MA40189: Topics in Bayesian statistics

Simon Shaw
s.shaw@bath.ac.uk
2018/19 Semester II

Contents

List of Figures

Introduction

Consider a problem where we wish to make inferences about a parameter θ given data x. In a classical setting the data is treated as if it is random, even after it has been observed, and the parameter is viewed as a fixed unknown constant. Consequently, no probability distribution can be attached to the parameter. Conversely in a Bayesian approach parameters, having not been observed, are treated as random and thus possess a probability distribution whilst the data, having been observed, is treated as being fixed.

Example 1 *Suppose that we perform n independent Bernoulli trials in which we observe x, the number of times an event occurs. We are interested in making inferences about θ, the probability of the event occurring in a single trial. Let's consider the classical approach to this problem.*

Prior to observing the data, the probability of observing x was

$$P(X = x \,|\, \theta) \quad = \quad \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{1}$$

This is a function of the (future) x, assuming that θ is known. If we know x but don't know θ we could treat (1) as a function of θ, $L(\theta)$, the likelihood function. We then choose the value which maximises this likelihood. The maximum likelihood estimate is $\frac{x}{n}$ with corresponding estimator $\frac{X}{n}$.

In the general case, the classical approach uses an estimate $T(x)$ for θ. Justifications for the estimate depend upon the properties of the corresponding estimator $T(X)$ (bias, consistency, ...) using its sampling distribution (given θ). That is, we treat the data as being random even though it is known! Such an approach can lead to nonsensical answers.

Example 2 *Suppose in the Bernoulli trials of Example 1 we wish to estimate θ^2. The maximum likelihood estimator*[1] *is $\left(\frac{X}{n}\right)^2$. However this is a biased estimator as*

$$\begin{aligned} E(X^2 \,|\, \theta) &= Var(X \,|\, \theta) + E^2(X \,|\, \theta) \\ &= n\theta(1-\theta) + n^2\theta^2 \\ &= n\theta + n(n-1)\theta^2. \end{aligned} \tag{2}$$

[1] Remember the invariance properties of maximum likelihood estimators. If $T(X)$ is the maximum likelihood estimator of θ then the maximum likelihood of $g(\theta)$, a function of θ, is $g(T(X))$.

Noting that $E(X \mid \theta) = n\theta$ *then (2) may be rearranged as*

$$E(X^2 \mid \theta) - E(X \mid \theta) \;=\; n(n-1)\theta^2 \quad \Rightarrow \quad E\{X(X-1) \mid \theta\} \;=\; n(n-1)\theta^2.$$

Thus, $\frac{X(X-1)}{n(n-1)}$ *is an unbiased estimator of* θ^2*. Suppose we observe* $x = 1$*. Then our estimate of* θ^2 *is 0: we estimate a chance as zero even though the event has occurred!*

Example 3 *Now let's consider two different experiments, both modelled as Bernoulli trials.*

1. *Toss a coin* n *times and observe* x *heads. Parameter* θ_c *represents the probability of tossing a head on a single trial.*

2. *Toss a drawing pin* n *times and observe that the pin lands facing upwards on* x *occasions.*

The maximum likelihood estimates for θ_c *and* θ_p *are identical and share the same properties. Is this sensible?*

I, and perhaps you do too, have lots of experience of tossing coins and these are well known to have propensities close to $\frac{1}{2}$*. Thus, even before I toss the coin on these* n *occasions, I have some knowledge about* θ_c*. (At the very least I can say something about where I think it will be and how confident I am in this location which could be viewed as specifying a mean and a variance for* θ_c*.) Equally, I have little knowledge about the tossing propensities of drawing pins - I don't really know much about* θ_p*. Shouldn't I take these differences into account somehow? The classical approach provides no scope for this as* θ_c *and* θ_p *are both unknown constants. In a Bayesian analysis*[2] *we can reflect our* ***prior*** *knowledge about* θ_c *and* θ_p *by specifying probability distributions for* θ_c *and* θ_p*.*

Let's think back to maximum likelihood estimation. We ask

> "what value of θ makes the data most likely to occur?"

Isn't this the wrong way around, what we are really interested in is

> "what value of θ is most likely given the data?"

In a classical analysis this question makes no sense. However, it can be answered in a Bayesian context. We specify a prior distribution $f(\theta)$ for θ and combine this with the likelihood $f(x \mid \theta)$ to obtain the posterior distribution for θ given x using Bayes' theorem,

$$\begin{aligned} f(\theta \mid x) &= \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta)\, d\theta} \\ &\propto f(x \mid \theta) f(\theta). \end{aligned}$$

Bayesian analysis is concerned with the distributions of θ and how they are changed in the light of new information (typically data). The distribution $f(x \mid \theta)$ is irrelevant to the Bayesian after X has been observed yet to the classicist it is the only distribution they can work with.

[2] See Example 5.

1 The Bayesian method

We shall adopt the notation $f(\cdot)$ (some authors use $p(\cdot)$) to represent the density function, irrespective of whether the random variable over which the distribution is specified is continuous or discrete. In general notation we shall make no distinction as to whether random variables are univariate or multivariate. In specific examples, if necessary, we shall make the dimension explicit.

1.1 Bayes' theorem

Let X and Y be random variables with joint density function $f(x, y)$. The **marginal distribution** of Y, $f(y)$, is the joint density function averaged over all possible values of X,

$$f(y) \quad = \quad \int_X f(x, y)\, dx. \tag{1.1}$$

For example, if Y is univariate and $X = (X_1, X_2)$ where X_1 and X_2 are univariate then

$$f(y) \quad = \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, y)\, dx_1 dx_2.$$

The **conditional distribution** of Y given $X = x$ is

$$f(y \,|\, x) \quad = \quad \frac{f(x, y)}{f(x)} \tag{1.2}$$

so that by substituting (1.2) into (1.1) we have

$$f(y) \quad = \quad \int_X f(y \,|\, x) f(x)\, dx.$$

which is often known as the **theory of total probability**. X and Y are independent if and only if

$$f(x, y) \quad = \quad f(x) f(y). \tag{1.3}$$

Substituting (1.2) into (1.3) we see that an equivalent result is that

$$f(y \,|\, x) \quad = \quad f(y)$$

so that independence reflects the notion that learning the outcome of X gives us no information about the distribution of Y (and vice versa). If Z is a third random variable then X and Y are **conditionally independent** given Z if and only if

$$f(x, y \,|\, z) \;=\; f(x \,|\, z) f(y \,|\, z). \tag{1.4}$$

Note that

$$\begin{aligned} f(y \,|\, x, z) &= \frac{f(x, y, z)}{f(x, z)} \\ &= \frac{f(x, y \,|\, z) f(z)}{f(x \,|\, z) f(z)} \\ &= \frac{f(x, y \,|\, z)}{f(x \,|\, z)} \end{aligned} \tag{1.5}$$

so, by substituting (1.4) into (1.5), an equivalent result to (1.4) is that

$$f(y \,|\, x, z) \;=\; f(y \,|\, z).$$

Thus, conditional independence reflects the notion that having observed Z then there is no further information about Y that can be gained by additionally observing X (and vice versa for X and Y).

Bayes' theorem[1] states that, for $f(x) > 0$,

$$\begin{aligned} f(y \,|\, x) &= \frac{f(x \,|\, y) f(y)}{f(x)} \\ &= \frac{f(x \,|\, y) f(y)}{\int_Y f(x \,|\, y) f(y)\, dy}. \end{aligned} \tag{1.6}$$

1.2 Bayes' theorem for parametric inference

Consider a general problem in which we have data x and require inference about a parameter θ. In a Bayesian analysis θ is unknown and viewed as a random variable. Thus, it possesses a density function $f(\theta)$. From Bayes' theorem[2], (1.6), we have

$$\begin{aligned} f(\theta \,|\, x) &= \frac{f(x \,|\, \theta) f(\theta)}{f(x)} \\ &\propto f(x \,|\, \theta) f(\theta). \end{aligned} \tag{1.7}$$

Colloquially (1.7) is

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

[1] Bayes' theorem is named after Thomas Bayes (c1701-1761).

[2] Bayes' theorem holds for any random variables. In a Bayesian analysis, θ is a random variable and so Bayes' theorem can be utilised for probability distributions concerning θ.

Most commonly, both the parameter θ and data x are continuous. There are cases when θ is continuous and x is discrete[3]. In exceptional cases θ could be discrete.

The Bayesian method comprises of the following principle steps

1. **Prior**

   Obtain the prior density $f(\theta)$ which expresses our knowledge about θ prior to observing the data.

2. **Likelihood**

   Obtain the likelihood function $f(x \mid \theta)$. This step simply describes the process giving rise to the data x in terms of θ.

3. **Posterior**

   Apply Bayes' theorem to derive posterior density $f(\theta \mid x)$ which expresses all that is known about θ after observing the data.

4. **Inference**

   Derive appropriate inference statements from the posterior distribution e.g. point estimates, interval estimates, probabilities of specified hypotheses.

1.3 Sequential data updates

It is important to note that the Bayesian method can be used sequentially. Suppose we have two sources of data x and y. Then our posterior for θ is

$$f(\theta \mid x, y) \quad \propto \quad f(x, y \mid \theta) f(\theta). \tag{1.8}$$

Now

$$\begin{aligned} f(x, y \mid \theta) f(\theta) \;=\; f(x, y, \theta) \;&=\; f(y \mid x, \theta) f(x, \theta) \\ &=\; f(y \mid x, \theta) f(\theta \mid x) f(x) \\ &\propto\; f(y \mid x, \theta) f(\theta \mid x) \end{aligned} \tag{1.9}$$

Substituting (1.9) into (1.8) we have

$$f(\theta \mid x, y) \quad \propto \quad f(y \mid x, \theta) f(\theta \mid x). \tag{1.10}$$

We can update first by x and then by y. Note that, in this case, $f(\theta \mid x)$ assumes the role of the prior (given x) and $f(y \mid x, \theta)$ the likelihood (given x).

[3] For example, the Bernoulli trials case considered in Example 1.

1.4 Conjugate Bayesian updates

Example 4 ***Beta-Binomial.*** *Suppose that* $X \mid \theta \sim Bin(n, \theta)$. *We specify a prior distribution for* θ *and consider* $\theta \sim Beta(\alpha, \beta)$ *for* $\alpha, \beta > 0$ *known*[4]. *Thus, for* $0 \leq \theta \leq 1$ *we have*

$$f(\theta) \quad = \quad \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{1.11}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ *is the beta function and* $E(\theta) = \frac{\alpha}{\alpha+\beta}$. *Recall that as*

$$\int_0^1 f(\theta)\, d\theta \quad = \quad 1$$

then

$$B(\alpha, \beta) \quad = \quad \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta. \tag{1.12}$$

Using Bayes' theorem, (1.7), the posterior is

$$\begin{aligned} f(\theta \mid x) \;\propto\; f(x \mid \theta) f(\theta) \quad &= \quad \binom{n}{x} \theta^x (1-\theta)^{n-x} \times \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \quad \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \quad \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}. \end{aligned} \tag{1.13}$$

So, $f(\theta \mid x) = c\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$ *for some constant* c *not involving* θ. *Now*

$$\int_0^1 f(\theta \mid x)\, d\theta \;=\; 1 \quad \Rightarrow \quad c^{-1} \;=\; \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, d\theta.$$

Notice that from (1.12) we can evaluate this integral so that

$$c^{-1} \;=\; \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, d\theta \;=\; B(\alpha+x, \beta+n-x)$$

whence

$$f(\theta \mid x) \quad = \quad \frac{1}{B(\alpha+x, \beta+n-x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$$

i.e. $\theta \mid x \sim Beta(\alpha+x, \beta+n-x)$.

Notice the tractability of this update: the prior and posterior distribution are both from the same family of distributions, in this case the Beta family. This is an example of conjugacy. The update is simple to perform: the number of successes observed, x, is added to α whilst the number of failures observed, $n-x$, is added to β.

[4] It is possible to consider α and/or β to be random variables and then specify prior distributions for them. This would be an example of a hierarchical Bayesian model.

Definition 1 *(Conjugate family)*
A class Π *of prior distributions is said to form a conjugate family with respect to a likelihood* $f(x \mid \theta)$ *if the posterior density is in the class of* Π *for all* x *whenever the prior density is in* Π*.*

In Example 4 we showed that, with respect to the Binomial likelihood, the Beta distribution is a conjugate family.

One major advantage of Example 4, and conjugacy in general, was that it was straightforward to calculate the constant of proportionality. In a Bayesian analysis we have $f(\theta \mid x) \propto f(x \mid \theta) f(\theta)$ so that $f(\theta \mid x) = c f(x \mid \theta) f(\theta)$ where c is a constant not involving θ. As $f(\theta \mid x)$ is a density function and thus integrates to unity we have

$$c^{-1} = \int_{\theta} f(x \mid \theta) f(\theta)\, d\theta.$$

In practice, it is not always straightforward to compute this integral and no closed form for it may exist.

We will now explore the effect of the prior and the likelihood on the posterior. To ease the comparison we will view the likelihood as a function of θ and consider the standardised likelihood $\frac{f(x \mid \theta)}{\int_{\theta} f(x \mid \theta)\, d\theta}$, that is scaled so that the area under the curve is unity. Note that this is only appropriate when the integral on the denominator is finite - there are instances when this is not the case. We shall use the Beta distribution as an example. The Beta distribution is an extremely flexible distribution on $(0, 1)$ and, as Figure 1.1 shows, careful choice of the parameters α, β can be used to create a wide variety of shapes for the prior density.

Example 5 *Recall Example 3. We consider the two experiments.*

1. *Toss a coin* n *times and observe* x *heads. Parameter* θ_c *represents the probability of tossing a head on a single trial.*

2. *Toss a drawing pin* n *times and observe that the pin lands facing upwards on* x *occasions.*

In both cases we have a Binomial likelihood and I judge that a Beta distribution can be used to model my prior beliefs about both θ_c *and* θ_p*. I have considerably more prior knowledge about* θ_c *than* θ_p*. In particular, I am confident that* $E(\theta_c) = \frac{1}{2}$ *and that the distribution is pretty concentrated around this point, so that the variance is small. If I model my beliefs about* θ_c *with a Beta distribution then* $E(\theta_c) = \frac{1}{2}$ *implies that I must take* $\alpha = \beta$*. I judge that the* $Beta(20, 20)$ *distribution adequately models my prior knowledge about* θ_c*. I have considerably less knowledge about* θ_p *but I have no reason to think the process is unfair so take* $E(\theta_p) = \frac{1}{2}$*. However, I reflect my lack of knowledge about* θ_p*, compared certainly to* θ_c*, with a large variance so that the distribution is less concentrated around the mean. I judge that the* $Beta(2, 2)$ *distribution adequately models my prior knowledge about* θ_p*.*

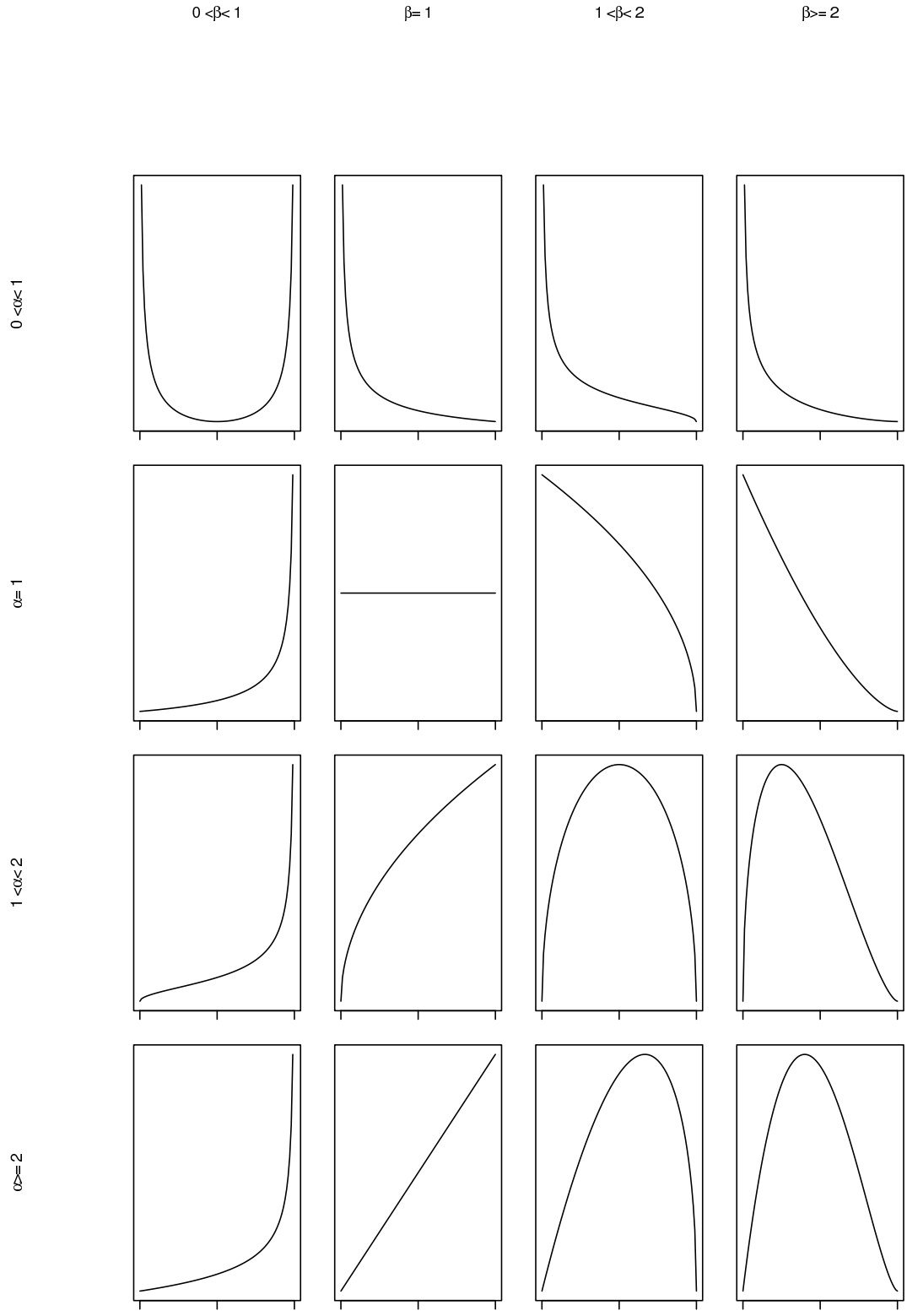
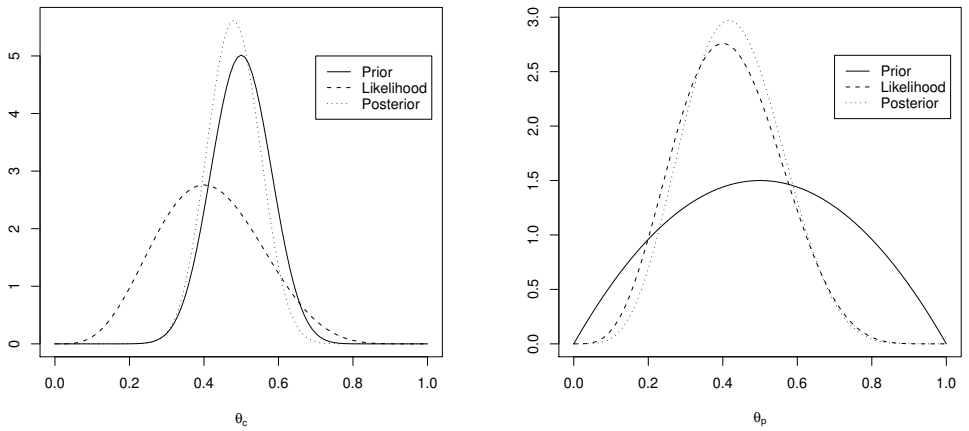


Figure 1.1: A plot of the Beta distribution, $Beta(\alpha, \beta)$, for varying choices of the parameters α, β



(a) Tossing coins. Prior $\theta_c \sim Beta(20, 20)$, posterior $\theta_c \,|\, x \sim Beta(24, 26)$.

(b) Tossing drawing pins. Prior $\theta_p \sim Beta(2, 2)$, posterior $\theta_p \,|\, x \sim Beta(6, 8)$.

Figure 1.2: Prior to posterior for Bernoulli trials model, $n = 10$, $x = 4$. The likelihoods are standardised.

Suppose that, in each case, I perform $n = 10$ tosses and observe $x = 4$ successes. We may use Example 4 to compute the posterior for θ_c and for θ_p.

	Prior	*Likelihood*	*Posterior*
θ_c	$Beta(20, 20)$	$x = 4, n = 10$	$Beta(24, 26)$
θ_p	$Beta(2, 2)$	$x = 4, n = 10$	$Beta(6, 8)$

In Figure 1.2 we plot the prior density, standardised likelihood and posterior density for both θ_c and θ_p. Notice how in Figure 1.2(a), the strong prior information is reflected by the posterior density closely following the prior density. The mode of the posterior is shifted to the left of the prior, towards the mode of the likelihood,reflecting that the maximum likelihood estimate ($\frac{4}{10}$) was smaller than the prior suggested. The posterior density is (slightly) taller and narrower than the prior reflecting the new information about θ_c from the data. In Figure 1.2(b), the weaker prior information can be seen by the fact that the posterior density now much more follows the (standardised) likelihood - most of our posterior knowledge about θ_p is coming from the data. Notice how the posterior density is much taller and narrower than the prior reflecting how the data resolves a lot of the initial uncertainty about θ_p.

