SMSTC (2018/19)

Statistics

Lecture 16: MCMC Methods 1: Introduction to Bayesian Methods

Valentin Popov
School of Mathematics and Statistics
University of St Andrews

www.smstc.ac.uk

Contents

16.1 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(s) $\boldsymbol{\theta}$ on which we wish to make inference, and a probability mechanism $f(\boldsymbol{x}|\boldsymbol{\theta})$ which determines the probability of observing different data, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, under different parameter values, $\boldsymbol{\theta}$. In the classical (also called "frequentist") approach, θ is considered to be some fixed, but unknown, constant. Inference is then based on the likelihood $f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{x}$ represents a sample of observations from the population. Thus, the classical approach looks at the distribution of the data given the parameter, to estimate the parameter(s) $\boldsymbol{\theta}$. For example, we may calculate the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$.

Conversely in the Bayesian paradigm, we no longer assume that the parameters have a fixed value, but consider $\boldsymbol{\theta}$ to be a random quantity. We then assume that $\boldsymbol{\theta}$ has some unknown distribution, which we wish to estimate. This distribution is denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, 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 $p(\boldsymbol{\theta})$, which represents our initial beliefs about the distribution of $\boldsymbol{\theta}$ *prior* to observing any data.[a]

In most situations, when we are trying to estimate the parameter(s) $\boldsymbol{\theta}$, we have some knowledge, or preconceptions, about the value of $\boldsymbol{\theta}$ 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 A. Then, let B_1 denote the event that it is a tree, and B_2 that it is a postman. Then, you reject event B_2 and accept event B_1, since,

$$\mathbb{P}(A|B_1) > \mathbb{P}(A|B_2).$$

Thus, we are essentially maximising the likelihood.

However, suppose that we entertain a third possibility, event B_3, that the object is a plastic replica tree. In this case it might well be that $\mathbb{P}(A|B_1) = \mathbb{P}(A|B_3)$, and yet you would still reject this hypothesis in favour of B_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, isolated 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

[a] We do not delve into the comparison of classical and Bayesian statistics at all here; instead see Barnett (2008) (see Bibliography for details).

choosing the point in parameter space that 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.

16.2 Bayes' Theorem

Suppose that we observe data, $\boldsymbol{x}$, and wish to estimate the parameters, $\boldsymbol{\theta}$. Then, Bayes' Theorem states that:

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

Here we have:

- $\pi(\boldsymbol{\theta}|\boldsymbol{x})$: the **posterior** distribution of the parameters, given the observed data $\boldsymbol{x}$;
- $f(\boldsymbol{x}|\boldsymbol{\theta})$: the **likelihood** of the data, given the parameter values;
- $p(\boldsymbol{\theta})$: the **prior** distribution of the parameters, specified *independently* of the data (thus the prior represents our initial beliefs about the parameter values, before observing any data); and
- $f(\boldsymbol{x})$: the normalisation constant (so that the posterior distribution is a valid probability function).

Note that we often express the normalisation constant in the form,

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

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.

Another interpretation of the posterior distribution is the following. Before we observe any data, we have our prior beliefs concerning the parameters (represented by the prior distribution). We then conduct an experiment and observe some data. We update our prior beliefs given the data that we have now observed, by combining the information contained within the data (via the likelihood) with our prior beliefs concerning the parameters, to form our updated beliefs in the form of the posterior distribution of the parameters.

16.3 Posterior distribution

The posterior distribution incorporates all the information concerning the parameter of interest, and so is the most informative description of the parameter. Often (when there are very few parameters – 2 or 3 at most) the full posterior distribution is displayed graphically, in order to illustrate the information in a readily interpretable way. However, in multiple dimensions, the posterior distribution significantly increases in complexity. As a result, often we may only be interested in the marginal posterior distribution of a single parameters, conditional on the observed data. For example, suppose that $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ and that we are only interested in θ_1, then,

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

(Note that this integration is often too complex to do in practice, however, we discuss how we can use computationally intensive methods to obtain an empirical estimate of this distribution later.)

Example

Suppose that we observe data $\boldsymbol{x} = \{x_1, \dots, x_n\}$, 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,

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

Then, since the posterior distribution of $\lambda|\boldsymbol{x}$ is of standard form, by inspection, the corresponding constant of proportionality is equal to,

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

Note that in this example, both the prior and posterior distributions for λ belonged to the same family (i.e. Gamma distribution), irrespective of the data. This is a special case, and when this occurs, the prior distribution is called a *conjugate* prior.

