theta^{(0)}$, a burn-in period is used where the first $n$ values are removed and discarded. Using a burn-in period will mitigate the choice in initial value. This allows for an arbitrary initial value.

Since the Metropolis-Hastings generates a Markov Chain, adjacent draws $\theta^{(k)}$ and $\theta^{(k+1)}$ are slightly autocorrelated. Correlation can be reduced by thinning, a process where 1 out of every $n$ draws are kept and the remainder are thrown away. Thinning will result in a smaller sample size so this will increase runtime to generate an adequate number of simulated values.
14.7 Special cases

14.7.1 Special cases: Symmetric chains

A symmetric chain is defined by its transition kernel, \( p \). A chain is said to be symmetric if \( p(\theta, \phi) = p(\phi, \theta) \), \( \forall (\phi, \theta) \) paired states. In Metropolis et al. (1953), the example of a chain is a symmetric chain with the properties on \( q \). If \( q \) depends on \( (\theta, \phi) \) only through the distance, \(|\phi - \theta|\), it is easy to see that \( q(\theta, \phi) = q(\phi, \theta) \).

A helpful property of symmetric chains comes directly from its acceptance criteria because if \( q(\theta, \phi) = q(\phi, \theta) \) then the acceptance criteria can be reduced as follows,

\[
\alpha(\theta, \phi) = \min \left\{ 1, \frac{f(\phi)q(\phi, \theta)}{f(\theta)q(\theta, \phi)} \right\} = \min \left\{ 1, \frac{f(\phi)}{f(\theta)} \right\}.
\]

Since acceptance no longer depends on \( q \), the computational time and complexity may be reduced.

14.7.2 Special cases: Random walk chains

Previously, a random walk has been defined as a Markov chain with \( \theta^{(j)} = \theta^{(j-1)} + \epsilon_j \), where \( \epsilon_j \) is an iid random variable with a distribution independent of the chain. Then \( \theta^{(j)} - \theta^{(j-1)} \) is also iid and has a distribution independent of the chain.

So far, the mixed normal sample has been simulated using a random walk chain with \( \epsilon_j \sim N(0, \sigma_{\text{tuning}}^2) \).

14.7.3 Special cases: Independence chains

An independence chain happens when the proposed transition does not depend on the current position, \( \theta \), of the chain. In this case, \( q(\theta, \phi) \) is the density evaluated at \( \phi \), so \( f(\phi) \). Although the transition kernel does not depend on the current position of the chain, the acceptance probability still does, so it is considered a Markov chain.

14.7.4 Independence chain example

Using the mixed normal example, we generate a sample using an independence chain. The histograms in Figure 14.3 represent a sample using the normal distribution for \( q \) with mean 0 and variance \( \sigma_{\text{tuning}}^2 = 3 \). The sample sizes for the histograms are 500, 5000, 50,000, and 500,000.

The trace plots in Figure 14.4 were generated from a sample using an independence chain Metropolis-Hastings. The flat regions on the trace plots correspond to areas of high rejection. When the proposed move is rejected, the chain remains and the position on the graph does not move vertically.
CHAPTER 14. METROPOLIS–HASTINGS ALGORITHM

Figure 14.3: Histograms of independence chain with variance for $\epsilon = 3$, initial value $\theta^{(0)} = 0$. Successive draws of 500, 5000, 50000, and 500000 samples.

Figure 14.4: Trace plot of independence chain with three different variances for $\epsilon$, 0.5, 3, and 10, initial value $\theta^{(0)} = 0$. 
14.8 Exercises


\[ f(\theta) = 0.9 f_N(\theta; 0, 1) + 0.1 f_N(\theta; 3.5, 0.5). \]

Sample from \( f(\theta) \) using either a random walk Metropolis or independence chain Metropolis algorithm. Use initial value \( \theta(0) = -5 \) or \( \theta(0) = -7 \). For a random walk Metropolis, the variance of \( \epsilon \) is \( \sigma^2 = 0.1^2 \) or \( \sigma^2 = 0.5^2 \). For an independence walk Metropolis, use a normal distribution with mean 0 and variance 1 or 9.