Notice that in our calculation of $f(\theta \,|\, x)$, see (1.13), we were only interested in quantities proportional to $f(\theta \,|\, x)$ and we identified that $\theta \,|\, x$ was a Beta distribution by recognising the form $\theta^{\text{“}\alpha\text{”}-1}(1-\theta)^{\text{“}\beta\text{”}-1}$.

Definition 2 *(Kernel of a density)*
For a random variable X with density function $f(x)$ if $f(x)$ can be expressed in the form

$cq(x)$ where c is a constant, not depending upon x, then any such $q(x)$ is a kernel of the density $f(x)$.

Example 6 *If $\theta \sim Beta(\alpha, \beta)$ then $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ is a kernel of the $Beta(\alpha, \beta)$ distribution.*

Spotting kernels of distributions can be very useful in computing posterior distributions.

Example 7 *In Example 4 we can find the posterior distribution $\theta \,|\, x$ by observing that (1.13) is a kernel of the $Beta(\alpha + x, \beta + n - x)$ distribution.*

Example 8 *Let X be normally distribution with unknown mean θ and known variance σ^2. It is judged that $\theta \sim N(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are known. For the prior we have*

$$\begin{aligned} f(\theta) &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} \qquad (1.14) \\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\mu_0\theta)\right\} \qquad (1.15) \end{aligned}$$

where (1.14) and (1.15) are both kernels of the normal distribution. For the likelihood we have

$$\begin{aligned} f(x \,|\, \theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \theta)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\theta^2 - 2x\theta)\right\} \qquad (1.16) \end{aligned}$$

where, since $f(x \,|\, \theta)$ represents the likelihood, in (1.16) we are only interested in θ. For the posterior, from (1.15) and (1.16), we have

$$\begin{aligned} f(\theta \,|\, x) &\propto \exp\left\{-\frac{1}{2\sigma^2}(\theta^2 - 2x\theta)\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\mu_0\theta)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left[\theta^2 - 2\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\theta\right]\right\} \qquad (1.17) \end{aligned}$$

We recognise (1.17) as a kernel of a normal distribution so that $\theta \,|\, x \sim N(\mu_1, \sigma_1^2)$ where

$$\begin{aligned} \sigma_1^2 &= \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \qquad (1.18) \\ \mu_1 &= \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \qquad (1.19) \end{aligned}$$

Notice that we can write (1.18) as

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}$$

so that the posterior precision[5] *is the sum of the data precision and the prior precision. From (1.19) we observe that the posterior mean is a weighted average of the data and the prior mean, weighted according to the corresponding precisions. The posterior mean will be closest to whichever of the prior and likelihood has the stronger information. For example, if the prior information is very weak, expressed by large* σ_0^2 *and thus small precision, then* μ_1 *will be close to the data* x*.*

We note that with respect to a normal likelihood (with known variance), the normal distribution is a conjugate family.

1.5 Using the posterior for inference

Our posterior beliefs are captured by a whole distribution $f(\theta \,|\, x)$. Typically we want to summarise this distribution. We might use

1. Graphs and plots of the shape of the distribution.
2. Measures of location such as the mean, median and mode.
3. Measures of dispersion such as the variance, quantiles e.g. quartiles, interquartile range.
4. A region that captures most of the values of θ.
5. Other relevant summaries e.g. the posterior probability that θ is greater than some value.

We shall focus upon the fourth of these points. In the discussion that follows we shall assume that θ is univariate: the results and definitions simply generalise for the case when θ is multivariate.

1.5.1 Credible intervals and highest density regions

Credible intervals (or posterior intervals) are the Bayesian analogue of confidence intervals. A $100(1-\alpha)\%$ confidence interval for a parameter θ is interpreted as meaning that with a large number of repeated samples[6], $100(1-\alpha)\%$ of the corresponding confidence intervals would contain the true value of θ. A $100(1-\alpha)\%$ credible interval is an actual interval that contains the parameter θ with probability $1-\alpha$.

[5] The precision is the reciprocal of the variance. The lower the variance the higher the precision.

[6] Repeated sampling is a cornerstone of the classical statistics. There are a number of issues with this e.g. the long-run repetitions are hypothetical, what do we do when we can only make a finite number of realisations or an event occurs once and once only.

Definition 3 *(Credible interval[7])*
A $100(1-\alpha)\%$ *credible interval* (θ_L, θ_U) *is an interval within which* $100(1-\alpha)\%$ *of the posterior distribution lies,*

$$P(\theta_L < \theta < \theta_U \,|\, x) \;=\; \int_{\theta_L}^{\theta_U} f(\theta \,|\, x)\, d\theta \;=\; 1-\alpha.$$

Notice that there are infinitely many such intervals. Typically we fix the interval by specifying the probability in each tail. For example, we frequently take

$$P(\theta < \theta_L \,|\, x) \;=\; \frac{\alpha}{2} \;=\; P(\theta > \theta_U \,|\, x).$$

We **can** explicitly state that there is a $100(1-\alpha)\%$ probability that θ is between θ_L and θ_U. Observe that we can construct similar intervals for any (univariate) random variable, in particular for the prior. In such a case the interval is often termed a prior credible interval.

Example 9 *Recall Example 5. We construct 95% prior credible intervals and credible intervals for the two cases of tossing coins and drawing pins. We elect to place 2.5% in each tail. The intervals may be found using R and the command* `qbeta(0.025, a, b)` *for* θ_L *and for* θ_U `qbeta(0.975, a, b)`.

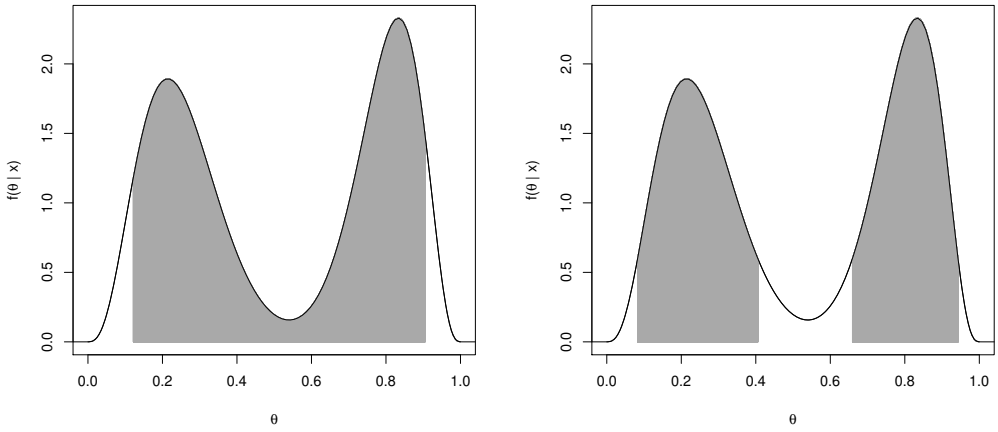
	Prior	*Prior credible*	*Posterior*	*(Posterior) credible*
θ_c	$Beta(20,20)$	$(0.3478, 0.6522)$	$Beta(24,26)$	$(0.3442, 0.6173)$
θ_p	$Beta(2,2)$	$(0.0950, 0.9057)$	$Beta(6,8)$	$(0.1922, 0.6842)$

Notice that the prior credible interval and (posterior) credible interval for θ_c *cover a similar location: this is not a surprise as, from Figure 1.2(a), we observed that the prior and posterior distributions for* θ_c *were very similar with the latter shifted slightly to the left in the direction of the likelihood. This feature is reflected in the interval* $(0.3442, 0.6173)$ *covering the left hand side of* $(0.3478, 0.6522)$*. The width of the credible interval is also narrower: 0.2731 compared to 0.3044 for the prior credible interval. This is an illustration of the data reducing our uncertainty about* θ_c.[8]

For θ_p *the prior credible interval has a width of 0.8107 reflecting the weak prior knowledge. Observation of the data shrinks the width of this interval considerably to 0.4920 which is still quite wide (wider than the initial prior credible interval for* θ_c*) which is not surprising: we had only weak prior information and the posterior, see Figure 1.2(b), closely mirrors the likelihood. We observed only ten tosses so we would not anticipate that the data would resolve most, or indeed much, of the uncertainty about* θ_p.

[7] The generalisation for multivariate θ is a region C contained within the probability space of θ which contains θ with probability $1-\alpha$.

[8] Note that, see the solution to Question Sheet One 5(b), for generic parameter θ and data x we have $Var(\theta) = Var(E(\theta \,|\, X)) + E(Var(\theta \,|\, X))$ so that $E(Var(\theta \,|\, X)) \leq Var(\theta)$: we expect the observation of data x to reduce our uncertainty about parameter θ.



(a) Shaded area denotes 90% credible interval for a bimodal distribution with 5% in each tail.

(b) Shaded areas denote 90% highest density region for the bimodal distribution.

Figure 1.3: Credible interval (with equal tail probabilities) and highest density region for a bimodal distribution.

Definition 4 *(Highest density region)*
The highest density region (HDR) is a region C that contains $100(1-\alpha)\%$ of the posterior distribution and has the property that for all $\theta_1 \in C$ and $\theta_2 \notin C$,

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

Thus, for a HDR the density insider the region is never lower than outside of the region. The HDR is the region with $100(1-\alpha)\%$ probability of containing θ with the shortest total length. For unimodal symmetric distributions the HDR and credible interval are the same. Credible intervals are generally easier to compute (particularly if we are doing so by placing $\frac{\alpha}{2}$ in each tail) and are invariant to transformations of the parameters. HDRs are more complex and there is a potential lack of invariance under parameter transformation. However, they may be more meaningful for multinomial distributions.

Example 10 *Figure 1.3 shows a 90% credible interval and HDR for a parameter θ whose posterior distribution*[9]*, given by $f(\theta \,|\, x)$, is bimodal. Notice that the credible interval, see Figure 1.3(a), is not a HDR. In particular, $\theta_1 = 0.5$ is in the credible interval and $\theta_2 = 0.95$ is not and $f(0.95 \,|\, x) > f(0.5 \,|\, x)$. Moreover, the credible interval may not be sensible. For example, it includes the interval $(0.5, 0.6)$ which has a small probability of containing θ. The HDR is shown in Figure 1.3(b). It consists of two disjoint intervals and thus captures the*

[9]The plotted density is a mixture of a $Beta(4, 12)$ and a $Beta(16, 4)$ with $f(\theta \,|\, x) = \frac{1}{2B(4,12)}\theta^3(1-\theta)^{11} + \frac{1}{2B(16,4)}\theta^{15}(1-\theta)^3$.

bimodality of the distribution. It has a total length of 0.61 compared to the total length of 0.78 for the credible interval.

2 Modelling

In our work so far we have made use of the concepts of parameters, likelihoods and prior distributions with little attention focused upon clarity or justification.

Example 11 *Bernoulli model, different interpretation of parameter.*

1. *Consider tossing coins. It is natural to think about the model of independent Bernoulli trials given θ_c but what exactly is θ_c and can we justify it? Intuitively we think of θ_c as representing the 'true probability of a head' but what does this mean?*

2. *Consider an urn[1] which contains balls of proportion θ_u of colour purple and $(1-\theta_u)$ of colour green. We sample with replacement from the urn, so a ball is drawn from the urn, its colour noted and returned to the urn which is then shaken prior to the next draw. Given θ_u we might model the draws of the balls identically to the tossing of coins but our intuitive understanding of θ_c and θ_u are not the same. For example, we can never determine θ_c but we could physically determine θ_u: we could smash the urn and count the balls[2].*

How can we reconcile these cases? The answer lies in a judgment of exchangeability which is the key result of this chapter.

2.1 Predictive distribution

Suppose that we want to make predictions about the values of future observations of data. The distribution of future data Z given observed data x is the **predictive distribution** $f(z \mid x)$. Notice that this distribution only depends upon z and x. If we have a parametric model then

$$f(z \mid x) \quad = \quad \int_\theta f(z \mid \theta, x) f(\theta \mid x)\, d\theta. \tag{2.1}$$

[1] Urn problems are widely used as thought experiments in statistics and probability. For an initial overview see urn problem.

[2] Other, more pacifistic, approaches are available e.g. just empty the urn.

If we judge that X and Z are conditionally independent[3] given θ then $f(z\,|\,\theta,x) = f(z\,|\,\theta)$ so that (2.1) becomes

$$f(z\,|\,x) \quad = \quad \int_\theta f(z\,|\,\theta) f(\theta\,|\,x)\,d\theta. \tag{2.2}$$

Example 12 *Consider a Beta-Binomial analysis, such as tossing coins. We judge that* $X\,|\,\theta \sim Bin(n,\theta)$ *with* $\theta \sim Beta(\alpha,\beta)$. *Then, from Example 4,* $\theta\,|\,x \sim Beta(\alpha+x, \beta+n-x)$. *For ease of notation, let* $a = \alpha+x$ *and* $b = \beta+n-x$. *Now consider a further random variable* Z *for which we judge* $Z\,|\,\theta \sim Bin(m,\theta)$ *and that* X *and* Z *are conditionally independent given* θ. *So, having tossed a coin* n *times, we consider tossing it a further* m *times. We seek the predictive distribution of* Z *given the observed* x. *As* X *and* Z *are conditionally independent given* θ *then, from (2.2),*

$$\begin{aligned} f(z\,|\,x) &= \int_\theta f(z\,|\,\theta) f(\theta\,|\,x)\,d\theta \\ &= \int_0^1 \binom{m}{z} \theta^z (1-\theta)^{m-z} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\,d\theta \\ &= \binom{m}{z} \frac{1}{B(a,b)} \int_0^1 \theta^{a+z-1}(1-\theta)^{b+m-z-1}\,d\theta \qquad (2.3) \\ &= \binom{m}{z} \frac{B(a+z, b+m-z)}{B(a,b)}. \qquad (2.4) \end{aligned}$$

Notice that

$$\frac{B(a+z,b+m-z)}{B(a,b)} \quad = \quad \frac{\Gamma(a+z)\Gamma(b+m-z)}{\Gamma(a+b+m)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

so that (2.4) can be expressed as

$$f(z\,|\,x) \quad = \quad c \binom{m}{z} \Gamma(a+z)\Gamma(b+m-z)$$

where the constant (i.e. not depending upon z*)*

$$c \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a+b+m)\Gamma(a)\Gamma(b)}.$$

$Z\,|\,x$ *is the* ***Binomial-Beta distribution*** *with parameters* a, b, m.

Notice that the derivation of this predictive distribution involved, see (2.3), the identification of a kernel of a $Beta(a+z, b+m-z)$ distribution. This should not have been a surprise. From (1.10) (with $y = z$ and using the conditional independence of X and Z given θ), $f(z\,|\,\theta)f(\theta\,|\,x) \propto f(\theta\,|\,z,x)$ and, from Example 4, $\theta\,|\,z,x \sim Beta(a+z, b+m-z)$.

[3] See (1.4) for the definition of this.

The predictive distribution can be difficult to calculate but predictive summaries are often easily available. In particular we can apply generalisations of the results[4] of Question Sheet One Exercise 5 to the predictive distribution.

Lemma 1 *If X, Z and θ are three random variables with X and Z conditionally independent given θ then*

$$E(Z \,|\, X) = E(E(Z \,|\, \theta) \,|\, X), \tag{2.5}$$
$$Var(Z \,|\, X) = Var(E(Z \,|\, \theta) \,|\, X) + E(Var(Z \,|\, \theta) \,|\, X). \tag{2.6}$$

Proof - From Question Sheet One Exercise 5, conditioning everywhere, we have

$$E(Z \,|\, X) = E(E(Z \,|\, \theta, X) \,|\, X), \tag{2.7}$$
$$Var(Z \,|\, X) = Var(E(Z \,|\, \theta, X) \,|\, X) + E(Var(Z \,|\, \theta, X) \,|\, X). \tag{2.8}$$

Now as X and Z are conditionally independent given θ then $E(Z \,|\, \theta, X) = E(Z \,|\, \theta)$ and $Var(Z \,|\, \theta, X) = Var(Z \,|\, \theta)$. Substituting these into (2.7) and (2.8) gives (2.5) and (2.6). □

Observe that $E(Z \,|\, \theta)$ and $Var(Z \,|\, \theta)$ are computed using $f(z \,|\, \theta)$ and are both functions of θ, $g_1(\theta)$ and $g_2(\theta)$ respectively say. We then obtain $E(g_1(\theta) \,|\, X)$, $Var(g_1(\theta) \,|\, X)$ and $E(g_2(\theta) \,|\, X)$ using $f(\theta \,|\, x)$. These were also the distributions we needed, see (2.2), to compute $f(z \,|\, x)$. The conditional independence of X and Z given θ means that any calculation involving X, Z and θ can be performed using calculations between X and θ only and Z and θ only: by exploiting independence structure we can reduce the dimensionality of our problems[5].

Example 13 *From Exercise 12 we consider the predictive expectation of Z given x. As $Z \,|\, \theta \sim Bin(m, \theta)$ then $E(Z \,|\, \theta) = m\theta$ whilst as $\theta \,|\, x \sim Beta(a, b)$ then $E(\theta \,|\, X) = \frac{a}{a+b}$. Hence, we have that*

$$\begin{aligned} E(Z \,|\, X) &= E(E(Z \,|\, \theta) \,|\, X) \\ &= E(m\theta \,|\, X) \\ &= m\frac{a}{a+b} = m\frac{\alpha + x}{\alpha + \beta + n}. \end{aligned}$$

The modelling in Examples 12 and 13 raises an interesting question. X and Z are not independent and nor would we expect them to be: if we don't know θ then we expect observing x to be informative for Z

[4] For random variables X and Y we have $E(X) = E(E(X \,|\, Y))$ and $Var(X) = Var(E(X \,|\, Y)) + E(Var(X \,|\, Y))$.

[5] This is the basis behind what is known as local computation and is often exploited in highly complex models, specifically in Bayesian networks.

e.g. When tossing a coin, if we don't know whether or not the coin is fair then an initial sequence of n tosses will be informative for a future sequence of m tosses.

Consider such a model from a classical perspective[6] We would view X and Z as comprising a random sample and of being independent and identically distributed. As we can see from the prediction of Z given x this is a slightly misleading statement: they are only independent **conditional** on the parameter θ.

2.2 Exchangeability

The concept of exchangeability, introduced by Bruno de Finetti[7] in the 1930s, is the basic modelling tool within Bayesian statistics. One of the key differences between the classical and Bayesian schools is that observations that are former would treat as independent are treated as exchangeable by the latter.

Definition 5 *(Finite exchangeability)*
The random variables $X_1, \ldots, X_n$ are judged to be finitely exchangeable if their joint density function satisfies

$$f(x_1, \ldots, x_n) \quad = \quad f(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

for all permutations π defined on the set $\{1, \ldots, n\}$.

Example 14 *X_1 and X_2 are finitely exchangeable if $f(x_1, x_2) = f(x_2, x_1)$. X_1, X_2 and X_3 are finitely exchangeable if $f(x_1, x_2, x_3) = f(x_1, x_3, x_2) = f(x_2, x_1, x_3) = f(x_2, x_3, x_1) = f(x_3, x_1, x_2) = f(x_3, x_2, x_1)$.*

In essence, exchangeability captures the notion that only the values of the observations matter and not the order in which they were obtained. The labels are uninformative. Exchangeability is a stronger statement than identically distributed but weaker than independence. Independent observations are exchangeable as

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} f(x_i).$$

However, exchangeable observations need not be independent.

Example 15 *Suppose that X_1 and X_2 have the joint density function*

$$f(x_1, x_2) \quad = \quad \frac{3}{2}(x_1^2 + x_2^2)$$

for $0 < x_1 < 1$ and $0 < x_2 < 1$. Then X_1 and X_2 are (finitely) exchangeable but they are not independent.

[6] There is an interesting side question as to how to perform prediction in this case as $f(z \mid x)$ does not explicitly depend upon any parameter θ.

[7] Bruno de Finetti (1906-1985).

Definition 6 *(Infinite exchangeability)*
The infinite sequence of random variables X_1, X_2, ... are judged to be infinitely exchangeable if every finite subsequence is judged finitely exchangeable.

A natural question is whether every finitely exchangeable sequence can be embedded into, or extended, to an infinitely exchangeable sequence.