□

Although the complete specification of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ (or a marginal distribution) 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 of the parameters may be given. We shall discuss in more detail a variety of summary statistics that are often used.

16.4 Summary statistics

For simplicity we assume that we are interested in the posterior distribution of a single parameter θ, given the observed data $\boldsymbol{x}$, denoted by $\pi(\theta|\boldsymbol{x})$. Thus, this could be a marginal distribution, or a problem with only a single parameter to be estimated. We discuss in more detail a variety of summary statistics that are often used.

16.4.1 Point estimates

There are a variety of different point estimates that are often used to describe the posterior distribution. In particular, the location of the distribution is often of interest, and so an "average" of the posterior distribution is provided. The most common are the posterior mean and median (though the mode is also sometimes used). (Note that each of these summary estimates have a decision theoretic justification. In particular, they minimise an expected loss function with respect to the posterior distribution, for different loss functions: quadratic loss function (mean); absolute error loss function (median) and zero-one loss function (mode).) Alternatively, the "spread" of the distribution is often summarised via the posterior variance or standard deviation. However, this does not provide any information regarding possible skewness, for example, of the posterior distribution. A more informative description is provided via interval estimates.

16.4.2 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. 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 16.1 *The interval* (a, b) *is defined as an* $100(1-\alpha)\%$ *credible interval if,*

$$\mathbb{P}_\pi(\theta \in [a,b]) = \int_a^b \pi(\theta|\boldsymbol{x})d\theta = 1-\alpha, \qquad 0 \leq \alpha \leq 1.$$

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 16.1.

PSfrag replacements

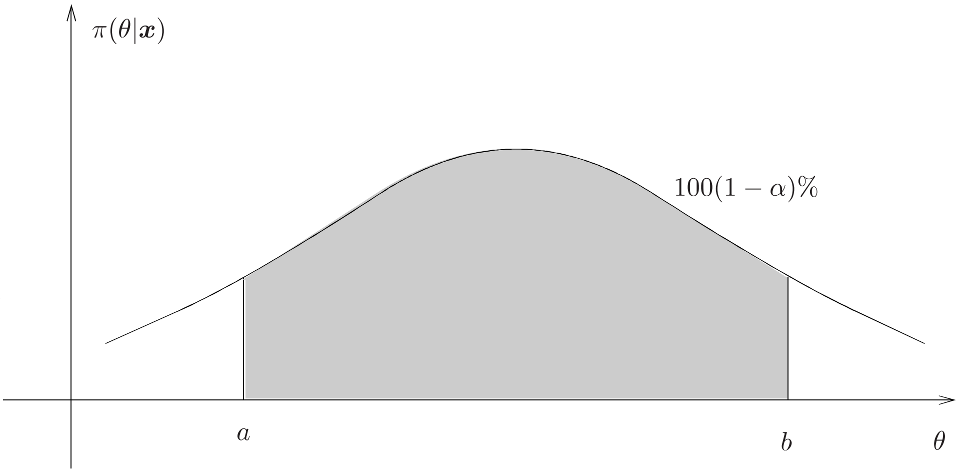


Figure 16.1: Typical $100(1-\alpha)\%$ credible interval.

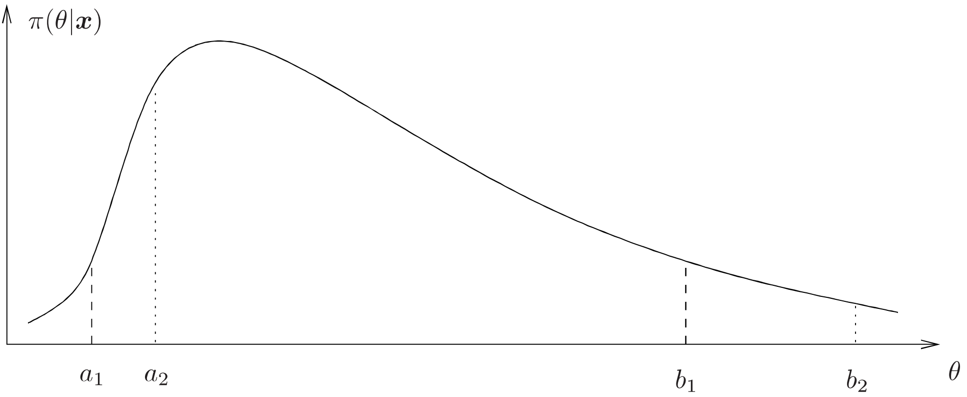
Note that a $100(1-\alpha)\%$ credible interval is not unique, since, in general, there will be many choices of a and b, such that, $\int_a^b \pi(\theta|\boldsymbol{x})d\theta = 1-\alpha$. For example, see Figure 16.2, where we have two $100(1-\alpha)\%$ credible intervals, $[a_1, b_1]$ and $[a_2, b_2]$.

