MATH11177: Bayesian Theory

Course Notes

Prof Ruth King and Dr Gordon Ross

Rev. Thomas Bayes (?)

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

Chapter 0

INTRODUCTION

0.1 Scope of the Course

This course focuses on the theory of Bayesian statistics. There has been a huge increase in the application of the Bayesian paradigm throughout all areas of statistics in recent years, and this has had a considerable impact in the analysis of complex data. The basic underlying principles will be discussed in detail within the course, which is broadly structured into two parts. The first part (Chapter 1: Bayesian Statistics) will describe the Bayesian paradigm, stating Bayes Theorem (the underlying principle that Bayesian inference hinges on), discuss associated issues and consider a range of examples. The latter part of the course (Chapter 2: Bayesian Computation) will look at the associated computational techniques and algorithms (given sufficient computing power) for applying a Bayesian approach to more complex inference problems. Two appendices are also provided within the printed notes as useful reference material. Appendix A provides a summary of common probability distributions that will be used in the course; and Appendix B briefly describes associated `R` distributional commands that may be useful in practice.

The aims of the course are as follows:

- To understand the Bayesian philosophy.
- To be able to specify a prior distribution and derive a posterior distribution (up to proportionality).
- To be able to derive additional related posterior distributions and quantities.
- To understand the theory of Markov Chain Monte Carlo (MCMC) and associated implementation issues.

Note that the course *Bayesian Data Analysis* in Semester 2 builds on the ideas in this course applying the techniques to real data problems using the specialist software BUGS.

0.2 Lecturers

The course will be co-lectured by:

Prof Ruth King - Chapter 1: Bayesian Statistics;

Dr Gordon Ross - Chapter 2: Bayesian Computation.

0.3 Timetable

Lectures:

Tuesday 12.10 in Ashworth Lecture Theatre 3

Friday 12.10 in JCMB room 1501

Workshops:

Friday 9.00 in even weeks in JCMB room 1206C starting in week 2.

The instructors for the workshops will be Prof Ruth King, Dr Gordon Ross, Dr Jonathan Gair, Mr Nicolo Margaritella and Mr Jack Baker.

0.4 Course Texts

The printed notes, associated lectures and associated worksheets contain all the information needed for this course. However, there is an ever increasing range of books, at a range of levels, that cover all or part of the material within this course that may be of interest to supplement the provided notes and/or provide additional examples. Personally I would recommend the following textbooks:

- *Bayesian Statistics: An Introduction* (4th edition). Lee. Wiley
- *Markov chain Monte Carlo: Stochastic simulation for Bayesian Inference* (2nd edition). Gamerman and Lopes. CRC Press
- *Bayesian Data Analysis* (3rd edition). Gelman, Carlin, Stern, Dunson, Vehtari and Rubin. CRC Press

(Earlier editions of these books are equally as good for our purposes).

0.5 Assessment

The course will be assessed by a combination of examination (90%) in the December diet and continuous assessment (10%). The coursework will consist of one assignment that will be given out around the start of week 5 with the associated deadline of the end of week 6. Feedback will be provided by the start of week 9.

Chapter 1

BAYESIAN STATISTICS

1.1 Introduction

The result on which Bayesian inference rests - Bayes Theorem - is uncontroversial. It is simply a result in elementary probability theory, by the Presbyterian minister Reverend Thomas Bayes (1701-61), though this work was only published posthumously in 1763, by his friend Richard Price.

The underlying concept for Bayesian inference essentially works as follows. We have some population parameter θ on which we wish to make inference. Given this parameter(s), θ, we have an associated probability mechanism $f(\boldsymbol{x}|\theta)$ corresponding to the joint probability density/mass function for the random variables $\boldsymbol{X}$ for which we observe the values $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. In the classical approach, θ is considered to be some fixed, but unknown, constant. Inference is then based on the likelihood $f(\boldsymbol{\theta}; \boldsymbol{x}) \equiv f(\boldsymbol{x}|\theta)$. In other words, the classical approach looks at the distribution of the data given the parameter, to estimate the parameter θ. For example, we may calculate the maximum likelihood estimator (MLE) of θ.

Conversely in the Bayesian paradigm, we no longer assume that the parameter(s) have a fixed value, but consider θ to be a random quantity. We then assume that θ has some unknown distribution, which we wish to estimate. This distribution is denoted by $\pi(\theta|\boldsymbol{x})$, and so we look at the distribution of the parameter, given the data. In many ways this is a more natural way to make inference, but we shall see that to achieve this we will have to specify a *prior probability distribution*, denoted by $p(\theta)$, which represents our initial beliefs about the distribution of θ *prior* to observing any data.

In most situations, when we are trying to estimate a parameter θ, we have some knowledge, or preconceptions, about the value of θ before we take into account the data that we observe. For example, suppose that you are working hard at your desk, and glance out of the window to see a large wooden looking object with branches covered in green things. You consider two alternatives: one that it is a tree, the other it is a postman. Obviously you choose that it is a tree, since the object does not look anything like a postman.

We can formulate the process here. Suppose that you denote the event that you see a wooden looking object with green things on by B. Then, let A_1 denote the event that it is a tree, and A_2 that it is a postman. Then, you reject event A_2 and accept event A_1, since,

$$\mathbb{P}(B|A_1) > \mathbb{P}(B|A_2).$$

In this case these probability values are the likelihood values evaluated at the two different parameter values (the object is a tree or a postman). We then decide on the value of the parameter that maximises the likelihood (i.e. it is a tree). Thus, we are maximising the likelihood.

However, suppose that we entertain a third possibility, event A_3, that the object is a plastic replica tree. In this case it might well be that $\mathbb{P}(B|A_1) = \mathbb{P}(B|A_3)$, and yet you would still reject this hypothesis in favour of A_1, i.e. it is a real tree. That is, even though the probability of seeing what you observe (a large wooden looking object with green bits on) is the same whether it is a real tree or a replica tree, your *prior* belief is that it is more likely to be a real tree, and you include this in your decision. However, this might change if, for example, you were working at a desk inside a replica tree factory. Then, your *prior* beliefs would reflect this additional information, and so you may conclude that what you see is a replica tree.

The essential point is that experiments are not abstract devices. Invariably we have some knowledge about the process being investigated before obtaining any data. It is sensible to include this into our inferential process, and Bayesian inference is the mechanism for drawing inference from this combined knowledge. It should be pointed out, however, that although the underlying probabilistic derivation of Bayes' Theorem is uncontroversial, the reliance on the prior beliefs is the main criticism of Bayesian statistics. Whilst advocates of the Bayesian approach see this reliance on a prior distribution as an advantage, opponents point out that using different prior beliefs will lead to different inferences. It is whether or not you see this as a good or bad thing that determines how acceptable you find the Bayesian approach.

In more mathematical terms; in classical statistics we obtain maximum likelihood estimates, by choosing the point in parameter space which maximises the likelihood surface. In Bayesian statistics, we average across the likelihood surface, rather than maximising. This averaging is weighted according to the prior distribution. However, in classical statistics, we often apply different weights to different pieces of information, thus the Bayesian approach is simply a method of incorporating that weighting procedure within a rigid mathematical framework.

1.2 Bayes' Theorem

There are several different ways that Bayes' Theorem can be written, dependent on whether the parameter(s) are discrete or continuous. For simplicity we shall consider the cases separately, beginning with the discrete case.

1.2.1 Discrete case

Let A and B denote possible events, such that $\mathbb{P}(B) > 0$. Then, Bayes' Theorem states that:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Proof:

By definition of conditional probability,

$$\begin{aligned} \mathbb{P}(A|B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \qquad \square \end{aligned}$$

The denominator in the expression for Bayes' Theorem is most often expressed in an alternative way. Suppose that we let A_i, denote a set of mutually exclusive and exhaustive events, for $i = 1, \ldots, n$. Then, by the law of total probability we can express $\mathbb{P}(B)$ in the form,

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Thus, an alternative expression for Bayes theorem is given by,

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

Example

A test for Hepatitis C is given to a population which is described to be possible carriers, although it is believed that only 10% are positive. However, the test itself is only 95% accurate for people who have Hepatitis C, and 80% accurate for those who do not have the disease. What are the probabilities that we obtain a false negative or a false positive? Assuming that we obtain a positive result, the individual is retested, using the same test and where the result is independent of the first. Given that their second test is again positive, what is the probability that they do not have Hepatitis C?

Let the event that an individual has Hepatitis C be denoted by A, and the event that their test result is positive be denoted by B. Then, we have that,

$$\begin{aligned} \mathbb{P}(A) &= 0.1; \\ \mathbb{P}(B|A) &= 0.95; \text{ and} \\ \mathbb{P}(B^c|A^c) &= 0.8. \end{aligned}$$

Then, we want,

$$\begin{aligned} \mathbb{P}(\text{false negative}) &= \mathbb{P}(A|B^c) \\ &= \frac{\mathbb{P}(B^c|A)\mathbb{P}(A)}{\mathbb{P}(B^c)} \\ &= \frac{\mathbb{P}(B^c|A)\mathbb{P}(A)}{\mathbb{P}(B^c|A)\mathbb{P}(A) + \mathbb{P}(B^c|A^c)\mathbb{P}(A^c)} \\ &= \frac{0.05 \times 0.1}{(0.05 \times 0.1) + (0.8 \times 0.9)} \\ &= 0.0069 \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{P}(\text{false positive}) &= \mathbb{P}(A^c|B) \\ &= \frac{\mathbb{P}(B|A^c)\mathbb{P}(A^c)}{\mathbb{P}(B)} \\ &= \frac{0.2 \times 0.9}{(0.2 \times 0.9) + (0.95 \times 0.1)} \\ &= 0.6545 \end{aligned}$$

Now, let C be the event that the second test is positive. Note that since the same test is applied, we have that,

$$\begin{aligned} \mathbb{P}(C) &= \mathbb{P}(B); \text{ and} \\ \mathbb{P}(C|A^c) &= \mathbb{P}(B|A^c). \end{aligned}$$

Then, using Bayes' Theorem, the probability that an individual does not have Hepatitis C, given that the second test is again positive, is given by,

$$\begin{aligned}
\mathbb{P}(A^c|B,C) &= \frac{\mathbb{P}(B,C|A^c)\mathbb{P}(A^c)}{\mathbb{P}(B,C)} \\
&= \frac{\mathbb{P}(B|A^c)\mathbb{P}(C|A^c)\mathbb{P}(A^c)}{P(B)\mathbb{P}(C)} \\
& \quad \text{since the events } B \text{ and } C \text{ are independent} \\
&= 0.476.
\end{aligned}$$

1.2.2 Continuous (single parameter) case

So far we have considered the discrete case. Bayes' Theorem is equally well defined for continuous parameters, and it is this specification of Bayes' Theorem that is most often used (though we only consider the single parameter case here).

Suppose that we have a parameter $\theta \in \Theta$, on which we wish to make inference. We then observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from some known probability density function (pdf) $f(\boldsymbol{x}|\theta)$, which is a function of the parameter value θ. Then Bayes' Theorem states that,

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)p(\theta)}{f(\boldsymbol{x})}$$

Here the term $p(\theta)$ is referred to as the **prior** density and $\pi(\theta|\boldsymbol{x})$ the **posterior** density. Essentially the prior distribution represents the initial beliefs concerning the parameters prior to any data being observed; whereas the posterior distribution represents an update of these beliefs, following the data $\boldsymbol{x}$ being observed. The information contained in the data on the parameter θ is represented by the term $f(\boldsymbol{x}|\theta)$, and is most usually called the **likelihood**. We can write the denominator of Bayes' Theorem as,

$$f(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

Thus, this term is independent of θ, the parameter of interest, and is simply equal to some constant. The term $f(\boldsymbol{x})^{-1}$ is usually referred to as the normalisation constant, and as we shall see is often found by inspection (i.e. by using the fact that the posterior density must integrate to one to be a valid probability density function). The term $f(\boldsymbol{x})^{-1}$ is not of interest in itself as it is independent of θ. Note that in many cases the integration within the normalisation constant may be analytically intractable, or tedious to calculate. More often Bayes' Theorem is quoted as,

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)p(\theta).$$

This formula forms the essential core of Bayesian inference. In practice we typically only need to calculate the posterior distribution up to proportionality.

Bayesian inference obeys what is commonly called the *likelihood principle*, which essentially means that for a given sample of data, $\boldsymbol{x}$, then any probability models that lead to the same likelihood function will yield the same inference for θ. This implies that the data only affects the posterior via the likelihood, $f(\boldsymbol{x}|\theta)$, and that the prior is independent of the data.

Reminder: Likelihood

Recall that the likelihood, $f(\boldsymbol{x}|\theta)$, is the joint probability mass/density function evaluated at the observed data values $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ given the parameter value θ. If the data, $x_1, \ldots, x_n$ are *independent* observations then we can write the likelihood in the form:

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

where $f(x_i|\theta)$ is the probability mass/density function of the datum x_i.

Example

Suppose that we observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, such that each $X_i \overset{iid}{\sim} Exp(\lambda)$. We place the following prior on λ, namely that,

$$\lambda \sim \Gamma(\alpha, \beta).$$

Then, the corresponding posterior distribution for λ is given by,

$$\begin{aligned}
\pi(\lambda|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\lambda)p(\lambda) \\
&= \prod_{i=1}^{n} \lambda \exp(-x_i\lambda) \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\lambda\beta) \\
&\propto \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \times \lambda^{\alpha-1} \exp(-\lambda\beta) \\
&= \lambda^{n+\alpha-1} \exp(-\lambda[n\bar{x} + \beta]) \\
&\qquad\qquad\qquad\qquad \text{where } \bar{x} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i \\
&\propto \frac{(n\bar{x}+\beta)^{n+\alpha}}{\Gamma(n+\alpha)} \lambda^{n+\alpha-1} \exp(-\lambda[n\bar{x}+\beta]) \\
\\
\Rightarrow \lambda|\boldsymbol{x} &\sim \Gamma(n+\alpha, n\bar{x}+\beta).
\end{aligned}$$

Note: The posterior distribution of $\lambda|\boldsymbol{x}$ is of standard form. By inspection, the corresponding normalising constant is given by,

$$f(\boldsymbol{x}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(n+\alpha)}{(n\bar{x}+\beta)^{n+\alpha}},$$

which ensures that the posterior distribution integrates to one.

Example

We observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, such that each $X_i \overset{iid}{\sim} Geom(\theta)$, where we wish to make inference on the parameter θ. Then, initially we must specify a prior on the parameter θ. Suppose that we set,

$$\theta \sim Beta(\alpha, \beta).$$

Then, the posterior distribution is given by,

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\theta)p(\theta) \\
&\propto \prod_{i=1}^{n} \theta(1-\theta)^{x_i-1} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{n+\alpha-1}(1-\theta)^{n\bar{x}-n+\beta-1} \\
&= \theta^{a-1}(1-\theta)^{b-1},
\end{aligned}$$

where $a = n + \alpha$ and $b = n\bar{x} - n + \beta$. Thus,

$$\theta|\boldsymbol{x} \sim Beta(a, b) = Beta(n+\alpha, n\bar{x}-n+\beta),$$

by inspection.

Note that in the previous examples that the posterior distribution belonged to the same family as the prior distribution (irrespective of the dataset). When this is the case the prior distribution is known as a **conjugate** prior.

1.3 Prior distributions

The specification of a prior on the unknown parameters of interest, before observing any data is controversial. Bayesians argue that the Bayesian approach allows the introduction of any external information that may be available in a formal and robust framework. Conversely, classicists/frequentists argue that the analysis of the data should be objective, and any results obtained should be purely on the evidence of the data, and not influenced by subjective priors that will generally differ between individuals. So, this raises the question of how should we assign the prior $p(\theta)$? Let us be clear from the beginning - there is no such thing as the single *correct choice* of $p(\theta)$ for a given problem. The actual choice of prior lies entirely with the statistician and the information and experience s/he has at the time - the prior is determined such that it reflects the prior information. Note that in general the prior specification is non-unique - several different prior specifications may represent the same prior information. It is up to the statistician and scientist to determine what prior to use.

1.3.1 Conjugate priors

Often, a particular distributional family is chosen for the prior, such that the corresponding posterior distribution of the parameter belongs to the same family, irrespective of the sample size and any value of the observations. This leads to the following definition.

Definition 1.1: *A family of distributions, $\mathcal{F}$ is conjugate to a family of sampling distributions, $\mathcal{P}$ if, whenever the prior belongs to the family, $\mathcal{F}$, then for any sample size and any value of observations, the posterior also belongs to the same family.*

Example

Suppose that we are interested in the probability that when we toss a coin that we obtain a head, denoted by p. Then, we toss a coin n times, and obtain x heads. Assuming that we place a $Beta(\alpha, \beta)$ prior on p, what is the posterior distribution for p?

Each coin toss is an independent Bernoulli experiment with probability p. Thus, we have that,

$$X|p \sim Bin(n, p).$$

Then, by Bayes' Theorem,

$$\begin{aligned}\pi(p|x) &\propto f(x|p)p(p)\\ &\propto p^x(1-p)^{n-x} \times p^{\alpha-1}, (1-p)^{\beta-1}\\ &= p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\\ &= p^{a-1}(1-p)^{b-1},\end{aligned}$$

where $a = x + \alpha$ and $b = n - x + \beta$. Then, by inspection, we have that,

$$p|x \sim Beta(a, b) \equiv Beta(x + \alpha, n - x + \beta).$$

Thus, the Beta distribution is a conjugate prior to the Binomial distribution.

Conjugate priors are often specified for statistical (and computational) convenience - with modern computational tools this is now less essential - however it may still have an influence on computational efficiency (see later in the course).

1.3.2 Non-informative/vague priors

An obvious question to ask is: what should we do if we do not have any prior information concerning the parameter of interest? Bayes himself suggested that when this is the case, the Uniform prior should be used, so that $p(\theta) = c$, for all θ. When this is the case, clearly we have that,

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta),$$

i.e. the posterior distribution is the same shape as the likelihood function. Note that in this case, the posterior mode of the distribution is equal to the MLE of the parameter. However, non-linear transformations of the parameter θ, denoted by $\phi = g(\theta)$, say, will result in a non-Uniform prior on this transformed parameter ϕ.

For example, suppose that we place a Uniform prior on $\theta \in [0, 1]$, so that $p(\theta) = 1$. The corresponding prior on $\phi = \theta^2$ is non-Uniform. Namely, we have that using the transformation of variables,

$$p(\phi) = 1. \left|\frac{d\theta}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}$$

Thus, in practice, priors should be specified on the parameter that the statistician is interested in within the analysis.

Jeffreys' prior

Jeffreys suggested a prior based on an invariance rule for one-to-one (bijective) transformations. Suppose that we are interested in the parameter θ, and specify $\phi = h(\theta)$, where h is a bijective function. Then, the prior for θ is the same as for ϕ, when the scale is transformed. Jeffreys' prior is given by,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})},$$

where $I(\theta|\boldsymbol{x})$ is the Fisher's Information, and is defined to be,

$$I(\theta|\boldsymbol{x}) = \mathbb{E}\left(\frac{d \log f(\boldsymbol{x}|\theta)}{d\theta}\right)^2.$$

Essentially, Fisher's information is an indicator of the amount of information supplied by data about an unknown parameter.

Under certain regularity conditions, Fisher's information can also be expressed in the form,

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}\left[\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\right].$$

Fisher's information is generally more easily calculated using this latter expression.

Then, suppose that we wish to consider the parameter ϕ, which is a transformation of the parameter θ, i.e. $\phi = h(\theta)$. If we specify,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})}$$

then,

$$p(\phi) \propto \sqrt{I(\phi|\boldsymbol{x})}.$$

Proof:

Consider Fisher's information as a function of the transformed variable ϕ which can be calculated using:

$$\begin{aligned} I(\theta|\boldsymbol{x}) &= \mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\theta)}{d\theta}\right)^2 \\ &= \mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\phi=h(\theta))}{d\phi}\times\left|\frac{d\phi}{d\theta}\right|\right)^2 \\ &= \left|\frac{d\phi}{d\theta}\right|^2\mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\phi)}{d\phi}\right)^2 \\ &= I(\phi|\boldsymbol{x})\left|\frac{d\phi}{d\theta}\right|^2. \end{aligned} \tag{1.1}$$

Then, the prior on the parameter θ is specified in the form:

$$p(\theta)\propto\sqrt{I(\theta|\boldsymbol{x})}$$

Using the transformation of variable rule, we have that,

$$\begin{aligned} p(\phi) &\propto \sqrt{I(\phi|\boldsymbol{x})\left|\frac{d\phi}{d\theta}\right|^2}\times\left|\frac{d\theta}{d\phi}\right| \\ &= \sqrt{I(\phi|\boldsymbol{x})}, \end{aligned}$$

as required.

□

Note: suppose that we have n independent observations $\boldsymbol{x}=\{x_1,\dots,x_n\}$ generated from the same distribution with pdf f. Then, it can be easily shown that Fisher's information is given by,

$$I(\theta|\boldsymbol{x}) = nI(\theta|x),$$

where $X\sim f$.

Jeffreys' prior can be extended to the case where there are several unknown parameters. Then, Fisher's information is defined as the matrix, with the element in row i and column j given by,

$$(I(\boldsymbol{\theta}|\boldsymbol{x}))_{i,j}=-\mathbb{E}\left(\frac{d^2\log f(\boldsymbol{x}|\boldsymbol{\theta})}{d\theta_i d\theta_j}\right).$$

Then, the prior is specified as,

$$p(\boldsymbol{\theta})\propto\sqrt{\det I(\boldsymbol{\theta}|\boldsymbol{x})}.$$

Example

Let X denote the number of defective items in a batch of n cream cakes, where each cake is defective with probability θ, independently of each other. Then $X\sim Bin(n,\theta)$ and,

$$\log f(x|\theta)=x\log\theta+(n-x)\log(1-\theta)+C,$$

so that,

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

Then,

$$\begin{aligned} I(\theta|x) &= -\mathbb{E}\left(\frac{d^2 \log f(x|\theta)}{d\theta^2}\right) \\ &= \mathbb{E}\left(\frac{X}{\theta^2}\right) + \mathbb{E}\left(\frac{(n-X)}{(1-\theta)^2}\right) \\ &= \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

So that, Jeffreys' prior for the probability parameter θ is,

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}.$$

In other words, $\theta \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$.