1. Sample from \( f(\theta) \) at least 10000 times and check the density distribution.

2. Sample from \( f(\theta) \) 1000 times and generate a trace plot.

3. Looking at the trace plot, is a burn-in period recommended?

4. Do you think there is bad mixing and why?
Chapter 15

Markov chains and applications in veterinary diagnostic testing

By:
Sonja Hartnack

15.1 Introduction to Markov chains

Markov chains are named after the Russian mathematician Andrei Andreivich Markov. Markov chains are at the basis of MCMC (Markov chain Monte Carlo) simulations and are applied in a variety of different fields. They designate random or stochastic processes that undergo transitions from one state to another on a state space.

A Markov chain possesses a property that is characterised as “memoryless”: the probability of the next state depends only on the current state and not on the sequence of events that preceded it.

More formally, the Markov property is defined as:

\[ \Pr(X_{n+1} = x|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x|X_n = x_n) \]

assuming that both conditional probabilities are defined and that

\[ \Pr(X_1 = x_1, \ldots, X_n = x_n) > 0. \]

For constructing a Markov chain it is necessary to specify its transition matrix designating the probabilities of transition between the different states. Transition matrices have all lines summing to one.

\[
P = \begin{pmatrix} A & \bar{A} \\ \bar{A} & \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix} \end{pmatrix}
\]
Here for the two discrete states, being either $A$ or $\bar{A}$, the transition matrix $P$ indicates that for being in $A$ there is a transition probability of 0.9 to stay in state $A$ and of 0.1 to switch to $\bar{A}$. Similarly, if being initially $\bar{A}$, there is a transition probability of 0.7 to switch to $A$ and of 0.3 to stay in $\bar{A}$.

If the initial (or starting) values of $A$ and $\bar{A}$ are known, typically named $S_0$, the initial state distribution, then by multiplying it with $P$, the next (or first) state distribution $S_1$ can be derived.

For example for $A = 0.2$ and $\bar{A} = 0.8$

$$S_0 \cdot P = S_1$$

$$(0.2 \quad 0.8) \cdot \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix} = \begin{pmatrix} 0.74 & 0.26 \end{pmatrix}$$

which indicates a huge “jump” from the probability of 0.2 of $A$ to 0.74 at $S_1$. For $S_2$ the probability of $A$ would be 0.848 and for $S_3$ 0.8696 and so on.

The initial or starting values are of minor importance, since after a number of steps, the Markov chain will eventually converge to a stationary distribution. A distribution $\pi$ is called stationary if $\pi = \pi \cdot P$, and thus will retain this distribution for all subsequent stages.

Considering MCMC, the property of a Markov chain of being ergodic, is important. This means that during numerous iterations, the chain will explore every point (or possible state) and will do so proportionally to its probability. To be considered ergodic, the Markov chain must be

1. irreducible: for every state there is a positive probability of moving to any other state;
2. aperiodic: the chain must not get trapped in cycles.

If ergodicity of a chain is established, important limiting theorems can be stated, i.e., Markov chain equivalents to the law of large numbers and to the central limit theorem.

Another important property of Markov chains is being reversible. For a stationary distribution $\pi$ and a transition matrix $P$, the reversibility condition can be written as (Gamerman and Lopes, 2006)

$$\pi(x)P(x, y) = \pi(y)P(y, x), \text{ for all } (x, y) \in S.$$ 

It can be interpreted as saying that the rate at which the system moves from $x$ to $y$ when in equilibrium, $\pi(x)P(x, y)$, is the same as the rate at which it moves from $y$ to $x$, $\pi(y)P(y, x)$. Reversible chains are useful, because constructing Markov chains with a given stationary distribution $\pi$ reduces to finding transition probabilities $P(x, y)$ satisfying the condition of reversibility.