Example 16 *Suppose that X_1, X_2, X_3 are three finitely exchangeable[8] events with*

$$\begin{aligned} P(X_1 = 0, X_2 = 1, X_3 = 1) &= P(X_1 = 1, X_2 = 0, X_3 = 1) \\ &= P(X_1 = 1, X_2 = 1, X_3 = 0) = \tfrac{1}{3} \end{aligned} \tag{2.9}$$

with all other combinations having probability 0. Is there an X_4 such that X_1, X_2, X_3, X_4 are finitely exchangeable? If there is then, for example,

$$P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = P(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 1). \tag{2.10}$$

Now, assuming X_1, X_2, X_3, X_4 are finitely exchangeable,

$$\begin{aligned} &P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) \\ &= P(X_1 = 0, X_2 = 1, X_3 = 1) - P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 1) \\ &= \frac{1}{3} - P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 1) \qquad (2.11) \\ &= \frac{1}{3} - P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 0) \qquad (2.12) \end{aligned}$$

where (2.11) follows from (2.9) and (2.12) from finite exchangeability. Note that

$$\begin{aligned} &P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 0) \\ &= P(X_1 = 1, X_2 = 1, X_3 = 1) - P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1) \\ &\leq P(X_1 = 1, X_2 = 1, X_3 = 1) = 0. \qquad (2.13) \end{aligned}$$

Substituting (2.13) into (2.12) gives

$$P(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = \frac{1}{3}. \tag{2.14}$$

However,

$$\begin{aligned} &P(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 1) \\ &= P(X_1 = 0, X_2 = 0, X_3 = 1) - P(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 0) \\ &\leq P(X_1 = 0, X_2 = 0, X_3 = 1) = 0. \qquad (2.15) \end{aligned}$$

So, from (2.14) and (2.15) we observe that (2.10) does not hold: a contradiction to the assumption of finite exchangeability of X_1, X_2, X_3, X_4.

[8] It is common notation to often just term a sequence as exchangeable and leave it to context as to whether this means finitely or infinitely exchangeable.

Example 15 thus shows that a finitely exchangeable sequence can not necessarily even be embedded into a larger finitely exchangeable sequence let alone an infinitely exchangeable one.

Theorem 1 *(Representation theorem for 0-1 random variables[9])*
Let $X_1, X_2, \ldots$ be a sequence of infinitely exchangeable 0-1 random variables (i.e. events). Then the joint distribution of $X_1, \ldots, X_n$ has an integral representation of the form

$$f(x_1, \ldots, x_n) \quad = \quad \int_0^1 \left\{ \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right\} f(\theta)\, d\theta$$

with $y_n = \sum_{i=1}^{n} x_i$ and $\theta = \lim_{n\to\infty} \frac{y_n}{n}$.

The interpretation of this theorem is of profound significance. It is as if

1. Conditional upon a random variable θ, the X_i are judged to be independent Bernoulli random variables.
2. θ itself is assigned a probability distribution $f(\theta)$.
3. $\theta = \lim_{n\to\infty} \frac{y_n}{n}$ so that $f(\theta)$ represents beliefs about the limiting value of the mean of the X_is.

So, conditional upon θ, $X_1, \ldots, X_n$ are a random sample from a Bernoulli distribution with parameter θ generating a parametrised joint sampling distribution

$$\begin{aligned} f(x_1, \ldots, x_n \,|\, \theta) \quad &= \quad \prod_{i=1}^{n} f(x_i \,|\, \theta) \\ &= \quad \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \quad \prod_{i=1}^{n} \theta^{y_n} (1-\theta)^{n-y_n} \end{aligned}$$

where the parameter is assigned a prior distribution $f(\theta)$.

> Theorem 1 provides a justification for the Bayesian approach of combining a likelihood and a prior.

Notice that θ is provided with a formal definition and the sentence "the true value of θ" has a well-defined meaning.

Theorem 2 *(General representation theorem - simplified form)*
If $X_1, X_2 \ldots$ is an infinitely exchangeable sequence of random variables then the joint distribution of $X_1, \ldots, X_n$ has an integral representation of the form

$$f(x_1, \ldots, x_n) \quad = \quad \int_\theta \left\{ \prod_{i=1}^{n} f(x_i \,|\, \theta) \right\} f(\theta)\, d\theta$$

[9] This is often termed de Finetti's theorem.

where θ is the limit as $n \to \infty$ of some function of the observations $x_1, \ldots, x_n$ and $f(\theta)$ is a distribution over θ.

Points to note.

1. In full generality θ is an unknown distribution function[10] (cdf) in the infinite dimensional space of all possible distribution functions so is, in effect, an infinite dimensional parameter.

2. Typically we put probability 1 to the event that θ lies in a family of distributions to obtain the familiar representation above.

3. The representation theorem is an existence theorem: it generally does not specify the model[11] $f(x_i \,|\, \theta)$ and it never specifies the prior $f(\theta)$.

   The crux is that infinitely exchangeable random variables (and not just events) may be viewed as being conditionally independent given a parameter θ and provide a justification for the Bayesian approach.

Example 17 *An extension of Example 8. Let $X_1, \ldots, X_n$ be a finite subset of a sequence of infinitely exchangeable random variables[12] which are normally distributed with unknown mean θ and known variance σ^2. Thus, from Theorem 2, we can assume that $X_i \,|\, \theta \sim N(\theta, \sigma^2)$ and that, conditional upon θ, the X_i are independent. Hence, specification of a prior distribution for θ will allow us to obtain the joint distribution of $X_1, \ldots, X_n$. Suppose that our prior beliefs about θ can be expressed by $\theta \sim N(\mu_0, \sigma_0^2)$ for known constants μ_0 and σ_0^2. We compute the posterior distribution of θ given $x = (x_1, \ldots, x_n)$. The likelihood is*

$$\begin{aligned} f(x \,|\, \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \theta)^2\right\} \\ &\propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(\theta^2 - 2x_i\theta)\right\} \\ &= \exp\left\{-\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta)\right\}. \end{aligned} \tag{2.16}$$

Notice that, when viewed as a function of θ, (2.16) has the form of a normal kernel (a kernel of the $N(\bar{x}, \frac{\sigma^2}{n})$). The prior

$$f(\theta)) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\mu_0\theta)\right\}$$

[10] Attempts to model θ in this way are closely related to the vibrant research area of Bayesian nonparametrics.

[11] Theorem 1 provides an example when the model is specified.

[12] This rather wordy description is often just shortened to 'exchangeable'.

so that the posterior

$$\begin{aligned} f(\theta \,|\, x) &\propto \exp\left\{-\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta)\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\mu_0\theta)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left[\theta^2 - 2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\sigma_0^2}\mu_0\right)\theta\right]\right\}. \quad (2.17) \end{aligned}$$

We recognise (2.17) as the kernel of a Normal density so that $\theta \,|\, x \sim N(\mu_n, \sigma_n^2)$ *where*

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \quad (2.18)$$

$$\mu_n = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\sigma_0^2}\mu_0\right). \quad (2.19)$$

Notice the similarity of these values with those in Example 8, (1.18) and (1.19). The posterior precision, $Var^{-1}(\theta \,|\, X)$*, is the sum of the data precision,* $Var^{-1}(\bar{X} \,|\, \theta)$*, and the prior precision,* $Var^{-1}(\theta)$*. The posterior mean,* $E(\theta \,|\, X)$*, is a weighted average of the data mean,* $\bar{x}$*, and the prior mean,* $E(\theta)$*, weighted according to the corresponding precisions. Observe that weak prior information is represented by a large prior variance. Letting* $\sigma_0^2 \to \infty$ *we note that* $\mu_n \to \bar{x}$ *and* $\sigma_n^2 \to \frac{\sigma^2}{n}$*: the familiar classical model.*

2.3 Sufficiency, exponential families and conjugacy

Definition 7 *(Sufficiency)*
A statistic $t(X)$ *is said to be sufficient for* X *for learning about* θ *if we can write*

$$f(x \,|\, \theta) = g(t, \theta)h(x) \quad (2.20)$$

where $g(t, \theta)$ *depends upon* $t(x)$ *and* θ *and* $h(x)$ *does not depend upon* θ *but may depend upon* x*.*

Equivalent statements to (2.20) are

1. $f(x \,|\, t, \theta)$ does not depend upon θ so that $f(x \,|\, t, \theta) = f(x \,|\, t)$.
2. $f(\theta \,|\, x, t)$ does not depend upon x so that $f(\theta \,|\, x, t) = f(\theta \,|\, t)$.

Sufficiency represents the notion that given $t(x)$ nothing further can be learnt about θ from additionally observing x: θ and x are conditionally independent given t.

Example 18 *In Example 17, we have, from (2.16) and reinstalling the constants of proportionality,*

$$\begin{aligned} f(x \,|\, \theta) &= \exp\left\{-\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta)\right\} \times \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right\} \\ &= g(\bar{x}, \theta)h(x). \end{aligned}$$

Hence, $\bar{X}$ *is sufficient for* $X = (X_1, \ldots, X_n)$ *for learning about* θ*.*

Definition 8 *(k-parameter exponential family)*
A probability density $f(x \mid \theta)$, $\theta = (\theta_1, \ldots, \theta_k)$, is said to belong to the k-parameter exponential family if it is of the form[13]

$$\begin{aligned} f(x \mid \theta) &= Ef_k(x \mid g, h, u, \phi, \theta) \\ &= \exp\left\{ \sum_{j=1}^{k} \phi_j(\theta) u_j(x) + g(\theta) + h(x) \right\} \end{aligned} \tag{2.21}$$

where $\phi(\theta) = (\phi_1(\theta), \ldots, \phi_k(\theta))$ and $u(x) = (u_1(x), \ldots, u_k(x))$. The family is regular if the sample space of X does not depend upon θ, otherwise it is non-regular[14].

Example 19 *Suppose that $X \mid \theta \sim Bernoulli(\theta)$. Then*

$$\begin{aligned} f(x \mid \theta) &= \theta^x (1-\theta)^{1-x} \\ &= \exp\left\{ x \log\theta + (1-x)\log(1-\theta) \right\} \\ &= \exp\left\{ \left(\log\frac{\theta}{1-\theta} \right) x + \log(1-\theta) \right\} \end{aligned}$$

Hence, this is the 1-parameter regular exponential family with $\phi_1(\theta) = \log\frac{\theta}{1-\theta}$, $u_1(x) = x$, $g(\theta) = \log(1-\theta)$ and $h(x) = 0$.

Proposition 1 *If $X_1, \ldots, X_n$ is an exchangeable sequence such that, given a regular k-parameter exponential family $Ef_k(\cdot \mid \cdot)$,*

$$f(x_1, \ldots, x_n) = \int_\theta \left\{ \prod_{i=1}^{n} Ef_k(x_i \mid g, h, u, \phi, \theta) \right\} f(\theta)\, d\theta$$

then $t_n = t_n(X_1, \ldots, X_n) = [n, \sum_{i=1}^n u_1(X_i), \ldots, \sum_{i=1}^n u_k(X_i)]$, $n = 1, 2, \ldots$ is a sequence of sufficient statistics.

Proof - From the representation, we have

$$\begin{aligned} f(x \mid \theta) &= \prod_{i=1}^{n} Ef_k(x_i \mid g, h, u, \phi, \theta) \\ &= \prod_{i=1}^{n} \exp\left\{ \sum_{j=1}^{k} \phi_j(\theta) u_j(x_i) + g(\theta) + h(x_i) \right\} \\ &= \exp\left\{ \sum_{j=1}^{k} \phi_j(\theta) \left(\sum_{i=1}^{n} u_j(x_i) \right) + n g(\theta) + \sum_{i=1}^{n} h(x_i) \right\} \\ &= \exp\left\{ \sum_{j=1}^{k} \phi_j(\theta) \left(\sum_{i=1}^{n} u_j(x_i) \right) + n g(\theta) \right\} \times \exp\left\{ \sum_{i=1}^{n} h(x_i) \right\} \\ &= \tilde{g}(t_n, \theta) \tilde{h}(x) \end{aligned}$$

[13] This will prove to be the most useful form for our purposes. In MA40092 you saw it expressed as $f(x \mid \theta) = \tilde{g}(\theta)\tilde{h}(x) \exp\left\{ \sum_{j=1}^{k} \phi_j(\theta) u_j(x) \right\}$ where $\tilde{g}(\theta) = \exp\{g(\theta)\}$ and $\tilde{h}(x) = \exp\{h(x)\}$.

[14] An example of this is the Uniform distribution, see Question 4 of Question Sheet Three.

where $\tilde{g}(t_n, \theta)$ is a function of t_n and θ and $\tilde{h}(x)$ a function of x. Hence t_n is sufficient for $X = (X_1, \ldots, X_n)$ for learning about θ. □

Example 20 *Let $X_1, \ldots, X_n$ be an exchangeable sequence with $X_i \,|\, \theta \sim Bernoulli(\theta)$. Then, from Example 19, with $x = (x_1, \ldots, x_n)$*

$$f(x \,|\, \theta) \quad = \quad \prod_{i=1}^{n} \exp\left\{ \left(\log \frac{\theta}{1-\theta} \right) x_i + \log(1-\theta) \right\}.$$

For this 1-parameter exponential family we have $u_1(x_i) = x_i$ and $\sum_{i=1}^{n} u_1(x_i) = \sum_{i=1}^{n} x_i$ so that, from Proposition 1, $t_n = (n, \sum_{i=1}^{n} X_i)$ is sufficient for $X_1, \ldots, X_n$ for learning about θ.

2.3.1 Exponential families and conjugate priors

If $f(x \,|\, \theta)$ is a member of a k-parameter exponential family, it is easy to observe that a conjugate prior can be found in the $(k+1)$-parameter exponential family. Regarding $f(x \,|\, \theta)$ as a function of θ, notice that we can express (2.21) as

$$f(x \,|\, \theta) \quad = \quad \exp\left\{ \sum_{j=1}^{k} u_j(x)\phi_j(\theta) + g(\theta) + h(x) \right\} \tag{2.22}$$

which can be viewed as an exponential family over θ. If we take as our prior the following $(k+1)$-parameter exponential family over θ

$$f(\theta) \quad = \quad \exp\left\{ \sum_{j=1}^{k} a_j\phi_j(\theta) + dg(\theta) + c(a,d) \right\} \tag{2.23}$$

where $a = (a_1, \ldots, a_k)$ and $c(a,d)$ is a normalising constant so that,

$$c(a,d) \quad = \quad -\log \int_{\theta} \exp\left\{ \sum_{j=1}^{k} a_j\phi_j(\theta) + dg(\theta) \right\} d\theta, \tag{2.24}$$

then, from (2.22) and (2.23), our posterior is

$$\begin{aligned} f(\theta \,|\, x) \quad &= \quad \exp\left\{ \sum_{j=1}^{k} u_j(x)\phi_j(\theta) + g(\theta) + h(x) \right\} \exp\left\{ \sum_{j=1}^{k} a_j\phi_j(\theta) + dg(\theta) + c(a,d) \right\} \\ &\propto \quad \exp\left\{ \sum_{j=1}^{k} [a_j + u_j(x)]\phi_j(\theta) + (d+1)g(\theta) \right\} \end{aligned} \tag{2.25}$$

Notice that, up to constants of proportionality, (2.25) has the same form as (2.23). Indeed if we let $\tilde{a}_j = a_j + u_j(x)$ and $\tilde{d} = d + 1$ then we can express the posterior distribution as

$$f(\theta \,|\, x) \quad = \quad \exp\left\{ \sum_{j=1}^{k} \tilde{a}_j\phi_j(\theta) + \tilde{d}g(\theta) + c(\tilde{a}, \tilde{d}) \right\} \tag{2.26}$$

where $\tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_k)$ and $c(\tilde{a}, \tilde{d})$ is a normalising constant, equivalent to (2.24) but with $\tilde{a}_j$ for a_j and $\tilde{d}$ for d. Thus, we have that the $(k+1)$-parameter exponential family is a conjugate family with respect to the k-parameter exponential family likelihood[15]. In this case, we talk about the natural conjugate prior.

Example 21 *We find the natural conjugate prior for $X \mid \theta \sim Bernoulli(\theta)$. From Example 19 we have that*

$$f(x \mid \theta) = \exp\left\{\left(x \log \frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\}$$

so that we take a prior of the form

$$\begin{aligned} f(\theta) &\propto \exp\left\{\left(a \log \frac{\theta}{1-\theta}\right) + d \log(1-\theta)\right\} \\ &= \left(\frac{\theta}{1-\theta}\right)^a (1-\theta)^d \\ &= \theta^a (1-\theta)^{d-a} \end{aligned}$$

which is a kernel of a Beta distribution. To obtain the familiar parametrisation we take $a = a(\alpha, \beta) = \alpha - 1$ and $d = d(\alpha, \beta) = \beta + \alpha - 2$. The likelihood has one parameter, θ, so the natural conjugate prior has two parameters, α and β.

We term the $(k+1)$-parameter exponential family as being the natural conjugate prior to the k-parameter exponential family likelihood as the $(k+1+s)$-parameter exponential family

$$f(\theta) = \exp\left\{\sum_{j=1}^{k} a_j \phi_j(\theta) + d g(\theta) + \sum_{r=1}^{s} e_r E_r(\theta) + c(a, d, e)\right\},$$

where $e = (e_1, \ldots, e_s)$ and $c(a, d, e)$ is the normalising constant, is also conjugate to the k-parameter exponential family likelihood.

From Example 21, we noted that the natural conjugate to the Bernoulli likelihood has two parameters α and β, the different possible values of these parameters indexing the specific member of the Beta family chosen as the prior distribution. In order to distinguish the parameters indexing the family of prior distributions from the parameters θ about which we wish to make inference we term the former **hyperparameters**. Thus, in the Bernoulli case, α and β are the hyperparameters. The general $(k+1+s)$-parameter exponential family conjugate prior has $k+1+s$ hyperparameters.

Conjugate priors are useful for a number of reasons. Firstly, they ease the inferential process following the observation of data in that the posterior is straightforward to calculate: we

[15] It should be clear, using a similar approach to Proposition 1, that this result is easily obtained for the case when $X_1, \ldots, X_n$ is exchangeable with $X_i \mid \theta$ a member of the k-parameter exponential family.

only need to update the hyperparameters[16]. Secondly, they ease the burden on the prior specification: specifying the prior distribution reduces to specifying the hyperparameters. This can be done by specifying either the hyperparameters directly or through a series of distributional summaries from which they can be inferred. For example, which can infer the two hyperparameters of a Beta prior from specifying the mean and the variance of the prior[17].

Notice that in order for a conjugate family to exist, the likelihood $\prod_{i=1}^{n} f(x_i \mid \theta)$ must involve only a finite number of different functions of $x = (x_1, \ldots, x_n)$ for n arbitrarily large. Thus, the likelihood must contains a finite number of sufficient statistics which implies (given regularity conditions) that the likelihood is a member of a regular exponential family[18]. Thus, (subject to these regularity conditions), only regular exponential families exhibit conjugacy[19].

2.4 Noninformative prior distributions

The posterior distribution combines the information provided by the data with the prior information. In many situations, the available prior information may be too vague to be formalised as a probability distribution or too subjective to be used in public decision making.

There has been a desire for prior distributions that are guaranteed to play a minimal part in the posterior distribution. Such priors are sometimes called 'reference priors' and the prior density is said be 'flat' or 'noninformative'. The argument proposed is to **"let the data speak for themselves"**.

Example 22 *Suppose that* $X \mid \theta \sim Bin(n, \theta)$ *so that*

$$f(x \mid \theta) \quad = \quad \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

The natural conjugate prior is the $Beta(\alpha, \beta)$ *distribution. If the hyperparameters* α *and* β *and chosen so that* $\alpha = \beta = 1$ *then the prior is* $Unif(0, 1)$ *so that*

$$f(\theta) \quad = \quad 1, \quad 0 \leq \theta \leq 1.$$

The prior suggests that we judge that θ *is equally likely to be anywhere between* 0 *and* 1 *and can be viewed as a judgement of ignorance. The use of this uniform prior distribution is*

[16] In the majority of cases, from Proposition 1 and (2.25), this will involve simple adjustments using the sufficient statistics of the likelihood.

[17] A further example of this is given in Question 2 of Question Sheet Three.

[18] This is the Pitman-Koopman-Darmois theorem.

[19] An example that breaks the regulatory conditions is the non-regular exponential family likelihood given by $U(0, \theta)$. From Question 4 of Question Sheet Three we find that this has a conjugate prior given by the Pareto distribution.

often referred to as Bayes' postulate. The posterior is

$$\begin{aligned} f(\theta \,|\, x) &\propto f(x \,|\, \theta) f(\theta) \\ &= f(x \,|\, \theta). \end{aligned}$$

The posterior density is thus

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

which is the scaled likelihood. In this case, we have $\theta \,|\, x \sim Beta(x+1, n-x+1)$.

The basic idea is to represent ignorance using uniform prior distributions. However, if the possible values of θ do not lie in interval for which both endpoints are finite then no such proper distribution (i.e. one that integrates to unity) exists. However, this may not be a problem if the scaled likelihood has a finite integral over θ.

Definition 9 *(Improper prior)*
The specification $f(\theta) = f^*(\theta)$ *is said to be an improper prior if*

$$\int_\theta f^*(\theta)\, d\theta = \infty.$$

Example 23 *Recall Example 17. For an exchangeable collection* $X = (X_1, \ldots, X_n)$ *let* $X_i \,|\, \theta \sim N(\theta, \sigma^2)$ *where* σ^2 *is known and* $\theta \sim N(\mu_0, \sigma_0^2)$ *for known constants* μ_0 *and* σ_0^2. *The posterior is* $\theta \,|\, x \sim N(\mu_n, \sigma_n^2)$ *where*

$$\begin{aligned} \frac{1}{\sigma_n^2} &= \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \\ \mu_n &= \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\sigma_0^2}\mu_0\right). \end{aligned}$$

Recall that, in Example 17, as $\sigma_0^2 \to \infty$ *then* $\mu_n \to \bar{x}$ *and* $\sigma_n^2 \to \frac{\sigma^2}{n}$ *which matches the scaled likelihood. As* $\sigma_0^2 \to \infty$ *the prior density* $f(\theta)$ *becomes flatter and flatter which we could view as becoming more and more uniform. Notice that we could obtain* $N(\bar{x}, \frac{\sigma^2}{n})$ *as our posterior distribution using the improper prior* $f(\theta) \propto 1$.