Note that although Jeffrey's prior has the desirable property that it is invariant to bijective transformations it can lead to "improper" priors. These are priors that are not "proper" in that they are not a valid probability density function as they do not integrate to one. For example, suppose that we are interested in the parameter $\mu \in \mathbb{R}$. For a given situation we may obtain Jeffrey's prior to be of the form $p(\mu) \propto 1$ for $-\infty < \mu < \infty$. Improper priors can be used but additional care is needed in the analysis as it can (but not always) lead to improper posterior distributions.

Alternative vague or non-informative prior distributions often have a reasonable mean for the distribution, but with a large variance parameter. Again, usually within a Bayesian analysis, several different priors may be considered, each of which may be described to be "vague" or "non-informative", and the sensitivity of the posterior on these priors investigated.

Informative priors

In many experiments or problems, experts will have an understanding of the system and hence some ideas as to what value the parameter(s) should take. This information needs to be independent of the observed data - and so it is useful for experts to specify their prior beliefs before conducting the given experiment. A Bayesian analysis of the subsequent data permits the formal inclusion of these prior beliefs within the analysis of the data within a robust framework via the specification of informative priors. These informative priors are specified that reflect the experts prior knowledge. However, in general an expert will not specify their prior beliefs in the form of a statistical distribution. Thus in order to form a prior distribution that reflects an experts prior knowledge a process of elicitation is used. Elicitation is the process of extracting prior knowledge in a suitable manner to permit the formulation of a suitable prior distribution that represents this information as accurately as possible. In general, the goal of this process is to identify and quantify the key aspects of prior knowledge, since the expert will find these easier to specify. It is, however, important to recognise that no matter how many summaries for a distribution we choose to elicit, they will not identify a complete distribution. For example, a continuous distribution for a parameter θ implies an infinite number of statements about this parameter. Once summaries have been elicited, there will be more than one distributions that fit those summaries. In practice, once we elicit some key aspects of prior knowledge, we fit them to a distribution. We typically discuss the distribution with the expert, for example, showing a plot of the distribution and discussing several aspects or consequences of the given prior. If the expert is happy with the feedback, we consider it to be the prior; else discuss why the distribution does not represent the experts beliefs and adjust the prior distribution accordingly and repeat.

Note that in order to elicit the information from the expert it is important to ask the right questions in understandable terms. For example, asking an expert for the median for the parameter θ may lead to some confusion. Instead we could ask what value of θ would they expect such it is equally likely that θ lies above the value as below. In addition it is useful to identify what information an expert is able to specify, for example, location (mean or median); quantiles (probabilities that the parameter is less than a given value); shape (symmetric or skewed). It is interesting to note that in general psychologists have found that experts are usually overconfident when it comes to assessing their own uncertainty, i.e. they provide distributions with unreasonably small variances.

1.4 Summarising posteriors

The posterior distribution incorporates all the information concerning the parameter of interest, and so is the most informative description of the parameter. Often the distribution is displayed graphically, in order to illustrate the information in a readily interpretable way. However, although the complete specification of the posterior pdf $\pi(\theta|\boldsymbol{x})$ is, for Bayesian statisticians, the desired end product, it may be somewhat of a sophisticated concept for a non-mathematical client, for example. Then, for convenience, a more easily understandable *summary* of the posterior distribution may be desirable. For example, the posterior mean and standard deviation may be given. We shall discuss in more detail a variety of summary statistics that are often used.

1.4.1 Point estimates

There are a variety of different point estimates that are often used to describe the posterior distribution. Some of these have a decision theoretic interpretation. We shall consider three in more detail, which give an indication of the "average" value of the distribution. However, first we shall presentl some decision theory.

Suppose that we wish to estimate the parameter $\theta \in \Theta$. Then, we define a loss function $L(\theta, a)$ to be the associated loss for the estimate a, when the true value is θ. The corresponding Bayes estimator is then chosen to minimise the expectation of the loss function with respect to the posterior distribution, i.e. the posterior expected loss. The corresponding estimate, $\hat{\theta}$, chosen under this rule is called the *Bayes estimate*. Mathematically, the Bayes estimate, $\hat{\theta}$ is defined such that,

$$\begin{aligned} \hat{\theta} &= \min_{a \in \Theta} \mathbb{E}_{\pi}[L(\theta, a)] \\ &= \min_{a \in \Theta} \left[\int_{\theta \in \Theta} L(\theta, a)\pi(\theta|\boldsymbol{x}) d\theta \right]. \end{aligned}$$

We now consider possible summary estimates of a posterior distribution in the context of loss functions. We shall consider three results (the first two of which are derived from the continuous case, the third for the discrete case), relating to point estimates of the posterior distribution.

Theorem 1.1: *The mean of the posterior distribution is the Bayes estimate with respect to the quadratic loss function.*

Proof: The quadratic loss function is given by,

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

Then, the corresponding Bayes estimate $\hat{\theta}$ is defined to be the estimate of θ such that it minimises the posterior expected loss, given by,

$$\mathbb{E}_{\pi}[L(\theta, a)] = \int_{\theta \in \Theta} (\theta - a)^2 \pi(\theta|\boldsymbol{x}) d\theta = g(\theta, a).$$

Differentiating with respect to a and equating to zero, we obtain a stationary point at,

$$\begin{aligned} a &= \frac{\int_{\theta \in \Theta} \theta \pi(\theta|\boldsymbol{x}) d\theta}{\int_{\theta \in \Theta} \pi(\theta|\boldsymbol{x}) d\theta} \\ &= \int_{\theta \in \Theta} \theta \pi(\theta|\boldsymbol{x}) d\theta, \end{aligned}$$

since $\int_{\theta \in \Theta} \pi(\theta|\boldsymbol{x}) d\theta = 1$, as $\pi(\theta|\boldsymbol{x})$ is a probability density function. To show that this is a minimum, taking the second derivative, we obtain,

$$\begin{aligned} \frac{d^2 g(\theta, a)}{da^2} &= 2 \int_{\theta \in \Theta} \pi(\theta|\boldsymbol{x}) d\theta \\ &= 2 > 0, \end{aligned}$$

and hence a minimum.

Thus, we have that,

$$\hat{\theta} = \int_{\theta \in \Theta} \theta \pi(\theta|\boldsymbol{x}) d\theta = \mathbb{E}_{\pi}(\theta),$$

i.e. the Bayes estimate is the posterior mean. □

Theorem 1.2: *The median of the posterior distribution is the Bayes estimate with respect to the absolute error loss function.*

Proof: The absolute error loss function is given by,

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

We shall consider the case where $\theta \in \mathbb{R}$. Then, we choose a, such that it minimises the posterior expected loss,

$$\int_{-\infty}^{\infty} |\theta - a| \pi(\theta|\boldsymbol{x}) d\theta = \int_{-\infty}^{a} (a - \theta) \pi(\theta|\boldsymbol{x}) d\theta + \int_{a}^{\infty} (\theta - a) \pi(\theta|\boldsymbol{x}) d\theta \tag{1.2}$$

We shall use the fact that,

$$\begin{aligned} \frac{d}{d\alpha} \int_{\alpha}^{\beta} f(x) dx &= \frac{d}{d\alpha}(F(\beta) - F(\alpha)) \\ &= -F'(\alpha) \\ &= -f(\alpha); \end{aligned}$$

and similarly,

$$\frac{d}{d\beta} \int_{\alpha}^{\beta} f(x) dx = f(\beta).$$

Then, differentiating equation (1.2) with respect to a and equating to zero, we obtain,

$$\int_{-\infty}^{a} \pi(\theta|\boldsymbol{x}) d\theta = \int_{a}^{\infty} \pi(\theta|\boldsymbol{x}) d\theta,$$

that is,

$$2\int_{-\infty}^{a} \pi(\theta|\boldsymbol{x})d\theta = \int_{-\infty}^{\infty} \pi(\theta|\boldsymbol{x})d\theta = 1.$$

Thus,

$$\int_{-\infty}^{a} \pi(\theta|\boldsymbol{x}) = \frac{1}{2},$$

and so $\hat{\theta} = a$ is the median of the posterior distribution. □

Theorem 1.3: *The mode of the posterior distribution is the Bayes estimate with respect to the zero-one loss function.*

Proof: We consider the discrete case only. The loss function is of the form,

$$L(\theta, a) = \begin{cases} 1 & \text{if } a \neq \theta; \\ 0 & \text{if } a = \theta, \end{cases}$$

and we wish to minimise,

$$\sum_{\theta\in\Theta} L(\theta, a)\pi(\theta|\boldsymbol{x}).$$

If we choose $a = \theta^*$, this expression becomes,

$$\sum_{\theta\in\Theta\backslash\theta^*} \pi(\theta|\boldsymbol{x}) = 1 - \pi(\theta^*|\boldsymbol{x}).$$

Thus to minimise this, we simply maximise $\pi(\theta^*|\boldsymbol{x})$. In other words we should take $\hat{\theta}$ to be the "most likely" value of θ in the posterior distribution. □

Example

Consider the earlier example where $X|p \sim Bin(n,p)$ with associated prior $p \sim Beta(\alpha, \beta)$. Then we showed that the posterior distribution for p is given by $p|x \sim Beta(x+\alpha, n-x+\beta)$. We assume that there is no prior information on p and specify $\alpha = \beta = 1$, i.e. a $U[0,1]$ prior on p. Figure 1.1 shows the corresponding posterior distribution of p for one replicate of the experiment when $n = 4, 32, 128, 512$. Note that in this simulation, we obtained $x = 3, 20, 61, 248$, respectively.

We can see from Figure 1.1 that as we obtain more information about the parameter, through repeated coin tosses, the posterior distribution becomes more and more peaked.

We can obtain further insight into the way in which the posterior distribution combines information from the data with that from the prior, by considering the form of the posterior mean, which we recall is the Bayes estimate corresponding to the assumption of quadratic loss. We have that,

$$\mathbb{E}_\pi(p) = \frac{x+\alpha}{n+\alpha+\beta}.$$

We can rewrite this expectation in the form,

$$\mathbb{E}_\pi(p) = \frac{(\alpha+\beta)\left(\frac{\alpha}{\alpha+\beta}\right) + n\left(\frac{x}{n}\right)}{n+\alpha+\beta},$$

which can be reformulated as,

$$(1-w)\left(\frac{\alpha}{\alpha+\beta}\right) + w\left(\frac{x}{n}\right),$$

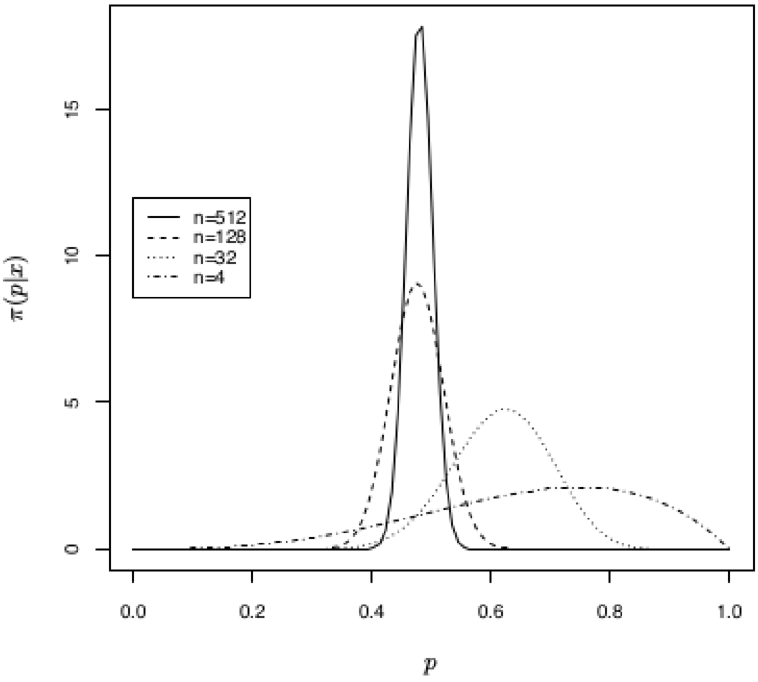


Figure. 1.1: The posterior distribution of p, the probability of obtaining a head when tossing a coin, for (i) $n = 4, x = 3$, (ii) $n = 32, x = 20$, (iii) $n = 128, x = 61$ and (iv) $n = 512, x = 248$, with a $U[0, 1]$ prior on p.

where $w = n/(n + \alpha + \beta)$. In other words, the Bayes estimate is a *weighted average* of the two quantities,

$$\frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \frac{x}{n}.$$

The first is the mean of the prior distribution and is the Bayes estimate we would use if we had no data (assuming quadratic loss). The latter is the "usual" classical estimate of p, derived via maximum likelihood or minimum variance unbiased estimation.

In an obvious sense, we see that our estimate is a combination of what the data tells us, x/n, and what we believed before observing any data, $\alpha/(\alpha + \beta)$. As the amount of data increases i.e. as n increases, more and more weight is placed on x/n; mathematically, in the limiting case, as $n \to \infty$, we have that $w \to 1$. Conversely, if we have no data, i.e. $n = 0$, then $w = 0$ and our only source of information on the parameter is contained within the prior.

Usually, we are not only interested in the Bayes estimate of the parameter, but the whole shape of the posterior distribution of the parameter. Often, we may also be interested in the "spread" of the distribution. Clearly, from Figure 1.1, as the number of coin tosses, n, increases, the precision of the posterior distribution for p increases, as we have more information on the parameter from the data. This can be seen formally, by considering the posterior variance for p,

$$Var_\pi(p) = \frac{(x+\alpha)(n-x+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)}$$

So, that in the limiting case, as $n \to \infty$, we have that $Var_\pi(p) \to 0$. Thus, irrespective of our prior beliefs represented by the prior parameters α and β, as the amount of information increases, the posterior distribution becomes more and more dominated by the data, and our posterior beliefs become more and more concentrated on a value of p tending to a value of x/n.

In general, since the posterior distribution is formed by combining the likelihood with the prior, there is a trade-off between the information contained in the data and the strength of the prior beliefs. Posterior distributions are often said to be "data-driven" if the likelihood dominates the posterior; and "prior-driven" if the prior dominates the posterior.

Example

Consider the above example, where we toss n coins and obtain x number of heads. Then, suppose that we have two different priors on the probability p:

$$\begin{aligned} p &\sim Beta(2,10) \\ p &\sim Beta(10,2). \end{aligned}$$

We toss the coin 5 times and obtain 3 heads, i.e. $n = 5$, and that $x = 3$. The priors, likelihood function and corresponding posterior distributions are given in Figure 1.2. Clearly, here we can see that the prior dominates the posterior distribution. I.e. the posterior distribution looks more like the prior than the likelihood.

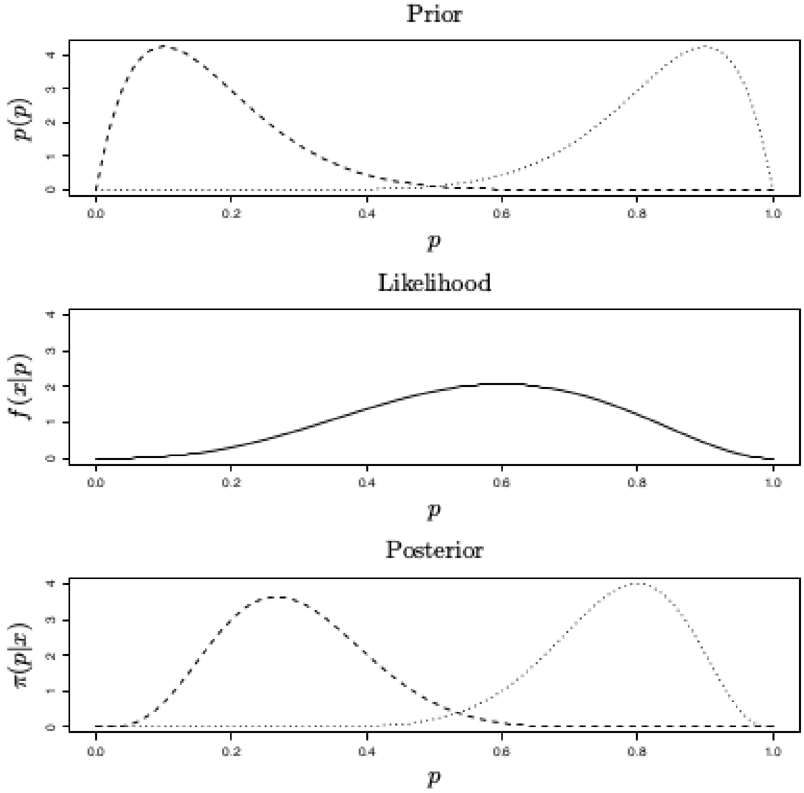


Figure. 1.2: The prior, likelihood and posterior distribution of p when $n = 5$, for prior 1: $p \sim Beta(2,10)$ (dashed line) and prior 2: $p \sim Beta(10,2)$ (dotted line).

However, suppose that we continue to toss the coin, so that we toss the coin a total of 500 times, and obtain 242 heads. The corresponding priors, likelihood and posterior distributions are given in Figure 1.3. Clearly, here the posterior distribution is dominated by the likelihood term, which contains the information contained in the data. Thus, the posterior distribution is data-driven, with little influence from the prior distribution.

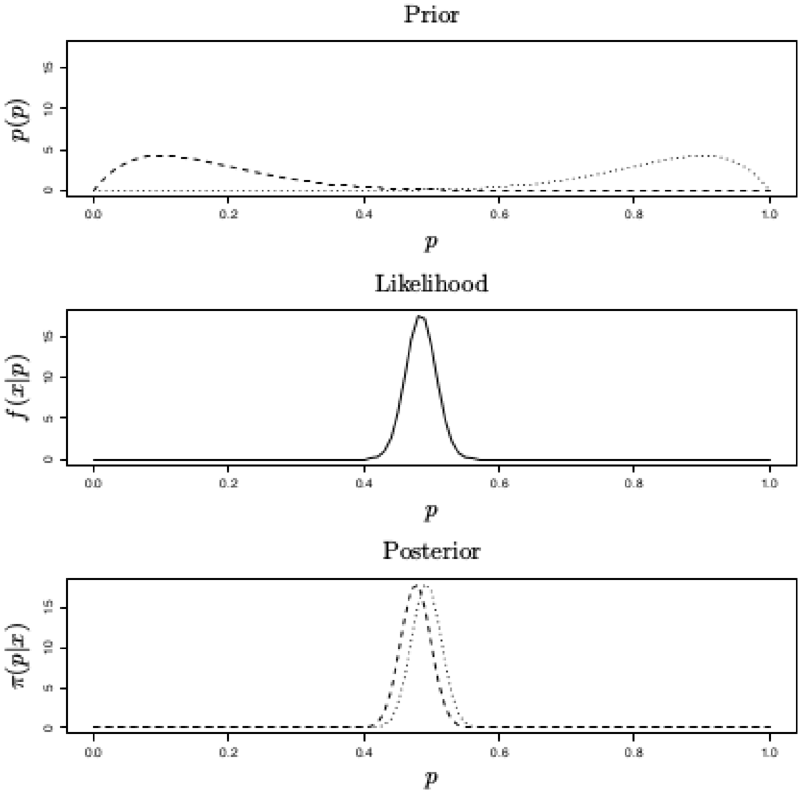


Figure. 1.3: The prior, likelihood and posterior distribution of p when $n = 500$, for prior 1: $p \sim Beta(2, 10)$ (dashed line) and prior 2: $p \sim Beta(10, 2)$ (dotted line).

Note that in typical Bayesian analyses, often a variety of different prior distributions will be used and the corresponding posterior distributions compared in order to see to what extent the priors influence the posterior distribution. This is often called a prior sensitivity analysis.

Within our results, we shall usually examine the posterior mean of the distribution, taking this point estimate to be a “reasonable” point estimate. Whatever form of point estimate we use, however, it provides a very poor description of the complete posterior distribution. A more adequate summary of the latter, and yet still readily understandable, is the idea of an interval estimate.

1.4.2 Interval estimates

Interval estimates give an estimate of the spread of the posterior distribution. These are analogous to confidence intervals within the classical case, and are called **credible intervals**. However, their interpretation is very different. Recall that a classical $100(1-\alpha)\%$ confidence interval is defined such that, if the data collection process is repeated again and again, then in the long run, $100(1-\alpha)\%$ of the confidence intervals formed would contain the (fixed) unknown parameter value. Conversely, the interpretation of the Bayesian $100(1-\alpha)\%$ credible interval is that this interval contains $100(1-\alpha)\%$ of the posterior distribution of the parameter. More formally, suppose that we are interested in the parameter θ, which has posterior distribution $\pi(\theta|\boldsymbol{x})$. Then, we make the following definition.

Definition 1.2: *The interval* (a, b) *is defined as an* $100(1-\alpha)\%$ *posterior credible interval if,*

$$\int_a^b \pi(\theta|\boldsymbol{x})d\theta = 1-\alpha, \qquad 0 \leq \alpha \leq 1.$$

Typical values of α are 0.1, 0.05, 0.01, and we speak of 90%, 95% and 99% credible intervals. The idea is illustrated in Figure 1.4.