To construct Markov chains, two main algorithms are utilised: Gibbs and Metropolis–Hastings algorithms. Gibbs sampling is a special case of Metropolis–Hastings sampling. In their famous paper from 1953, Metropolis et al. (1953) showed how to construct a Markov chain with stationary distribution $\pi$ such that $\pi(x) = p_x$, $x \in S$. Here the first step to obtaining the stationary distribution of a Markov chain is to prove that the probabilities of a distribution satisfy the reversibility condition.
15.2 Diagnostic test accuracies

Diagnostic tests in veterinary medicine are of paramount importance since they are used in virtually all problem-solving activities. Veterinarians need diagnostic tests to obtain a diagnosis and eventually a prognosis for an individual patient. Diagnostic tests are also at the basis of classifying herds as being exposed, infected or free of a specific pathogen. Being certified as “officially-free” of a specific pathogen may be linked to a higher sanitary status in QS-Systems with a financial advantage or even to lifting restrictions after disease outbreaks. All surveillance and monitoring activities in the context of animal disease control relevant for international trade rely on the (known) performance of diagnostic tests. Similarly, for conducting import risk assessments, knowledge of diagnostic tests’ performance in the field is essential. Risk factor studies assessing the relevance of specific risk factors for infection, development or progression of disease, recovery, death or else encompass always diagnostic tests to define the status of interest. In consequence, uncertainty about the performance of a diagnostic test may lead to a flawed risk factor analysis. Therefore, “understanding of the principles of test evaluation and interpretation are basic to many of our activities” (Dohoo et al., 2009).

Diagnostic sensitivity and specificity are defined as the ability of a diagnostic test (i.e., an ELISA, an enzyme-linked immunosorbent assay) to correctly classify the exposed and unexposed individuals, respectively. Sensitivity and specificity are conditional probabilities, thus conditioning on the true disease status

\[
Se = \Pr(T^+|D^+)
\]

and

\[
Sp = \Pr(T^-|D^-).
\]

15.3 No gold standard models

15.3.1 Frequentist approach

In the absence of a true gold standard, estimating sensitivity and specificity from cross-tabulated test results from two diagnostic tests becomes impossible since only three parameters can be estimated, but five parameters (two sensitivities, two specificities and one prevalence) need to be estimated.

In 1980 Hui and Walter introduced a model allowing the estimation of sensitivity and specificity of two tests, based on cross-tabulated results from two populations with different prevalences (Hui and Walter, 1980). This applies when the following three assumptions are fulfilled (Toft et al., 2005).

1. The population is divided into two or more populations in which two or more tests are evaluated.

2. Sensitivity and specificity are the same in all populations.

3. The tests are conditionally independent given the disease status.
Figure 15.1: Hui–Walter paradigm: In the absence of a gold standard, test results from two diagnostic tests cannot be correctly classified in one contingency table. If — based on the Hui–Walter paradigm — the two tests are applied to individuals from two populations with different prevalences, estimation of the unknown parameters might be possible.

After Hui and Walter was named a paradigm, with $S$ indicating the number of populations and $R$ the number of tests:

$$S \geq \frac{R}{(2^R - 1)}.$$

If this condition is fulfilled, any combination of $S$ and $R$ may allow to estimate $Se$ and $Sp$, e.g., $(2T, 2P), (3T, 1P), (4T, 1P), \ldots$.

A Maximum Likelihood (ML) approach to estimate the values of the unknown parameters (sensitivities, specificities and prevalences) can be utilised. Here the value of the unknown parameters are estimated from the sample data such that the values chosen maximize the probability of obtaining the observed data. For each of the four cells in the $2 \times 2$ table, the likelihood contributions are determined as the probability of observing data in each cell conditional on the parameters, raised to the power of the observed frequency for that cell.