2.4.1 Jeffreys' prior

Suppose we take $f(\theta) \propto 1$ to represent ignorance about a parameter θ. If we consider the parameter $\lambda = \frac{1}{\theta}$ then $f(\lambda) \propto \frac{1}{\lambda^2}$ which is not uniform. Although we attempt to express ignorance about θ we do not have ignorance about $\frac{1}{\theta}$!

> Can we calculate prior distributions that are invariant to the choice of parameterisation?

The answer is yes if we use Jeffreys' prior,[20]

$$f(\theta) \quad \propto \quad |I(\theta)|^{\frac{1}{2}} \tag{2.27}$$

where $I(\theta)$ is the **Fisher information matrix** and $|\cdot|$ represents the determinant of a matrix. Recall that if $\theta = (\theta_1, \ldots, \theta_k)$ is a k-dimensional parameter then $I(\theta)$ is the $k \times k$ matrix with (i, j)th entry

$$\begin{aligned}(I(\theta))_{ij} &= E\left\{\frac{\partial}{\partial\theta_i}\log f(x\,|\,\theta)\frac{\partial}{\partial\theta_j}\log f(x\,|\,\theta)\,\middle|\,\theta\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x\,|\,\theta)\,\middle|\,\theta\right\}\end{aligned}$$

Example 24 *For an exchangeable collection $X = (X_1, \ldots, X_n)$ let $X_i\,|\,\theta \sim N(\theta, \sigma^2)$ where σ^2 is known. We find Jeffreys' prior for θ.*

$$\begin{aligned}f(x\,|\,\theta) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}.\end{aligned}$$

Hence, as θ is univariate,

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f(x\,|\,\theta) &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right\} \\ &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\left(-\frac{n}{2}\log 2\pi\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\right\} \\ &= \frac{\partial}{\partial\theta}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)\right\} \\ &= -\frac{n}{\sigma^2}\end{aligned}$$

so that the Fisher information is

$$\begin{aligned}I(\theta) &= -E\left(-\frac{n}{\sigma^2}\,|\,\theta\right) \\ &= \frac{n}{\sigma^2}.\end{aligned} \tag{2.28}$$

Substituting (2.28) into (2.27) we see that, in this case, Jeffreys' prior is

$$f(\theta) \;=\; \sqrt{\frac{n}{\sigma^2}} \;= \propto 1$$

which is the improper prior we suggested in Example 23.

[20] Jeffreys' prior is named after Harold Jeffreys (1891-1989).

Recall how to transform the distribution of random variables. If $X = (X_1, \ldots, X_p)$ is a p-dimensional random variable with density function $f_X(x)$ and $Y = (Y_1, \ldots, Y_p)$ is a p-dimensional random variable with density function $f_Y(y)$ then if $Y = g(X)$

$$f_Y(y) \quad = \quad |\det J(x,y)| f_X(g^{-1}(y))$$

where $J(x,y)$ is the $p \times p$ Jacobian matrix with (i,j)th element

$$(J(x,y))_{ij} \quad = \quad \frac{\partial x_i}{\partial y_j}$$

where we write $x_i = h_i(y)$ and $y_j = g_j(x)$, $\det J(x,y)$ the determinant of the Jacobian[21] and $|\cdot|$ denotes the modulus.

Example 25 *(Not given in lectures but just to remind you how to transform variables) Consider $X = (X_1, X_2)$ denoting Cartesian coordinates and $Y = (Y_1, Y_2)$ denoting polar coordinates. We have that $Y = (\sqrt{X_1^2 + X_2^2}, \tan^{-1}(\frac{X_2}{X_1}))$. We wish to find the density $f_Y(y)$ by transforming the density $f_X(x)$. We have $y_1 = g_1(x) = \sqrt{x_1^2 + x_2^2}$, $y_2 = g_2(x) = \tan^{-1}(\frac{x_2}{x_1})$. The inverse transformation is*

$$\begin{aligned} x_1 \quad = \quad h_1(y) \quad &= \quad y_1 \cos y_2, \\ x_2 \quad = \quad h_2(y) \quad &= \quad y_1 \sin y_2. \end{aligned}$$

Thus, the Jacobian

$$J(x,y) \quad = \quad \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} \quad = \quad \begin{pmatrix} \cos y_2 & -y_1 \sin y_2 \\ \sin y_2 & y_1 \cos y_2 \end{pmatrix}$$

with $\det J(x,y) = y_1$. Hence,

$$f_Y(y) \quad = \quad |\det J(x,y)| f_X(g^{-1}(y)) \quad = \quad y_1 f_X(y_1 \cos y_2, y_1 \sin y_2).$$

Let's consider transformations of Jeffreys' prior. For simplicity of exposition we'll consider the univariate case[22]. For a univariate parameter θ, Jeffreys' prior $f_\theta(\theta) \propto \sqrt{I(\theta)}$. Consider a univariate transformation $\phi = g(\theta)$ (e.g. $\phi = \log\theta$, $\phi = \frac{1}{\theta}$, ...). There are two possible ways to obtain Jeffreys' prior for ϕ.

1. Obtain Jeffreys' prior for θ and transform $f_\theta(\theta)$ to $f_\phi(\phi)$.
2. Transform the data immediately and obtain $f_\phi(\phi)$ using Jeffreys' prior for ϕ.

We'll show the two approaches are identical. Transforming $f_\theta(\theta)$ to $f_\phi(\phi)$ we have

$$f_\phi(\phi) \quad = \quad \left|\frac{\partial\theta}{\partial\phi}\right| f_\theta(g^{-1}(\phi)) \quad \propto \quad \left|\frac{\partial\theta}{\partial\phi}\right| \sqrt{I(\theta)}. \tag{2.29}$$

[21] This is often called the Jacobian determinant or even just the Jacobian.

[22] The multivariate case is similar, only, for ϕ and θ multivariate, (2.30) becomes $I(\phi) = J(\theta,\phi) I(\theta) J(\theta,\phi)^T$ so that $|I(\phi)| = |J(\theta,\phi)|^2 |I(\theta)|$ and the result follows.

(e.g. $\phi = \log\theta$ gives $\theta = e^{\phi}$, $\frac{\partial\theta}{\partial\phi} = e^{\phi}$ so that $f_{\phi}(\phi) = e^{\phi} f_{\theta}(e^{\phi})$; $\phi = \frac{1}{\theta}$ gives $\theta = \frac{1}{\phi}$, $\frac{\partial\theta}{\partial\phi} = -\frac{1}{\phi^2}$ so that $f_{\phi}(\phi) = \frac{1}{\phi^2} f_{\theta}(\frac{1}{\phi})$.) We now consider finding Jeffreys' prior for ϕ directly. We have

$$\begin{aligned} I(\phi) &= E\left\{\left(\frac{\partial}{\partial\phi}\log f(x\,|\,\phi)\right)^2 \middle| \phi\right\} \\ &= E\left\{\left(\frac{\partial\theta}{\partial\phi}\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right)^2 \middle| \theta\right\} \\ &= \left(\frac{\partial\theta}{\partial\phi}\right)^2 E\left\{\left(\frac{\partial}{\partial\theta}\log f(x\,|\,\theta)\right)^2 \middle| \theta\right\} \\ &= \left(\frac{\partial\theta}{\partial\phi}\right)^2 I(\theta). \end{aligned} \tag{2.30}$$

Hence, Jeffreys prior for ϕ is

$$f_{\phi}(\phi) \propto \sqrt{I(\phi)} = \left|\frac{\partial\theta}{\partial\phi}\right| \sqrt{I(\theta)} \tag{2.31}$$

so that (2.29) and (2.31) are identical. Jeffreys' prior has the property that the prior is invariant in that, whatever scale we choose to measure the unknown parameter the same prior results when the scale is transformed. To quote Jeffreys[23]

> any arbitrariness in the choice of parameters could make no difference to the results.

2.4.2 Some final remarks about noninformative priors

A number of objections can be made to noninformative priors. One major objection to Jeffreys' prior is that it depends upon the form of the data whereas the prior should only reflect the prior information and not by influenced by what data are to be collected. For example, on Question 4 of Question Sheet Four we show that Jeffreys' priors for the binomial and negative binomial likelihoods are different which leads to the violation of the likelihood principle[24]. The likelihood principle states that the likelihood contains all the information about the data x so that two likelihoods contain the same information if they are proportional.

The use of improper priors may appear as a convenient tool to minimise the role of the prior distribution. However, the posterior is not always guaranteed to be proper. This is particularly the case if the likelihood, viewed as a function of θ, does not have a non-zero finite integral when integrated over θ.

Example 26 *Consider $X\,|\,\theta \sim Bin(n,\theta)$ and the improper prior $f(\theta) \propto \theta^{-1}(1-\theta)^{-1}$. In this case, θ is often said to have the improper $Beta(0,0)$ density. The posterior is*

$$f(\theta\,|\,x) \propto \theta^{x}(1-\theta)^{n-x} \times \theta^{-1}(1-\theta)^{-1} = \theta^{x-1}(1-\theta)^{n-x-1}$$

[23] see Jeffreys, H.D. (1961). *Theory of Probability*, 3rd ed. Oxford: University Press.

[24] The adoption of the likelihood principle is controversial and has caused much debate. Classical statistics violates the likelihood principle but Bayesian statistics (using proper prior distributions) does not.

which looks like a kernel of a $Beta(x, n-x)$ *density so that* $\theta \,|\, x \sim Beta(x, n-x)$*. Notice that, in this case, we have*

$$E(\theta \,|\, x) \quad = \quad \frac{x}{n}$$

so that the posterior mean is equal to the maximum likelihood estimate which provides a motivation for the use of this improper prior. However, if $x = 0$ *(or* $n = 1$*) then the posterior is improper.*

3 Computation

Practical implementation of Bayesian methods requires substantial computation. This is required, essentially, to calculate summaries of the posterior distribution. So far, we have worked with prior distributions and likelihoods of a sufficiently convenient form to simplify the construction of the posterior. In practice,

1. we need to be able to work with much more complex models which actually relate to real life problems.
2. A "good" Bayesian will specify prior distributions that accurately reflect the prior information rather than use a prior of a mathematically convenient form.

The Bayesian needs computational tools to calculate a variety of posterior summaries from distributions that are

1. mathematically complex
2. often high-dimensional

Typically, we shall be concerned with focusing on the posterior

$$f(\theta \,|\, x) \quad = \quad cg(\theta)$$

where $g(\theta) = f(x \,|\, \theta) f(\theta)$ and c is the normalising constant. In many cases c will be unknown because the integral cannot be carried out analytically.

Example 27 *Suppose that $X \,|\, \theta \sim N(\theta, \sigma^2)$ where σ^2 is a known constant and the prior for θ is judged to follow a t-distribution with ν degrees of freedom. Hence,*

$$f(\theta) \quad = \quad \frac{1}{\sqrt{\nu} B(\frac{1}{2}, \frac{1}{2}\nu)} \left(1 + \frac{\theta^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}$$

The posterior distribution for $\theta \,|\, x$ is

$$f(\theta \,|\, x) \quad = \quad \frac{c}{\sqrt{2\pi\sigma\nu} B(\frac{1}{2}, \frac{1}{2}\nu)} \left(1 + \frac{\theta^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}$$

where

$$c^{-1} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma\nu}B(\frac{1}{2},\frac{1}{2}\nu)} \left(1+\frac{\theta^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\} d\theta. \quad (3.1)$$

However, we can not perform the integral in (3.1) analytically.

In practice, many computational techniques do not explicitly require c but compute summaries of $f(\theta \mid x)$ from $g(\theta)$.

3.1 Normal approximations

We consider approximating the posterior distribution by a normal distribution. The approach utilises a Taylor series expansion of $\log f(\theta \mid x)$ about the mode $\tilde{\theta}$. The approximation is most reasonable when $f(\theta \mid x)$ is unimodal and roughly symmetric.

Assume that θ is univariate. The one-dimensional Taylor series for a generic function $h(\theta)$ about θ_0 is

$$h(\theta) = h(\theta_0) + (\theta-\theta_0)\left.\frac{\partial}{\partial\theta}h(\theta)\right|_{\theta=\theta_0} + \cdots + \frac{(\theta-\theta_0)^r}{r!}\left.\frac{\partial^r}{\partial\theta^r}h(\theta)\right|_{\theta=\theta_0} + \cdots$$

Taking $h(\theta) = \log f(\theta \mid x)$ and $\theta_0 = \tilde{\theta}$ we have

$$\begin{aligned}\log f(\theta \mid x) &= \log f(\tilde{\theta} \mid x) + (\theta-\tilde{\theta})\left.\frac{\partial}{\partial\theta}\log f(\theta \mid x)\right|_{\theta=\tilde{\theta}} \\ &\quad + \frac{(\theta-\tilde{\theta})^2}{2}\left.\frac{\partial^2}{\partial\theta^2}\log f(\theta \mid x)\right|_{\theta=\tilde{\theta}} + \text{h.o.t.} \quad (3.2)\end{aligned}$$

where h.o.t. denotes higher order terms. Now as $\tilde{\theta}$ is the mode of $f(\theta \mid x)$ then it is a maximum of $\log f(\theta \mid x)$ so that

$$\left.\frac{\partial}{\partial\theta}\log f(\theta \mid x)\right|_{\theta=\tilde{\theta}} = 0. \quad (3.3)$$

Letting

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

denote the observed information[1] then, using this and (3.3), (3.2) becomes

$$\log f(\theta \mid x) = \log f(\tilde{\theta} \mid x) - \frac{I(\tilde{\theta} \mid x)}{2}(\theta-\tilde{\theta})^2 + \text{h.o.t.}$$

Hence, approximately,

$$f(\theta \mid x) \propto \exp\left\{-\frac{I(\tilde{\theta} \mid x)}{2}(\theta-\tilde{\theta})^2\right\}$$

[1] Note that as $\tilde{\theta}$ is a maximum of $\log f(\theta \mid x)$ then $I(\tilde{\theta} \mid x)$ is positive.

which is a kernel of a $N(\tilde{\theta}, I^{-1}(\tilde{\theta} \,|\, x))$ density so that, approximately, $\theta \,|\, x \sim N(\tilde{\theta}, I^{-1}(\tilde{\theta} \,|\, x))$.

The approximation can be generalised for the case when θ is multivariate. Suppose that θ is the $p \times 1$ vector $(\theta_1, \ldots, \theta_p)^T$ with posterior density $f(\theta \,|\, x)$. Let the posterior mode be $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_p)^T$ and the $p \times p$ observed information matrix be $I(\theta \,|\, x)$ where

$$(I(\theta \,|\, x))_{ij} - \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\theta \,|\, x)$$

then, approximately, $\theta \,|\, x \sim N_p(\tilde{\theta}, I^{-1}(\tilde{\theta} \,|\, x))$, the multivariate normal distribution (of dimension p) with

$$f(\theta \,|\, x) \quad = \quad \frac{1}{(2\pi)^{\frac{p}{2}} |I^{-1}(\tilde{\theta} \,|\, x)|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\theta - \tilde{\theta})^T I^{-1}(\tilde{\theta} \,|\, x)(\theta - \tilde{\theta}) \right\}.$$

Example 28 *Suppose that $X_1, \ldots, X_n$ are exchangeable and $X_i \,|\, \theta \sim Po(\theta)$. Find a Normal approximation to the posterior distribution when we judge a uniform (improper) prior is appropriate for θ.*

We take $f(\theta) \propto 1$ so that

$$f(\theta \,|\, x) \quad \propto \quad \prod_{i=1}^{n} \theta^{x_i} e^{-\theta} \quad = \quad \theta^{n\bar{x}} e^{-n\theta}$$

which we recognise as a kernel of a $Gamma(n\bar{x}+1, n)$ distribution thus $\theta \,|\, x \sim Gamma(n\bar{x}+1, n)$. A $Gamma(\alpha, \beta)$ distribution has mode equal to $\frac{\alpha-1}{\beta}$ so that the mode of $\theta \,|\, x$ is $\tilde{\theta} = \frac{n\bar{x}+1-1}{n} = \bar{x}$. The observed information is

$$I(\theta \,|\, x) \quad = \quad -\frac{\partial^2}{\partial\theta^2} \left\{ n\bar{x} \log\theta - n\theta + \log \frac{n^{n\bar{x}+1}}{\Gamma(n\bar{x}+1)} \right\} \quad = \quad \frac{n\bar{x}}{\theta^2}$$

so that $I(\tilde{\theta} \,|\, x) = I(\bar{x} \,|\, x) = \frac{n}{\bar{x}}$. Thus, approximately, $\theta \,|\, x \sim N(\bar{x}, \frac{\bar{x}}{n})$.

3.2 Posterior sampling

If we can draw a sample of values from the posterior distribution then we can use the properties of the sample (e.g. mean, variance, quantiles, ...) to estimate properties of the posterior distribution.

3.2.1 Monte Carlo integration

Suppose that we wish to estimate $E\{g(X)\}$ for some function $g(x)$ with respect to a density $f(x)$. Thus,

$$E\{g(X)\} \quad = \quad \int_X g(x) f(x)\, dx. \tag{3.4}$$

For example, to compute the posterior mean of $\theta \,|\, x$ we have

$$E(\theta \,|\, X) \;=\; \int_\theta \theta f(\theta \,|\, x)\, d\theta.$$

Monte Carlo[2] integration involves the following:

1. sample N values $x_1, \ldots, x_N$ independently[3] from $f(x)$
2. calculate $g(x_i)$ for each x_i
3. estimate $E\{g(X)\}$ by the sample mean of the $g(x_i)$,

$$g_N(x) \;=\; \frac{1}{N}\sum_{i=1}^{N} g(x_i).$$

Note that the corresponding estimator $g_N(X)$ is an unbiased estimator of $E\{g(X)\}$ as

$$\begin{aligned} E\{g_N(X)\} &= \frac{1}{N}\sum_{i=1}^{N} E\{g(X_i)\} \\ &= \sum_{i=1}^{N} E\{g(X)\} \;=\; E\{g(X)\}. \end{aligned}$$

The variance of the estimator is

$$\begin{aligned} Var\{g_N(X)\} &= \frac{1}{N^2}\sum_{i=1}^{N} Var\{g(X_i)\} \\ &= \frac{1}{N^2}\sum_{i=1}^{N} Var\{g(X)\} \;=\; \frac{1}{N} Var\{g(X)\}. \end{aligned}$$

3.2.2 Importance sampling

Suppose that we can not sample from $f(x)$ but can generate samples from some other distribution $q(x)$ which is an approximation to $f(x)$.

> e.g. $q(x)$ might be the approximation to $f(x)$ obtained by a Normal approximation about the mode.

We can use samples from $q(x)$ using the method of **importance sampling**. Let us consider the problem of estimating $E\{g(X)\}$ for some function $g(x)$ with respect to a density $f(x)$. Then we may rewrite (3.4) as

$$\begin{aligned} E\{g(X)\} &= \int_X \frac{g(x)f(x)}{q(x)} q(x)\, dx \\ &= E\left\{ \frac{g(X)f(X)}{q(X)} \,\middle|\, X \sim q(x) \right\} \end{aligned} \tag{3.5}$$

[2] The term Monte Carlo was introduced by John von Neumann (1903-1957) and Stanisław Ulam (1909-1984)) as a code word for the simulation work they were doing for the Manhattan Project during World War II.

[3] We shall later see how obtaining an independent sample can be difficult.

where (3.5) makes clear that the expectation is calculation with respect to the density $q(x)$. If we draw a random sample $x_1, \ldots x_N$ from $q(x)$ then we may approximate $E\{g(X)\}$ by

$$\hat{I} \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i) f(x_i)}{q(x_i)}.$$

Notice that

$$\begin{aligned} E(\hat{I} \,|\, X \sim q(x)) \quad &= \quad \frac{1}{N} \sum_{i=1}^{N} E\left(\frac{g(X_i) f(X_i)}{q(X_i)} \,\middle|\, X \sim q(x) \right) \\ &= \quad \frac{1}{N} \sum_{i=1}^{N} E\left(\frac{g(X) f(X)}{q(X)} \,\middle|\, X \sim q(x) \right) \quad = \quad E\{g(X)\} \end{aligned}$$

so that $\hat{I}$ is an unbiased estimator of $E\{g(X)\}$. In a similar fashion we obtain that

$$Var(\hat{I} \,|\, X \sim q(x)) \quad = \quad \frac{1}{N} Var\left(\frac{g(X) f(X)}{q(X)} \,\middle|\, X \sim q(x) \right).$$

Hence, the variance of $\hat{I}$ will depend both on the choice of N and the approximation $q(x)$. Given a choice of approximating distributions from which we can sample we might choose the approximation which minimises the variance of $\hat{I}$.

3.3 Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) is a general technique that has revolutionised practical Bayesian statistics. It is a tool that generates samples from complex multi-dimensional posterior distributions in cases where the distribution is analytically intractable.