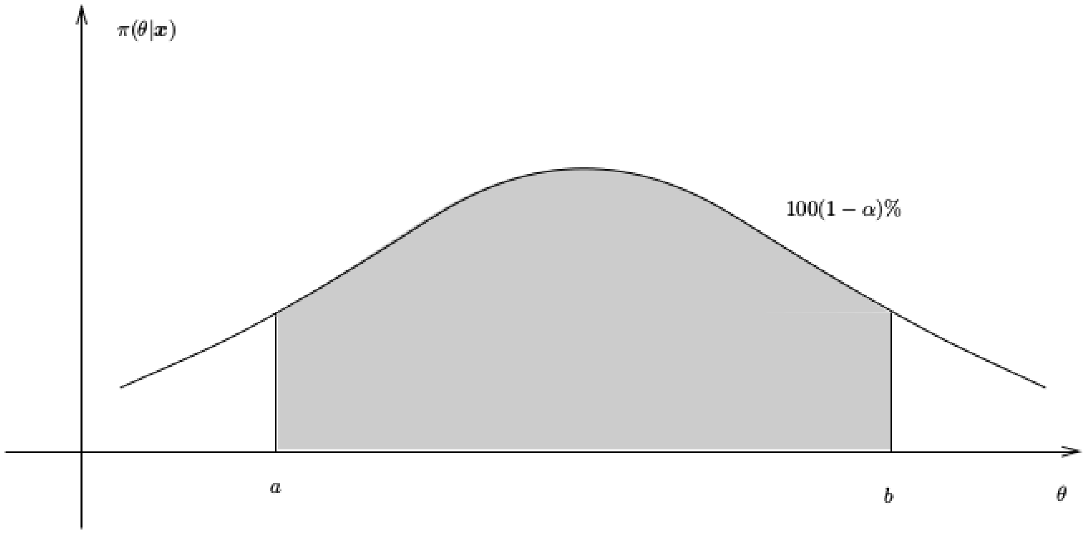


Figure. 1.4: Typical $100(1-\alpha)\%$ credible interval.

Note that a $100(1-\alpha)\%$ credible interval is not unique, since, in general, there will be many choices of a and b, such that, $\int_a^b \pi(\theta|\boldsymbol{x})d\theta = 1-\alpha$. For example, see Figure 1.5, where we have two $100(1-\alpha)\%$ credible intervals, $[a_1, b_1]$ and $[a_2, b_2]$.

Often, a symmetric $100(1-\alpha)\%$ credible interval (a, b) is used. This credible interval is unique, and is defined such that,

$$\int_{-\infty}^a \pi(\theta|\boldsymbol{x})d\theta = \frac{\alpha}{2} = \int_b^\infty \pi(\theta|\boldsymbol{x})d\theta.$$

I.e. the credible interval such that a corresponds to the lower $\frac{\alpha}{2}$ quantile, and b the upper $1-\frac{\alpha}{2}$ quantile of the posterior distribution $\pi(\theta|\boldsymbol{x})$.

Consider Figure 1.5, again. Both of the intervals contain $100(1-\alpha)\%$ of the distribution, so that, in betting terms, there is no objection to either of them. However, what about communicating information about θ? The interval $[a_1, b_1]$ is clearly more informative, since, for a given α, a shorter interval represents a "tighter" inference. This motivates the following refinement of a credible interval.

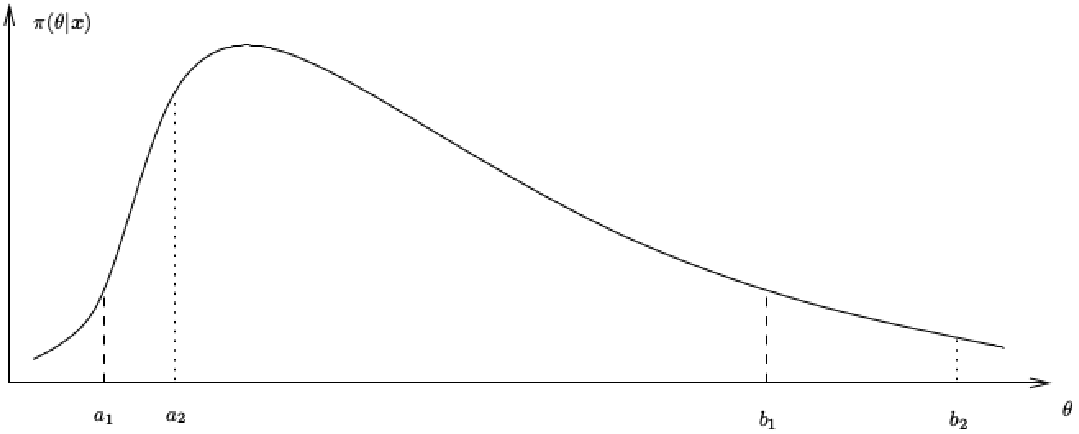


Figure. 1.5: Two alternative $100(1-\alpha)\%$ credible intervals.

Definition 1.3: *The interval* $[a, b]$ *is a* $100(1-\alpha)\%$ ***highest posterior density interval*** *(HPDI) if:*

1. $[a, b]$ *is a* $100(1-\alpha)\%$ *credible interval; and*

2. *for all* $\theta' \in [a, b]$ *and* $\theta'' \notin [a, b]$, $\pi(\theta'|\boldsymbol{x}) \geq \pi(\theta''|\boldsymbol{x})$.

In an obvious sense, this is the required definition for the "shortest possible" interval having a given credible level $1-\alpha$, and essentially centres the interval around the mode, in the uni-model case. Clearly, if the distribution is symmetrical about the mean, such as the Normal distribution, the $100(1-\alpha)\%$ symmetric credible interval is identical to the $100(1-\alpha)\%$ HPDI. In the case where the posterior distribution is multi-modal, the corresponding HPDI may consist of several disjoint intervals.

Note that suppose that we have a symmetric credible interval $[a, b]$ for a given parameter θ. Then, if we consider a bijective (monotonic) transformation of the parameters, such as $g(\theta)$, the corresponding symmetric credible interval is given by $[g(a), g(b)]$, However, this is not always true for HPDI's. The corresponding HPDI on $g(\theta)$ is $[g(a), g(b)]$, if and only if, g is a linear transformation; else, the HPDI needs to be recalculated for $g(\theta)$.

With the relatively recent increase in computational power, the use of simulation has become increasingly popular in obtaining summary estimates of posterior distributions. If we are able to obtain a sample from the posterior distribution (using for example `R`), then we are able to use this sample to estimate corresponding summary statistics. For example, the mean of the posterior distribution can be estimated via the mean of the corresponding sample from the posterior distribution (this is known as a Monte Carlo estimate). Any number of posterior summary statistics may be of interest. For example, suppose that we are interested in the posterior probability that the parameter of interest, θ has a value greater than 10. Then, we can estimate $\mathbb{P}_\pi(\theta > 10)$ by simply calculating the proportion of the sample from the posterior distribution for which the parameter value is greater than 10. These (and other) ideas will be discussed in greater detail in the latter half of the course. Here we are simply providing a sample of the types of summary estimates that we may be interested in, and giving a taster of how they may be calculated.

Usually, if describing a posterior distribution via summary statistics, a variety of point and interval estimates are given to encapsulate the main properties of the distribution.

1.5 Hypothesis testing

The underlying principle within hypothesis testing is that we wish to test some hypothesis H_0, say, against an alternative hypothesis H_1. We refer to H_0 as the *null hypothesis* and H_1 as the *alternative hypothesis.* The hypotheses are statements concerning the parameter(s) of interest. Typically, suppose that we are interested in the parameter $\theta \in \Theta$. Then, our hypotheses are of the form,

$$\begin{aligned} H_0 &: \theta \in \Theta_0; \qquad \text{and} \\ H_1 &: \theta \in \Theta_1, \end{aligned}$$

where Θ_1 and Θ_2 are disjoint and exhaustive subsets of the parameter space Θ.

Then, within the Bayesian context, we need to place a prior on the parameter space. We let,

$$p_i = \mathbb{P}(\theta \in \Theta_i),$$

for $i = 0, 1$, such that $p_0 + p_1 = 1$. Then, the prior odds for the null hypothesis, H_0, against the alternative hypothesis, H_1 are p_0/p_1. Therefore, the prior odds specifies your prior beliefs concerning which of the two hypotheses are more likely, before observing any data. If $p_0 > p_1$ we would favour the null hypothesis; if $p_0 < p_1$, then we would favour the alternative hypothesis; and if $p_0 \approx p_1$, we regard both hypotheses as roughly equally likely within the prior specification.

Once we have specified the prior, we observe data $\boldsymbol{x}$, and wish to update our prior beliefs concerning the hypotheses. We do this by calculating the corresponding *posterior odds*, given by,

$$\frac{\mathbb{P}(\theta \in \Theta_0|\boldsymbol{x})}{\mathbb{P}(\theta \in \Theta_1|\boldsymbol{x})}.$$

If the posterior odds are greater than one, we would favour the null hypothesis; if the posterior odds are less than one, we favour the alternative hypothesis; if they are equal each hypothesis is equally likely *a posteriori.* Note that the posterior odds give a quantitative comparison between hypotheses H_0 and H_1.

A further statistic that is often used is the *Bayes factor.* This is simply defined to be the ratio of posterior odds to prior odds. For example, the Bayes factor for H_0 against H_1 is denoted by B_{01} and given by,

$$B_{01} = \frac{\mathbb{P}(\theta \in \Theta_0|\boldsymbol{x})/\mathbb{P}(\theta \in \Theta_1|\boldsymbol{x})}{p_0/p_1}.$$

Kass and Raftery (1995) suggested the following 'rule of thumb' for Bayes factors.

Bayes Factor	Interpretation
< 3	No evidence of H_0 over H_1;
> 3	Positive evidence for H_0;
> 20	Strong evidence for H_0;
> 150	Very strong evidence for H_0.

This general guideline is commonly used in the interpretation of Bayes Factors. Note that given the prior probabilities of the hypotheses, the Bayes factor and posterior model probabilities are interchangable (they are simply functions of each other).

1.5.1 Simple hypotheses

Suppose that we wish to test,

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1.$$

Then, using Bayes Theorem, the posterior probability that hypothesis H_i is true is given by,

$$\mathbb{P}(\theta = \theta_i|\boldsymbol{x}) = \pi(\theta_i|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta_i)p_i,$$

for $i = 0, 1$. The corresponding constant of proportionality is given by,

$$\left(\sum_{j=0}^{1} f(\boldsymbol{x}|\theta_j)p_j\right)^{-1}.$$

Then, in the case of simple hypotheses, the Bayes factor is simply equal to the likelihood ratio, i.e.,

$$B_{01} = \frac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\theta_1)}$$

and hence is based solely on the data.

1.5.2 Composite hypotheses

Suppose that θ is continuous, and that we wish to test,

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

Then, if we let the corresponding prior for θ be denoted by $p(\theta)$, the corresponding posterior probability for hypothesis $i = 0, 1$, is given by,

$$\begin{aligned}
\mathbb{P}(\theta \in \Theta_i|\boldsymbol{x}) &= \int_{\theta \in \Theta_i} \pi(\theta|\boldsymbol{x}) d\theta \\
&= \frac{1}{f(\boldsymbol{x})} \int_{\theta \in \Theta_i} f(\boldsymbol{x}|\theta)p(\theta) d\theta \\
&\qquad \text{using Bayes' Theorem} \\
&= \frac{1}{f(\boldsymbol{x})} p_i \int_{\theta \in \Theta_i} f(\boldsymbol{x}|\theta)p_i(\theta) d\theta,
\end{aligned}$$

where p_i denotes the prior probability of the hypothesis H_i, such that $\theta \in \Theta_i$, and,

$$p_i(\theta) = \frac{p(\theta)}{p_i},$$

can be regarded as the prior density restricted to Θ_i, renormalised to give a probability density over Θ_i. Then, the corresponding posterior odds is given by,

$$\frac{\mathbb{P}(\theta \in \Theta_0|\boldsymbol{x})}{\mathbb{P}(\theta \in \Theta_1|\boldsymbol{x})} = \frac{p_0 \int_{\theta \in \Theta_0} f(\boldsymbol{x}|\theta)p_0(\theta) d\theta}{p_1 \int_{\theta \in \Theta_1} f(\boldsymbol{x}|\theta)p_1(\theta) d\theta}.$$

The corresponding Bayes' factor is given by,

$$B_{01} = \frac{\int_{\theta \in \Theta_0} f(\boldsymbol{x}|\theta)p_0(\theta) d\theta}{\int_{\theta \in \Theta_1} f(\boldsymbol{x}|\theta)p_1(\theta) d\theta},$$

which is the ratio of "weighted" likelihoods of the densities $p_i(\theta)$. Thus, the Bayes factor is not solely dependent on the data, but also on the corresponding prior placed on the parameter.

Example

Suppose that $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, where each $X_i \overset{iid}{\sim} N(\mu, 1)$. We observe data $\boldsymbol{x}$:

$$3.4,\ 2.9,\ 3.0,\ 3.5,\ 3.3,\ 3.7,\ 2.7,\ 3.9,\ 2.7,\ 2.9,$$

and wish to test the simple hypothesis:

$$H_0 : \mu = 3, \quad \text{vs} \quad H_1 : \mu = 3.5.$$

What is the corresponding Bayes factor of H_0 against H_1? We have that,

$$\mathbb{P}(\mu = 3|\boldsymbol{x}) \quad = \quad \frac{p_0 f(\boldsymbol{x}|\mu = 3)}{f(\boldsymbol{x})}.$$

Now,

$$\begin{aligned} f(\boldsymbol{x}|\mu = 3) &= \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - 3)^2}{2}\right) \\ &= \frac{1}{(2\pi)^5} \exp\left(-\frac{2}{2}\right) \\ &= \frac{\exp(-1)}{(2\pi)^5}. \end{aligned}$$

Similarly, we have that,

$$f(\boldsymbol{x}|\mu = 3.5) = \frac{\exp(-1.25)}{(2\pi)^5}$$

Thus, we have that,

$$\begin{aligned} B_{01} &= \frac{f(\boldsymbol{x}|\mu = 3)}{f(\boldsymbol{x}|\mu = 3.5)} \\ &= 1.28. \end{aligned}$$

Thus there is only slight evidence to support model H_0 over H_1.

1.6 Prediction

1.6.1 Prior predictive distributon

Suppose that we wish to make inference about a random vector $\boldsymbol{X}$, with pdf $f(\boldsymbol{x}|\theta)$, with unknown parameter $\theta \in \Theta$. If our best estimate of θ is represented by the prior, $p(\theta)$, then the corresponding predictive pdf of $\boldsymbol{X}$ is given by,

$$f(\boldsymbol{x}) = \int_{\theta \in \Theta} f(\boldsymbol{x}|\theta) p(\theta) d\theta.$$

In other words, we take the expected pdf, where the expectation is with respect to the prior density $p(\theta)$. The term $f(\boldsymbol{x})$ is called the *prior predictive distribution.* (Recall that this is also the denominator in the expression for Bayes' Theorem). Essentially, we are weighting the pdf with our best estimate for θ, and since we have not observed any data, that is equal to our prior distribution for θ.

Example

Suppose that X is a random variable, such that,

$$X \sim Exp(\lambda),$$

and that our prior beliefs on λ are described by $\Gamma(\alpha, \beta)$ distribution. Then, the prior predictive distribution is given by,

$$\begin{aligned}
f(x) &= \int_{-\infty}^{\infty} f(\boldsymbol{x}|\lambda)p(\lambda)d\lambda \\
&= \int_{0}^{\infty} \lambda \exp(-\lambda x) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(\beta+x)^{\alpha+1}}\int_0^\infty \frac{(\beta+x)^{\alpha+1}}{\Gamma(\alpha+1)}\lambda^\alpha \exp(-\lambda(x+\beta))d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(\beta+x)^{\alpha+1}},
\end{aligned}$$

since the integral is over a $\Gamma(\alpha+1, \beta+x)$ pdf, and integrates to one.

1.6.2 Posterior predictive distributon

Suppose that we now observe data $\boldsymbol{x}$, and wish to predict future observations $\boldsymbol{y}$, from the same process, assuming that conditional on the parameter θ, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Then, the *posterior predictive distribution* for $\boldsymbol{Y}$ is given by,

$$\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{x}) &= \int_\Theta f(\boldsymbol{y}|\boldsymbol{x},\theta)\pi(\theta|\boldsymbol{x})d\theta \\
&= \int_\Theta f(\boldsymbol{y}|\theta)\pi(\theta|\boldsymbol{x})d\theta,
\end{aligned}$$

since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent, given θ. Thus, we are now weighting the corresponding pdf for $\boldsymbol{Y}$ with our current (posterior) beliefs for θ having already observed data $\boldsymbol{x}$.

Example

Suppose that the number of calls to a telephone switchboard in z minutes has a $Poisson(\lambda z/10)$ distribution, where $\lambda > 0$ is unknown. Being an enthusiastic Bayesian research graduate, the operator forms the following prior on λ, (from working in a similar telephone switchboard),

$$\lambda \sim Exp(10).$$

Being reasonably bored, they decide to calculate their prior predictive distribution for the number of calls that they receive in the first 10 minutes of work, denoted by X.

Then, the prior predictive distribution is given by,

$$\begin{aligned}
f(x) &= \int_{-\infty}^{\infty} f(x|\lambda)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^x}{x!}\exp(-\lambda) \times 10\exp(-10\lambda)d\lambda \\
&= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}}\int_0^\infty \frac{11^{x+1}}{\Gamma(x+1)}\lambda^x \exp(-11\lambda)d\lambda \\
&= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}} \\
&= \frac{10}{11^{x+1}}, \qquad \text{since } \Gamma(x+1) = x! \text{ as } x \text{ is a positive integer.}
\end{aligned}$$

The operator then takes x calls within the first 10 minutes, and wants a 20 minute break. They decide to calculate the posterior predictive distribution for the number of calls that they will receive in this time, from which they can calculate the probability of no calls being missed in this time, i.e. receiving no calls in these twenty minutes. First they calculate the corresponding posterior distribution for λ, given that they have received a total of x calls in the first 10 minutes,

$$\begin{aligned}\pi(\lambda|x) &\propto f(x|\lambda)p(\lambda)\\ &= \frac{\lambda^x}{x!}\exp(-\lambda)\times 10\exp(-10\lambda)\\ &\propto \lambda^x\exp(-11\lambda),\end{aligned}$$

so that,

$$\lambda|x \sim \Gamma(x+1, 11).$$

Let Y denote the number of calls they receive in 20 minutes. Then,

$$Y|\lambda \sim Poisson(2\lambda).$$

The corresponding predictive posterior distribution is given by,

$$\begin{aligned}f(y|x) &= \int_{-\infty}^{\infty} f(y|\lambda)\pi(\lambda|x)d\lambda\\ &= \int_0^{\infty}\frac{(2\lambda)^y\exp(-2\lambda)}{y!}\times\frac{11^{x+1}}{\Gamma(x+1)}\lambda^x\exp(-11\lambda)d\lambda\\ &= \frac{11^{x+1}}{13^{x+y+1}}\frac{2^y}{y!}\frac{\Gamma(x+y+1)}{\Gamma(x+1)}\int_0^{\infty}\frac{13^{x+y+1}}{\Gamma(x+y+1)}\lambda^{x+y}\exp(-13\lambda)d\lambda\\ &= \frac{11^{x+1}}{13^{x+y+1}}\frac{2^y}{y!}\frac{\Gamma(x+y+1)}{\Gamma(x+1)}.\end{aligned}$$

So that the predictive posterior probability of no calls within the next 20 minutes is,

$$f(0|x) = \left(\frac{11}{13}\right)^{x+1}.$$

Then, for example, if they have not had more than 3 calls in the last 10 minutes, their posterior predictive probability of not missing a call in their break exceeds 1/2.

1.7 Bayes' Theorem (multivariate)

Bayes' Theorem is easily generalised to the multi-parameter case. Suppose that we have a set of parameters $\boldsymbol{\theta}$ on which we wish to make inference, and that we observe data $\boldsymbol{x}$. Then the (joint) posterior distribution for $\boldsymbol{\theta}$ is given by,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\boldsymbol{x})},$$

where,

$$f(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Since this integration becomes increasingly more complex in higher dimensions, Bayes' Theorem is most often quoted in the form,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The prior is then specified jointly over all parameters. However, often in practice, the parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ are assumed to be independent of each other, *a priori*, so that,

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\theta_i).$$

In multi-dimensions the posterior distribution significantly increases in complexity. However, often we may only be interested in the marginal distribution of a single parameter conditional on the data. For example, suppose that we are only interested in θ_1, then,

$$\pi(\theta_1|\boldsymbol{x}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{x}) d\theta_2, \ldots, d\theta_n.$$

This integration is often too complex to do in practice, however, the latter part of the course will show how we may be able to obtain summary statistics of a marginal posterior distribution, such as this, using an alternative method called Markov chain Monte Carlo (MCMC).

Note that the ideas presented for the single parameter case are all directly generalised to the multi-parameter case, for example, the concepts of sufficiency and conjugacy. Posterior credible intervals, however, now generalise to posterior credible regions with dimension equal to the number of parameters. For example, in the two parameter case, where $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, the $100(1-\alpha)\%$ posterior credible interval is now a two-dimensional region, R, such that,

$$\mathbb{P}((\theta_1, \theta_2) \in R|\boldsymbol{x}) = 1 - \alpha.$$

Clearly, a computer is needed in order to plot these more complex regions. Often, in practice, the marginal posterior density intervals may be calculated for a single parameter, rather than the more complex higher-dimensional density regions for the full set of parameters.

1.8 Examples: Bayesian Inference For Normal Distributions

Within this section, we shall initially assume that we observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, such that $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$. We consider the two different cases, relating to whether μ or σ^2 are unknown, and obtain inference on the parameters which are unknown. We shall the extend the case to the multivariate case such that the observed data are now observations from a Multivariate normal distribution, where again the mean is unknown.

1.8.1 Single parameter problems

Case i) Unknown μ, known σ^2

We need to first place a prior on the unknown mean, μ. Suppose that we specify the prior,

$$\mu \sim N(\phi, \tau^2).$$

Then, the corresponding posterior distribution can be obtained by applying Bayes' Theorem:

$$\begin{aligned}
\pi(\mu|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu)p(\mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right) \\
&\propto \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right) \\
&\propto \exp\left(-\frac{(n\mu^2-2n\bar{x}\mu)}{2\sigma^2}\right) \exp\left(-\frac{(\mu^2-2\mu\phi)}{2\tau^2}\right) \\
&= \exp\left(-\frac{\mu^2(\tau^2 n+\sigma^2)-2\mu(\tau^2 n\bar{x}+\sigma^2\phi)}{2\sigma^2\tau^2}\right).
\end{aligned}$$

Then, by completing the square, we can identify the posterior distribution for μ to be,

$$\mu|\boldsymbol{x} \sim N\left(\frac{\tau^2 n\bar{x}+\sigma^2\phi}{\tau^2 n+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n+\sigma^2}\right).$$

Thus, the normal prior on μ is a conjugate prior. We can also note that $\bar{x}$ is sufficient in the normal case. It is also clear from the form of the posterior distribution, that the posterior mean is a mixture of the prior mean (ϕ) and the classical MLE for the mean ($\bar{x}$), as we can write,

$$\frac{\tau^2 n\bar{x}+\sigma^2\phi}{\tau^2 n+\sigma^2} = w\bar{x}+(1-w)\phi,$$

where,

$$w = \frac{\tau^2 n}{\tau^2 n+\sigma^2}.$$

The value of the prior variance, τ^2, specifies the informativeness of the prior. A small variance, implies a "tight" prior distribution around ϕ for the parameter μ; whereas a large prior variance suggests that there is little information contained in the prior concerning the parameter value, with a relatively flat prior distribution. This can be clearly seen, if we consider the posterior distribution for μ.