For example for two diagnostic tests named $T_1$ and $T_2$ the probabilities of the four different options of binary test results ($+, +, -, -, -)$ could be modelled as follows

$$
\begin{align*}
Pr(T_1^+, T_2^+) &= \left[ \text{Pre} \cdot Se_1 \cdot Se_2 + (1 - \text{Pre}) \cdot (1 - Sp_1) \cdot (1 - Sp_2) \right]^{a_i} \\
Pr(T_1^+, T_2^-) &= \left[ \text{Pre} \cdot Se_1 \cdot (1 - Se_2) + (1 - \text{Pre}) \cdot (1 - Sp_1) \cdot Sp_2 \right]^{b_i} \\
Pr(T_1^-, T_2^+) &= \left[ \text{Pre} \cdot (1 - Se_1) \cdot Se_2 + (1 - \text{Pre}) \cdot Sp_1 \cdot (1 - Sp_2) \right]^{c_i} \\
Pr(T_1^-, T_2^-) &= \left[ \text{Pre} \cdot (1 - Se_1) \cdot (1 - Se_2) + (1 - \text{Pre}) \cdot Sp_1 \cdot Sp_2 \right]^{d_i}.
\end{align*}
$$

The likelihood function for the unknown parameters based on the overall data is obtained by multiplying the likelihood contributions across the $i$ populations (or summing the loglikelihood
15.3. NO GOLD STANDARD MODELS

contributions) of the number of independently sampled populations. After the (log)likelihood function for the six unknown parameters has been formulated, first and second derivatives are needed to obtain the maximum likelihood estimates (MLEs) for the sensitivities, specificities and prevalences as well as the variance-covariance matrix. Since an exact analytical solution will not exist, numerical iteration methods such as Newton–Raphson or Expectation-Maximisation (EM) algorithm are needed to estimate the six model parameters.

15.3.2 Criticism of the Hui-Walter paradigm

Assumption 1 of distinct prevalences is necessary for the Hui-Walter model because otherwise, the data can be collapsed into a single $2 \times 2$ table with only three degrees of freedom for estimation. Whereas Enoe et al. (2000) state that the “MLE exhibited little bias even when the difference in prevalence was as little as 0.01”, Toft et al., 2005 mention that a difference of at least 0.1 is needed. In case of rare diseases, it might be difficult to find populations with prevalences higher than 0.1.

Regarding assumption 2, if the accuracies would differ two populations, this would add four additional parameters to be estimated, while there are only three additional degrees of freedom.

Assumption 3 about conditional independence of diagnostic test accuracies was first questioned by Vacek (1985) indicating that not not considering these might lead to overestimating prevalence and diagnostic test accuracies. Conditional independence implies that given an animal is diseased (or not) the probability of positive (or negative) outcomes for $T_1$ is the same regardless of the known outcome for $T_2$.

Conditional dependence, in contrast, implies that

$$\Pr(T_1^+ | T_2^+) \neq \Pr(T_1^+ | T_2^-)$$

and / or

$$\Pr(T_1^- | T_2^-) \neq \Pr(T_1^- | T_2^+)$$.

Seen from a biological perspective, conditional dependency between two diagnostic tests could occur if both tests are based on the same biological principle. The specificities of for example two ELISAs might be conditionally dependent because they are both affected by the same cross-reacting agent.

15.3.3 Bayesian approach

Incorporating also conditional dependencies between diagnostic sensitivities ($\gamma_{Se}$) and specificities ($\gamma_{Sp}$) (Georgiadis et al., 2003) leads to the following equations

$$\Pr(T_1^+, T_2^+) = [\text{Pre} \cdot (\text{Se}_1 \cdot \text{Se}_2 + \gamma_{Se12}) + (1 - \text{Pre}) \cdot ((1 - \text{Sp}_1) \cdot (1 - \text{Sp}_2) + \gamma_{Sp12})]^a$$

$$\Pr(T_1^+, T_2^-) = [\text{Pre} \cdot (\text{Se}_1 \cdot (1 - \text{Se}_2) - \gamma_{Se12}) + (1 - \text{Pre}) \cdot ((1 - \text{Sp}_1) \cdot \text{Sp}_2 - \gamma_{Sp12})]^b$$