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 16.2, 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



PSfrag replacements

Figure 16.2: Two alternative $100(1-\alpha)\%$ credible intervals.

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.

Definition 16.2 *The interval* $[a, b]$ *is a* $100(1-\alpha)\%$ ***highest posterior density interval*** *(HPDI) if:*

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

(ii) 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 $g(\theta)$.

Note that the ideas presented for the single parameter case are all directly generalisable to the multi-parameter case. For example, point estimates can be obtained for correlations between parameters; posterior credible intervals 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.

16.5 Prior specification

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. 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, how should we assign the prior $p(\theta)$? Let us be clear from the beginning - there is no such thing as the *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.

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 16.3 *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,

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

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

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

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

Consider the particular case where we specify our prior beliefs, such that $\alpha = \beta = 1$, i.e. a $U[0, 1]$ prior on p. Then, Figure 16.3 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 16.3 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. 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,

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

PSfrag replacements

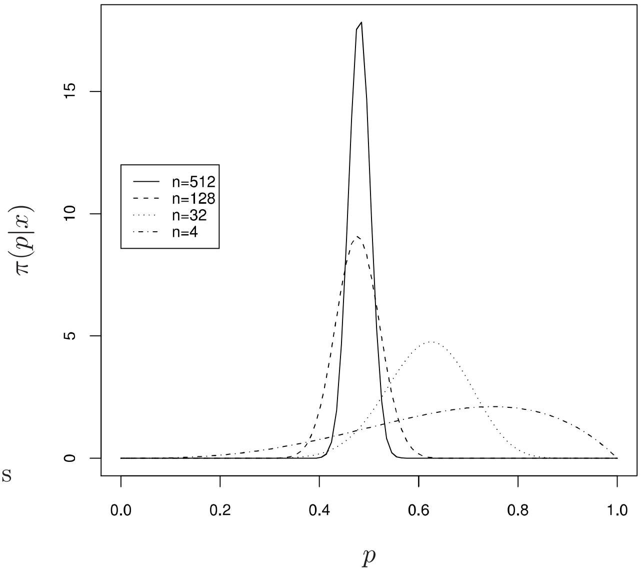


Figure 16.3: The posterior distribution of p, the probability of obtaining a head when tossing a coin, for (i) $n = 4, x = 3$, (ii) $n = 32, x = 20$, (iii) $n = 128, x = 61$ and (iv) $n = 512, x = 248$, with a $U[0, 1]$ prior on p.

where $w = n/(n + \alpha + \beta)$. In other words, the posterior mean 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 we would use if we had no data; 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 location of the distribution, 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 16.3, 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,

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

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.

□

Notes

16–1. 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. In many cases we may not have any strong prior information relating to the parameters of interest. Typically, in these circumstances, we would wish to specify a prior that did not strongly influence the posterior distribution, resulting in a "data-driven" distribution.

16–2. In the multi-parameter case, we define the joint prior distribution over the set of parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$, denoted by $p(\boldsymbol{\theta})$. However, (particularly when there is no strong prior information), we may specify independent prior distributions on each of the parameters. The prior can then be expressed in the form,

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

16–3. Informative priors aim to reflect information available to the analyst that is gained independently of the data being studied. The prior information then needs to be translated into the form of a prior distribution. Typically, a prior elicitation process is used which involves choosing a suitable family of prior distributions for each parameter and then attempting to find parameters for that prior that accurately reflects the information available. This is often an iterative process. For example, the analyst might begin by getting some idea of the range of plausible parameter values and use this to get suitable prior parameters. Plots can be made and summary statistics calculated which can then be shown to the expert to see if they are consistent with their beliefs. Any mismatches can be used to alter the prior parameters and the process is repeated.