Initially, consider the case for τ^2 small. Then, in the limiting case, as $\tau^2 \to 0$, we have that the mean of the distribution tends to ϕ (i.e. the prior mean for μ), with corresponding variance,

$$\frac{\sigma^2\tau^2}{\tau^2 n+\sigma^2} = \frac{\tau^2}{n\tau^2/\sigma^2+1}.$$

Then, by the Binomial Theorem, we have,

$$(n\tau^2/\sigma^2+1)^{-1} = 1-\frac{n\tau^2}{\sigma^2}+O(\tau^4).$$

(Recall the Binomial Theorem that for $|x|<1$, then $(1+x)^{-1} = \sum_{k=0}^{\infty}(-1)^k x^k$)).

Substituting back into our expression for variance, we obtain,

$$\begin{aligned}
\frac{\sigma^2\tau^2}{\tau^2 n+\sigma^2} &= \tau^2(1-\frac{n\tau^2}{\sigma^2}+O(\tau^4)) \\
&= \tau^2+O(\tau^4).
\end{aligned}$$

So that as $\tau^2 \to 0$, clearly, the posterior variance tends to τ^2. Thus, the prior dominates the posterior distribution.

Conversely, consider τ^2 large, so that we have a vague prior on the parameter. Then, again in the limiting case $\tau^2 \to \infty$, the posterior mean for μ tends to $\bar{x}$. Additionally, the variance tends to σ^2/n. This can be compared to the result in classical statistics, where the sampling distribution of

$\bar{X}$ is normal with mean μ and variance σ^2/n. However, these statements must not be confused, and although we may have the same answer, the question is different between the classical and Bayesian approaches. For a Bayesian, we apply the probability statement to the parameter μ; for a classicist, the probability statement is applied to the statistic $\bar{X}$.

Example

Data, $\boldsymbol{x}$, are collected on the length of time that it takes students to answer a particular question within an examination. The data collected (in minutes) are:

$$36,\ 67,\ 44,\ 39,\ 56,\ 65,\ 43,\ 49.$$

It is assumed that each random variables $X_1, \ldots, X_{10} \overset{iid}{\sim} N(\mu, \sigma^2)$, where the mean is unknown, and is to be estimated. The variance σ^2 can be taken to represent the variability in the ability of students: small σ^2 would represent a fairly homogeneous set of students, whereas a large σ^2 would represent a heterogeneous mix of students of varying abilities. Here, from previous examinations, we know that $\sigma^2 = 100$.

The corresponding prior for the mean is specified in the form,

$$\mu \sim N(\phi, \tau^2).$$

The mean, $\phi = 45$, is the length of time the examiner expects a student to take answering the question. The variance, τ^2 is to be specified by the expert (i.e. the examiner).

The corresponding distribution for μ is given by,

$$\mu|\boldsymbol{x} \sim N\left(\frac{399\tau^2 + 4500}{8\tau^2 + 100}, \frac{100\tau^2}{8\tau^2 + 100}\right).$$

Figure 1.6 gives the corresponding posterior distribution for μ, under two different priors: (a) $\tau^2 = 100$ and (b) $\tau^2 = 1$.

The corresponding posterior mean for μ can be expressed as a weighted average of the prior mean and classical MLE, as shown above. So that in this case the mean can be expressed in the form,

$$\frac{8\tau^2}{8\tau^2 + 100}\bar{x} + \frac{100}{8\tau^2 + 100}\phi.$$

So that, when the prior variance on the parameter μ is equal to 100, we obtain mixture weights of $\frac{8}{9}$ and $\frac{1}{9}$, on the MLE and prior, respectively, and thus a posterior mean for μ of 49.33. Alternatively, when we consider the variance equal to 1, the corresponding weights are 0.074 and 0.926, giving a posterior mean of 45.36.

Case ii) Known μ, unknown σ^2

As usual, we need to place a prior on the unknown parameter σ^2. We specify,

$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

In other words $1/\sigma^2 \sim \Gamma(\alpha, \beta)$. Then, the corresponding posterior distribution is given by,

$$\begin{aligned}
\pi(\sigma^2|\boldsymbol{x}) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \times (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\left(\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta\right)}{\sigma^2}\right) \\
\Rightarrow \sigma^2|\boldsymbol{x} &\sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta\right).
\end{aligned}$$

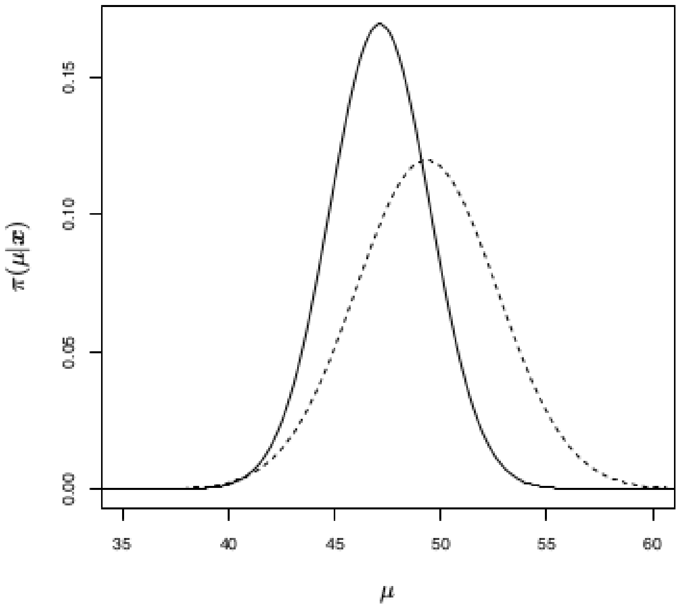


Figure. 1.6: The posterior distribution of μ, for (a) $\tau^2 = 100$ (dashed line) and (b) $\tau^2 = 1$ (solid line).

Consider again the posterior pdf for σ^2:

$$\pi(\sigma^2|\boldsymbol{x}) \propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\left(\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2+\beta\right)}{\sigma^2}\right).$$

Writing $z^2 = \sum_{i=1}^n (x_i-\mu)^2$, we note that z^2 is simply a constant, so that,

$$\pi(\sigma^2|\boldsymbol{x}) \quad \propto \quad (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{z^2+2\beta}{2\sigma^2}\right)$$

Then, we can see that setting $a = \frac{n+2\alpha}{2}$ and $b = \frac{z^2+2\beta}{2}$ we have that,

$$\sigma^2|\boldsymbol{x} \sim \Gamma^{-1}(a,b) = \Gamma^{-1}\left(\frac{n+2\alpha}{2}, \frac{z^2+2\beta}{2}\right).$$

Alternatively, we can rewrite the posterior distribution as,

$$\pi(\sigma^2|\boldsymbol{x}) \quad \propto \quad \left(\frac{\sigma^2}{z^2+2\beta}\right)^{-((n+2\alpha)/2+1)} \exp\left(-\frac{1}{2\sigma^2/(z^2+2\beta)}\right).$$

Then, comparing this expression to the pdf of a χ_ν^{-2} distribution, we see that,

$$\left.\frac{\sigma^2}{z^2+2\beta}\right| \boldsymbol{x} \sim \chi^{-2}_{n+2\alpha}.$$

alternatively, we may write,

$$\sigma^2|\boldsymbol{x} \sim (z^2+2\beta)\chi^{-2}_{n+2\alpha} \tag{1.3}$$

Note that, in general, if,

$$\theta \sim \Gamma^{-1}\left(\frac{c}{2}, \frac{d}{2}\right),$$

then,

$$\theta \sim d\chi_c^{-2}.$$

Example

Suppose that we observe data $\boldsymbol{x}$, given by,

$$2.1,\ 4.2,\ 6,\ 4.5,\ 3.5,\ 2.1,\ 4.4,\ 3.2,\ 3.7,\ 3.9$$

where $\mu = 4$. We place a $\Gamma^{-1}(0.5, 0.5)$ prior on σ^2.

Then, what is the corresponding distribution for σ^2? We have that,

$$z^2 = \sum_{i=1}^{n}(x_i - \mu)^2 = 17.46.$$

So that,

$$\sigma^2|\boldsymbol{x} \sim \Gamma^{-1}(5.5, 9.23).$$

Thus, the posterior mean for σ^2 is 2.051, with standard deviation 1.202.

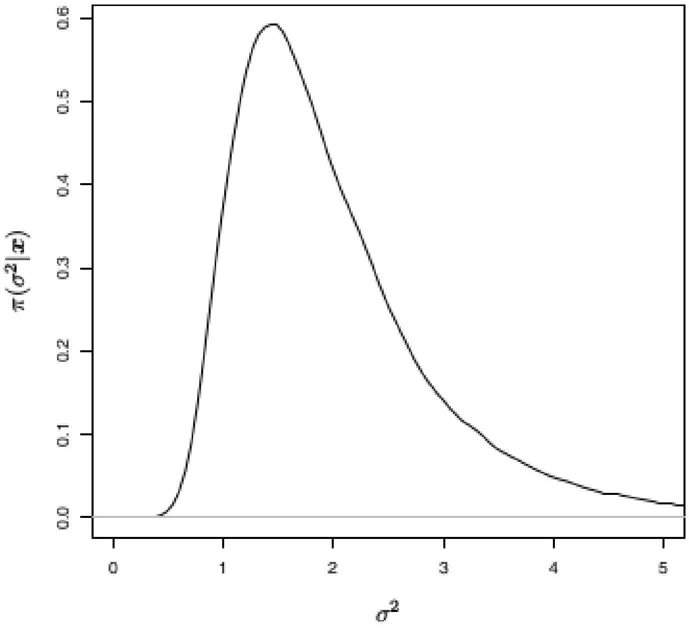


Figure. 1.7: Posterior distribution for σ^2.

Alternative parameterisation

Suppose that the parameter of interest is the precision $\tau = 1/\sigma^2$ and that we place a $\Gamma(\alpha, \beta)$ prior on τ. This is the same as placing a $\Gamma^{-1}(\alpha, \beta)$ distribution on σ^2, i.e. the same prior as in the above example. This can be easily seen via a transformation of variables. Suppose that $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, and $\tau = 1/\sigma^2$. Then, using a transformation of variables, the corresponding distribution on τ, is given by,

$$\begin{aligned} p_\tau(\tau) &= p_{\sigma^2}(\sigma^2(\tau))\left|\frac{d\sigma^2}{d\tau}\right| \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\tau}\right)^{-(\alpha+1)} \exp\left(-\beta\tau\right)\frac{1}{\tau^2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}\exp\left(-\beta\tau\right), \end{aligned}$$

so that, $\tau \sim \Gamma(\alpha, \beta)$.

Then, since the posterior distribution for σ^2 is also Γ^{-1}, the corresponding posterior for τ is Γ, and in particular,

$$\tau|\boldsymbol{x} \sim \Gamma\left(\frac{n}{2}+\alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\beta\right)$$

This can be shown in the usual way, by combining the likelihood with the corresponding Gamma prior on τ, and is left as an exercise.

Following the analogous argument above, it can be shown that,

$$\tau|\boldsymbol{x} \sim \frac{1}{(z^2+2\beta)}\chi^2_{n+2\alpha}.$$

This is again left as an exercise.

Then, for the data given above, we have that,

$$\tau^2|\boldsymbol{x} \sim \Gamma(5.5, 9.23),$$

so that the posterior mean of τ^2 is 0.596, with standard deviation 0.254.

1.8.2 Multi-parameter problem

Case iii) Multivariate Normal

Suppose that we observe random variables $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ from a Multivariate Normal distribution with known (symmetric) covariance matrix Σ and unknown mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, which we wish to estimate. Then, we write $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. (Note that if $\Sigma = \sigma^2 I$, then the random variables, $X_1, \ldots, X_p$ are independent Normal random variables, with mean $\mu_1, \ldots, \mu_p$, respectively). The pdf for $\boldsymbol{X}$, given parameters $\boldsymbol{\mu}$, is defined to be,

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

We specify the prior,

$$\boldsymbol{\mu} \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma_p).$$

Then, the corresponding posterior distribution for $\boldsymbol{\mu}$ is given by,

$$\begin{aligned}
\pi(\boldsymbol{\mu}|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \\
&= \frac{1}{\sqrt{2\pi}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \times \frac{1}{\sqrt{2\pi}|\Sigma_p|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta})^T\Sigma_p^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta})\right) \\
&\propto \exp\left(-\frac{1}{2}[\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x} - \boldsymbol{x}^T\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\Sigma_p^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma_p^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}^T\Sigma_p^{-1}\boldsymbol{\mu}]\right) \\
&= \exp\left(-\frac{1}{2}[\boldsymbol{\mu}^T(\Sigma^{-1}+\Sigma_p^{-1})\boldsymbol{\mu} - \boldsymbol{\mu}^T(\Sigma^{-1}\boldsymbol{x}+\Sigma_p^{-1}\boldsymbol{\theta}) - (\boldsymbol{x}^T\Sigma^{-1}+\boldsymbol{\theta}^T\Sigma_p^{-1})\boldsymbol{\mu}]\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_\pi)^T\Sigma_\pi^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_\pi)\right),
\end{aligned}$$

where,

$$\boldsymbol{\mu}_\pi = \Sigma_\pi(\Sigma^{-1}\boldsymbol{x}+\Sigma_p^{-1}\boldsymbol{\theta}); \quad \text{and} \quad \Sigma_\pi^{-1} = \Sigma^{-1}+\Sigma_p^{-1}.$$

Exercise: check.

Thus,

$$\boldsymbol{\mu}|\boldsymbol{x} \sim N(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi).$$

The Multivariate Normal prior on the mean vector $\boldsymbol{\mu}$ is a conjugate prior.

1.9 Motivating Examples

Throughout this section we have generally considered conjugate priors and particular vague priors. These have all resulted in standard posterior distributions for the parameters of interest. However, what happens if we wish to use a different prior for a given parameter(s), resulting in a more complex posterior distribution, or where the likelihood is complex, and there is no prior of standard form which results in a standard posterior distribution for the parameters. Consider a simple example. Suppose that we have a random variables, $\boldsymbol{X} = \{X_1, \dots, X_n\}$ which are independent and Normally distributed with mean μ and variance σ^2 (assumed to be known). Then, we may wish to specify a prior on μ of the form,

$$\mu \sim \log N(\phi, \tau^2).$$

(This means that $\log \mu \sim N(\phi, \tau^2)$).

The corresponding posterior distribution for μ is given by,

$$\pi(\mu|\boldsymbol{x}) \propto \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right)\frac{1}{\mu}\exp\left(-\frac{(\log\mu-\phi)^2}{2\tau^2}\right).$$

This posterior distribution is clearly non-standard, so that alternative (numerical) approaches need to be implemented in order to obtain any inference on the parameter μ. For example, in this case the distribution can be plotted within Maple. However, in general, as the number of dimensions increases the visual representation of the posterior distribution becomes increasingly difficult. Calculating a posterior marginal distribution of a particular parameter is also often difficult due to the complex integration needed (which may be analytically intractable), resulting in the need for approximations. As the number of dimensions increases, alternative approaches need to be considered.

Consider a relatively simple two-dimensional case, where we have data $x_1, \dots, x_n$ such that each X_i are independent and Normally distributed with mean μ and (unknown) variance σ^2. We specify independent priors:,

$$\mu \sim N(\phi, \tau^2); \qquad \text{and} \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

The corresponding joint posterior distribution is of the form,

$$\begin{aligned}\pi(\mu,\sigma^2|\boldsymbol{x}) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\times\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right)\times\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(\sigma^2)^{-(\alpha+1)}\exp\left(-\frac{\beta}{\sigma^2}\right)\\ &\propto (\sigma^2)^{-(n/2+\alpha+1)}\exp\left(-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{\beta}{\sigma^2}\right)\exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right),\end{aligned}$$

after algebra. We note that the posterior conditional distributions for μ and σ^2 are of standard form, namely,

$$\begin{aligned}\mu|\boldsymbol{x},\sigma^2 &\sim N\left(\frac{\tau^2 n\bar{x}+\sigma^2\phi}{\tau^2 n+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n+\sigma^2}\right);\\ \sigma^2|\boldsymbol{x},\mu &\sim \Gamma^{-1}\left(\frac{n}{2}+\alpha, \frac{1}{2}\sum_{i=1}^n(x_i-\mu)^2+\beta\right),\end{aligned}$$

(see Section 1.8.1 for the algebra). Thus, although the joint posterior distribution is of a complex form, the (uni-dimensional) posterior conditional distributions of the individual parameters are of standard form. So, what would happen if we repeatedly simulate values for μ and σ^2 from the posterior conditional distributions of the parameters? We will answer this in the following section of the course where we utilise this kind of property in obtaining inference on more complex posterior distributions.

Chapter 2

BAYESIAN COMPUTATION

In the last chapter we described the ideas behind Bayesian statistics and saw a number of (simple) examples. However, real statistical problems are typically significantly more complex, most notably in terms of the number of unknown parameters to be estimated, resulting in high dimensional posterior distributions of complex and often of non-standard form. This was essentially the reason why the classical approach dominated statistics last century. However, the increase in computational power, coupled with computational algorithms introduced to the statistical literature around 1990 has made Bayesian analyses feasible (and often easier than alternative classical approaches!). In this latter part of the course we describe these computational algorithms. (In the further course MATH11175 Bayesian Data Analysis you will use these techniques via (freely-available) computer software for conducting Bayesian analyses.)

2.1 Monte Carlo Integration

In many circumstances, we are faced with the problem of evaluating an integral which is too complex to calculate explicitly. With particular reference to Bayesian inference, posterior distributions are typically summarised using statistics such as the mean and/or variance. Such posterior summary statistics require integration of the posterior density (which is often analytically intractable, particularly for high-dimensional posterior distributions). For example, we may wish to estimate the posterior (marginal) expectation of a parameter θ, given observed data $\boldsymbol{x}$:

$$\mathbb{E}_\pi(\theta) = \int \theta \pi(\theta|\boldsymbol{x}) d\theta.$$

We can use the simulation technique of *Monte Carlo integration* to obtain an estimate of a given integral (and hence posterior expected value). The method is based upon drawing observations from the distribution of the variable of interest and simply calculating the empirical estimate of the expectation. For example, given a sample of observations, $\theta^1, \dots, \theta^n \sim \pi(\theta|\boldsymbol{x})$, we can estimate the expectation by,

$$\frac{1}{n}\sum_{i=1}^{n} \theta^i.$$

For independent samples, the Law of Large Numbers ensures that

$$\frac{1}{n}\sum_{i=1}^{n} \theta^i \to \mathbb{E}_\pi(\theta) \quad \text{as } n \to \infty.$$

Independent sampling from $\pi(\theta|\boldsymbol{x})$ may be difficult, however this result still holds if we generate our samples, not independently, but via some other method (although this may be less effective than independently drawn samples in that larger sample sizes are needed to obtain the same level of accuracy).

In other words we estimate the posterior mean by the sample mean of observations taken from the posterior distribution. This is *Monte Carlo integration.* The idea extends directly to any function of θ, denoted by $f(\theta)$. For example, suppose that we wish to calculate the posterior mean of $f(\theta)$. Given a sample of observations, $\theta^1, \ldots, \theta^n \sim \pi(\theta|\boldsymbol{x})$, we can estimate the posterior mean of $f(\theta)$ by

$$\frac{1}{n}\sum_{i=1}^{n} f(\theta^i).$$

Similarly, we estimate the posterior variance, $\text{Var}_\pi(\theta)$, by the sample variance of observations taken from the posterior distribution, i.e.

$$\frac{1}{n-1}\left[\sum_{i=1}^{n}(\theta^i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\theta^i\right)^2\right].$$

Finally we note that we can obtain (marginal) density plots of the parameters of interest by using standard software. For example, suppose that in `R` the sample values of the parameters are stored in the vector `theta`, we can obtain a density plot using the command `plot(density(theta),type="l")`.

Example

Suppose that $\theta \sim N(0,1)$. Directly we have that $\mathbb{E}(\theta) = 0$ and $\text{Var}(\theta) = 1$. However, for the purposes of illustration we will use Monte Carlo integration to estimate these posterior summary statistics. We will use the function `rnorm` in `R` to independently simulate observations from this distribution and calculate the sample mean and standard deviation (SD) for different numbers of random deviates simulated (i.e. different values of n). We repeat this 3 times for each value of n. The corresponding sample means and variances I obtained were:

Repetition number	1		2		3	
n	Mean	SD	Mean	SD	Mean	SD
10	-0.39	1.26	-0.36	1.32	-0.09	0.87
100	0.36	0.88	-0.23	0.94	0.10	0.95
1000	-0.10	1.01	-0.01	1.01	0.02	0.98
10000	-0.01	1.00	0.00	0.99	0.00	1.00
100000	0.00	1.00	0.00	1.00	0.00	1.00

The above table illustrates the idea of *Monte Carlo error.* Monte Carlo error essentially measures the variation we would expect to see if multiple replications of the experiment are conducted. Monte Carlo error decreases with increasing sample size (i.e. n). Formally, it can be shown that ergodic averages satisfy the central limit theorem. For example, using the central limit theorem, we have the result that for θ values drawn independently from the distribution π,

$$\overline{\theta} \sim N\left(\mathbb{E}_\pi(\theta), \frac{\sigma^2}{n}\right).$$

The term σ^2/n is the Monte Carlo variance, or more commonly, $\sqrt{\sigma^2/n}$ is the Monte Carlo error. For the above simulations, I obtained Monte Carlo errors of 0.21, 0.07, 0.02, 0.009 and 0.004 for the increasing values of n for repetition 1.