$$\Pr(T_1^-, T_2^+) = [\text{Pre} \cdot ((1 - \text{Se}_1) \cdot \text{Se}_2 - \gamma_{Se12}) + (1 - \text{Pre}) \cdot \text{Sp}_1 \cdot (1 - \text{Sp}_2) - \gamma_{Sp12})]^c$$

$$\Pr(T_1^-, T_2^-) = [\text{Pre} \cdot ((1 - \text{Se}_1) \cdot (1 - \text{Se}_2) + \gamma_{Se12}) + (1 - \text{Pre}) \cdot (\text{Sp}_1 \cdot \text{Sp}_2 + \gamma_{Sp12})]^d.$$
15.4 Veterinary applications

In the following a few veterinary examples utilising MCMC simulations in OpenBUGS or JAGS to estimate diagnostic test accuracies in the absence of a true gold standard are given.

15.4.1 Bovine tuberculosis

Bovine Tuberculosis elimination programs are based on test-and-slaughter strategies. No diagnostic test used in the control of Bovine Tuberculosis (bTB), a zoonotic disease, is a perfect gold standard having a diagnostic sensitivity and specificity of 100%. To find relevant data to assess diagnostic accuracy of one variant of the tuberculin test, the Single Intradermal Comparative Cervical Tuberculin Test (SICCT) used in the diagnosis of bTB a systematic literature search was performed (Hartnack and Torgerson, 2012). Subsequently, to estimate diagnostic test accuracies of the SICCT, a Bayesian latent class analysis has been undertaken by using the data from (Liebana et al., 2008). Under the assumption that all bacteriologically confirmed M. bovis are truly M. bovis, we set the specificity of culture equal to 100%. The presence of conditional dependencies between tests was checked by assessing separately the impact of pairs of covariance terms (conditional on a subject being disease positive or disease negative, beta distribution Be(1,1)), compared to a covariance term set to 0 on the other estimates. Model selection was performed by monitoring the Deviance Information Criterion (DIC) and by the effective number of parameters ($p_D$) in the fitted model (Spiegelhalter et al., 2002) where a lower DIC and a higher $p_D$ indicated a better model fit. Models were fitted using Markov Chain Monte Carlo sampling in the software Open Bugs. Model diagnostics was performed by visually checking the convergence of three independent chains and by using the usual Gelman–Rubin diagnostics.

Specificity of gross lesion detection and histopathology were estimated as 90% and 97% respectively. Estimates of the sensitivity of the Single Intradermal Comparative Cervical Tuberculin Test (SICCT), gross lesion detection, histopathology and bacteriological isolation were 80%, 89%, 93% and 98% respectively.

15.4.2 Brachyspira

For swine dysentery, which is caused by Brachyspira hyodysenteriae infection and is an economically important disease in intensive pig production systems worldwide, a perfect or error-free diagnostic test (“gold standard”) is not available. In contrast to risk factor studies in food animals, where adjustment for within group correlations is both usual and required for good statistical practice, diagnostic test evaluation studies rarely take such clustering aspects into account, which can result in misleading results. The aim of the present study was to estimate test accuracies of a PCR originally designed for use as a confirmatory test, displaying a high diagnostic specificity, and cultural examination for B. hyodysenteriae Hartnack et al. (2014).

This estimation was conducted based on results of 239 samples from 103 herds originating from routine diagnostic sampling. We used a Bayesian latent class modelling approach comprising of a hierarchical beta-binomial approach with gamma hyperpriors. This allowed prevalence across individual herds to vary as herd level random effect. The estimated diagnostic sensitivity of
15.4. VETERINARY APPLICATIONS

PCR (95% credible intervals) and culture were 73.2% (62.3 to 82.9) and 88.6% (74.9 to 99.3), respectively. The estimated specificity of the PCR was 96.2% (90.9 to 99.8).