A Markov chain is a random process with the property that, conditional upon its current value, future values are independent of the past. Under certain conditions such a chain will converge to a stationary distribution so that eventually values may be treated as a sample from the stationary distribution.

The basic idea behind MCMC techniques for Bayesian statistics is:

1. Construct a Markov chain that has the required posterior distribution as its stationary distribution. This is sometimes referred to as the target distribution.

2. From some starting point, generate sequential values from the chain until (approximate) convergence. (One of the difficulties of the approach is knowing when this has occurred.)

3. Continue generating values from the chain which are now viewed as a sample of values from the posterior distribution. (As the chain has converged, we are now sampling from the stationary distribution.)

4. Use the resulting sample (from the stationary distribution) to estimate properties of the posterior distribution.

3.3.1 Useful results for Markov chains

We will consider only discrete time Markov chains (simulation is itself discrete) and summarise only the important definitions required for MCMC.

A **Markov chain** is a discrete time stochastic process $\{X_0, X_1, \ldots\}$ with the property that the distribution of X_t given all previous values of the process only depends upon X_{t-1}. That is,

$$P(X_t \in A \,|\, X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) \quad = \quad P(X_t \in A \,|\, X_{t-1} = x_{t-1})$$

for any set A.

The probability of a transition, or jump, from x_{t-1} at time $t-1$ to X_t at time t is given by the **transition kernel**. If X_t is discrete then the transition kernel is a transition matrix with (i, j)th element

$$P_{ij} \quad = \quad P(X_t = j \,|\, X_{t-1} = i) \tag{3.6}$$

which represents the probability of moving from state i to state j. In the continuous case, if

$$P(X_t \in A \,|\, X_{t-1} = x_{t-1}) \quad = \quad \int_A q(x \,|\, x_{t-1})\, dx \tag{3.7}$$

then the transition kernel is $q(x \,|\, x_{t-1})$. Notice that it is common to see the transition kernel written as $q(x_{t-1}, x)$ to indicate the transition from x_{t-1} to x; $q(x_{t-1}, x)$ does not represent the joint density of x_{t-1} and x.

Definition 10 *(Irreducible)*
A Markov chain is said to be irreducible if for every i, j there exists k such that

$$P(X_{t+k} = j \,|\, X_t = i) \quad > \quad 0$$

that is, all states can be reached from any other state in a finite number of moves.

Definition 11 *(Periodic/Aperiodic)*
A state i is said to be periodic with period d_i if starting from state i the chain returns to it within a fixed number of steps d_i or a multiple of d_i.

$$d_i \quad = \quad gcd\{t : P(X_t = i \,|\, X_0 = i) > 0\}$$

where gcd is the greatest common divisor. If $d_i = 1$ then the state i is said to be aperiodic.

Irreducible Markov chains have the property that all states have the same period. A Markov chain is called **aperiodic** if some (and hence all) states are aperiodic.

Definition 12 *(Recurrent/Positive Recurrent)*
Let τ_{ii} be the time of the first return to state i

$$t_{ii} \quad = \quad \min\{t > 0 : X_t = i \,|\, X_0 = i\}$$

A state i is recurrent if $P(\tau_{ii} < \infty) = 1$ and positive recurrent if $E(\tau_{ii}) < \infty$.

Thus, a state i is recurrent if the chain will return to state i with probability 1 and positive recurrent if, with probability 1, it will return in a finite time. An irreducible Markov chain is positive recurrent if some (and hence all) states i are positive recurrent.

Definition 13 *(Ergodic)*
A state is said to be ergodic if it is aperiodic and positive recurrent. A Markov chain is ergodic if all of its states are ergodic.

Definition 14 *(Stationary)*
A distribution π is said to be a stationary distribution of a Markov chain with transition probabilities P_{ij}, see (3.6), if

$$\sum_{i \in S} \pi_i P_{ij} \quad = \quad \pi_j \quad \forall j \in S$$

where S denotes the state space.

In matrix notation if P is the matrix of transition probabilities and π the vector with ith entry π_i then the stationary distribution satisfies

$$\pi \quad = \quad \pi P$$

These definitions have assumed that the state space of the chain is discrete. They are naturally generalised to the case when the state space is continuous. For example if $q(x \,|\, x_{t-1})$ denotes the transition kernel, see (3.7), then the stationary distribution π of the Markov chain must satisfy

$$\pi(x) \quad = \quad \int \pi(x_{t-1}) q(x \,|\, x_{t-1}) \, dx_{t-1}$$

Theorem 3 *(Existence and uniqueness)*
Each irreducible and aperiodic Markov chain has a unique stationary distribution π.

Theorem 4 *(Convergence)*
Let X_t be an irreducible and aperiodic Markov chain with stationary distribution π and arbitrary initial value $X_0 = x_0$. Then

$$P(X_t = x \,|\, X_0 = x_0) \quad \to \quad \pi(x)$$

as $t \to \infty$.

Theorem 5 *(Ergodic)*
Let X_t be an ergodic Markov chain with limiting distribution π. If $E\{g(X) \,|\, X \sim \pi(x)\} < \infty$ then the sample mean converges to the expectation of $g(X)$ under π,

$$P\left\{\frac{1}{N}\sum_{i=1}^{N} g(X_i) \to E\{g(X) \,|\, X \sim \pi(x)\}\right\} \quad = \quad 1.$$

Consequence of these theorems:

> If we can construct an ergodic Markov chain θ_t which has the posterior distribution $f(\theta \,|\, x)$ as the stationary distribution $\pi(\theta)$ then, starting from an initial point θ_0, if we run the Markov chain for long enough, we will sample from the posterior.
>
> - for large t, $\theta_t \sim \pi(\theta) = f(\theta \,|\, x)$
> - for each $s > t$, $\theta_s \sim \pi(\theta) = f(\theta \,|\, x)$
> - the ergodic averages converge to the desired expectations under the target distribution.

3.3.2 The Metropolis-Hastings algorithm

Suppose that our goal is to draw samples from some distribution $f(\theta \,|\, x) = cg(\theta)$ where c is the normalising constant which may not be known or very difficult to compute. The Metropolis-Hastings[4] algorithm provides a way of sampling from $f(\theta \,|\, x)$[5] without requiring us to know c.

Let $q(\phi \,|\, \theta)$ be an arbitrary transition kernel: that is the probability of moving, or **jumping**, from current state θ to proposed new state ϕ, see (3.7). This is sometimes called the **proposal distribution**. The following algorithm will generate a sequence of values $\theta^{(1)}$, $\theta^{(2)}$, ... which form a Markov chain with stationary distribution given by $f(\theta \,|\, x)$.

Algorithm 1 *The Metropolis-Hastings Algorithm.*

1. *Choose an arbitrary starting point $\theta^{(0)}$ for which $f(\theta^{(0)} \,|\, x) > 0$.*
2. *At time t*
   (a) *Sample a **candidate point** or **proposal**, θ^*, from $q(\theta^* \,|\, \theta^{(t-1)})$, the proposal distribution.*

[4] The original algorithm was developed by Nicholas Metropolis (1915-1999) in 1953 and generalised by W. Keith Hastings (1930-2016)) in 1970. Metropolis, aided by the likes of Mary Tsingou (1928-) and Klára Dan von Neumann (1911-1963), was involved in the design and build of the MANIAC I computer.

[5] The algorithm can be used to sample from any distribution $f(\theta)$. In this course, our interest centres upon sampling from the posterior distribution.

(b) *Calculate the* ***acceptance probability***

$$\alpha(\theta^{(t-1)}, \theta^*) = \min\left(1, \frac{f(\theta^* \mid x) q(\theta^{(t-1)} \mid \theta^*)}{f(\theta^{(t-1)} \mid x) q(\theta^* \mid \theta^{(t-1)})}\right). \tag{3.8}$$

(c) *Generate* $U \sim U(0,1)$.

(d) *If* $U \leq \alpha(\theta^{(t-1)}, \theta^*)$ ***accept*** *the proposal and set* $\theta^{(t)} = \theta^*$. *Otherwise,* ***reject*** *the proposal and set* $\theta^{(t)} = \theta^{(t-1)}$. *(We thus accept the proposal with probability* $\alpha(\theta^{(t-1)}, \theta^*)$*.)*

3. *Repeat step 2.*

Notice that if $f(\theta \mid x) = cg(\theta)$ then, from (3.8),

$$\begin{aligned}\alpha(\theta^{(t-1)}, \theta^*) &= \min\left(1, \frac{cg(\theta^*) q(\theta^{(t-1)} \mid \theta^*)}{cg(\theta^{(t-1)}) q(\theta^* \mid \theta^{(t-1)})}\right) \\ &= \min\left(1, \frac{g(\theta^*) q(\theta^{(t-1)} \mid \theta^*)}{g(\theta^{(t-1)}) q(\theta^* \mid \theta^{(t-1)})}\right)\end{aligned}$$

so that we do not need to know the value of c in order to compute $\alpha(\theta^{(t-1)}, \theta^*)$ and hence utilise the Metropolis-Hastings algorithm to sample from $f(\theta \mid x)$.

If the proposal distribution is symmetric, so $q(\phi \mid \theta) = q(\theta \mid \phi)$ for all possible ϕ, θ, then, in particular, we have $q(\theta^{(t-1)} \mid \theta^*) = q(\theta^* \mid \theta^{(t-1)})$ so that the acceptance probability (3.8) is given by

$$\alpha(\theta^{(t-1)}, \theta^*) = \min\left(1, \frac{f(\theta^* \mid x)}{f(\theta^{(t-1)} \mid x)}\right). \tag{3.9}$$

The Metropolis-Hastings algorithm with acceptance probability (3.8) replaced by (3.9) is the **Metropolis algorithm** and was the initial version of this MCMC approach before later being generalised by Hastings. In this context it is straightforward to interpret the acceptance/rejection rule.

1. If the proposal **increases** the posterior density, that is $f(\theta^* \mid x) > f(\theta^{(t-1)} \mid x)$, we always move to the new point θ^*.

2. If the proposal **decreases** the posterior density, that is $f(\theta^* \mid x) < f(\theta^{(t-1)} \mid x)$, then we move to the new point θ^* with probability equal to the ratio of the new to current posterior density.

Example 29 *Suppose that we want to sample from* $\theta \mid x \sim N(0,1)$ *using the Metropolis-Hastings algorithm and we let the proposal distribution* $q(\phi \mid \theta)$ *be the density of the* $N(\theta, \sigma^2)$ *for some* σ^2. *Thus,*

$$\begin{aligned}q(\phi \mid \theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\phi - \theta)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \phi)^2\right\} \\ &= q(\theta \mid \phi).\end{aligned}$$

The proposal distribution is symmetric and so the Metropolis-Hastings algorithm reduces to the Metropolis algorithm. The algorithm is performed as follows.

1. *Choose a starting point* $\theta^{(0)}$ *for which* $f(\theta^{(0)} \mid x) > 0$. *As* $\theta \mid x \sim N(0,1)$ *this will hold for any* $\theta^{(0)} \in (-\infty, \infty)$.

2. *At time t*

   (a) *Sample* $\theta^* \sim N(\theta^{(t-1)}, \sigma^2)$.

   (b) *Calculate the acceptance probability*

$$\begin{aligned} \alpha(\theta^{(t-1)}, \theta^*) &= \min\left(1, \frac{f(\theta^* \mid x)}{f(\theta^{(t-1)} \mid x)}\right) \\ &= \min\left(1, \frac{(2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\theta^*)^2\}}{(2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\theta^{(t-1)})^2\}}\right). \end{aligned}$$

   (c) *Generate* $U \sim U(0,1)$.

   (d) *If* $U \leq \alpha(\theta^{(t-1)}, \theta^*)$ *accept the move,* $\theta^{(t)} = \theta^*$. *Otherwise reject the move,* $\theta^{(t)} = \theta^{(t-1)}$.

3. *Repeat step 2.*

Example 30 *We shall now illustrate the results of Example 29 using a simulation exercise in R. The function* `metropolis` *(courtesy of Ruth Salway) uses the Metropolis-Hastings algorithm to sample from* $N(mu.p, sig.p^2)$ *with the proposal* $N(theta[t\text{-}1], sig.q^2)$.

```
metropolis = function (start=1, n=100, sig.q=1, mu.p=0, sig.p=1) {

#starting value theta_0 = start
#run for n iterations

  #set up vector theta[] for the samples
  theta=rep(NA,n)
  theta[1]=start

  #count the number of accepted proposals
  accept=0

  #Main loop
  for (t in 2:n) {

    #pick a candidate value theta* from N(theta[t-1], sig.q)
    theta.star = rnorm(1, theta[t-1], sqrt(sig.q))
```

```
    #generate a random number between 0 and 1:
    u = runif(1)

    # calculate acceptance probability:
    r = (dnorm(theta.star,mu.p,sig.p)) / (dnorm(theta[t-1],mu.p,sig.p) )

    a=min(r,1)

    #if u<a we accept the proposal; otherwise we reject
    if (u<a) {
        #accept
            accept=accept+1
            theta[t] = theta.star
    }
    else {
        #reject
        theta[t] = theta[t-1]
        }
  #end loop
  }
```

We illustrate two example runs from of the algorithm. Firstly, in Figures 3.1 - 3.3, for $\theta \,|\, x \sim N(0,1)$ *where the proposal distribution is normal with mean* θ *and chosen variance* $\sigma^2 = 1$ *and the starting point is* $\theta^{(0)} = 1$ *and secondly, in Figures 3.4 - 3.6, for* $\theta \,|\, x \sim N(0,1)$ *where the proposal distribution is normal with mean* θ *and chosen variance* $\sigma^2 = 0.36$ *and the starting point is* $\theta^{(0)} = 1$*. In each case we look at the results for nine, 100 and 5000 iterations from the sampler.*

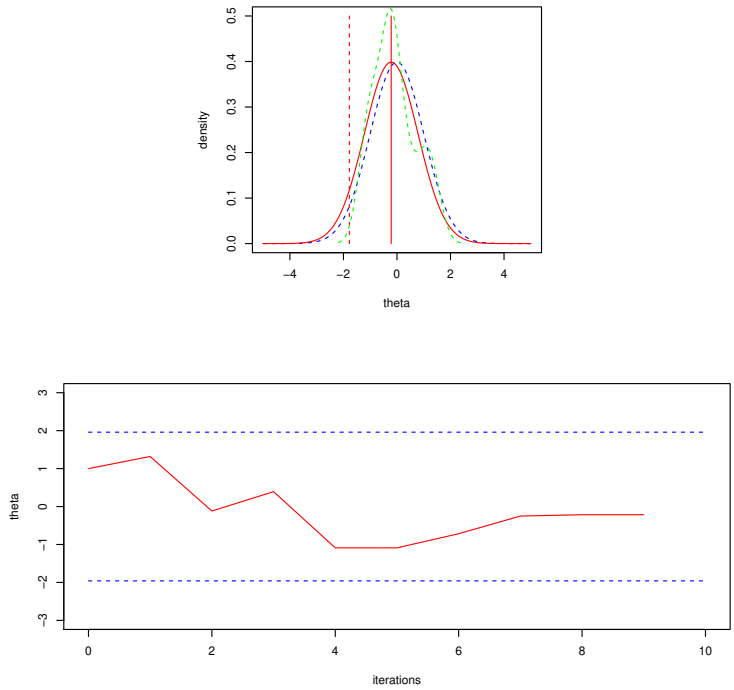


Figure 3.1: Nine iterations from a Metropolis-Hastings sampler for $\theta \,|\, x \sim N(0,1)$ where the proposal distribution is normal with mean θ and chosen variance $\sigma^2 = 1$ and the starting point is $\theta^{(0)} = 1$. The top plot shows the target distribution, $N(0,1)$, the proposal distribution for $\theta^{(9)}$ which is $N(-0.216, 1)$, the proposal $\theta^* = -1.779$ and the current observed density. The move is rejected; the bottom plot is the trace plot of $\theta^{(t)}$.

The following table summarises the stages of the algorithm for the iterations given in Figure 3.1.

t	1	2	3	4	5	6	7	8	9
θ^*	1.319	−0.118	0.392	−1.089	−1.947	−0.716	−0.251	−0.216	−1.779
Accept/Reject?	*Accept*	*Accept*	*Accept*	*Accept*	*Reject*	*Accept*	*Accept*	*Accept*	*Reject*
$\theta^{(t)}$	1.319	−0.118	0.392	−1.089	−1.089	−0.716	−0.251	−0.216	−0.216

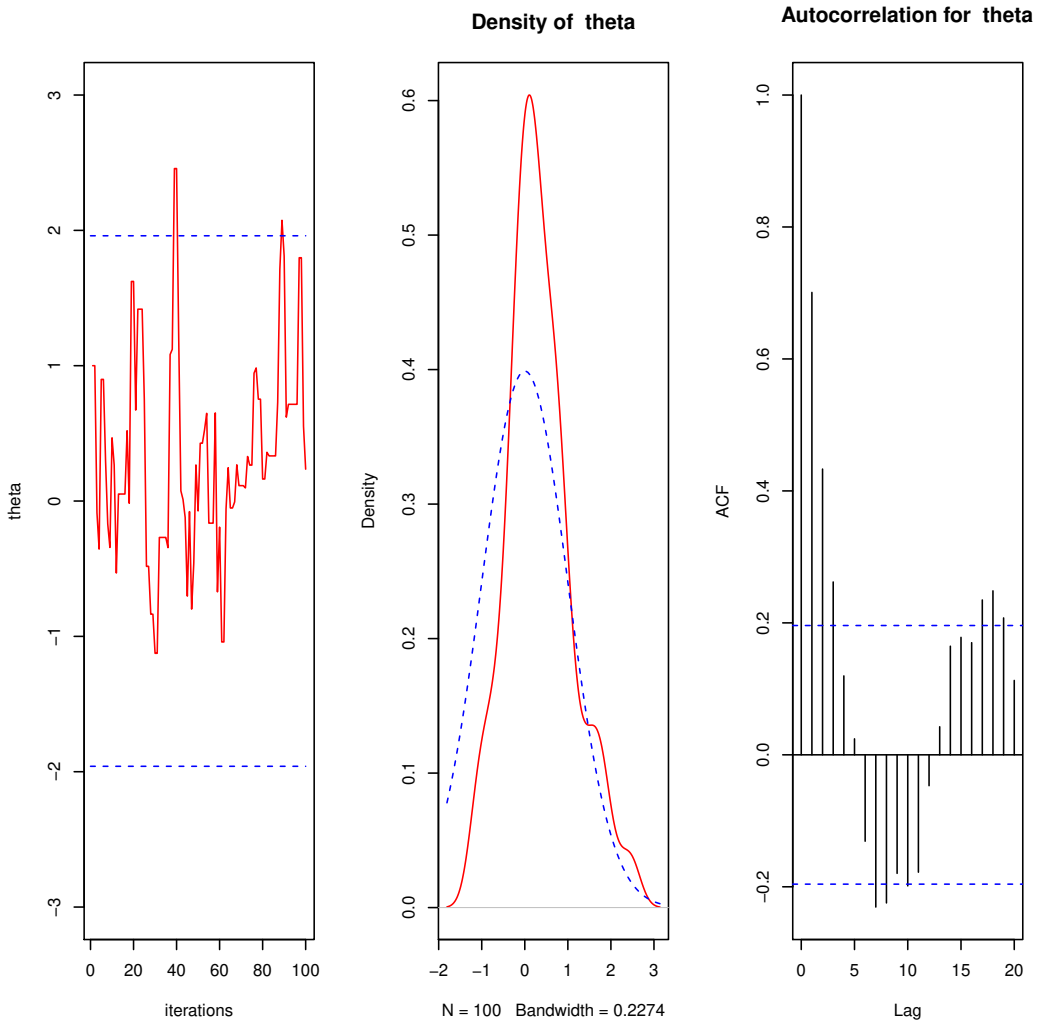


Figure 3.2: 100 iterations from a Metropolis-Hastings sampler for $\theta \,|\, x \sim N(0,1)$ where the proposal distribution is normal with mean θ and chosen variance $\sigma^2 = 1$ and the starting point is $\theta^{(0)} = 1$. The first plot shows the trace plot of $\theta^{(t)}$, the second the observed density against the target density and the third the autocorrelation plot. The sample mean is 0.3435006 and the sample variance is 0.5720027. 3% of points were observed to be greater than 1.96 and the acceptance rate is 0.66.