The above example concentrates on obtaining estimates of the posterior mean (and variance) of a distribution. However, any number of posterior summary statistics may be of interest. For example, suppose that we are interested in the posterior probability that the parameter of interest, θ, has a value greater than 1. Then we can estimate $\mathbb{P}(\theta > 1|\boldsymbol{x})$ by simply calculating the proportion of the sampled values from the posterior distribution for which the parameter value is greater than 1.

Thus, we can replace the integration problem by a sampling problem, but this creates two new problems. In general, $\pi(\theta|\boldsymbol{x})$ represents a high-dimensional and complex distribution from which samples would usually be difficult to obtain. In addition, large sample sizes are often required and so powerful computers are needed to generate these samples. So, how do we obtain a potentially large sample from the posterior distribution, when in general, this will be very complex and often high-dimensional? We consider a number of "direct" sampling techniques to obtain independent samples from the posterior distribution, before the most common method applied in practice which obtains a dependent sample via the use of a Markov chain.

2.2 Direct Sampling

It is sometimes possible (for simple, and usually contrived, situations) that we can sample from the posterior distribution directly. This is typically the case for standard (or simple) posterior distributions, and hence before the development of computers and associated computational algorithms, was one of the reasons for the (historical) use of conjugate priors. We describe a number of sampling approaches for obtaining independent samples from the posterior distribution.

2.2.1 Method of Inversion

This is the simplest of all procedures and is nothing more than a straightforward application of the probability integral transform: if $X \sim F$, where F is the corresponding cumulative distribution function, then $F(X) \sim U[0,1]$. Suppose that we wish to simulate a continuous random variable X with cumulative distribution function

$$F(x) = \mathbb{P}(X \leq x).$$

Suppose also that the inverse function $F^{-1}(u)$ is well defined for $0 \leq u \leq 1$. Then we can use the following algorithm to sample from F.

```
Step 1.  Generate $U \sim U[0,1]$.
Step 2.  Let $X = F^{-1}(U)$.
```

Proof:

If $X = F^{-1}(U)$, then what is the distribution of X?

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x),$$

but since F is the cumulative distribution function of a continuous random variable, F is a strictly monotonic, increasing and continuous function of x. Hence,

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)).$$

But, as U is a $U[0,1]$ random variable, so that,

$$\mathbb{P}(U \leq F(x)) = F(x),$$

i.e.,

$$\mathbb{P}(X \leq x) = F(x),$$

where $X = F^{-1}(U)$. □

Example

Let $X \sim Exp(\lambda)$, $(\lambda > 0)$ then

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0,$$

and

$$F(x) = \int_0^x \lambda e^{-\lambda u} du = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

We need to find the function F^{-1}. Let

$$F(x) = u = 1 - e^{-\lambda x},$$

then

$$x = -\frac{1}{\lambda} \ln(1-u).$$

So, we can set

$$X = F^{-1}(U) = -\frac{1}{\lambda} \ln(1-U), \text{ where } U \sim U[0,1].$$

Note that if $U \sim U[0,1]$ then $(1-U) \sim U[0,1]$ so that we can write, $X = -\frac{1}{\lambda} \ln U$.

2.2.2 Rejection sampling

Suppose that we wish to generate observations $\theta^1, \theta^2, \ldots, \theta^n$ from the posterior distribution $\pi(\theta|\boldsymbol{x})$. One way to do this is to enclose the density function within a rectangular box and generate points uniformly at random over this region - see Figure 2.1.

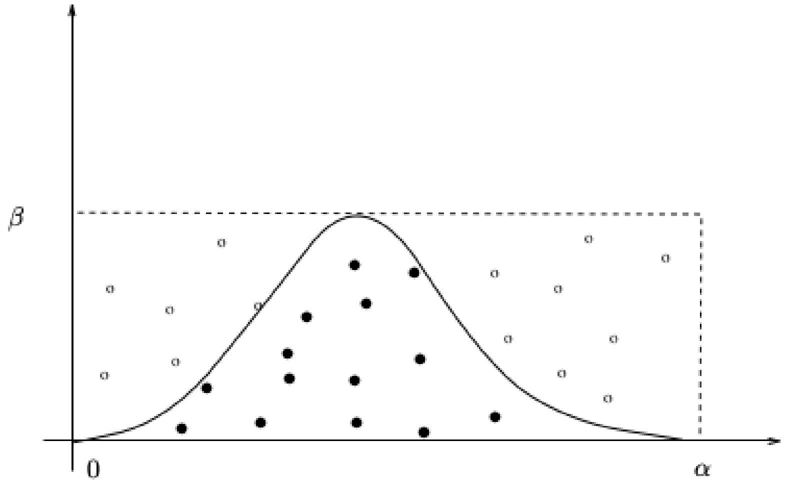


Figure. 2.1: Simulation by rejection sampling using a rectangular envelope box.

Any points outside the density function are rejected and any underneath are accepted. We take the abscissa (or x-coordinate) of the accepted points to be our required random number from the given distribution.

Thus, we use the following algorithm. For $0 \leq \theta \leq \alpha$ and $0 \leq \pi(\theta|\boldsymbol{x}) \leq \beta$.

```
Step 1.  Generate $\theta \sim U[0, \alpha]$.
Step 2.  Generate $Y \sim U[0, \beta]$.
Step 3.  Accept $\theta$ if $Y \leq \pi(\theta|\boldsymbol{x})$, else if $Y > \pi(\theta|\boldsymbol{x})$ go back to Step 1 and repeat.
```

Notes:

1. Simulating $\theta \sim U[0, \alpha]$ is equivalent to simulating $V \sim U[0, 1]$ and setting $\theta = V\alpha$.

2. We can also replace steps 2 and 3 above by simulating $W \sim U[0, 1]$ and accepting the simulated value of θ if $W \leq \frac{\pi(\theta|\boldsymbol{x})}{\beta}$.

There are a number of problems associated with this method.

1. A rectangle cannot be used when $\pi(\theta|\boldsymbol{x})$ has an infinite range, and

2. The probability of rejection can become quite large.

A way of overcoming these problems is to *envelope* the posterior density function, $\pi(\theta|\boldsymbol{x})$ *not* by a rectangle, but some some other (more general) curve, $g(\theta)$. $g(\theta)$ is often some multiple of a second p.d.f. $h(\theta)$ i.e., $g(\theta) = Mh(\theta)$, from which it easy to sample.

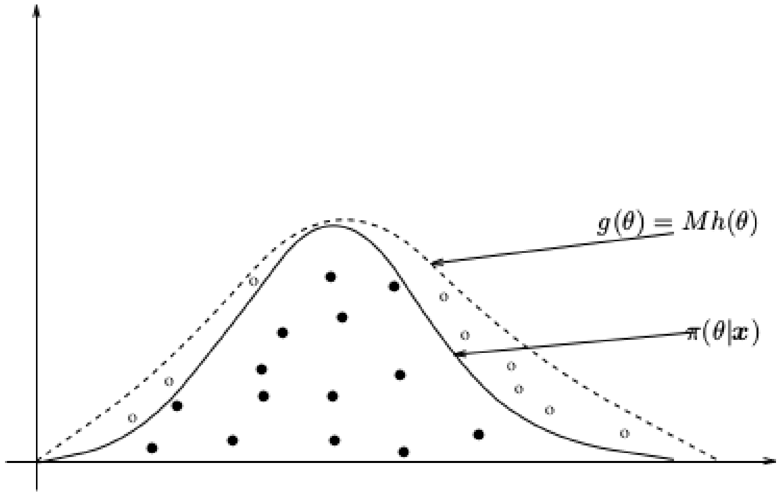


Figure. 2.2: Simulation by rejection sampling, with general envelope function g.

If we let $g(\theta) = Mh(\theta)$, where $M \geq 1$, then we can use the following algorithm.

```
Step 1.  Simulate $\theta \sim h(\theta)$.
Step 2.  Generate $Y \sim U[0, g(\theta)]$.
Step 3.  Accept $\theta$ as a realisation from $\pi(\theta|\boldsymbol{x})$ if and only if $Y \leq \pi(\theta|\boldsymbol{x})$.
```

Notes:

1. We can again replace steps 2 and 3 above by simulating $W \sim U[0,1]$ and accepting the simulated value of θ if $W \leq \frac{\pi(\theta|\boldsymbol{x})}{g(\theta)}$.

Why does this work (i.e. proof)? By definition, we need to show that

$$\mathbb{P}(\theta \leq z \mid \theta \text{ is accepted}) = \int_{-\infty}^{z} \pi(\theta|\boldsymbol{x}) d\theta.$$

(Recall the definition of a pdf).

Now we have that, using Bayes' theorem,

$$\mathbb{P}(\theta \leq z \mid \theta \text{ is accepted}) = \frac{\mathbb{P}(\theta \leq z \text{ and } \theta \text{ is accepted})}{\mathbb{P}(\theta \text{ is accepted})}.$$

We consider the numerator and denominator in turn, starting with the denominator. From Figure 2.2 we have that,

$$\begin{aligned}
\mathbb{P}(\theta \text{ is accepted}) &= \text{proportion of area under } g(\theta) \text{ that lies under } \pi(\theta|\boldsymbol{x}) \\
&= \frac{\int_{-\infty}^{\infty} \pi(\theta|\boldsymbol{x}) d\theta}{\int_{-\infty}^{\infty} g(\theta) d\theta} \\
&= \frac{1}{\int_{-\infty}^{\infty} M h(\theta) d\theta} \qquad \text{(since the pdf } \pi \text{ integrates to unity)} \\
&= \frac{1}{M} \qquad \text{(since the pdf } g \text{ integrates to unity).}
\end{aligned}$$

Further, for the numerator, using the analogous argument,

$$\mathbb{P}(\theta \leq z \text{ and } \theta \text{ is accepted}) = \frac{1}{M} \int_{-\infty}^{z} \pi(\theta|\boldsymbol{x}) d\theta.$$

Substituting back into the formula above we obtain,

$$\mathbb{P}(\theta \leq z \mid \theta \text{ is accepted}) = \int_{-\infty}^{z} \pi(\theta|\boldsymbol{x}) d\theta,$$

so that accepted values have pdf π. □

Note that the probability of rejection is given by

$$\begin{aligned}
\mathbb{P}(\text{reject}) &= 1 - \mathbb{P}(\text{accept}) \\
&= 1 - \frac{1}{M}.
\end{aligned}$$

Therefore, we would like M to be as close to 1 as possible, subject to $M \geq 1$. But, we also need that $g(\theta) = Mh(\theta) \geq \pi(\theta|\boldsymbol{x})$, for all θ, since g must "envelope" π. Therefore,

$$M \geq \frac{\pi(\theta|\boldsymbol{x})}{h(\theta)} \; \forall \theta,$$

so that the optimal M is simply

$$M^* = \sup_{\theta} \left(\frac{\pi(\theta|\boldsymbol{x})}{h(\theta)} \right),$$

where this maximum is finite.

Example

Suppose that we wish to sample from a $Beta(3,2)$ distribution. The support for the distribution is $[0,1]$ and so we can consider a rectangular sampling distribution (i.e. $h \equiv U[0,1]$). Given the sampling distribution, we calculate the optimal value of M, to be the smallest value of M such that,

$$M \geq \frac{f(\theta)}{h(\theta)} = \frac{\Gamma(5)}{\Gamma(3)\Gamma(2)}\theta^2(1-\theta).$$

The maximum is obtained when $\theta = \frac{2}{3}$ giving the value of $M = \frac{16}{9} = 1.78$. (Check!). The rejection sampling algorithm is:

Step 1. Simulate $\theta \sim U[0,1]$.
Step 2. Generate $Y \sim U[0,1.78]$.
Step 3. Accept θ as a realisation from $Beta(3,2)$ if and only if $Y \leq \frac{\Gamma(5)}{\Gamma(3)\Gamma(2)}\theta^2(1-\theta)$.

In a simulation of 100 points, (shown in Figure 2.3) 58 points are accepted, with a sample average of 0.633. The acceptance probability is $1/M = 0.56$ and the expectation of the distribution is known to be 0.6.

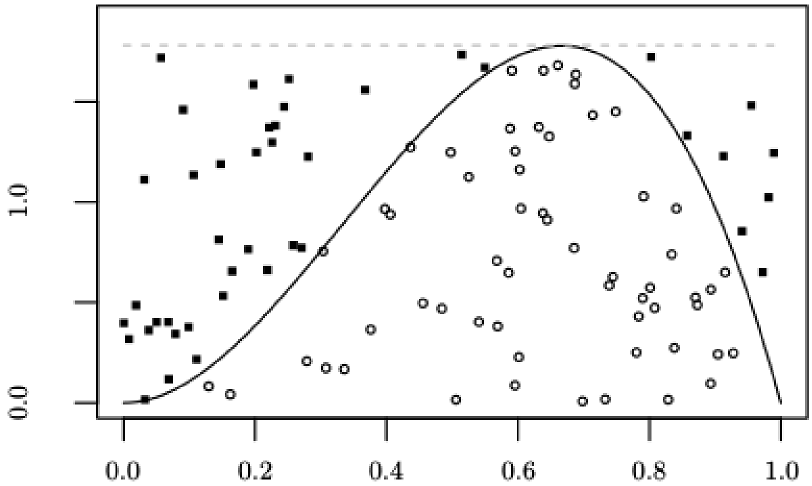


Figure. 2.3: Illustration of a rejection method for the $Beta(3,2)$ distribution. The ×'s are rejected and the ∘'s are accepted. The x-coordinates of the ∘'s are realisations of the random variable with the $Beta(3,2)$ density function.

Note that rejection sampling can be extended to the case where the normalisation constant is unknown (as for a posterior distribution). Suppose that we write the posterior distribution as,

$$\pi(\theta|\boldsymbol{x}) = \frac{f_1(\theta|\boldsymbol{x})}{f(\boldsymbol{x})},$$

where $f_1(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)\pi(\theta)$. Let $g(\theta) = Mh(\theta)$ so that $g(\theta)$ envelopes $f_1(\theta|\boldsymbol{x})$. Then use the algorithm,

Step 1. Simulate $\theta \sim h(\theta)$.
Step 2. Generate $Y \sim U[0, g(\theta)]$.
Step 3. Accept θ as a realisation from $\pi(\theta|\boldsymbol{x})$ if and only if $Y \leq f_1(\theta|\boldsymbol{x})$.

(Note - this works since we can essentially incorporate the normalisation constant into the M term). In general, rejection sampling can be very wasteful and is only really feasible in one or two dimensions. In addition, it can be difficult to identify a suitable "g" function (as we ideally want this to be of similar shape to π for efficiency).

2.2.3 Importance sampling

We once more wish to obtain a sample from the posterior distribution $\pi(\theta|\boldsymbol{x})$, which we assume is difficult to do directly. However, suppose that we can easily sample from some other distribution $g(\theta)$ (such that if $\pi(\theta|\boldsymbol{x}) > 0$ then $g(\theta) > 0$ - in other words g has at least the same support as π). We initially consider the case where both π and g are normalised densities (i.e. we know the normalisation constants). Suppose further that we are interested in obtaining an estimate of $\mathbb{E}_\pi(f(\theta))$. Let $\theta^1, \theta^2, \ldots, \theta^n$ be a sample from g and define "importance" weights

$$w(\theta^i) = \frac{\pi(\theta^i|\boldsymbol{x})}{g(\theta^i)}.$$

Then, we can estimate $\mathbb{E}_\pi(f(\theta))$ by

$$\hat{\theta}_g = \frac{1}{n}\sum_{i=1}^{n} w(\theta^i) f(\theta^i).$$

The advantage of this method is that we can use it for any densities provided that they are continuous and have the same support. In addition, it can be used even when the normalisation constant for π is unknown using,

$$\hat{\theta}_g = \frac{\sum_{i=1}^{n} w(\theta^i) f(\theta^i)/n}{\sum_{i=1}^{n} w(\theta^i)/n}$$

However there are several disadvantages. Firstly, the variance of $\hat{\theta}_g$ can be very large leading to unrealistic estimates. Secondly, without a normalisation constant for π (or g), the numerator for $\hat{\theta}_g$ tends to $c\mathbb{E}_\phi[f(\theta|\boldsymbol{x})]$ and the denominator to c. This can make the variance infinite.

Note that importance sampling can be used to reduce the variance of estimates. However the optimal importance sampling density varies with different functions of interest.

Example

Suppose that we wish to estimate the probability $\mathbb{P}(\theta > 2)$, where θ follows a Cauchy distribution, with known density

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}, \qquad \theta \in \mathbb{R}$$

so we require

$$\int_2^\infty \pi(\theta) d\theta = \int_{-\infty}^{\infty} I(\theta > 2)\pi(\theta) d\theta,$$

where I denotes the indicator function. We could simulate from the Cauchy distribution directly, but the variance of the ergodic average in this case, is very large. Alternatively, we observe that, for large θ, $\pi(\theta)$ is similar in behaviour to the density

$$g(\theta) = 2/\theta^2 \quad \theta > 2.$$

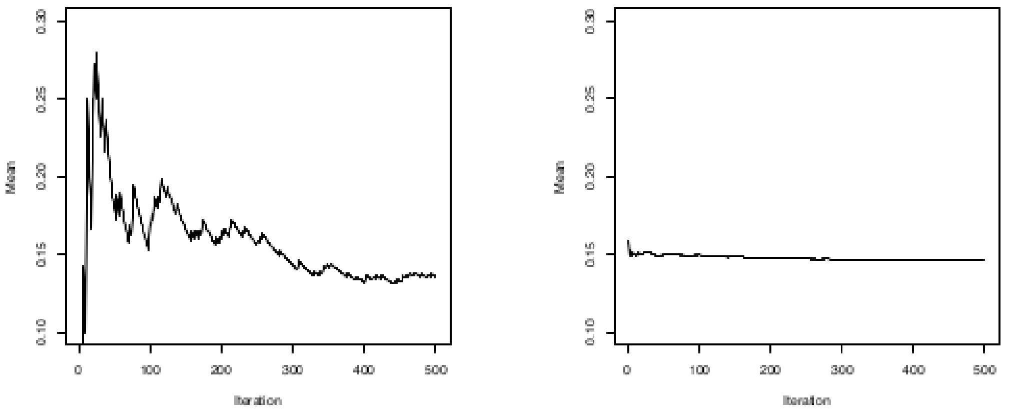
We can simulate from this distribution directly using the method of inversion. Let $U^i \sim U(0,1)$ and set $\theta^i = 2/U^i$ for $i = 1, \ldots, n$ (you should check this!). Note that g does not have the same support as π, but g does have the same support as $I(\theta > 2)\pi(\theta)$ and so we can still use importance sampling. Suppose that we sample $\theta^1, \ldots, \theta^n$ from g. We define importance sampling weights,

$$w_i = \frac{\pi(\theta^i)}{g(\theta^i)} = \frac{(\theta^i)^2}{2\pi(1+(\theta^i)^2)}.$$

Then, since each $\theta^i > 2$ we have that $f(\theta^i) = I(\theta^i > 2) = 1$ for all i. Thus, our estimator becomes:

$$\hat{\theta}_g = \frac{1}{n}\sum_{i=1}^{n} \frac{(\theta^i)^2}{2\pi(1+(\theta^i)^2)},$$

where $\theta^i = 2/U^i$, for $U^i \sim U[0,1]$ which can be easily coded in, for example, `R`.



(a) (b)

Figure. 2.4: Estimate of $\mathbb{P}(\theta > 2)$ (a) simulating directly from the Cauchy distribution and (b) using importance sampling.

The exact value of $\mathbb{P}(\theta > 2)$ is $0.5 - \pi^{-1}\tan 2 = 0.1476$. Figure 2.4(a) plots the estimated value of $\mathbb{P}(\theta > 2)$ obtained by using the `rcauchy()` command in `R` as a function of n. Figure 2.4(b) provides the corresponding plot of estimated values of $\mathbb{P}(\theta > 2)$ using the above importance sampling algorithm. Clearly, the reduction in variability is substantial!

2.2.4 Sampling importance resampling (SIR)

Sampling importance resampling in its simplest form is a simple extension of the importance sampling algorithm above. In particular, simulate random variables, $\theta^1, \ldots, \theta^n$, from some arbitrary probability density function $g(\theta)$, with (at least) the same support as the posterior distribution of interest, π. For each simulated random deviate from g, calculate the weight,

$$\omega_i = \frac{\pi(\theta^i|\boldsymbol{x})}{g(\theta^i)}$$

for $i = 1, \ldots, n$. (This is the importance sampling part). These weights are normalised using,

$$w_i = \frac{\omega_i}{\sum_{j=1}^{n}\omega_j}.$$

We independently resample with replacement the n simulated θ values, where the probability of simulating θ^i is given by w_i (the resampling part). Let the set of resampled parameter values be denoted by $\phi^1, \ldots, \phi^n$, which can then be used to obtain Monte Carlo estimates of summary statistics of interest. For example, to calculate an estimate of the posterior mean use,

$$\frac{1}{n}\sum_{i=1}^{n} \phi^i.$$

This approach of sampling importance resampling is a special form of particle filtering, which is very commonly used within economic time series data. The sampling importance sampling technique is performed sequentially over each step of the time series, though we do not consider this case here.

One problem with this sampling importance resampling type of approach can be that of *particle depletion*, where only a few simulated θ values contribute the majority of the weights. In other words suppose that we order the simulated θ values in decreasing order of their weights, and denote these by $\theta_{(1)}, \ldots, \theta_{(n)}$ where $\theta_{(j)}$ denotes the θ value with jth largest weight, with corresponding weight $w_{(j)}$. Then particle depletion occurs when $\sum_{j=1}^{n} w_{(j)}$ is close to one for n relatively small, so that only a relatively few θ values are resampled. Particle depletion leads to poor estimates of summary statistics with poor precision (i.e. large Monte Carlo error). The performance of the algorithm can be improved (as for standard importance sampling and rejection sampling) by a good choice of proposal distribution g.