A prior distribution that does not contain strong information relating to the parameter(s) is often referred to an non-informative or vague. However, the specification of such a prior is not necessarily always as straightforward as it would first appear.

16.5.1 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, as in the case of the HPDI, 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 $\theta \in [0, 1]$. Then 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.

16.5.2 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 Information, and is defined to be,

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

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.

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

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

where $X \sim f$. This means that we only need to consider a single observed datum when calculating Jeffrey' prior, simplifying the algebra.

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

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

Jeffreys' prior is then specified as,

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

where $|\cdot|$ denotes the determinant of the matrix.

Alternative vague or non-informative prior distributions often have a reasonable mean for the distribution, but with a large variance parameter.

16.5.3 Hierarchical priors

An alternative approach in the specifying of a vague prior is to adopt a hierarchical prior structure. In this case, the prior parameters are themselves assumed to have a specified (hyper-)prior distribution. For example, suppose that we wish to specify a prior on the parameter α. Using the idea suggested above, we may specify a prior with reasonable mean, but with a large variance values, such as $\alpha \sim N(0, 1000)$. Alternatively, we could specify a hierarchical prior, where we might set $\alpha \sim N(0, \sigma^2)$ where $\sigma^2 \sim \Gamma^{-1}(a, b)$ (this is a conjugate prior for σ^2). Specifying such a prior dilutes the influence on the posterior of any prior assumptions made and essentially creates random effects of the model parameters (see Section 20.3). The influence of the prior can always be assessed via a sensitivity analysis: 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. This is essentially a "prior sensitivity analysis".

16.5.4 Prior sensitivity

A prior sensitivity analysis should always be performed within a Bayesian analysis (irrespective of whether informative or vague priors have been specified). In the simplest form this essentially means considering a variety of different priors, and the corresponding impact on the posterior distribution of the parameter(s).

Example

Reconsider the previous 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 16.4. 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.

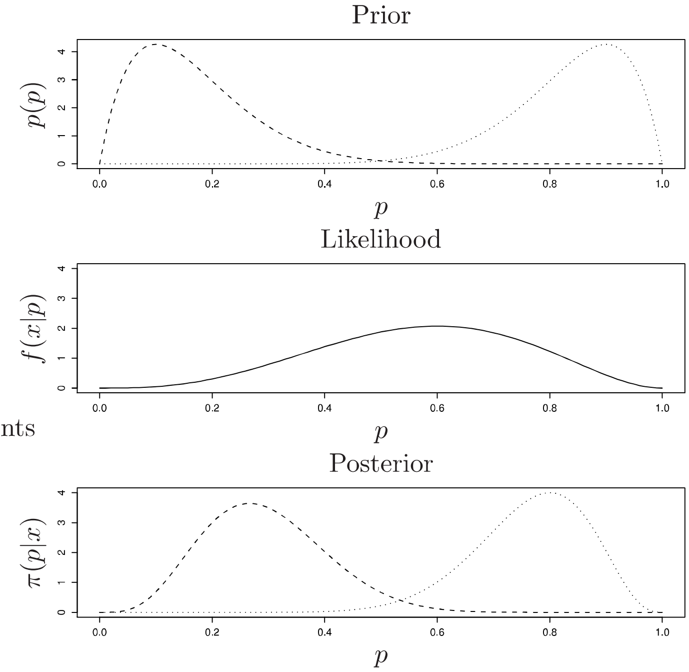
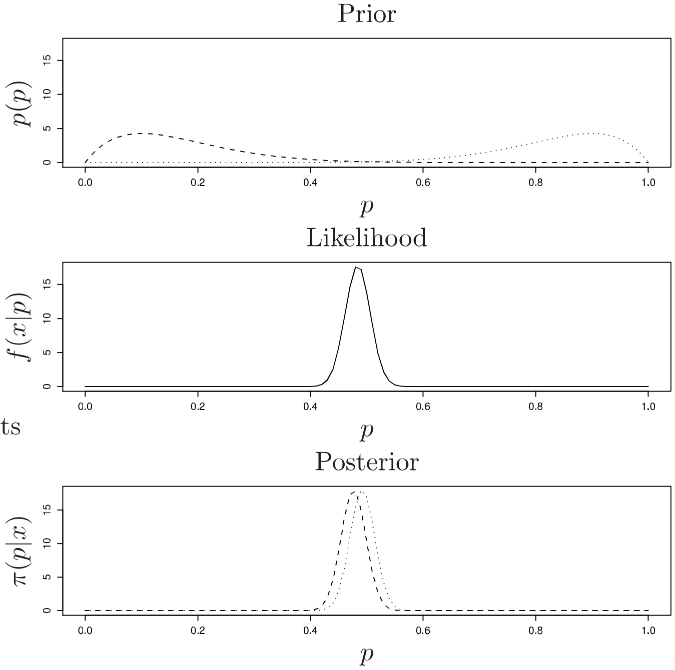