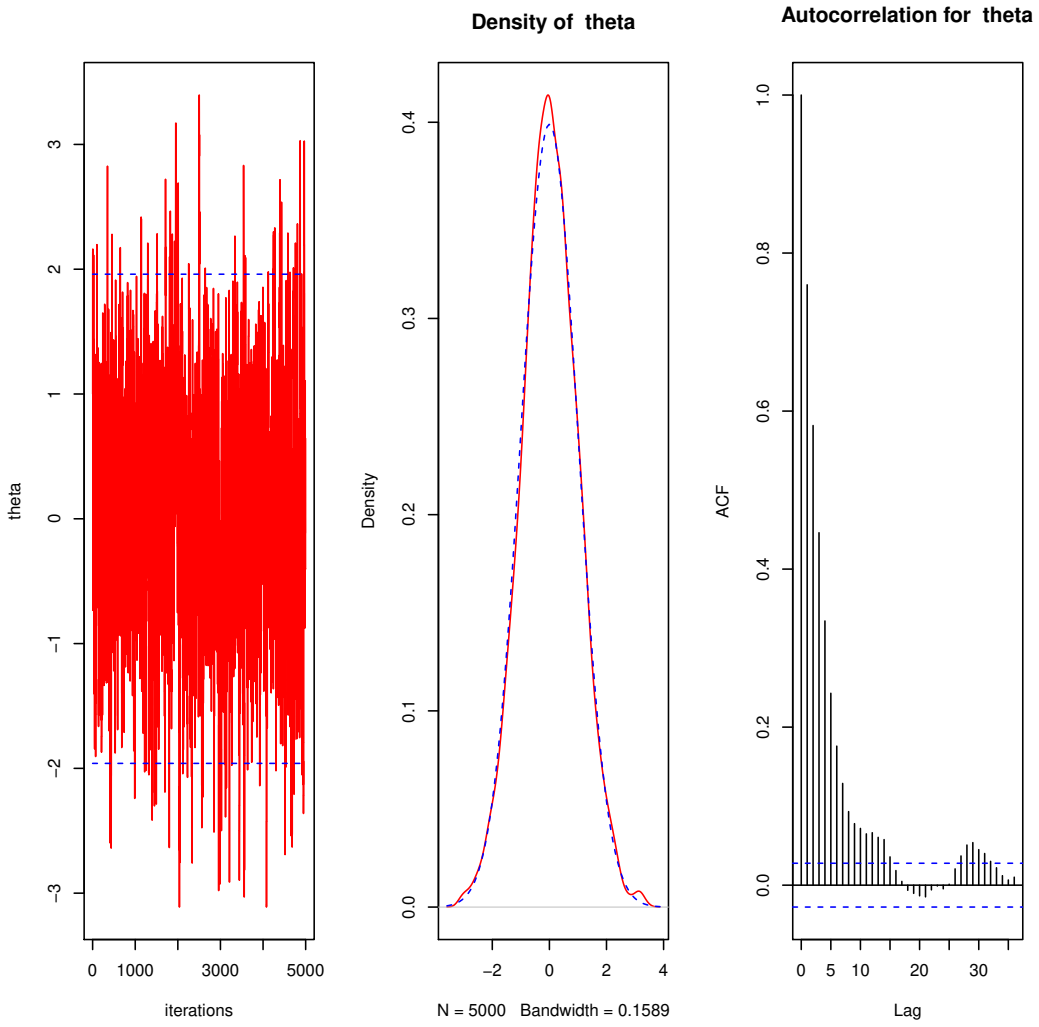


Figure 3.3: 5000 iterations from a Metropolis-Hastings sampler for $\theta \,|\, x \sim N(0,1)$ where the proposal distribution is normal with mean θ and chosen variance $\sigma^2 = 1$ and the starting point is $\theta^{(0)} = 1$. The first plot shows the trace plot of $\theta^{(t)}$, the second the observed density against the target density and the third the autocorrelation plot. The sample mean is 0.01340934 and the sample variance is 0.984014. 2.76% of points were observed to be greater than 1.96 and the acceptance rate is 0.7038.

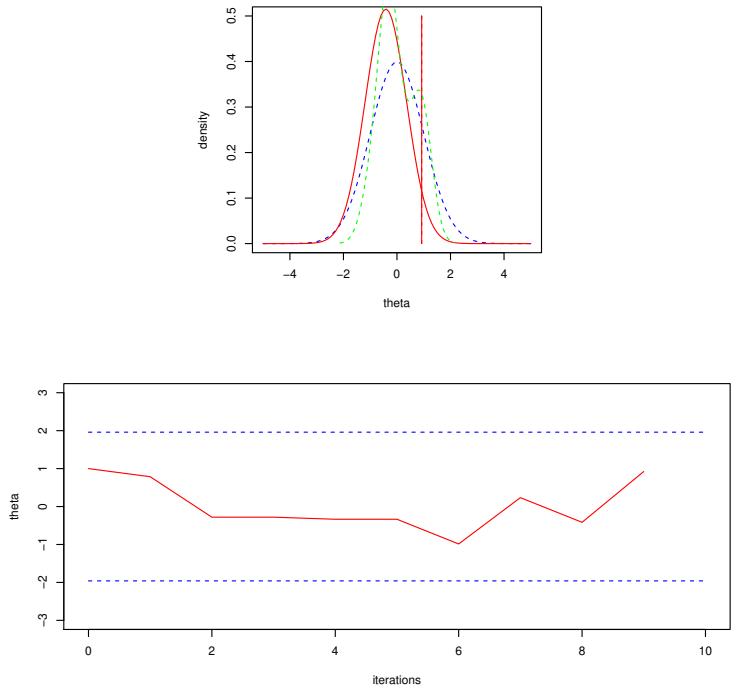


Figure 3.4: Nine iterations from a Metropolis-Hastings sampler for $\theta \,|\, x \sim N(0,1)$ where the proposal distribution is normal with mean θ and chosen variance $\sigma^2 = 0.36$ and the starting point is $\theta^{(0)} = 1$. The top plot shows the target distribution, $N(0,1)$, the proposal distribution for $\theta^{(9)}$ which is $N(-0.416, 0.36)$, the proposal $\theta^* = 0.923$ and the current observed density. The move is accepted; the bottom plot is the trace plot of $\theta^{(t)}$.

The following table summarises the stages of the algorithm for the iterations given in Figure 3.4.

t	1	2	3	4	5	6	7	8	9
θ^*	0.786	-0.280	-1.092	-0.335	-1.169	-0.987	0.234	-0.416	0.923
Accept/Reject?	*Accept*	*Accept*	*Reject*	*Accept*	*Reject*	*Accept*	*Accept*	*Accept*	*Accept*
$\theta^{(t)}$	0.786	-0.280	-0.280	-0.335	-0.335	-0.987	0.234	-0.416	0.923

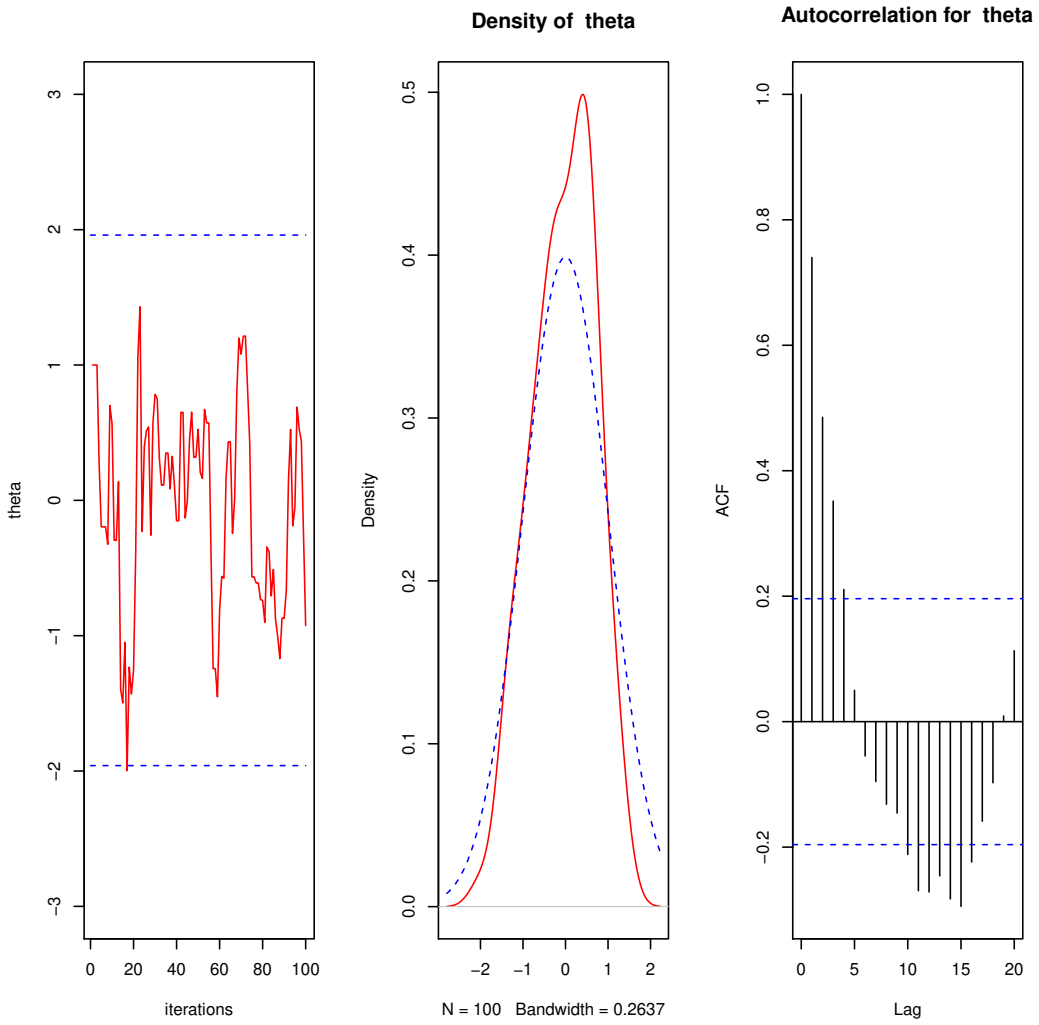


Figure 3.5: 100 iterations from a Metropolis-Hastings sampler for $\theta \,|\, x \sim N(0,1)$ where the proposal distribution is normal with mean θ and chosen variance $\sigma^2 = 0.36$ and the starting point is $\theta^{(0)} = 1$. The first plot shows the trace plot of $\theta^{(t)}$, the second the observed density against the target density and the third the autocorrelation plot. The sample mean is -0.04438661 and the sample variance is 0.5414882. 0% of points were observed to be greater than 1.96 and the acceptance rate is 0.83. Notice that, by comparing with Figure 3.2, reducing σ^2, that is we are proposing smaller jumps, increases the acceptance rate but reduces the mixing.

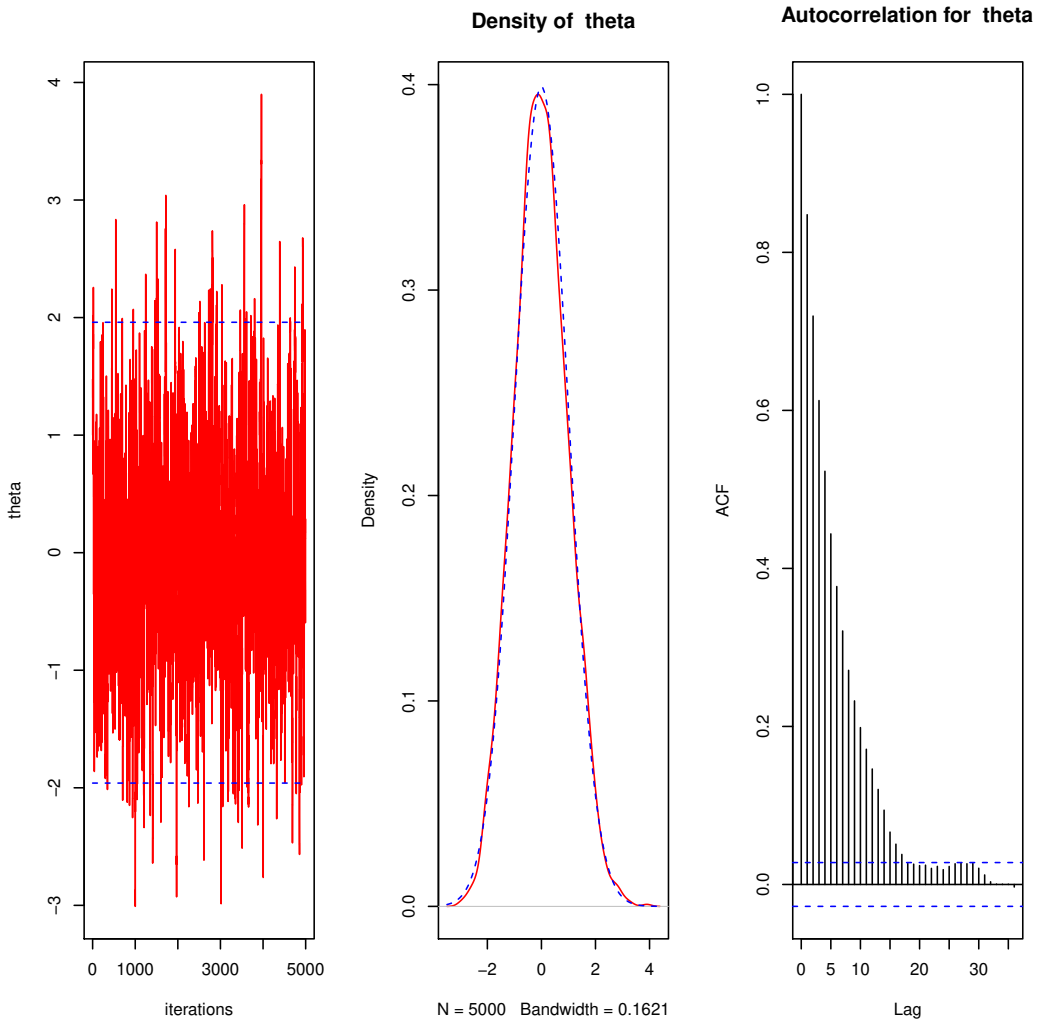


Figure 3.6: 5000 iterations from a Metropolis-Hastings sampler for $\theta \,|\, x \sim N(0,1)$ where the proposal distribution is normal with mean θ and chosen variance $\sigma^2 = 0.36$ and the starting point is $\theta^{(0)} = 1$. The first plot shows the trace plot of $\theta^{(t)}$, the second the observed density against the target density and the third the autocorrelation plot. The sample mean is -0.008158827 and the sample variance is 0.9792388. 2.42% of points were observed to be greater than 1.96 and the acceptance rate is 0.7894. Notice that, by comparing with Figure 3.3, reducing σ^2, that is we are proposing smaller jumps, increases the acceptance rate but reduces the mixing.

3.3.3 Gibbs sampler

The aim of the Gibbs[6] sampler is to make sampling from a high-dimensional distribution more tractable by sampling from a collection of more manageable smaller dimensional distributions.

Gibbs sampling is a MCMC scheme where the transition kernel is formed by the full conditional distributions. Assume that the distribution of interest is $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. Note that

1. Typically we will want $\pi(\theta) = f(\theta \,|\, x)$.
2. The components θ_i can be a scalar, a vector or a matrix. For simplicity of exposition you may regard them as scalars.

Let θ_{-i} denote the set $\theta \setminus \theta_i$ and suppose that the full conditional distributions $\pi_i(\theta_i) = \pi(\theta_i \,|\, \theta_{-i})$ $i = 1, \ldots, d$ are available and can be sampled from.

Algorithm 2 *The Gibbs sampler.*

1. *Choose an arbitrary starting point* $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$ *for which* $\pi(\theta^{(0)}) > 0$.
2.
   - *Obtain* $\theta_1^{(t)}$ *from conditional distribution* $\pi(\theta_1 \,|\, \theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)})$
   - *Obtain* $\theta_2^{(t)}$ *from conditional distribution* $\pi(\theta_2 \,|\, \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)})$

   $\vdots$

   - *Obtain* $\theta_p^{(t)}$ *from conditional distribution* $\pi(\theta_p \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)})$

   $\vdots$

   - *Obtain* $\theta_d^{(t)}$ *from conditional distribution* $\pi(\theta_d \,|\, \theta_1^{(t)}, \ldots, \theta_{d-1}^{(t)})$
3. *Repeat step 2.*

The algorithm is run until convergence is reached[7]. When convergence is reached, the resulting value $\theta^{(j)}$ is a realisation (or draw) from $\pi(\theta)$.

Example 31 *Suppose that the joint distribution of* $\theta = (\theta_1, \theta_2)$ *is given by*

$$f(\theta_1, \theta_2) \quad \propto \quad \binom{n}{\theta_1} \theta_2^{\theta_1 + \alpha - 1} (1 - \theta_2)^{n - \theta_1 + \beta - 1}$$

for $\theta_1 \in [0, 1, \ldots, n]$ *and* $0 \leq \theta_2 \leq 1$. *Hence,* θ_1 *is discrete and* θ_2 *is continuous. Suppose that we are interested in calculating some characteristic of the marginal distribution of* θ_1. *The Gibbs sampler will allow us to generate a sample from this marginal distribution in the following way.*

[6] Josiah Willard Gibbs (1839-1903).

[7] We'll look at more detail into strategies for judging this is a couple of lectures time.

The conditional distributions are

$$\begin{aligned} \theta_1 \,|\, \theta_2 &\sim Bin(n, \theta_2) \\ \theta_2 \,|\, \theta_1 &\sim Beta(\theta_1 + \alpha, n - \theta_1 + \beta) \end{aligned}$$

Assuming the hyperparameters n, α and β are known then it is straightforward to sample from these distributions and we now generate a "Gibbs sequence" of random variables.

- ***Step 1:*** *Specify an initial value $\theta_1^{(0)}$ and either specify an initial $\theta_2^{(0)}$ or sample it*[8] *from $\pi(\theta_2 \,|\, \theta_1^{(0)})$, the $Beta(\theta_1^{(0)} + \alpha, n - \theta_1^{(0)} + \beta)$ distribution, to obtain $\theta_2^{(0)}$.*
- ***Step 2.1.(a):*** *Obtain $\theta_1^{(1)}$ from sampling from $\pi(\theta_1 \,|\, \theta_2^{(0)})$, the $Bin(n, \theta_2^{(0)})$ distribution.*
- ***Step 2.1.(b):*** *Obtain $\theta_2^{(1)}$ from sampling from $\pi(\theta_2 \,|\, \theta_1^{(1)})$, the $Beta(\theta_1^{(1)} + \alpha, n - \theta_1^{(1)} + \beta)$ distribution.*
- ***Step 2.2.(a):*** *Obtain $\theta_1^{(2)}$ from sampling from $\pi(\theta_1 \,|\, \theta_2^{(1)})$, the $Bin(n, \theta_2^{(1)})$ distribution.*
- ***Step 2.2.(b):*** *Obtain $\theta_2^{(2)}$ from sampling from $\pi(\theta_2 \,|\, \theta_1^{(2)})$, the $Beta(\theta_1^{(2)} + \alpha, n - \theta_1^{(2)} + \beta)$ distribution.*

We continue run through the algorithm to obtain the sequence of values

$$\theta_1^{(0)}, \theta_2^{(0)}, \theta_1^{(1)}, \theta_2^{(1)}, \theta_1^{(2)}, \theta_2^{(2)}, \ldots, \theta_1^{(k)}, \theta_2^{(k)}$$

which may be expressed as $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)}$ where $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$. The sequence is such that the distribution of $\theta_1^{(k)}$ tends to $f(\theta_1)$ as $k \to \infty$ and, similarly, the distribution of $\theta_2^{(k)}$ tends to $f(\theta_2)$ as $k \to \infty$.

To allow the sequence to converge to the stationary distribution, we first remove what are termed the "burn-in" samples. Suppose we judge this to be after time b then we discard $\{\theta_1^{(t)}, \theta_2^{(t)} : t \leq b\}$ and the values $\{\theta_1^{(t)}, \theta_2^{(t)} : t > b\}$ may be considered a sample from the required distribution. So, for summaries of θ_1 we use the sample $\{\theta_1^{(t)} : t > b\}$.