All direct sampling algorithms suffer from the problem of dimensionality. These methods can be generally implemented to obtain posterior estimates of summary statistics of interest in one dimension (without too many problems) but become significantly more difficult (and generally impossible) to implement efficiently in higher dimensions. More general methods are needed. We now consider the most common approach for implementing Bayesian analyses and obtaining inference on the parameters of interest.

2.3 Markov chain Monte Carlo

2.3.1 Basic Idea

A Markov chain is simply a stochastic sequence of numbers where each value in the sequence depends *only* upon the last. In other words suppose we have a sequence of numbers $\theta^0, \theta^1, \theta^2, \ldots, \theta^n$ then θ^1 is only a function of θ^0; θ^2 is only a function of θ^1; $\ldots$; θ^n is only a function of θ^{n-1}. We let θ^0 be chosen to be equal to some arbitrary value. Thus, we can simulate a Markov chain by generating a new state of the chain, say θ^{n+1}, from some density, dependent only on θ^n:

$$\theta^{n+1} \sim \mathcal{K}(\theta^n, \theta^{n+1}) \quad \left(\equiv \mathcal{K}(\theta^{n+1}|\theta^n)\right).$$

We call $\mathcal{K}$ the transition kernel for the chain. The transition kernel uniquely describes the dynamics of the chain.

Under certain conditions (that the chain is aperiodic and irreducible) the distribution over the states of the Markov chain will converge to a *stationary* distribution (in this course we shall always assume that these conditions are met). The stationary distribution is *independent* of the initial starting values specified for the chains. Our aim is to construct a Markov chain such that the stationary distribution is equal to the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. If we can do this (which we obviously can or I would not be describing such a situation!), we can run the Markov chain until the stationary distribution is reached, so that realisations of the chain can be regarded as a *dependent* sample from the posterior distribution of interest (thus note that we need to discard the first portion of a Markov chain). We are then able to use this sample from the latter part of the chain, after it has converged, to obtain Monte Carlo estimates of the parameters of interest and/or plot their corresponding density function (see §2.1). Since we are combining a Markov chain with Monte Carlo integration this method is called Markov chain Monte Carlo (MCMC). The beauty of MCMC (as we shall see) is that the updating of the states in the Markov chain remains relatively simple, using standard techniques, irrespective of the complexity of the posterior distribution.

Clearly several questions immediately arise for such a technique:

1. How do we construct such a Markov chain?

2. Even if we can construct such a Markov chain with the correct stationary distribution, how long do we need to run the chain until the stationary distribution of interest has been reached?

3. How many samples do we need from the posterior distribution so that we accurately estimate the quantites of interest? (this should have already occurred to you for direct sampling methods!).

We will consider the latter two questions, before focussing on the first question (which clearly we need to know the answer to in order to use this method!).

2.3.2 Run Lengths

There are two issues to be considered when determining the number of iterations of the Markov chain (i.e. how many values to simulate):

(i) the time required for the Markov chain to reach the stationary distribution (i.e. for the chain to converge), and

(ii) the post-convergence sample size required for suitably small Monte Carlo errors.

Issue (i): Burn-in

We need to discard the realisations of the Markov chain before the chain has converged to the stationary distribution. The initial observations that we discard are referred to as the *burn-in.* The simplest method to determine the length of the burn-in period is to look at trace plots - these are simply the value of the parameter at each iteration of the Markov chain. It is often possible to see the individual parameters converging from their starting position to values based around a constant mean (i.e. the mean of the posterior distribution). For example, consider Figure 2.5, clearly the earliest values of the Markov chain do not look like the later values - these early values are dependent on the starting value, and hence would be discarded as burn-in. By eye we might suggest a suitable burn-in of around 200 iterations. Note it is always best to be conservative with regard to the burn-in and err on the side of overestimation to ensure that convergence has been achieved when obtaining the sample to be used to form Monte Carlo estimates of the parameters of interest.

This use of a trace plot is often a fairly efficient method, but it is not robust. For example, an ad hoc interpretation of the first trace plot in Figure 2.6 might suggest that the chain had converged after around 500 iterations. However, when the chain is run for longer, it is clear that the chain has not converged within these first 500 iterations.

Another early technique (sometimes referred to as the “thick-pen” technique) involves running two (or more chains) started at very different starting values and plotting the output on a single

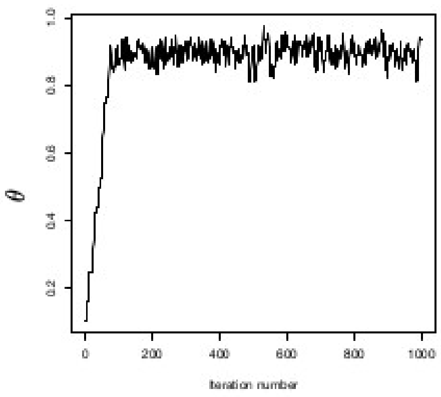


Figure. 2.5: A single MCMC trace plot.

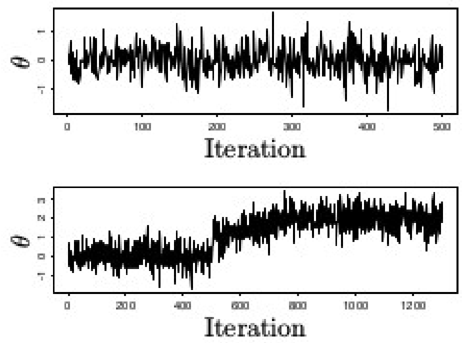


Figure. 2.6: MCMC sample paths.

graph. A "thick pen" is then taken and run over either of the trace plots from one of the simulated Markov chains. When the pen touches both lines of the plot it could be concluded that the chains had converged. For example, consider Figure 2.7 where we run two chains and plot the values from each Markov chain (ie. trace plot) on the same axes. Using this technique we might once again suggest a burn-in of at least 200.

This idea motivated many of the more formal (and mathematical) techniques for assessing convergence to the stationary distribution via the assessment of multiple replications starting from over-dispersed starting points. Essentially, this means running the Markov chains several times (from different starting points) and checking that given a suitable burn-in period the posterior estimates of all of the chains are essentially the same, providing evidence that no major nodes have been missed. The most common approach is the Brooks-Gelman-Rubin (BGR) method. There are various implementations of this diagnostic procedure, all based upon the idea of using an analysis of variance technique to check whether there are any differences in the posterior estimates obtained from the different replications. In order to implement this procedure at least two chains need to be simulated. The simplest implementation for a chain containing $2n$ iterations is to discard the first n iterations and take the ratio of the width of the empirical 80% credible interval obtained from all chains combined

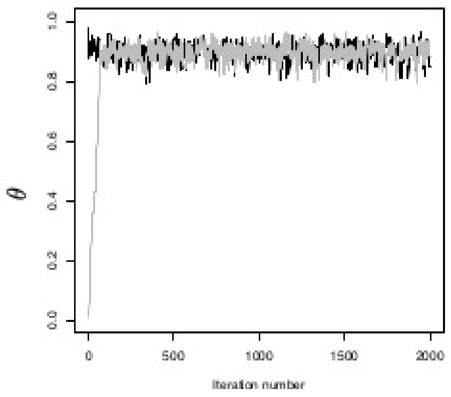


Figure. 2.7: Two MCMC trace plots from independent chains starting at different values

after the burn-in, with the corresponding mean within-chain 80% interval width, i.e., set

$$\hat{R} = \frac{\text{width of 80\% credible interval of pooled chains}}{\text{mean of width of 80\% credible interval of individual chains}}.$$

Convergence is assumed when these are roughly equal, implying that all chains have roughly equal variability, so that the $\hat{R} \approx 1$. The $\hat{R}$ value is plotted in Figure 2.8 for the two chains presented in Figure 2.7 for increasing value of n. Looking at this plot (and being conservative) we might suggest that convergence is achieved by around iteration 1000.

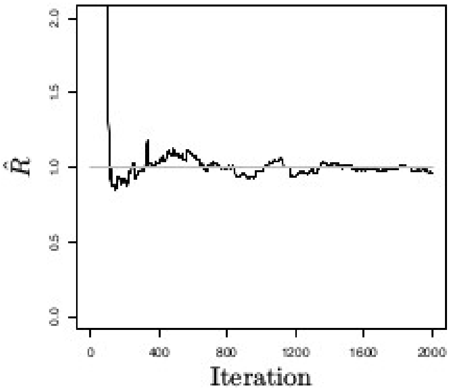
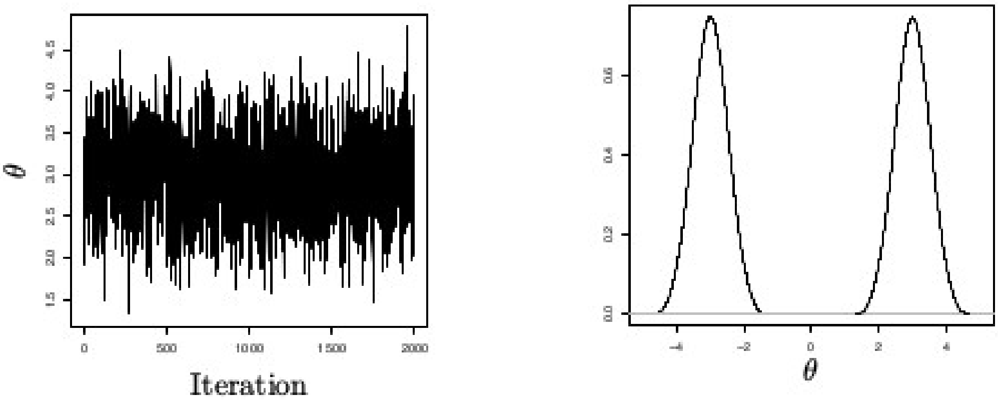


Figure. 2.8: BGR statistic for the trace plots provided in Figure 2.7.

Alternative techniques compare within-chain and between-chain variances rather than interval widths, but the principle remains the same. Note however that no convergence diagnostic can prove that the chain has converged, they can only identify when the chain has not converged. For example, consider the trace plot in Figure 2.9(a). This chain appears to converge very quickly, but the posterior distribution for this distribution is given in Figure 2.9(b). Starting several Markov chains in different starting values would quickly identify the bimodality in this example (but it can be more difficult to identify in high dimensional space).



(a) (b)

Figure. 2.9: (a) MCMC trace plot and (b) the underlying distribution.

Issue (ii): Monte Carlo error

To consider the issue of the number of iterations we need following the burn-in to obtain an accurate estimate of the summary statistics, we once more return to the idea of Monte Carlo error (see §2.1). Recall that if the samples are drawn *independently*, the sample mean, $\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta^i$ satisfies,

$$\overline{\theta} \sim N\left(\mathbb{E}_\pi(\theta), \frac{\sigma^2}{n}\right),$$

where $\sigma^2 = \text{Var}_\pi(\theta)$. However if the samples are *dependent* (as for MCMC), we also need to account for the covariance between successive samples, and

$$\sigma^2 = \text{Var}_\pi(\theta) + 2\sum_{k=2}^{\infty}\text{Cov}(\theta^1, \theta^k).$$

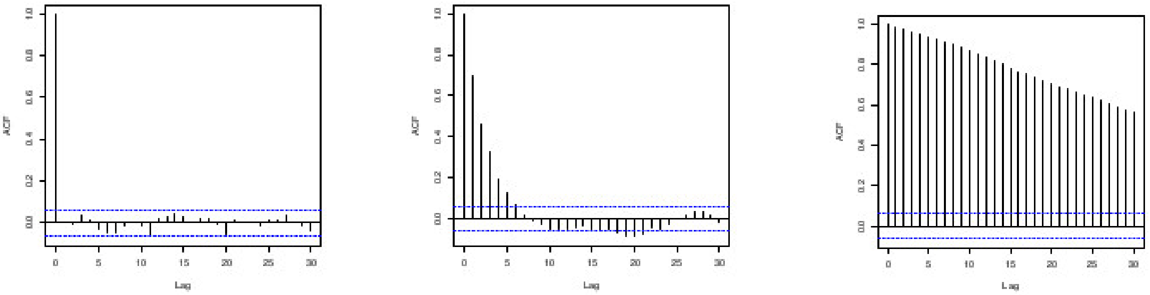
Estimating the covariance is non-trival. Thus, an alternative approach called *batching* is most often used, that estimates the variance σ^2 by using approximately independent samples. The idea involves dividing the chain into m distinct batches each of length T, so that $n = mT$ and where it is assumed that T is "large" leading to reasonably reliable sample mean estimates for each batch. Let $\overline{\theta}_1, \dots, \overline{\theta}_m$ denote the sample means for each batch. We then treat the $\overline{\theta}_1, \dots, \overline{\theta}_m$ as approximately independent. The batch means estimate of σ^2 is given by,

$$\hat{\sigma}^2 = \frac{T}{m-1}\sum_{i=1}^{m}(\overline{\theta}_i - \overline{\theta})^2.$$

Thus an estimate of the Monte Carlo error is,

$$\sqrt{\frac{\hat{\sigma}^2}{n}}.$$

The performance of the Markov chain, in terms of exploring the parameter space and hence level of Monte Carlo error is often initially performed by eye from trace plots. Chains that quickly explore the full range of plausible parameter values will have lower Monte Carlo error than chains that only slowly move over the set of plausible values. This leads to assessing the performance of the Markov chain via the *autocorrelation function* (ACF). This is simply defined to be the correlation between the given parameter value in the Markov chain separated by j iterations. The term $j > 1$ is usually referred to as *lag*. Mathematically, suppose that we are interested in parameter θ, that takes value θ^t at iteration t of the Markov chain. The autocorrelation of the parameter at lag j is simply defined to be $cor(\theta^t, \theta^{t+j})$. This is typically calculated for values $j = 1, \dots, j_{max}$ and plotted on a graph. Example ACF plots are provided in Figure 2.10.



(a) (b) (c)

Figure. 2.10: Sample ACF plots representing (a) ideal mixing; (b) typical good mixing; (c) poor mixing.

Note that the autocorrelation function is always equal to 1 for the value $j = 0$, since $cor(\theta^t, \theta^t) = 1$. Ideally, for efficient Markov chains (as in Figures 2.10(a) and (b)), there should be a fast decrease in the value of the autocorrelation function as the lag increases. In other words, in the ACF plot, this would be represented by a sharp gradient at low values of j. This would imply that there is little relationship between values of the Markov chain within a small number of iterations. Conversely, poorly mixing chains will typically have a very shallow gradient in the ACF plot, with high autocorrelation values for even relatively large values of j (say, $j \geq 20$, as in Figure 2.10(c)).

Finally, we note that Monte Carlo error can be reduced via the process of *thinning*. This involves simply taking every kth realisation (e.g. every 10th iteration) of the Markov chain and discarding the rest. This clearly reduces the autocorrelation of the MCMC sample being used to obtain posterior summary statistics of interest. The discarded values (although possibly very highly dependent on the previous value in the Markov chain) still provides information concerning the posterior distribution. Thus thinning should only be used if there are issues relating to the storage and/or memory allocation of the large number of sampled values.

We now describe the most common algorithms for simulating a Markov chain with a given stationary distribution.

2.3.3 Gibbs Sampler

The Gibbs sampler works as follows. Given a vector variable $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$ with distribution $\pi(\boldsymbol{\theta})$, the Gibbs sampler uses the set of full conditionals of π to sample indirectly from the full posterior distribution. (Within our Bayesian context, π is the posterior distribution of interest but we have dropped the conditioning on the data, $\boldsymbol{x}$, for notational simplicity).

Let $\pi(\theta_i|\boldsymbol{\theta}_{(i)})$ denote the induced full conditional of θ_i, given the values of the other components $\boldsymbol{\theta}_{(i)} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$, $i = 1, \dots, k$, $1 \leq k \leq p$ (and the data). Then, we initially set arbitrary starting values $\boldsymbol{\theta}^0 = (\theta_1^0, \dots, \theta_k^0)$. Given the Markov chain is in state $\boldsymbol{\theta}^t$ at iteration t of the Markov chain, the Gibbs sampler successively makes random drawings from the full conditional distributions $\pi(\theta_i|\boldsymbol{\theta}_{(i)})$, $i = 1, \dots, k$ as follows:

$$\begin{array}{lll}
\theta_1^{t+1} & \text{is sampled from} & \pi(\theta_1|\theta_2^t, \dots, \theta_k^t) \\
\theta_2^{t+1} & \text{is sampled from} & \pi(\theta_2|\theta_1^{t+1}, \theta_3^t, \dots, \theta_k^t) \\
. & . & . \\
. & . & . \\
. & . & . \\
\theta_i^{t+1} & \text{is sampled from} & \pi(\theta_i|\theta_j^{t+1},\ j < i \text{ and } \theta_j^t, j > i) \\
. & . & . \\
. & . & . \\
. & . & . \\
\theta_k^{t+1} & \text{is sampled from} & \pi(\theta_k|\theta_1^{t+1}, \dots, \theta_{k-1}^{t+1}).
\end{array}$$

This completes a transition from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$. Iteration of the full cycle of random variate generations from each of the full conditionals in turn, produces a sequence $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^t, \dots, \boldsymbol{\theta}^T$. The values $\boldsymbol{\theta}^0, \dots, \boldsymbol{\theta}^N$ are discarded as burn-in, for some suitable value of N (see §2.3.2), and the values $\boldsymbol{\theta}^{N+1}, \dots, \boldsymbol{\theta}^T$ can be used to obtain Monte Carlo estimates of interest.

The transition probability for going from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$ is given by

$$\mathcal{K}_G(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}) = \prod_{i=1}^{k} \pi(\theta_i^{t+1}|\theta_j^{t+1},\ j < i \text{ and } \theta_j^t,\ j > i), \tag{2.1}$$

and has stationary distribution π.

Thus, in two dimensions a typical trajectory of the Gibbs sampler may look something like that given in Figure 2.11.

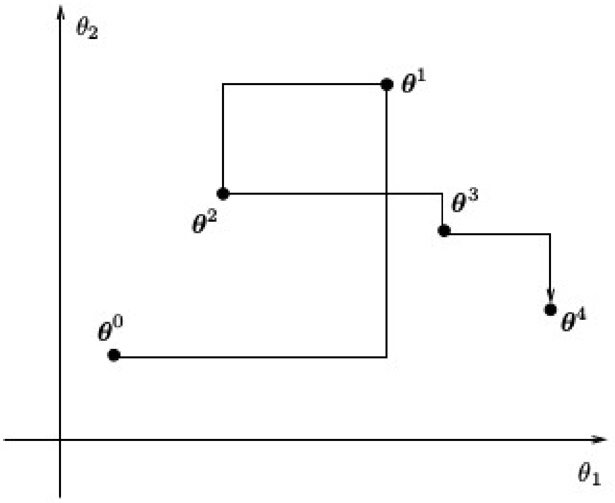


Figure. 2.11: Typical Gibbs sampler sample path.

Example

Consider observed data $x_1, \ldots, x_n$ such that each $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We specify independent priors:

$$\mu \sim N(\phi, \tau^2); \qquad \text{and} \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

Then we have that,

$$\begin{aligned} \mu|\boldsymbol{x}, \sigma^2 &\sim N\left(\frac{\tau^2 n\bar{x} + \sigma^2\phi}{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n + \sigma^2}\right); \\ \sigma^2|\boldsymbol{x}, \mu &\sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta\right), \end{aligned}$$

(see §1.8.1 for each single parameter case - recall that for the posterior conditional distributions we condition on all other parameters, and so treat them as if they were fixed). Thus, setting initial parameter values for μ^0 and $(\sigma^2)^0$, we would implement the Gibbs sampler by simulating,

$$\begin{aligned} \mu^{t+1}|\boldsymbol{x}, (\sigma^2)^t &\sim N\left(\frac{\tau^2 n\bar{x} + (\sigma^2)^t\phi}{\tau^2 n + (\sigma^2)^t}, \frac{(\sigma^2)^t\tau^2}{\tau^2 n + (\sigma^2)^t}\right); \\ (\sigma^2)^{t+1}|\boldsymbol{x}, \mu^{t+1} &\sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu^{t+1})^2 + \beta\right). \end{aligned}$$

In other words:

Step 1. Set initial parameter value for μ and σ^2 denoted by $(\mu, \sigma^2)^0 = \{\mu^0, (\sigma^2)^0\}$.

Step 2. Conditional on the current parameter value, $(\sigma^2)^t$, generate a new value for μ, from the posterior conditional distribution,

$$\mu^{t+1}|\boldsymbol{x}, (\sigma^2)^t \sim N\left(\frac{\tau^2 n\bar{x} + (\sigma^2)^t\phi}{\tau^2 n + (\sigma^2)^t}, \frac{(\sigma^2)^t\tau^2}{\tau^2 n + (\sigma^2)^t}\right).$$

Step 3. Conditional on the newly updated parameter value, μ_{t+1}, generate a new value for σ^2, from the posterior conditional distribution,

$$(\sigma^2)^{t+1}|\mu^{t+1}, \boldsymbol{x} \sim \Gamma\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu^{t+1})^2 + \beta\right).$$

Step 4. Increase t by one and return to step 2, until T iterations have been performed.

Note that the ordering of the parameters is unimportant so we could reverse the order of updating each parameters and use the updating scheme,

$$\begin{aligned} (\sigma^2)^{t+1}|\boldsymbol{x}, \mu^t &\sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu^t)^2 + \beta\right) \\ \mu^{t+1}|\boldsymbol{x}, (\sigma^2)^{t+1} &\sim N\left(\frac{\tau^2 n\bar{x} + (\sigma^2)^{t+1}\phi}{\tau^2 n + (\sigma^2)^{t+1}}, \frac{(\sigma^2)^{t+1}\tau^2}{\tau^2 n + (\sigma^2)^{t+1}}\right). \end{aligned}$$

Remember that within an iteration, we always use the "current" value of the other parameters - this corresponds to the updated value (i.e. the $(t+1)$th value) if we have already updated the parameter within the iteration of the Markov chain OR the value of the parameter in the previous iteration (i.e. the tth value).

For example, setting prior parameters $\phi = 0$, $\tau = 1$, $\alpha = 0.1$ and $\beta = 0.01$ and simulating $x_1, \ldots, x_{10}$ from a standard normal distribution (so the true values are $\mu = 0$ and $\sigma^2 = 1$), Figure 2.12 provides the corresponding (marginal) trace plots of a Gibbs sampler for parameters μ and σ^2 for 100 iterations.