Figure 16.4: The prior, likelihood and posterior distribution of p when $n = 5$, for prior 1: $p \sim Beta(2, 10)$ (dashed line) and prior 2: $p \sim Beta(10, 2)$ (dotted line).

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 16.5. 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.

We emphasise that in typical Bayesian analyses, 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 called a prior sensitivity analysis.



PSfrag replacements

Figure 16.5: 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).

16.6 Appendix - Common distributions

This appendix gives the form of the pdf and summary statistics for common distributions.

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 $\theta \sim Geom(p)$	$p \in [0, 1]$	$f(\theta) = p(1-p)^{\theta-1}$ $\theta = 1, 2, \ldots$	$\mathbb{E}(\theta) = 1/p$ $Var(\theta) = (1-p)/p^2$
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 $n \in \mathbb{N}$ $p_i \in [0, 1]$; $\sum_{i=1}^k p_i = 1$	$f(\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^k \theta_i!} \prod_{i=1}^k p_i^{\theta_i}$ $\theta_i = 0, 1, \ldots, n$; $\sum_{i=1}^k \theta_i = n$	$\mathbb{E}(\theta_j) = np_j$ $Var(\theta_i) = np_i(1-p_i)$

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 $\sigma > 0$	$f(\theta) = \frac{\exp\left(-(\theta-\mu)^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}}$ $\infty < \theta < \infty$	$\mathbb{E}(\theta) = \mu$ $Var(\theta) = \sigma^2$
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 $\alpha > 0$ rate $\beta > 0$	$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = \alpha/\beta$ $Var(\theta) = \alpha/\beta^2$
Inverse Gamma $\theta \sim \Gamma^{-1}(\alpha, \beta)$	shape $\alpha > 0$ rate $\beta > 0$	$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\exp(-\beta/\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = \beta/(\alpha-1)$, for $\alpha > 1$ $Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, $\alpha > 2$
Chi-squared $\theta \sim \chi^2_\nu$	dof $\nu > 0$	$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{\frac{\nu}{2}-1}\exp(-\theta/2)$ $\theta > 0$ (same as $\Gamma\left(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right)$)	$\mathbb{E}(\theta) = \nu$ $Var(\theta) = 2\nu$
Inverse Chi-squared $\theta \sim \chi^{-2}_\nu$	dof $\nu > 0$	$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{-\left(\frac{\nu}{2}+1\right)}\exp(-1/2\theta)$ $\theta > 0$ (same as $\Gamma^{-1}\left(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right)$)	$\mathbb{E}(\theta) = \frac{1}{\nu-2}$ $Var(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}$
Student-t $\theta \sim t_\nu(\mu, \sigma^2)$	dof $\nu > 0$ location μ scale $\sigma^2 > 0$	$f(\theta) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}}\left(1 + \frac{(\theta-\mu)^2}{\nu\sigma^2}\right)^{-\frac{(\nu+1)}{2}}$ $\infty < \theta < \infty$	$\mathbb{E}(\theta) = \mu$ $Var(\theta) = \frac{\nu\sigma^2}{(\nu-2)}$
Dirichlet $\theta \sim Dir(\alpha_1, \ldots, \alpha_k)$	$\alpha_i > 0$; $\alpha_0 \equiv \sum_{i=1}^k \alpha_i$	$f(\theta) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k \theta_i^{\alpha_i-1}$ $\theta_i > 0$; $\sum_{i=1}^k \theta_i = 1$	$\mathbb{E}(\theta_i) = \frac{\alpha_i}{\alpha_0}$ $Var(\theta_i) = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$

(dof = degrees of freedom)