Example 32 *We shall now illustrate the results of Example 31 using a simulation exercise in R. The function* `gibbs` *uses the Gibbs sampler to sample from the conditionals $x1 \,|\, x2 \sim Bin(n, x2)$, $x2 \,|\, x1 \sim Beta(x1 + \alpha, n - x1 + \beta)$.*

```
gibbs = function(n=1000 , m=10 , a = 1, b= 1, start1=5 , start2=0.5 )
 {
hbreaks <- c(0:(m +1)) - 0.5
xpos <- c(0:m)
```

[8] We can do this because we have $d = 2$ variables. In the general case, we could specify $\theta_1^{(0)}, \ldots, \theta_{d-1}^{(0)}$ and then sample $\theta_d^{(0)}$ from $\pi(\theta_d \,|\, \theta_1^{(0)}, \ldots, \theta_{d-1}^{(0)})$.

```
ypos <- c(0:m)
for (i in 1:(m+1))
{
ypos[i] <- choose(m, xpos[i])*beta(xpos[i]+a, m-xpos[i]+b)/beta(a,b)
}
x <- seq(0,1,length=100)
y <- dbeta(x,a,b)

  x1 <- c(1:n)
  x2 <- c(1:n)
  x1[1] <- start1
  x2[1] <- start2
  par( mfrow=c( 3,2 ) )
  for ( i in 2:n )
   {
    x1[i] <- rbinom( 1 , m, x2[i-1])
    x2[i] <- rbeta(1, a + x1[i], m - x1[i] + b)

  plot( x1[1:i] , type="l" , xlab="Iteration" ,xlim=c(1,n),ylab="x1",
main="Trace plot for x1")
  plot( x2[1:i] , type="l" , xlab="Iteration",xlim=c(1,n),ylab="x2",
main="Trace plot for x2" )
  plot( x1[1:i] , x2[1:i] , type="p" , pch="." ,cex=2,xlab="x1",ylab="x2",
main="Plot of each (x1,x2)")
  plot( x1[1:i] , x2[1:i] , type="l" ,col="red",xlab="x1",ylab="x2",
main = "Component-by-component updating")
hist(x1[1:i],prob=T,breaks=hbreaks,xlab="x1",
main="Histogram of x1 with target density")
lines(xpos,ypos)
hist(x2[1:i],prob=T,xlab="x2",main="Histogram of x2 with target density")
lines(x,y,col="black")
 }
}
```

We illustrate two example runs from of the algorithm. Firstly, in Figures 3.7 - 3.8, for $n = 10$ *and* $\alpha = \beta = 1$ *and secondly, in Figures 3.9 - 3.10, for* $n = 10$ *and* $\alpha = 2$ *and* $\beta = 3$ *In each case we look at the results for fifty and 1500 iterations from the sampler.*

Notice that the Gibbs sampler for $\theta = (\theta_1, \ldots, \theta_d)$ can be viewed as a special case of the Metropolis-Hastings algorithm where each iteration t consists of d Metropolis-Hastings steps each with an acceptance probability of 1. We shall show this in question 4 of Question Sheet

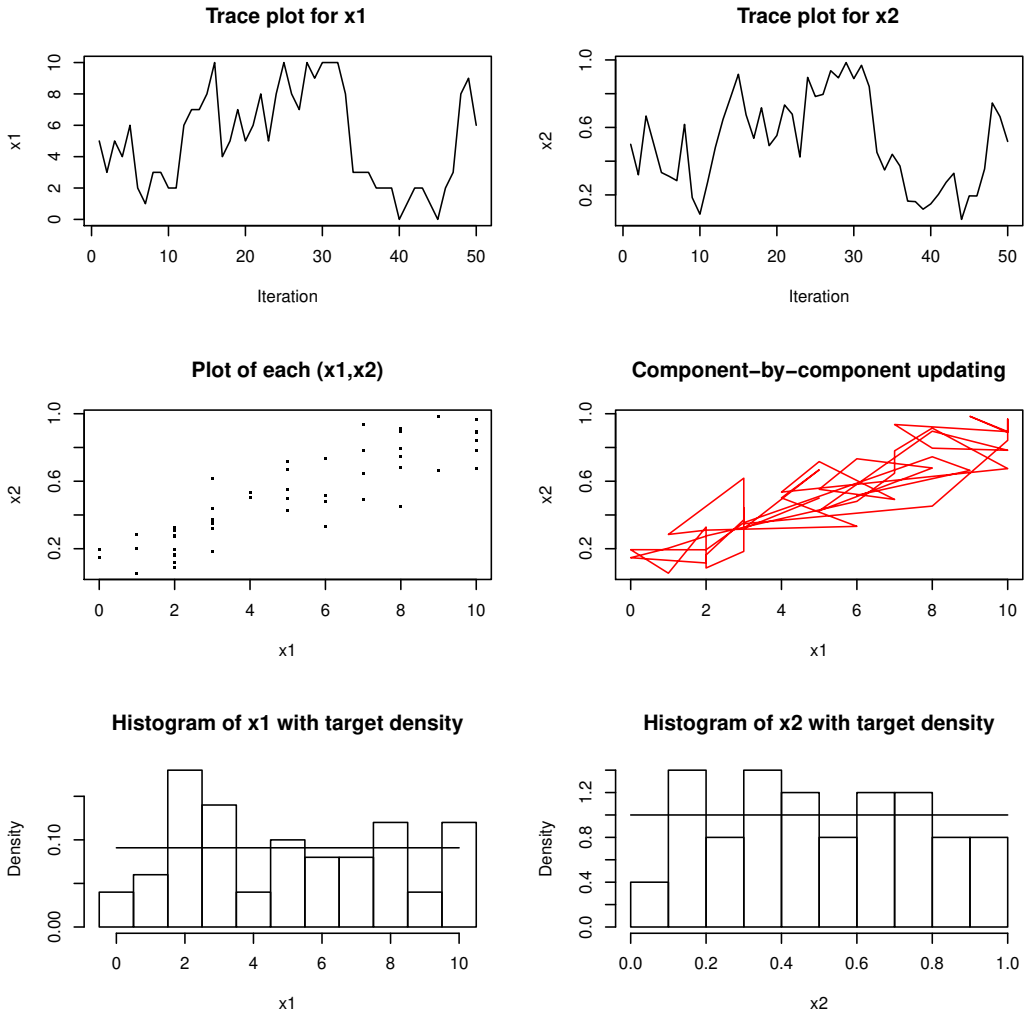


Figure 3.7: Fifty iterations from a Gibbs sampler for $x_1 \,|\, x_2 \sim Bin(n, x_2)$, $x_2 \,|\, x_1 \sim Beta(x_1 + \alpha, n - x_1 + \beta)$ where $n = 10$ and $\alpha = \beta = 1$. The marginal distributions are $x_1 \sim Beta - binomial(n, \alpha, \beta)$ and $x_2 \sim Beta(\alpha, \beta)$. For $\alpha = \beta = 1$, x_1 is the discrete uniform on $\{0, 1, \ldots, n\}$ and $x_2 \sim U(0, 1)$.

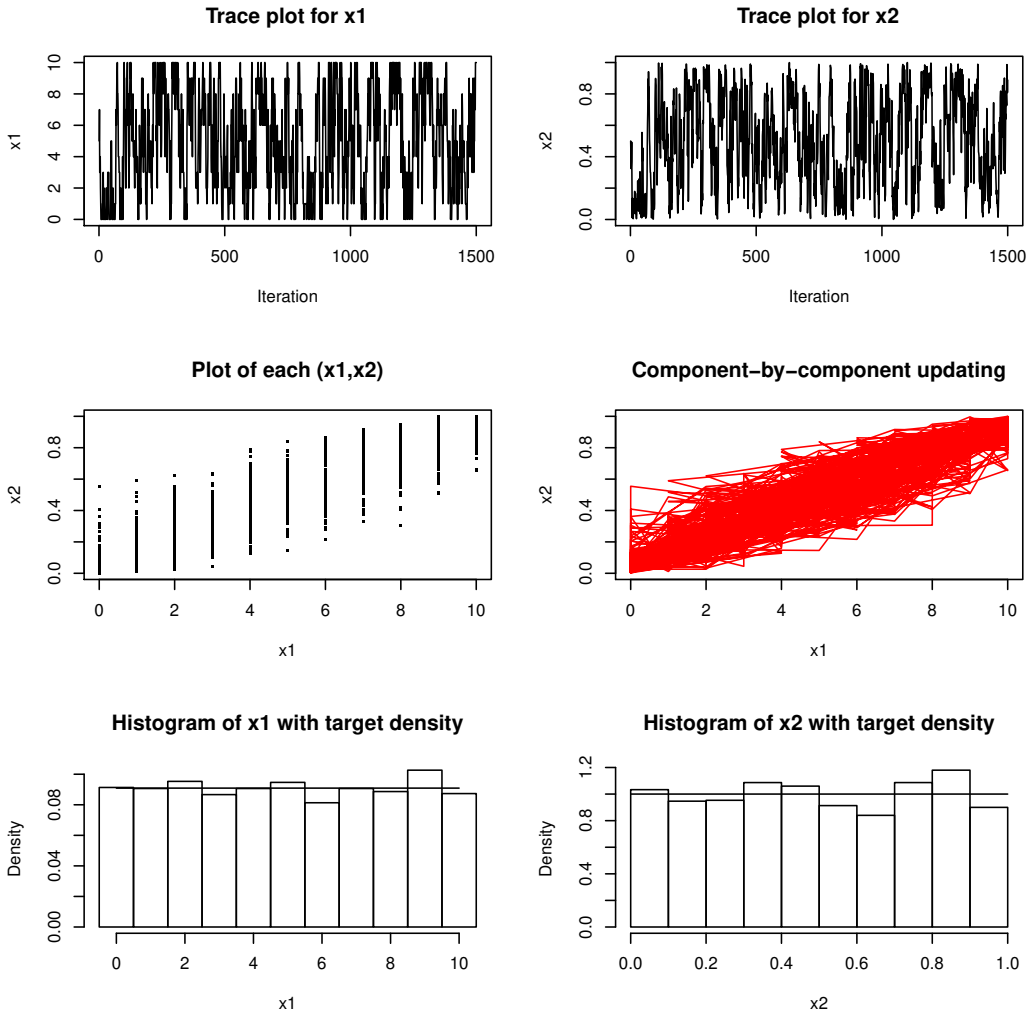


Figure 3.8: 1500 iterations from a Gibbs sampler for $x_1 \,|\, x_2 \sim Bin(n, x_2)$, $x_2 \,|\, x_1 \sim Beta(x_1 + \alpha, n - x_1 + \beta)$ where $n = 10$ and $\alpha = \beta = 1$. The marginal distributions are $x_1 \sim Beta - binomial(n, \alpha, \beta)$ and $x_2 \sim Beta(\alpha, \beta)$. For $\alpha = \beta = 1$, x_1 is the discrete uniform on $\{0, 1, \ldots, n\}$ and $x_2 \sim U(0, 1)$.

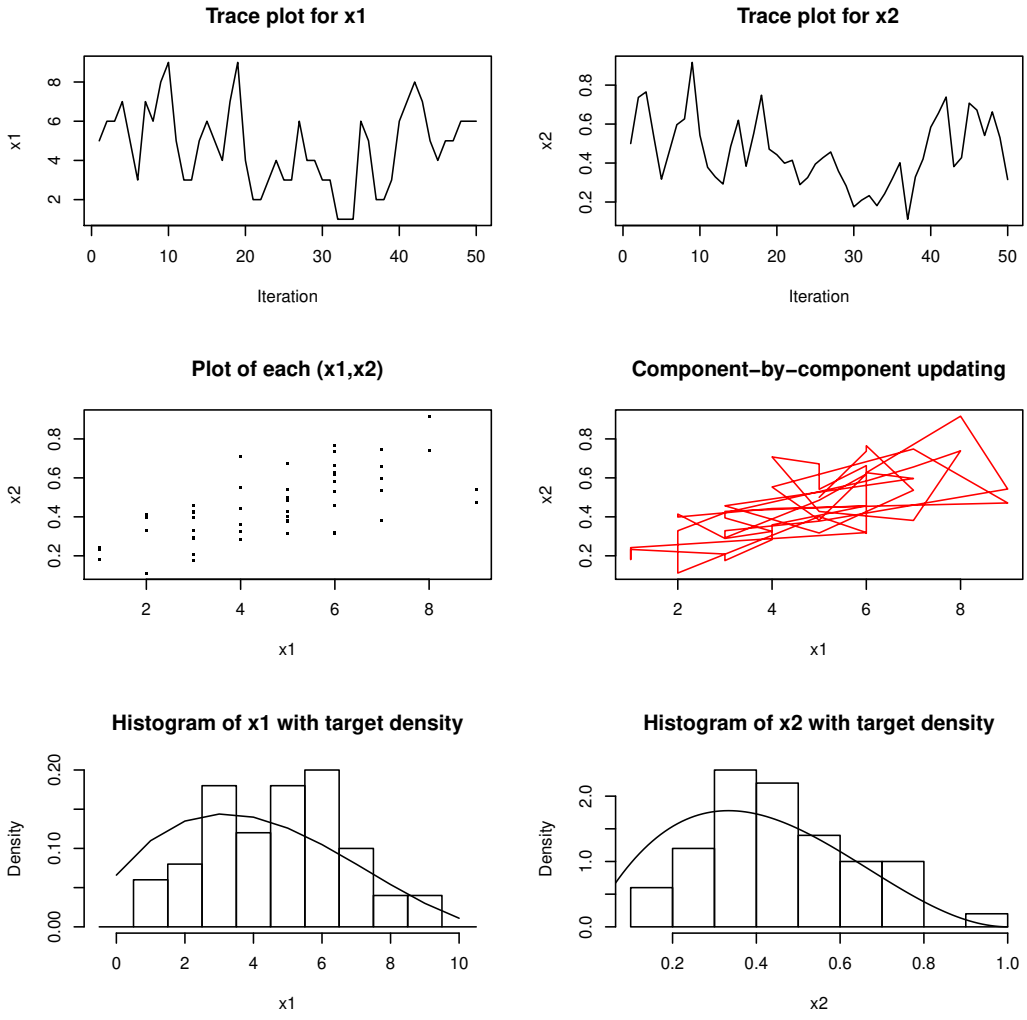


Figure 3.9: Fifty iterations from a Gibbs sampler for $x_1 \,|\, x_2 \sim Bin(n, x_2)$, $x_2 \,|\, x_1 \sim Beta(x_1 + \alpha, n - x_1 + \beta)$ where $n = 10$, $\alpha = 2$ and $\beta = 3$. The marginal distributions are $x_1 \sim Beta - binomial(n, \alpha, \beta)$ and $x_2 \sim Beta(\alpha, \beta)$.

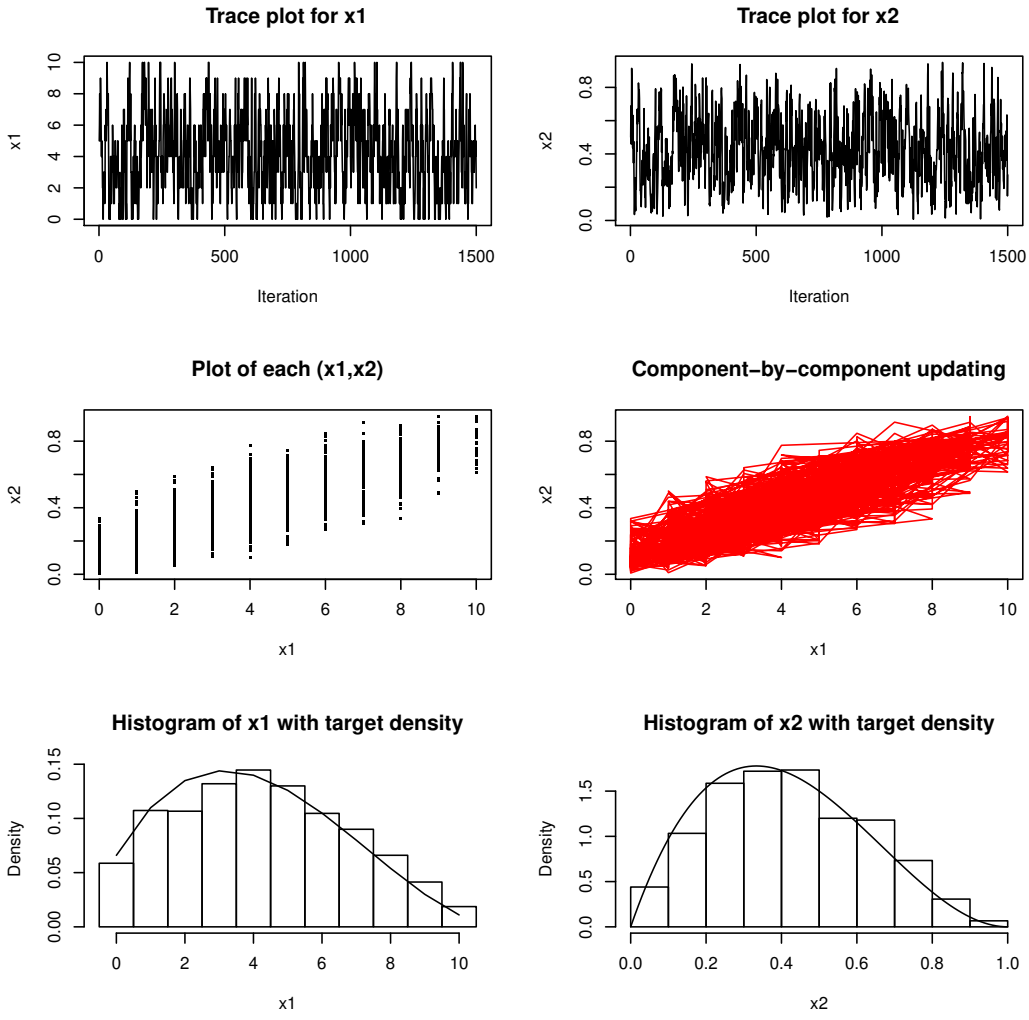


Figure 3.10: 1500 iterations from a Gibbs sampler for $x_1 \,|\, x_2 \sim Bin(n, x_2)$, $x_2 \,|\, x_1 \sim Beta(x_1 + \alpha, n - x_1 + \beta)$ where $n = 10$, $\alpha = 2$ and $\beta = 3$. The marginal distributions are $x_1 \sim Beta - binomial(n, \alpha, \beta)$ and $x_2 \sim Beta(\alpha, \beta)$.

Eight.

3.3.4 A brief insight into why the Metropolis-Hastings algorithm works

A topic of major importance when studying Markov chains is determining the conditions under which there exists a stationary distribution and what that stationary distribution is. In MCMC methods, the reverse approach is taken. We have a stationary distribution π (typically a posterior distribution from which we wish to sample) and we want to find the transition probabilities of a Markov chain which has π as the stationary distribution.

In this section we shall motivate why the Metropolis-Hastings works. For ease of understanding we shall work in the discrete setting. The case when the state space of the Markov chain is continuous follows very similarly, largely by the usual approach of replacing sums by integrals.

Consider a discrete Markov chain X_t with transition probabilities, see (3.6),

$$P_{\theta\phi} \quad = \quad P(X_t = \phi \,|\, X_{t-1} = \theta)$$

for all θ, $\phi \in S$, the state space of the chain. To make the link between the discrete and continuous case more explicit we shall let

$$P_{\theta\phi} = q_1(\phi \,|\, \theta).$$

Suppose that the chain has a stationary distribution, for example from Theorem 3 a sufficient condition is that the chain is irreducible and aperiodic, π. From Definition 14[9], the stationary distribution $\pi = \{\pi(\theta) : \theta \in S\}$ satisfies

$$\pi(\phi) \quad = \quad \sum_{\theta \in S} \pi(\theta) q_1(\phi \,|\, \theta). \tag{3.10}$$

In the continuous case we have

$$\pi(\phi) \quad = \quad \int_{\theta} \pi(\theta) q_1(\phi \,|\, \theta)\, d\theta$$

highlighting the replacement of sums by integrals.

Notice that, as the chain must move somewhere,

$$\sum_{\phi \in S} q_1(\phi \,|\, \theta) \quad = \quad 1. \tag{3.11}$$

If the equation

$$\pi(\theta) q_1(\phi \,|\, \theta) \quad = \quad \pi(\phi) q_1(\theta \,|\, \phi) \tag{3.12}$$

[9] Think of $\pi_i = \pi(\theta)$, $\pi_j = \pi(\phi)$ and $P_{ij} = P_{\theta\phi} = q_1(\phi \,|\, \theta)$.

is satisfied for all θ, ϕ then the $q_1(\phi \,|\, \theta)$ satisfy **time reversibility** or **detailed balance**. If (3.12) holds then the $q_1(\phi \,|\, \theta)$ are the transition probabilities of a Markov chain with stationary distribution π since

$$\begin{aligned} \sum_{\theta \in S} \pi(\theta) q_1(\phi \,|\, \theta) &= \sum_{\theta \in S} \pi(\phi) q_1(\theta \,|\, \phi) && (3.13) \\ &= \pi(\phi) \sum_{\theta \in S} q_1(\theta \,|\, \phi) \\ &= \pi(\phi) && (3.14) \end{aligned}$$

which is (3.10), where (3.13) follows from (3.12) and (3.14) from (3.11).

We now demonstrate how, by considering detailed balance, we can derive the Metropolis-Hastings algorithm. Let $q(\phi \,|\, \theta)$ denote the proposal density of a scheme where we wish to sample from $\pi(\theta)$. Typically, $q(\cdot)$ will be easy to sample from and the sufficient conditions for the existence of a stationary distribution (aperiodicity and irreducibility), $\pi(\theta)$, are mopped up in $q(\cdot)$. In a Bayesian context, we will usually wish to sample from the posterior so that $\pi(\theta) = f(\theta \,|\, x)$. We propose a move from θ to ϕ with probability

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

The transition probabilities of the Markov chain created using the Metropolis-Hastings algorithm are

$$q_1(\phi \,|\, \theta) = \begin{cases} q(\phi \,|\, \theta)\alpha(\theta, \phi) & \phi \neq \theta \\ 1 - \sum_{\phi \neq \theta} q(\phi \,|\, \theta)\alpha(\theta, \phi) & \phi = \theta. \end{cases}$$

(There are two ways for $q_1(\theta \,|\, \theta)$: either we reject a proposed move to $\phi \neq \theta$ or we sample $\theta = \phi$.) We now show that $q_1(\phi \,|\, \theta)$ satisfies detailed balance. For $\phi \neq \theta$ we have

$$\begin{aligned} \pi(\theta) q_1(\phi \,|\, \theta) &= \pi(\theta) q(\phi \,|\, \theta) \alpha(\theta, \phi) \\ &= \pi(\theta) q(\phi \,|\, \theta) \min\left(1, \frac{\pi(\phi) q(\theta \,|\, \phi)}{\pi(\theta) q(\phi \,|\, \theta)}\right) \\ &= \min\left(\pi(\theta) q(\phi \,|\, \theta), \pi(\phi) q(\theta \,|\, \phi)\right) \\ &= \pi(\phi) q(\theta \,|\, \phi) \min\left(1, \frac{\pi(\theta) q(\phi \,|\, \theta)}{\pi(\phi) q(\theta \,|\, \phi)}\right) \\ &= \pi(\phi) q(\theta \,|\, \phi) \alpha(\phi, \theta) \\ &= \pi(\phi) q_1(\theta \,|\, \phi). \end{aligned}$$

Thus, using (3.12), $\pi(\theta)$ is the stationary distribution of the Markov chain created using the Metropolis-Hastings algorithm with proposal distribution $q(\phi \,|\, \theta)$.