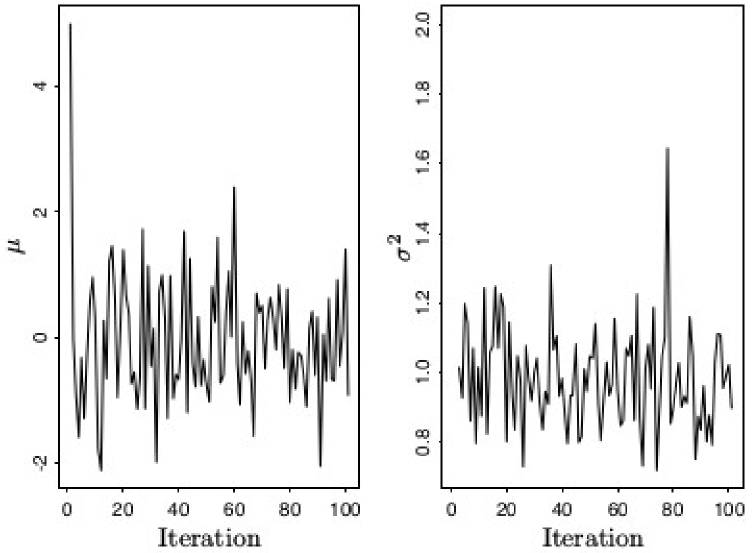


Figure. 2.12: Output from the Gibbs sampler, showing the sample paths for the variables μ and σ^2.

Conceptually, the Gibbs sampler appears to be a rather straightforward algorithmic procedure. Ideally, each of the conditionals will be of the form of a standard distribution and suitable prior specification often ensures that this is the case (for example, use of conjugate priors). However, in the cases where one or more of the conditionals is non-standard there are many ways to sample from univariate conditionals (e.g. direct sampling methods) however many of these algorithms are generally computationally intensive and inefficient. An alternative (and standard) approach is to use the Metropolis-Hastings algorithm for non-standard posterior conditionals.

2.3.4 Metropolis-Hastings Algorithm

A general way to construct MCMC samplers is as a form of generalised rejection sampling, where values are drawn from approximate distributions and “corrected” in order that, asymptotically, they behave as random observations from the target distribution. This is the motivation for methods such as the Metropolis-Hastings algorithm which sequentially draws candidate observations from a distribution, conditional only upon the last observation, thus inducing a Markov chain. The most important aspect of such algorithms is not the Markov property, but the fact that the approximating candidate distributions are improved at each step in the simulation.

The method commonly known as the Metropolis-Hastings algorithm, is based upon the observation that given a Markov chain with transition density $\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and exhibiting detailed balance for π i.e.,

$$\pi(\boldsymbol{\theta})\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi})\mathcal{K}(\boldsymbol{\phi}, \boldsymbol{\theta}), \tag{2.2}$$

the chain has stationary density, $\pi(\cdot)$.

The candidate generating density (or proposal density) typically depends upon the current state of the chain, and we denote it by $q(\boldsymbol{\phi}|\boldsymbol{\theta}^t)$. The choice of proposal density is essentially arbitrary. However, in general, the induced chain will not satisfy the reversibility condition of (2.2), so we introduce an acceptance function $\alpha(\boldsymbol{\theta}^t, \boldsymbol{\phi})$. We then accept the candidate observation, and set $\boldsymbol{\theta}^{t+1} = \boldsymbol{\phi}$, with probability $\alpha(\boldsymbol{\theta}^t, \boldsymbol{\phi})$; else if the candidate observation is rejected, the chain remains at $\boldsymbol{\theta}^t$, so that $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$.

It can be shown that the optimal form for the acceptance function, in the sense that suitable candidates are rejected least often and computational efficiency is maximised, is given by

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\phi}) = \min\left(1, \frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\theta}^t|\boldsymbol{\phi})}{\pi(\boldsymbol{\theta}^t)q(\boldsymbol{\phi}|\boldsymbol{\theta}^t)}\right),$$

with transition kernel given by

$$\mathcal{P}_H(\boldsymbol{\theta}, A) = \int_A \mathcal{K}_H(\boldsymbol{\theta}, \boldsymbol{\phi})d\boldsymbol{\phi} + r(\boldsymbol{\theta})I_A(\boldsymbol{\theta}),$$

where

$$\mathcal{K}_H(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\phi}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}),$$

$$r(\boldsymbol{\theta}) = 1 - \int q(\boldsymbol{\phi}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})d\boldsymbol{\phi}, \tag{2.3}$$

such that $I_A(\boldsymbol{\theta})$ denotes the indicator function for $\boldsymbol{\theta} \in A$.

Now $\mathcal{K}_H$ satisfies the reversibility condition of (2.2), implying that the kernel, $\mathcal{P}_H$ also preserves detailed balance for π.

The Metropolis-Hastings method can be written algorithmically as follows.

Step 1. Set initial parameter value for $\boldsymbol{\theta}$ denoted by $\boldsymbol{\theta}^0$.

Step 2. Given the current position, $\boldsymbol{\theta}^t = \boldsymbol{\theta}$, generate a new value, $\boldsymbol{\phi}$, from the distribution $q(\boldsymbol{\phi}|\boldsymbol{\theta})$.

Step 3. Calculate

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\left(1, \frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\theta}|\boldsymbol{\phi})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\phi}|\boldsymbol{\theta})}\right).$$

Step 4. With probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})$, set $\boldsymbol{\theta}^{t+1} = \boldsymbol{\phi}$, else set $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}$.

Step 5. Increase t by one and return to step 1 until T iterations have been performed.

To obtain a set of sampled values from π, discard $\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^N$ as burn-in, for some suitable

value of N, and consider the values $\boldsymbol{\theta}^{N+1}, \ldots, \boldsymbol{\theta}^T$ as a (dependent) sample from π which can be used to obtain Monte Carlo estimates of parameters of interest.

Important notes:

1. We only need to know π up to proportionality, since any constants of proportionality cancel in the numerator and denominator of the calculation of α.

2. The performance of the MCMC algorithm is dependent on the choice of the proposal distribution q. If q is chosen poorly, then the number of rejections may be high, so that the efficiency of the procedure can be low; conversely if q is chosen such that only very small moves are proposed, it may take a very long time for the Markov chain to traverse the set of plausible posterior parameter values (see below example).

(Simple) Example

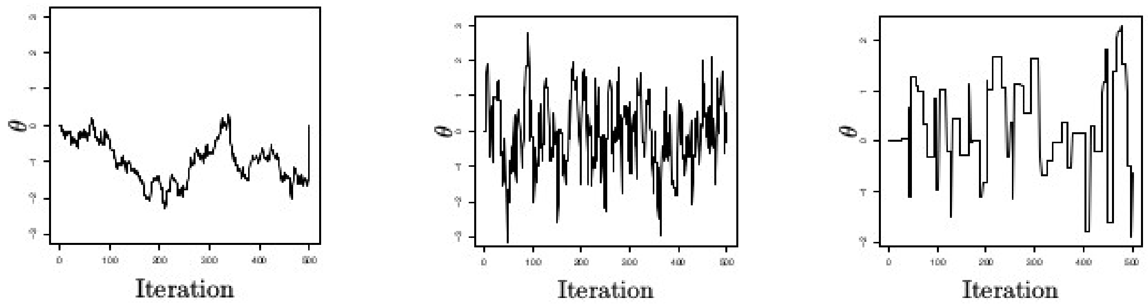
Suppose that we are interested in sampling from the standard normal distribution, and that we choose to use a proposal distribution of the form

$$q(\phi|\theta) \sim N(\theta, \sigma^2),$$

where σ^2 is to be specified. Then the acceptance probability is given by

$$\alpha(\theta, \phi) = \exp\left(-\frac{1}{2}(\phi^2 - \theta^2)\right). \tag{2.4}$$

Figure 2.13 plots the resulting output for $\sigma^2 = 0.1$, 1 and 100.

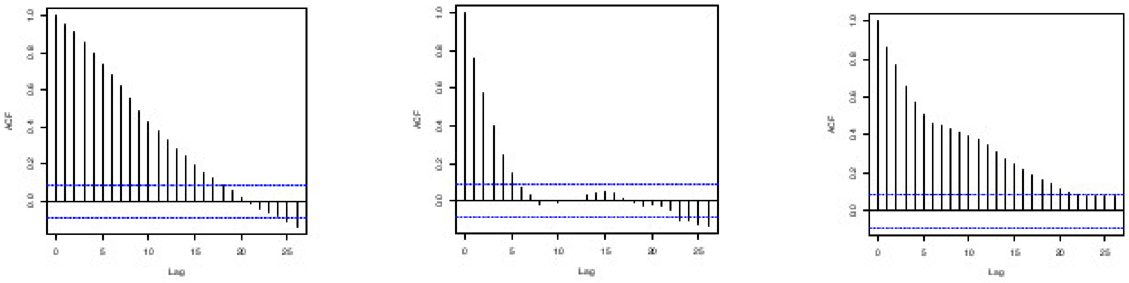


(a) (b) (c)

Figure. 2.13: Sample paths for Metropolis-Hastings algorithms with (a) $\sigma^2 = 0.1$, (b) $\sigma^2 = 1$ and (c) $\sigma^2 = 100$.

Notice that, in this case, the acceptance function is independent of σ, but that the value of σ has a significant impact upon the acceptance rate of the chain. This is because the proposal with $\sigma^2 = 1$ is much closer to the target distribution, so that more sensible candidates are generated and subsequently accepted. However, the proposal with $\sigma^2 = 100$, generates candidate observations too far out in the tail to come from the target distribution and these are subsequently rejected. Conversely, the proposal with $\sigma^2 = 0.1$ generates candidates very similar to the current value that are typically accepted, but means that it takes a long time to move over the parameter space. The movement around the parameter space is often referred to as "mixing" (see also Improving Performance)

Trace plots are a useful graphical tool for (informally) assessing the mixing of the Markov chain. In addition to trace plots ACF plots (see §2.3.2) are often used in conjunction with trace plots to assess mixing. For the above trace plots, Figure 2.14. provides the corresponding ACF plots for both $\sigma^2 = 0.1$ and $\sigma^2 = 100$ appear fairly similar. Thus, ACF plots on their own do not provide enough information to explain why a Markov chain may be experiencing poor mixing - this could be a result of (at least) a high rejection probability (as for $\sigma^2 = 100$) or always very small "step" sizes (as for $\sigma^2 = 0.1$).



(a) (b) (c)

Figure. 2.14: ACF plots for Metropolis-Hastings algorithms with (a) $\sigma^2 = 0.1$, (b) $\sigma^2 = 1$ and (c) $\sigma^2 = 100$.

Special Cases

There are a number of special cases (or "flavours") of the Metropolis-Hastings algorithm. We consider a number of these next.

Metropolis Algorithm

This (original) special case specifies the candidate generating function as a symmetric function, i.e., $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\phi})$. The acceptance function reduces to

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\left(1, \frac{\pi(\boldsymbol{\phi})}{\pi(\boldsymbol{\theta})}\right). \tag{2.5}$$

Random Walk Metropolis

If $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = f(|\boldsymbol{\phi} - \boldsymbol{\theta}|)$ for some arbitrary density f, then the kernel driving the chain is a random walk, since the candidate observation is of the form $\boldsymbol{\phi} = \boldsymbol{\theta}^t + \boldsymbol{z}$, where $\boldsymbol{z} \sim f$. There are many common choices for f, including the uniform distribution on the unit disk, or a multivariate normal or t-distribution. Note that these are symmetric, so that the acceptance probability is of the simple form given in (2.5).

The Independence Sampler

If $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = f(\boldsymbol{\phi})$, then the candidate observation is drawn *independently* of the current state of the chain. In this case, the acceptance probability can be written as

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\left(1, \frac{w(\boldsymbol{\phi})}{w(\boldsymbol{\theta})}\right),$$

where $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})/f(\boldsymbol{\theta})$. (This is the importance weight function that would be used in importance sampling given observations generated from f).

Single-updates and the Gibbs Sampler

The Metropolis-Hastings algorithm need not update all variables in the sample space simultaneously, it can be used in stages, updating variables one-at-a-time, much like the Gibbs Sampler. This is commonly called the single-update Metropolis-Hastings algorithm. This algorithm can be described as follows. Let the initial state be denoted by $\boldsymbol{\theta}^0 = (\theta_1^0, \dots, \theta_k^0)$. At iteration t we cycle through each of the $\theta_1, \dots, \theta_k$ parameters in turn and propose to update the parameter value. Consider parameter θ_p and set $\boldsymbol{\theta}_p^t = (\theta_1^{t+1}, \theta_{p-1}^{t+1}, \theta_p^t, \theta_{p+1}^t, \dots, \theta_k^t)$.

Step 1 Propose new candidate value $\phi_p \sim q(\phi_p|\boldsymbol{\theta}_p^t)$, and set $\boldsymbol{\phi}_p = (\theta_1^{t+1}, \theta_{p-1}^{t+1}, \phi_p, \theta_{p+1}^t, \dots, \theta_k^t)$ (i.e. $\boldsymbol{\phi}_p$ denotes the vector of parameter values replacing parameter θ_p^t with the candidate value for ϕ_p).

Step 2 Accept the candidate value with probability $\min(1, A)$, where,

$$A = \frac{\pi(\boldsymbol{\phi}_p|\boldsymbol{x})q(\theta_p^t|\boldsymbol{\phi}_p)}{\pi(\boldsymbol{\theta}_p^t|\boldsymbol{x})q(\phi_p|\boldsymbol{\theta}_p^t)}.$$

We note that this acceptance probability simplifies to,

$$A = \frac{\pi(\phi_p|\boldsymbol{x}, \boldsymbol{\theta}_{(p)}^t)q(\theta_p^t|\boldsymbol{\phi}_p')}{\pi(\theta_p^t|\boldsymbol{x}, \boldsymbol{\theta}_{(p)}^t)q(\phi_p|\boldsymbol{\theta}_p^t)}.$$

where $\boldsymbol{\theta}_{(p)}^t = (\theta_1^{t+1}, \dots, \theta_{p-1}^{t+1}, \theta_{p+1}^t, \dots, \theta_k^t)$. This simplification is obtained by noting that $\pi(\boldsymbol{\phi}_p|\boldsymbol{x}) = \pi(\phi_p|\boldsymbol{x}, \boldsymbol{\theta}_{(p)}^t)\pi(\boldsymbol{\theta}_{(p)}^t|\boldsymbol{x})$ and similarly for $\pi(\boldsymbol{\theta}_p^t|\boldsymbol{x})$.

If the candidate value is accepted, set $\theta_p^{t+1} = \phi_p$; else if the move is rejected set $\theta_p^{t+1} = \theta_p^t$.

We note that, in fact, the Gibbs Sampler is a special case of the single-update Metropolis-Hastings algorithm. Suppose that in the single-update Metropolis-Hastings algorithm, we break each iteration of the algorithm into k steps, and let q_j denote a proposal for candidates in the jth co-ordinate direction, so that

$$q_j(\boldsymbol{\phi}|\boldsymbol{\theta}) = \begin{cases} \pi(\phi_j|\boldsymbol{\theta}_{(j)}) & \boldsymbol{\phi}_{(j)} = \boldsymbol{\theta}_{(j)}, \quad j = 1, \dots, k. \\ 0 & \text{else} \end{cases}$$

With this proposal, the acceptance probability at the jth step is given by

$$\alpha_j(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\left(1, f(\boldsymbol{\theta}, \boldsymbol{\phi})\right),$$

where,

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \frac{\pi(\boldsymbol{\phi})q_j(\boldsymbol{\theta}|\boldsymbol{\phi})}{\pi(\boldsymbol{\theta})q_j(\boldsymbol{\phi}|\boldsymbol{\theta})} \\ &= \frac{\pi(\boldsymbol{\phi})/\pi(\phi_j|\boldsymbol{\theta}_{(j)})}{\pi(\boldsymbol{\theta})/\pi(\theta_j|\boldsymbol{\phi}_{(j)})} \\ &= \frac{\pi(\boldsymbol{\phi})/\pi(\phi_j|\boldsymbol{\phi}_{(j)})}{\pi(\boldsymbol{\theta})/\pi(\theta_j|\boldsymbol{\theta}_{(j)})}, \quad \text{since } \boldsymbol{\phi}_{(j)} = \boldsymbol{\theta}_{(j)} \\ &= \frac{\pi(\boldsymbol{\phi}_{(j)})}{\pi(\boldsymbol{\theta}_{(j)})}, \\ &\quad \text{by definition of conditional probability for } \boldsymbol{\theta} = (\theta_j, \boldsymbol{\theta}_{(j)}) \\ &= 1, \text{ since } \boldsymbol{\phi}_{(j)} = \boldsymbol{\theta}_{(j)}. \end{aligned}$$

Thus, at each step, the only possible jumps are to parameter vectors $\boldsymbol{\phi}$, that match $\boldsymbol{\theta}$ on all components other than the jth, and these are automatically accepted.

In other words we have the proposal density given by,

$$q(\boldsymbol{\phi}|\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j([\boldsymbol{\phi}_{<j}, \boldsymbol{\theta}_{\geq j}], [\boldsymbol{\phi}_{\leq j}, \boldsymbol{\theta}_{>j}]),$$

where

$$\boldsymbol{\theta}_{>j} = (\boldsymbol{\theta}_i, \;\; i > j),$$

$$\boldsymbol{\phi}_{<j} = (\boldsymbol{\phi}_i, \;\; i < j)$$

and, since $\alpha_j(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1 \quad \forall j$, the transition distribution can, by simple manipulation, be re-written in the same form as the Gibbs transition density given in (2.1).

Improving performance

The performance of the MCMC algorithm depends jointly on the target distribution of interest and the updating algorithm used. The target distribution is typically fixed (e.g. posterior distribution of interest), so that the performance of the algorithm can be changed only through the proposals used in the updating algorithm. With the exception of the Gibbs sampler, most MCMC updates require a degree of *pilot-tuning* to obtain a chain with good mixing properties. In practice, this often involves adjusting the relevant proposal variances to obtain a Metropolis-Hastings acceptance rate of 20-40% (Gelman *et al*, 1996). This can often be achieved by implementing a pilot run of the MCMC algorithm, for 1000 iterations, say, calculating the mean acceptance rate for each parameter and adjusting the proposal variance accordingly to obtain a mean acceptance rate in the given interval.

In addition, if parameters are highly correlated with each other (this is usually easy to assess from an initial pilot-run - for example calculating the correlation of the parameters), multi-parameter updates can be used, proposing to update a number of parameters simultaneously. This is often referred to as *blocking*, since parameters are updated in “blocks”. In practice this is most commonly done via the Metropolis-Hastings algorithm, since the joint posterior conditional distribution of the parameters is typically non-standard.

2.3.5 Comparison of Gibbs sampler and Metropolis-Hastings algorithm

The Gibbs sampler is usually the “default” updating algorithm when the posterior conditional distribution is of standard form. This is because it can be seen as “efficient” since the parameter is always updated (i.e. with probability 1) and there are many computer programs (including `R`) that includes efficient intrinsic functions for simulating from a wide range of standard distributions. In addition, since the Gibbs sampler is conditional on the other parameters, it can be viewed as an “adaptive” updating algorithm in that the parameters of the conditional distribution are usually a function of the other parameters in the model and so change as the other parameter values change within the Markov chain. In other words the conditional distribution of the parameter will change dependent on the current state of the Markov chain.

However, the Metropolis-Hastings algorithm has the advantage over the Gibbs Sampler, in that it is not necessary to know (or recognise) all of the conditional distributions, we need only simulate from q, which we can choose arbitrarily. However, if q is poorly chosen, then the mixing of the Markov chain can be slow, so that the efficiency of the procedure can be low. Thus, the choice of q typically involves some *pilot-tuning* for the parameters within the proposal distribution.

Typically, Gibbs updates are more computationally intensive than Metropolis-Hastings updates. Gibbs updates require simulating from (possibly) complex distributions, whereas the Metropolis-Hatings algorithm involves simulating from simpler distributions which is computationally faster (the Metropolis-Hastings only needs to simulate from the proposal distribution q and a $U[0, 1]$ distribution to perform the accept/reject step). However, typically time is taken when implementing the Metropolis-Hastings algorithm to perform pilot-tuning to make the algorithm efficient, whereas no such pilot-tuning is required for the Gibbs step. If parameters are highly correlated *a posteriori*, using a single-update MCMC algorithm (either Gibbs or Metropolis-Hastings) will often perform poorly. This can be solved by block updates (i.e. updating more than one parameter within a single step). Block-updates are typically easier to perform within the Metropolis-Hastings algorithm, due to the arbitrary nature of the proposal distribution q (we can define q to be any multivariate distribution, such as the multivariate Normal distribution).

In general the "default" MCMC algorithm would update parameters with standard posterior conditional distributions via the Gibbs sampler; whereas parameters with non-standard posterior conditionals are updated using the Metropolis-Hastings algorithm. The real exception comes when parameters are highly correlated, in which case (unless the joint posterior conditional distribution of these parameters are of standard form, particularly multivariate Normal) a block Metropolis-Hastings algorithm would typically be used.

2.3.6 Example - Rats

We consider an example corresponding to data from an experiment on the weight of rats over time.

Data

The data correspond to the weight of $N = 30$ rats on $T = 5$ different days. The form of the data are provided in Table 2.1. (The full dataset is provided on *Learn*) as a data file.

	Day (x_j)				
Rat	8	15	22	29	36
1	151	199	246	283	320
2	145	199	249	293	354
3	147	214	263	312	328
4	155	200	237	272	297
⋮					
30	153	200	244	286	324

Table. 2.1: The weight of each rat at days 8, 15, 22, 29 and 36.

Model

Let Y_{ij} denote the weight of rat $i = 1, \dots, 30$ at time $j = 1, \dots, 5$. We assume a very simple linear model, where weight is linearly regressed on time, and

$$Y_{ij} | \alpha, \beta, \sigma^2 \sim N(\alpha + \beta z_j, \sigma^2),$$

where α, β and σ^2 are parameters to be estimated; and z_j denotes the jth (normalised) time given by

$$z_j = \frac{x_j - \text{mean}(x)}{\text{sd(x)}}.$$

Note that this means that $\sum_{j=1}^{5} z_j = 0$.