15.4.3 Echinococcus

The diagnosis of canine echinococcosis, a zoonosis, can be a challenge in surveillance studies because there is no perfect gold standard that can be used routinely. We utilised a set of faecal and purge samples used previously to explore the epidemiology of canine echinococcosis on the Tibetan plateau. Previously only the purge results were reported and analysed in a largely deterministic way. In the present study, additional diagnostic tests of copro-PCR and copro-antigen ELISA were undertaken on the faecal samples. This enabled a Bayesian analysis in a latent-class model to examine the diagnostic performance of a genus specific copro-antigen ELISA, species-specific copro-PCR and arecoline purgation (Hartnack et al., 2013). Potential covariates including coinfection with Taenia, age and sex of the dog were also explored. The dependence structure of these diagnostic tests could also be analysed. Principle findings: The most parsimonious result, indicated by deviance-information criteria, suggested that co-infection with Taenia spp. was a significant covariate with the Echinococcus infection. The copro-PCRs had estimated sensitivities of 89% and 84% respectively for the diagnoses of Echinococcus multilocularis and E. granulosus. The specificities for the copro-PCR were estimated at 93 and 83% respectively. Copro-antigen ELISA had sensitivities of 55 and 57% for the diagnosis of E. multilocularis and E. granulosus and specificities of 71 and 69% respectively. Arecoline purgation with an assumed specificity of 100% had estimated sensitivities of 76% and 85% respectively. Significance: This study also shows that incorporating diagnostic uncertainty, in other words assuming no perfect gold standard, and including potential covariates like sex or Taenia co-infection into the epidemiological analysis may give different results than if the diagnosis of infection status is assumed to be deterministic and this approach should therefore be used whenever possible.

15.4.4 Criticism of the Deviance information criterion (DIC)

The DIC is widely used for model selection in Bayesian latent class analysis for veterinary diagnostic tests. One difficulty encountered in relying on DIC for model selection is that is not always possible to decide, simply based on the DIC values, which model is the best one. Whereas the BUGS program started in 1989, and in 1992 a program that “could fit arbitrarily complex Bayesian models using Markov chain Monte Carlo methods” (Spiegelhalter et al., 2014) was available, it took until 1996 that Spiegelhalter suggested a first measure of fit, the posterior mean deviance. This was replaced in 2002 by Spiegelhalter et al., 2002 and widely used in veterinary Bayesian latent class analyses. Responding to several critical remarks, Spiegelhalter et al present their view in a paper titled “The deviance information criterion: 12 years on”. Asking “What is wrong with DIC?” a number of aspects are described by:

1. $p_D$ (the effective number of parameters) is not invariant to reparameterization;

2. lack of consistency;

3. DIC is not based on a proper predictive criterion;
4. DIC has a weak theoretical justification.

The concluding remarks of Spiegelhalter et al. (2014) for using DIC are:

“Nevertheless there remain reasonable criticisms of its derivation and use. DIC has stimulated rich developments, and we eagerly await routine implementation of worthy alternatives.”
Bibliography


Ando, T. (2010). *Bayesian Model Selection and Statistical Modeling*. CRC Press. 61, 65


Curtis, S. M. (2015). *mcmcplots: Create Plots from MCMC Output*. R package version 0.4.2. 101


Contributor Index

Aktaa, Rasha ................................................................. 71
Elriedy, Amr .............................................................. 77
Flury, Roman ............................................................... 85
Hamza, Tasnim ......................................................... 37
Hartnack, Sonja ......................................................... 119
Hediger, Michael ..................................................... 47
Kaufmann, Marco ................................................... 29
Li, Ke ................................................................. 21
Meier, Andrea .......................................................... 1
Panunzi, Silvia ........................................................ 61
Payne, Megan ......................................................... 111
Reeve, Kelly ............................................................ 7
Schleiniger, Marco .................................................. 15
Schuster, Thimo ....................................................... 93
Steffen, Verena ....................................................... 101