(This linear regression model is not actually a very realistic model for these data!).

The corresponding likelihood function is given by:

$$\begin{aligned} f(\boldsymbol{y}|\alpha,\beta,\sigma^2) &= \prod_{i=1}^{N}\prod_{j=1}^{T}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_{ij}-\alpha-\beta z_j)^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-\frac{NT}{2}}\exp\left(-\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}(y_{ij}-\alpha-\beta z_j)^2}{2\sigma^2}\right), \end{aligned}$$

where $N = 30$ and $T = 5$.

Note: Within Bayesian analyses it is important to normalise the data corresponding to the explanatory variable(s). This simply means we take each set of explanatory variables, subtract the sample mean from each value and divide by the sample standard deviation. This is important in terms of the specification of the priors on the parameters for comparability and interpretability.

Prior distribution

We specify the independent priors:

$$\alpha \sim N(\mu_\alpha, \sigma_\alpha^2); \qquad \beta \sim N(\mu_\beta, \sigma_\beta^2); \qquad \sigma^2 \sim \Gamma^{-1}(a, b).$$

The corresponding prior density is given by,

$$\begin{aligned} p(\alpha,\beta,\sigma^2) &= p(\alpha)p(\beta)p(\sigma^2) \\ &= \frac{1}{2\pi\sigma_\alpha^2}\exp\left(-\frac{(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2}\right) \\ &\quad \times\frac{1}{2\pi\sigma_\beta^2}\exp\left(-\frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right)\times\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right). \end{aligned}$$

Note that we will assume that we do not have any prior information and hence specify $\mu_\alpha = \mu_\beta = 0$; $\sigma_\alpha^2\sigma_\beta^2 = 10^5$ and $a = b = 0.001$. However, it is often useful to calculate the posterior distribution in the general case, so that it is simple to change the prior parameters, for example, when conducting a prior sensitivity analysis.

Posterior distribution

The posterior distribution is given by,

$$\begin{aligned} \pi(\alpha,\beta,\sigma^2|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\alpha,\beta,\sigma^2)p(\alpha,\beta,\sigma^2) \\ &\propto (\sigma^2)^{-\frac{NT}{2}}\exp\left(-\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}(y_{ij}-\alpha-\beta z_j)^2}{2\sigma^2}\right) \\ &\quad \times\exp\left(-\frac{(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2}\right)\exp\left(-\frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right)(\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right). \end{aligned}$$

Posterior conditional distributions

We can calculate the posterior conditional distributions for the parameters as follows:

$$\begin{aligned}
\alpha|\boldsymbol{y},\beta,\sigma^2 &\sim N\left(\frac{NT\bar{y}\sigma_\alpha^2+\mu_\alpha\sigma^2}{NT\sigma_\alpha^2+\sigma^2},\frac{\sigma^2\sigma_\alpha^2}{NT\sigma_\alpha^2+\sigma^2}\right)\\
\beta|\boldsymbol{y},\alpha,\sigma^2 &\sim N\left(\frac{\sum_{ij}z_jy_{ij}\sigma_\beta^2+\mu_\beta\sigma^2}{N\sum_j z_j^2\sigma_\beta^2+\sigma^2},\frac{\sigma^2\sigma_\beta^2}{N\sum_j z_j^2\sigma_\beta^2+\sigma^2}\right)\\
\sigma^2|\boldsymbol{y},\alpha,\beta &\sim \Gamma^{-1}\left(\frac{NT}{2}+a,\frac{1}{2}\sum_{ij}(y_{ij}-\alpha-\beta z_j)^2+b\right).
\end{aligned}$$

Recall that $N = 30$ and $T = 5$. Exercise: check.

Note: In this case the posterior conditional for α is independent of β; and the posterior conditional for β is independent of α. In other words, the parameters α and β are *a posteriori* independent, conditional on σ^2. This is because we normalise (or simply centre) the explanatory variable, so that $\sum_{j=1}^5 z_j = 0$. This does not hold in general (i.e. for unnormalised explanatory variables).

MCMC algorithm

The posterior conditional distributions are all of standard form so that we can use the Gibbs sampler. In particular we use the following algorithm:

Step 1. Set initial parameter value for α, β and σ^2 denoted by $(\alpha,\beta,\sigma^2)^0=\{\alpha^0,\beta^0,(\sigma^2)^0\}$.

Step 2. Conditional on the current parameter values, $\{\beta^t(\sigma^2)^t\}$, generate a new value for α, from the posterior conditional distribution,

$$\alpha^{t+1}|\boldsymbol{y},\beta^t,(\sigma^2)^t\sim N\left(\frac{NT\bar{y}\sigma_\alpha^2+\mu_\alpha(\sigma^2)^t}{NT\sigma_\alpha^2+(\sigma^2)^t},\frac{(\sigma^2)^t\sigma_\alpha^2}{NT\sigma_\alpha^2+(\sigma^2)^t}\right)$$

Step 3. Conditional on the newly updated parameter value, α_{t+1} and current parameter $(\sigma^2)^t$ generate a new value for β, from the posterior conditional distribution,

$$\beta^{t+1}|\boldsymbol{y},\alpha^{t+1},(\sigma^2)^t\sim N\left(\frac{\sum_{ij}z_jy_{ij}\sigma_\beta^2+\mu_\beta(\sigma^2)^t}{N\sum_j z_j^2\sigma_\beta^2+(\sigma^2)^t},\frac{(\sigma^2)^t\sigma_\beta^2}{N\sum_j z_j^2\sigma_\beta^2+(\sigma^2)^t}\right).$$

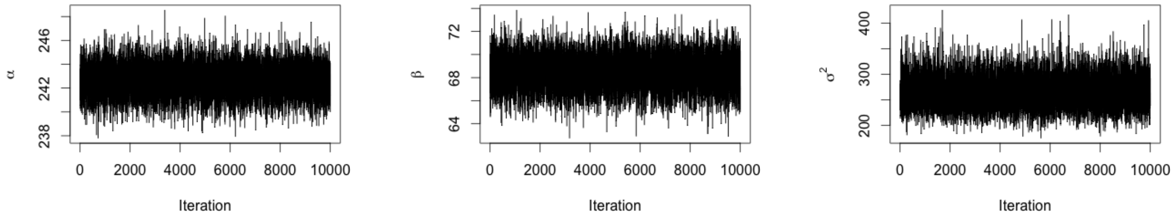
Step 4. Conditional on the newly updated parameter values, α_{t+1} and β^{t+1} generate a new value for σ^2, from the posterior conditional distribution,

$$(\sigma^2)^{t+1}|\boldsymbol{y},\alpha^{t+1},\beta^{t+1}\sim\Gamma^{-1}\left(\frac{NT}{2}+a,\frac{1}{2}\sum_{ij}(y_{ij}-\alpha^{t+1}-\beta^{t+1}z_j)^2+b\right).$$

Step 5. Increase t by one and return to step 2, until T iterations have been performed.

Finally we need to provide the initial parameter values for all the parameters. In this example, we specify initial values: $\alpha^0 = 0$, $\beta^0 = 0$, and $\sigma^2 = 1$.

The simulations are run for 10,000 iterations (this took about 4 seconds on my laptop). The corresponding trace plots of the parameters α, β, σ^2 and τ are provided in Figure 2.15.

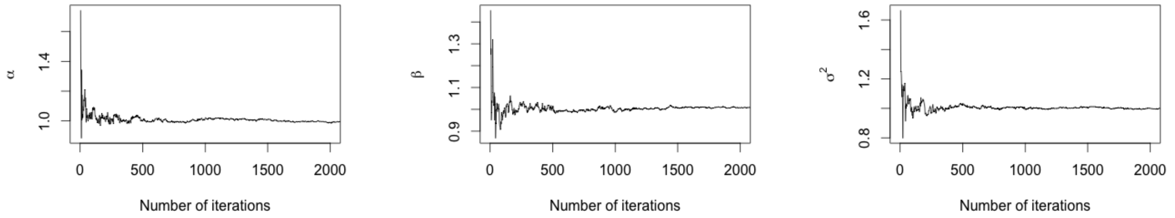


(a) (b) (c)

Figure. 2.15: Raw trace plots for the parameters (a) α, (b) β, and (c) σ^2.

Convergence Diagnostics

There is no real discernible burn-in from the plot. However, in order to check the convergence we run a second chain, with starting values $\alpha^0 = -100$, $\beta^0 = 100$, $\sigma^2 = 0.1$ (in practice we would typically use multiple starting values). The corresponding BGR statistic is calculated. See Figure 2.16 for the BGR statistic plotted against iteration number. Clearly we can see that the BGR quickly converges to a value around 1 for all of the parameters. Thus the burn-in of 1000 iterations seems adequate.



(a) (b) (c)

Figure. 2.16: BGR statistic for the parameters (a) α, (b) β, and (c) σ^2 for two Gibbs samplers.

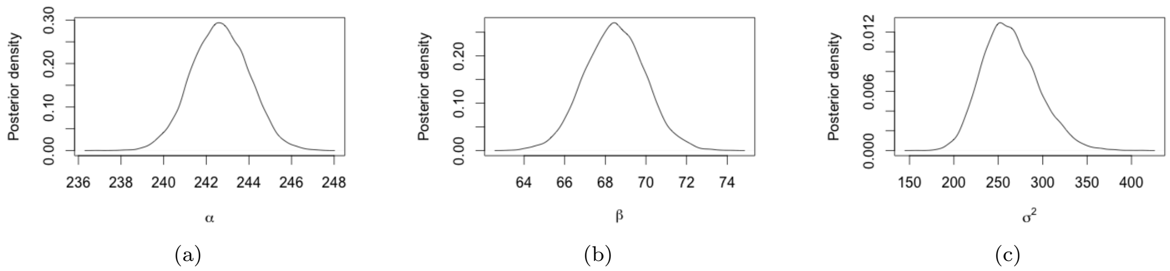
Results

The corresponding (default) summary statistics are given in Table 2.2. The corresponding estimates of the marginal posterior densities for each of the parameters are provided in Figure 2.17.

Parameter	Mean	SD	95% symmetric CI
α	242.6	1.32	(240.1, 245.3)
β	68.5	1.48	(65.5, 71.4)
σ^2	263.7	31.5	(209.9, 331.8)

Table. 2.2: Posterior summary statistics for the rats example with a normal linear regression model for weight.

We note that α can be interpreted as the intercept term (after the weights have been normalised) and β the slope of the linear regression. In particular, since β is clearly positive, there is strong



(a) (b) (c)

Figure. 2.17: Posterior density estimates for the parameters (a) α, (b) β, and (c) σ^2.

evidence that there is a positive relationship between the weight of the rat and time, so that the weight of the rat increases with time (which we assume is linear in nature). In addition we have that the 95% symmetric credible interval for β does not contain the value of 0 which would correspond to no (linear) relationship between weight and time. Thus an *ad-hoc* interpretation of this output might conclude that there is evidence that time is important in this regression (there are more formal methods for model selection but the techniques for calculating these are outside the remit of this course - recall tthe concept of hypothesis testing and Bayes Factors in Section 1.5).

Prior Sensitivity Analysis

We conduct a prior sensitivity analysis by rerunning the MCMC iterations using different priors on each of the parameters. In particular we specify the following priors,

$$\alpha \sim N(0, 10^3); \qquad \beta \sim N(0, 10^3); \qquad \sigma \sim U[0, 100].$$

We have changed the prior distribution on σ^2 by specifying a Uniform prior on σ (as opposed to the an Inverse Gamma prior distribution on σ^2).

Once more convergence is extremely swift and the corresponding posterior summary statistics are provided in Table 2.3

Parameter	Mean	SD	95% symmetric CI
α	242.6	1.3	(240.1, 245.2)
β	68.5	1.49	(65.5, 71.4)
σ^2	266.2	31.5	(210.8, 336.2)

Table. 2.3: Posterior summary statistics for the rats example with a normal linear regression model for weight.

These are again very similar to the previous posterior summary statistics. Thus we would conclude that the posterior distribution is data-driven.

Appendix A

PROBABILITY DISTRIBUTIONS

This appendix gives the form of the pmf/pdf and summary statistics for common distributions, which are frequently used within statistical problems.

A.1 Discrete distributions

Distribution	Parameters	Mass function	Mean and variance
Binomial $\theta \sim Bin(n,p)$	sample size $n \in \mathbb{N}$ $p \in [0,1]$	$f(\theta) = \left(\begin{array}{c} n \\ \theta \end{array}\right) p^\theta (1-p)^{n-\theta}$ $\theta = 0, 1, \ldots, n$	$\mathbb{E}(\theta) = np$ $Var(\theta) = np(1-p)$
Poisson $\theta \sim Poisson(\lambda)$	rate $\lambda > 0$	$f(\theta) = \lambda^\theta \exp(-\lambda)(\theta!)^{-1}$ $\theta = 0, 1, 2, \ldots$	$\mathbb{E}(\theta) = \lambda$ $Var(\theta) = \lambda$
Geometric $\theta \sim Geom(p)$	$p \in [0,1]$	$f(\theta) = p(1-p)^{\theta-1}$ $\theta = 0, 1, \ldots$	$\mathbb{E}(\theta) = 1/p$ $Var(\theta) = (1-p)/p^2$
Negative Binomial $\theta \sim Neg\text{-}Bin(\alpha,\beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$	$f(\theta) = \left(\begin{array}{c} \theta + \alpha - 1 \\ \alpha - 1 \end{array}\right) \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^\theta$ $\theta = 0, 1, 2, \ldots$	$\mathbb{E}(\theta) = \alpha/\beta$ $Var(\theta) = \frac{\alpha}{\beta^2}(\beta+1)$
Multinomial $\boldsymbol{\theta} \sim MN(n, \boldsymbol{p})$	sample size $n \in \mathbb{N}$ $p_i \in [0,1]$; $\sum_{i=1}^k p_i = 1$	$f(\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^k \theta_i!} \prod_{i=1}^k p_i^{\theta_i}$ $\theta_i = 0, 1, \ldots, n$; $\sum_{i=1}^k \theta_i = n$	$\mathbb{E}(\theta_j) = np_j$ $Var(\theta_i) = np_i(1-p_i)$

A.2 Continuous distributions

Distribution	Parameters	Density function	Mean and variance
Uniform $\theta \sim U[a,b]$	$b > a$	$f(\theta) = 1/(b-a)$ $\theta \in [a,b]$	$\mathbb{E}(\theta) = (a+b)/2$ $Var(\theta) = (b-a)^2/12$
Normal $\theta \sim N(\mu, \sigma^2)$	location μ scale $\sigma > 0$	$f(\theta) = \frac{\exp\left(-(\theta-\mu)^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}}$ $\infty < \theta < \infty$	$\mathbb{E}(\theta) = \mu$ $Var(\theta) = \sigma^2$
log Normal $\theta \sim \log N(\mu, \sigma^2)$	μ $\sigma > 0$	$f(\theta) = \frac{\exp\left(-(\log\theta-\mu)^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}\theta}$ $0 \leq \theta < \infty$	$\mathbb{E}(\theta) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ $Var(\theta) = \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)$
Beta $\theta \sim Beta(\alpha, \beta)$	$\alpha > 0$ $\beta > 0$	$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\theta \in [0,1]$	$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$ $Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Exponential $\theta \sim Exp(\lambda)$	$\lambda > 0$	$f(\theta) = \lambda\exp(-\lambda\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = 1/\lambda$ $Var(\theta) = 1/\lambda^2$
Gamma $\theta \sim \Gamma(\alpha, \beta)$	shape $\alpha > 0$ scale $\beta > 0$	$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = \alpha/\beta$ $Var(\theta) = \alpha/\beta^2$
Inverse Gamma $\theta \sim \Gamma^{-1}(\alpha, \beta)$	shape $\alpha > 0$ scale $\beta > 0$	$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\exp(-\beta/\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = \beta/(\alpha-1)$, for $\alpha > 1$ $Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, $\alpha > 2$
Chi-squared $\theta \sim \chi^2_\nu$	df $\nu > 0$ (deg. of freedom)	$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{\frac{\nu}{2}-1}\exp(-\theta/2)$ $\theta > 0$ (same as $\Gamma\left(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right)$)	$\mathbb{E}(\theta) = \nu$ $Var(\theta) = 2\nu$
Inverse Chi-squared $\theta \sim \chi^{-2}_\nu$	df $\nu > 0$ (deg. of freedom)	$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{-\left(\frac{\nu}{2}+1\right)}\exp(-1/2\theta)$ $\theta > 0$ (same as $\Gamma^{-1}\left(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right)$)	$\mathbb{E}(\theta) = \frac{1}{\nu-2}$ $Var(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}$
Dirichlet $\theta \sim Dir(\alpha_1, \ldots, \alpha_k)$	$\alpha_i > 0$; $\alpha_0 \equiv \sum_{i=1}^k \alpha_i$	$f(\theta) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k \theta_i^{\alpha_i-1}$ $\theta_i > 0$; $\sum_{i=1}^k \theta_i = 1$	$\mathbb{E}(\theta_i) = \frac{\alpha_i}{\alpha_0}$ $Var(\theta_i) = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$

Appendix B

DISTRIBUTIONAL COMMANDS IN R

R has a suite of commands related to distributions. These are exceptionally useful for many reasons - within this course this is primarily for calculating probabilities or quantiles of different distributions and simulating from different distributions. We will describe the general structure of commands in relation to the normal distribution, namely, $X \sim N$. There are four basic commands:

- `dnorm` ← probability mass/density function (pmf/pdf);
- `pnorm` ← cumulative distribution function (cdf);
- `qnorm` ← inverse cumulative distribution function;
- `rnorm` ← generates random deviates.

The first letter ("`d`", "`p`", "`q`", "`r`") corresponds to the particular distributional function to be implemented and the latter part of the command to the distribution. Each of these commands have input parameters corresponding to the value(s) associated with what we wish to evaluate or simulate and the associated parameters of the distribution of interest. For example consider $X \sim N(10, 25)$. Note that the associated input parameters for the normal distribution are the mean (μ) and *standard deviation* (σ) (i.e. NOT the variance, σ^2).

To evaluate the pdf at $x = 12$:

```
> dnorm(12,10,5)
[1] 0.07365403
```

To calculate the cdf at $x = 12$ (i.e. $\mathbb{P}(X \leq 12)$:

```
> pnorm(12,10,5)
[1] 0.6554217
```

To output the 2.5%, 50% and 97.5% quantiles (for the 95% symmetric credible interval and median):

```
> qnorm(c(0.025,0.5,0.975),10,5)
[1]  0.2001801 10.0000000 19.7998199
```

To simulate 1000 random deviates and place them in a vector `vec`:

```
> vec <- rnorm(1000,10,5)
```

Recall that the help function can be used to find out more about the commands in `R`.

Similar commands can be used for other distributions and we briefly outline some of these here for other common distributions using the "`d`" command with the others commands following similarly. Use the help function in `R` (or a Google search - which happens to use a Bayesian approach!) to find other other distributions as needed. We provide an example where for the given random variable X the corresponding pdf of the distribution is evaluated at the value `x`.

t distribution

Input parameter - degrees of freedom, e.g. for $X \sim t_{10}$ use `dt(10,x)`.

F distribution

Input parameters - degrees of freedom e.g. for $X \sim F_{1,2}$ use `df(10,1,2)`.

$Beta$ distribution

Input parameters - α and β e.g. for $X \sim Beta(1,2)$ use `dbeta(10,1,2)`.

Γ distribution

Input parameters - shape and scale, e.g. for $X \sim \Gamma(1,2)$ use `dgamma(1,2,x)`.

Note that there is an Γ^{-1} distribution in the `R` package `MCMCpack` (for evaluating the pdf and simulating random deviates; other packages also exist). However recall that if $X \sim \Gamma(\alpha,\beta)$, then this means that $X^{-1} \sim \Gamma^{-1}(\alpha,\beta)$. Thus obtaining probabilities, quantiles and random variables for the Γ^{-1} distribution can be obtained by using this relationship. For example, if $X \sim \Gamma^{-1}(1,2)$ and we want to calculate the $\mathbb{P}(X \geq 1.5)$ we can use that $\mathbb{P}(X \geq 1.5) = \mathbb{P}(X^{-1} \leq 0.75)$ where $X^{-1} \sim \Gamma(1,2)$. In `R`:

```
> pgamma(0.75,1,2)
[1] 0.7768698
```

Similarly to find the 95% symmetric credible interval for $X \sim \Gamma^{-1}(\alpha,\beta)$ we can find the lower and upper 2.5% quantiles for $X^{-1} \sim \Gamma(\alpha,\beta)$ and take the reciprocal in order to obtain the corresponding upper and lower 2.5% quantiles of X. For example for $X \sim \Gamma^{-1}(1,2)$ we initially calculate the quantiles for X^{-1} in `R` using:

```
> qgamma(c(0.025,0.975),1,2)
[1] 0.0126589 1.8444397
```

And then take their reciprocals:

```
> 1/c(0.0126589,1.8444397)
[1] 78.9958053  0.5421701
```

Thus the 95% symmetric credible interval is (0.54, 79.0).

Finally to simulate 1000 random values from $X \sim \Gamma^{-1}(1,2)$, we initially simulate values from $\Gamma(1,2)$ and take their reciprocals.

```
> temp <- rgamma(1000, 1, 2)
> vec <- 1/temp
```

Here `vec` is a vector containing 1000 values random drawn from the $\Gamma^{-1}(1,2)$ distribution.

Note that the above does not work for evaluating the pdf! (You can see this easily by looking at the corresponding pdfs of the Γ and Γ^{-1} distributions).

χ^2 distribution

Input parameter - degrees of freedom, e.g. for $X \sim \chi^2_{10}$ use `dchisq(10,x)`.

Note that there is an inverse χ^2 distribution in the `R` package `geoR`. However recall that if $X \sim \chi^2_\eta$, then $X^{-1} \sim \chi^{-2}_\eta$. Thus obtaining probabilities, quantiles and random variables for the χ^{-2} distribution can be obtained by using this relationship (as described above for the Γ and Γ^{-1} distributions).