3.3.5 Efficiency of the algorithms

The efficiency of the Metropolis-Hastings algorithm depends upon a "good" choice of proposal density $q(\phi \,|\, \theta)$. In practice we want

1. A high acceptance probability: we don't want to repeatedly sample points we reject. Practical experience suggests that a "good" $q(\phi \,|\, \theta)$ will be close to $f(\theta \,|\, x)$ but slightly heavier in the tails.

2. To explore the whole distribution: we want a good coverage of all possible values of θ and not to spend too long in only a small area of the distribution.

A chain that does not move around very quickly is said to be **slow mixing**.

The Gibbs sampler accepts all new points so we don't need to worry about the acceptance probability. However, we may experience slow mixing if the parameters are highly correlated. Ideally we want the parameters to be as close to independence as possible as sampling from conditionals reduces to sampling directly from the desired marginals.

Example 33 *Suppose that we want to use a Gibbs sampler to sampler from the posterior* $f(\theta \,|\, x)$ *where* $\theta = (\theta_1, \ldots, \theta_d)$. *For the Gibbs sampler we sample from the conditionals* $f(\theta_i \,|\, \theta_{-i}, x)$ *for* $i = 1, \ldots, d$ *where* $\theta_{-i} = \theta \setminus \theta_i$. *If the* θ_i *are independent then*

$$f(\theta \,|\, x) \quad = \quad \prod_{i=1}^{d} f(\theta_i \,|\, x)$$

and $f(\theta_i \,|\, \theta_{-i}, x) = f(\theta_i \,|\, x)$. *The Gibbs sampler will sample from the posterior marginal densities.*

In some cases, it may be worthwhile to deal with reparameterisations of θ, for example using a reparameterisation which makes the parameters close to independent.

3.3.6 Using the sample for inference

The early values of a chain, before convergence, are called the **burn-in**. The length of burn-in will depend upon the rate of convergence of the Markov chain and how close to the posterior we need to be. In general, it is difficult to obtain estimates of the rate of convergence and so analytically determining the length of burn-in required is not feasible. A practitioner uses times series plots of the chain as a way of judging convergence. Suppose that we judge convergence to have been reached after b iterations of an MCMC algorithm have been performed. We discard the observations $\theta^{(1)}, \ldots, \theta^{(b)}$ and work with the observations $\{\theta^{(t)} : t > b\}$ which are viewed as being a sample from the stationary distribution of the Markov chain which is typically the posterior distribution.

4 Decision theory

Any situation in which choices are to made amongst two or more possible alternative courses of action is a decision problem. If we are certain of the consequences of these choices, or **actions**, then making a decision is relatively straightforward. You can simply write down all of the possible options and choose the decision which you like the best. When the consequences are uncertain then the problem is much less straightforward and consequently much more interesting. We study how decisions **ought**[1] to be made in these circumstances, in short what is the optimal decision?

Let $\mathcal{D}$ be the class of possible decisions. For each $d \in \mathcal{D}$ let Θ be the set of relevant events which affect the result of choosing d. It is often helpful to think of each $\theta \in \Theta$ as describing a state of nature so the actual value denotes the "true state of nature".

Having chosen d we need a way of assessing the consequence of how good or bad the choice of decision d was under the event θ. This measurement is called your **utility**.

We will largely restrict our attention to statistical decision theory where we regard inference as a decision problem. We can typically think of each d as representing a method of estimating θ so that the utility measures how good or bad the estimation procedure is.

It is clear that the utility will depend upon context. For example, in some cases it may be much worse to overestimate a parameter θ than to underestimate it whilst in others in may be equally serious to under or over estimate the parameter. Thus, the optimal decision will also depend upon context.

4.1 Utility

You can think about the pair (θ, d) as defining your status, S, and denote the utility for this status by $u(S)$. We now consider how such a utility can be obtained.

Suppose that $S_1, \ldots, S_n$ constitute a collection of n statuses. We shall assume that you can compare these statuses. So, for any two statuses S_i, S_j we write

- $S_j \prec^\star S_i$, you prefer S_i to S_j, if you would pay an amount of money (however small) in order to swap S_j for S_i.

[1] The approach is prescriptive (that is how people should behave rationally in making decisions which satisfy some criteria) rather than descriptive (that is, studying decisions that people actually make).

- $S_j \sim^\star S_i$, you are indifferent between S_i and S_j, if neither $S_j \prec^\star S_i$ or $S_i \prec^\star S_j$ hold.
- $S_j \preceq^\star S_i$, S_i is at least as good as S_j, if one of $S_j \prec^\star S_i$ or $S_j \sim^\star S_i$ holds.

Notice that this gives us a framework for comparing anything.

Example 34 *A bakery has five types of cake available and I assert that*

$$\textit{fruit cake} \prec^\star \textit{carrot cake} \prec^\star \textit{banana cake} \prec^\star \textit{chocolate cake} \prec^\star \textit{cheese cake}.$$

Thus, I would be willing to pay to exchange a fruit cake for a carrot cake and then pay again to exchange the carrot cake for a banana cake and so on.

We make two assumptions about our preferences over statuses. Suppose that $S_1, S_2, \ldots, S_n$ constitute a collection of n statuses. We assume

1. (COMPARABILITY) For any S_i, S_j exactly one of $S_i \prec^\star S_j$, $S_j \prec^\star S_i$, $S_i \sim^\star S_j$ holds.
2. (COHERENCE) If $S_i \prec^\star S_j$ and $S_j \prec^\star S_k$ then $S_i \prec^\star S_k$.

Comparability ensures that we can express a preference between any two rewards.

Example 35 *Suppose that I didn't have coherence over my preferences for cakes. For example, consider that I assert carrot cake $\prec^\star$ banana cake and banana cake $\prec^\star$ chocolate cake but that chocolate cake $\prec^\star$ carrot cake. Then I would pay money to swap from carrot cake to banana cake and then from banana cake to chocolate cake. I am then are willing to pay to switch the chocolate cake for a carrot cake. I am back in my original position, but I have spent money to maintain this status quo. I am a* ***money pump****.*

The consequence of these assumptions is that, for a collection of n statuses $S_1, S_2, \ldots, S_n$, there is a labelling $S_{(1)}, S_{(2)}, \ldots, S_{(n)}$ such that

$$S_{(1)} \preceq^\star S_{(2)} \preceq^\star \cdots \preceq^\star S_{(n)}.$$

This is termed a **preference ordering** for the statuses. In particular, there is a worst state $S_{(1)}$ and a best state $S_{(n)}$. Notice that these need not necessary be unique.

In many situations, we are not certain as to which state will occur. This can be viewed as a **gamble**. We write

$$G \quad = \quad p_1 S_1 +_g p_2 S_2 +_g \cdots +_g p_n S_n$$

for the gamble that returns S_1 with probability p_1, S_2 with probability p_2, ..., S_n with probability p_n. We make two assumptions to ensure that our gambles are coherently compared.

1. If $S_j \preceq^\star S_i$, $p < q$ then $pS_i +_g (1-p)S_j \preceq^\star qS_i +_g (1-q)S_j$.

2. If $S_j \preceq^\star S_i$ then $pS_j +_g (1-p)S_k \preceq^\star pS_i +_g (1-p)S_k$ for any S_k.

Gambles provide the link between probability, preference and utility.

Definition 15 *(Utility)*
A utility function $u(\cdot)$ on gambles $G = p_1S_1 +_g p_2S_2 +_g \cdots +_g p_nS_n$ over statuses $S_1, S_2, \ldots, S_n$ assigns a real number $u(G)$ to each G subject to the following conditions

1. *Let G_i, G_j be any two gambles. If $G_j \prec^\star G_i$ then $u(G_j) < u(G_i)$, and if $G_j \sim^\star G_i$ then $u(G_j) = u(G_i)$.*

2. *For any $p \in [0,1]$ and any statuses A, B,*

$$u(pA +_g (1-p)B) \quad = \quad pu(A) + (1-p)u(B).$$

- Condition 1. says that utilities agree with preferences, so you choose the gamble with the highest utility.

- Condition 2. says that, for the generic gamble $G = p_1S_1 +_g p_2S_2 +_g \cdots +_g p_nS_n$, $u(G) = p_1u(S_1) + p_2u(S_2) + \cdots + p_nu(S_n)$. Hence, $u(G) = E\{u(G)\}$.

  i.e. Expected utility of a gamble = Actual utility of that gamble.

- Conditions 1. and 2. combined imply that we **choose the gamble with the highest expected utility**. So, **if** we can specify a utility function over statuses, we can solve any decision problem by choosing the decision which maximises expected utility.

Notice that a utility function over an ordered set of statuses $S_{(1)} \preceq^\star S_{(2)} \preceq^\star \cdots \preceq^\star S_{(n)}$ is often constructed by setting $u(S_{(1)}) = 0$ and $u(S_{(n)}) = 1$ and, for each $1 < i < n$, defining $u(S_{(i)})$ to be the probability p such that

$$S_{(i)} \sim^\star (1-p)S_{(1)} +_g pS_{(n)}.$$

Thus, p is the probability where you are indifferent between a guaranteed status of $S_{(i)}$ and a gamble with gives status $S_{(n)}$ with probability p and $S_{(1)}$ with probability $(1-p)$. p is often termed the indifference probability for status $S_{(i)}$. A utility function is unique up to a positive linear transformation.

4.2 Statistical decision theory

For any parameter value θ and decision d we have $u(\theta, d)$, the utility of choosing d when θ is the true value. We shall define loss to be

$$L(\theta, d) \quad = \quad -u(\theta, d). \tag{4.1}$$

A statistical decision has a number of ingredients.

1. The possible values of the parameter: Θ, the parameter space.

2. The set of possible decisions: $\mathcal{D}$, the decision space.

3. The probability distribution on Θ, $\pi(\theta)$. For example,

   (a) this could be a prior distribution, $\pi(\theta) = f(\theta)$.

   (b) this could be a posterior distribution, $\pi(\theta) = f(\theta \,|\, x)$ following the receipt of some data x.

   (c) this could be a posterior distribution $\pi(\theta) = f(\theta \,|\, x, y)$ following the receipt of some data x.

4. The loss function $L(\theta, d)$.

From (4.1), the decision which maximises the expected utility is the one which minimises the expected loss. Thus, we choose d to minimise

$$\rho(\pi, d) \quad = \quad \int_\theta L(\theta, d)\pi(\theta)\, d\theta \tag{4.2}$$

the risk of d under $\pi(\theta)$. The decision problem is completely specified by $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$.

Definition 16 *(Bayes rule and Bayes risk)*
The Bayes risk $\rho^*(\pi)$ minimises the expected loss,

$$\rho^*(\pi) \quad = \quad \inf_{d \in \mathcal{D}} \rho(\pi, d)$$

with respect to $\pi(\theta)$. A decision $d^* \in \mathcal{D}$ for which $\rho(\pi, d^*) = \rho^*(\pi)$ is a Bayes (decision) rule against $\pi(\theta)$.

The Bayes rule may not be unique, and in weird cases it might not exist. Typically, we solve $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ by finding $\rho^*(\pi)$ and (at least one) d^*.

Example 36 *Quadratic Loss. We consider the loss function*

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

From (4.2), the risk of decision d is

$$\begin{aligned} \rho(\pi, d) &= E\{L(\theta, d) \,|\, \theta \sim \pi(\theta)\} \\ &= E_{(\pi)}\{(\theta - d)^2\} \\ &= E_{(\pi)}(\theta^2) - 2dE_{(\pi)}(\theta) + d^2, \end{aligned}$$

where $E_{(\pi)}(\cdot)$ is a notational device to define the expectation computed using the distribution $\pi(\theta)$. Differentiating with respect to d we have

$$\frac{\partial}{\partial d}\rho(\pi, d) \quad = \quad -2E_{(\pi)}(\theta) + 2d.$$

So, the Bayes rule $d^* = E_{(\pi)}(\theta)$. *The corresponding Bayes risk is*

$$\begin{aligned}\rho^*(\pi) \;=\; \rho(\pi, d^*) &= E_{(\pi)}(\theta^2) - 2d^* E_{(\pi)}(\theta) + (d^*)^2 \\ &= E_{(\pi)}(\theta^2) - 2E^2_{(\pi)}(\theta) + E^2_{(\pi)}(\theta) \\ &= E_{(\pi)}(\theta^2) - E^2_{(\pi)}(\theta) \\ &= Var_{(\pi)}(\theta)\end{aligned}$$

where $Var_{(\pi)}(\theta)$ *is the variance of* θ *computed using the distribution* $\pi(\theta)$.

1. *If* $\pi(\theta) = f(\theta)$, *a prior for* θ, *then the Bayes rule of an immediate decision is* $d^* = E(\theta)$ *with corresponding Bayes risk* $\rho^* = Var(\theta)$.

2. *If we observe sample data* x *then the Bayes rule given this sample information is* $d^* = E(\theta \,|\, X)$ *with corresponding Bayes risk* $\rho^* = Var(\theta \,|\, X)$ *as* $\pi(\theta) = f(\theta \,|\, x)$.

Typically we can solve $[\Theta, \mathcal{D}, f(\theta), L(\theta, d)]$, the immediate decision problem, and solve $[\Theta, \mathcal{D}, f(\theta \,|\, x), L(\theta, d)]$, the decision problem after sample information. Often, we may be interested in the risk of the sampling procedure, before observing the sample, to decide whether or not to sample. For each possible sample, we need to specify which decision to make. We have a decision function

$$\delta : X \to \mathcal{D}$$

where X is the data or sample information. Let Δ be the collection of all decision functions, so $\delta \in \Delta \Rightarrow \delta(x) \in \mathcal{D}\ \forall x \in X$. The risk of decision function δ is

$$\begin{aligned}\rho(f(\theta), \delta) &= \int_x \int_\theta L(\theta, \delta(x)) f(\theta, x)\, d\theta dx \\ &= \int_x \int_\theta L(\theta, \delta(x)) f(\theta \,|\, x) f(x)\, d\theta dx \\ &= \int_x \left\{ \int_\theta L(\theta, \delta(x)) f(\theta \,|\, x)\, d\theta \right\} f(x)\, dx \\ &= \int_x E\{L(\theta, \delta(x)) \,|\, X\} f(x)\, dx \qquad (4.3)\end{aligned}$$

where, from (4.2), $E\{L(\theta, \delta(x)) \,|\, X\} = \rho(f(\theta \,|\, x), \delta(x))$, the posterior risk. We want to find the Bayes decision function δ^* for which

$$\rho(f(\theta), \delta^*) \;=\; \inf_{\delta \in \Delta} \rho(f(\theta), \delta).$$

From (4.3), as $f(x) \geq 0$, δ^* may equivalently be found as

$$\rho(f(\theta), \delta^*) \;=\; \inf_{\delta \in \Delta} E\{L(\theta, \delta(x)) \,|\, X\}, \qquad (4.4)$$

the posterior risk. The corresponding risk of the sampling procedure is

$$\rho^*_n \;=\; E[E\{L(\theta, \delta^*(x)) \,|\, X\}]. \qquad (4.5)$$

So, from (4.4), the Bayes decision function is the Bayes rule of the decision problem $[\Theta, \mathcal{D}, f(\theta \mid x), L(\theta, d)]$ considered as a function of (random) x whilst, from (4.5), the Bayes risk of the sampling procedure is the expected value of the Bayes risk of the decision problem $[\Theta, \mathcal{D}, f(\theta \mid x), L(\theta, d)]$ considered as a function of (random) x.

Example 37 *Suppose that we wish to estimate the parameter, θ, of a Poisson distribution. Our prior for θ is Gamma(α, β). The loss function, for estimate d and value θ, is*

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

1. *Find the Bayes rule and Bayes risk of an immediate decision.*
2. *Find the Bayes rule and Bayes risk if we take a sample of size n.*
3. *Find the Bayes risk of the sampling procedure.*

We consider the decision problem $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$. Relative to distribution π the expected loss is

$$\begin{aligned} E_{(\pi)}\{L(\theta, d)\} &= E_{(\pi)}\{\theta(\theta - d)^2\} \\ &= E_{(\pi)}(\theta^3 - 2d\theta^2 + d^2\theta) \\ &= E_{(\pi)}(\theta^3) - 2dE_{(\pi)}(\theta^2) + d^2 E_{(\pi)}(\theta) \end{aligned}$$

Differentiating with respect to d we find

$$\frac{\partial}{\partial d} E_{(\pi)}\{L(\theta, d)\} = -2E_{(\pi)}(\theta^2) + 2dE_{(\pi)}(\theta)$$

so that the Bayes rule is

$$d^* = \frac{E_{(\pi)}(\theta^2)}{E_{(\pi)}(\theta)} \tag{4.6}$$

with corresponding Bayes risk

$$\begin{aligned} \rho^*(\pi) &= E_{(\pi)}(\theta^3) - 2d^* E_{(\pi)}(\theta^2) + (d^*)^2 E_{(\pi)}(\theta) \\ &= E_{(\pi)}(\theta^3) - 2\frac{E_{(\pi)}(\theta^2)}{E_{(\pi)}(\theta)} E_{(\pi)}(\theta^2) + \left\{ \frac{E_{(\pi)}(\theta^2)}{E_{(\pi)}(\theta)} \right\}^2 E_{(\pi)}(\theta) \\ &= E_{(\pi)}(\theta^3) - \frac{E^2_{(\pi)}(\theta^2)}{E_{(\pi)}(\theta)}. \end{aligned} \tag{4.7}$$

We now consider the immediate decision by solving the decision problem $[\Theta, \mathcal{D}, f(\theta), L(\theta, d)]$. As $\theta \sim$ Gamma(α, β) then

$$\begin{aligned} E(\theta^k) &= \int_0^\infty \theta^k \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}\, d\theta \\ &= \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} \times \frac{\beta^\alpha}{\beta^{\alpha+k}} \int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha + k)} \theta^{\alpha+k-1} e^{-\beta\theta}\, d\theta \qquad (4.8) \\ &= \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} \times \frac{\beta^\alpha}{\beta^{\alpha+k}} \qquad (4.9) \end{aligned}$$

provided that $\alpha + k > 0$ so that the integral in (4.8) is of the density of a Gamma$(\alpha + k, \beta)$ distribution. Now, from (4.9), for $k = 1, 2, 3$,

$$E(\theta) = \frac{\Gamma(\alpha+1)\beta^\alpha}{\Gamma(\alpha)\beta^{\alpha+1}} = \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} = \frac{\alpha}{\beta} \tag{4.10}$$

$$E(\theta^2) = \frac{\Gamma(\alpha+2)\beta^\alpha}{\Gamma(\alpha)\beta^{\alpha+2}} = \frac{(\alpha+1)\alpha\Gamma(\alpha)}{\beta^2\Gamma(\alpha)} = \frac{(\alpha+1)\alpha}{\beta^2} \tag{4.11}$$

$$E(\theta^3) = \frac{\Gamma(\alpha+3)\beta^\alpha}{\Gamma(\alpha)\beta^{\alpha+3}} = \frac{(\alpha+2)(\alpha+1)\alpha\Gamma(\alpha)}{\beta^3\Gamma(\alpha)} = \frac{(\alpha+2)(\alpha+1)\alpha}{\beta^3} \tag{4.12}$$

Substituting (4.10) and (4.11) into (4.6), the Bayes rule of the immediate decision is

$$d^* = \frac{(\alpha+1)\alpha}{\beta^2} \times \frac{\beta}{\alpha} = \frac{\alpha+1}{\beta}. \tag{4.13}$$

Substituting (4.10)- (4.12) into (4.7), the Bayes risk of the immediate decision is

$$\rho^*(f(\theta)) = \frac{(\alpha+2)(\alpha+1)\alpha}{\beta^3} - \frac{(\alpha+1)\alpha}{\beta^2} \times \frac{\alpha+1}{\beta} = \frac{\alpha(\alpha+1)}{\beta^3}. \tag{4.14}$$

We now consider the problem after observing a sample of size n by solving $[\Theta, \mathcal{D}, f(\theta \,|\, x), L(\theta, d)]$ where $x = (x_1, \ldots, x_n)$. As $\theta \sim$ Gamma(α, β) and $X_i \,|\, \theta \sim$ Po(θ) then[2] *$\theta \,|\, x \sim$ Gamma$(\alpha + \sum_{i=1}^n x_i, \beta + n)$. We can exploit this conjugacy to observe that the Bayes rule and Bayes risk after sampling can be found from substituting $\alpha + \sum_{i=1}^n x_i$ for α and $\beta + n$ for β in (4.13) and (4.14) to obtain*

$$d^* = \frac{\alpha + \sum_{i=1}^n x_i + 1}{\beta + n} \tag{4.15}$$

as the Bayes rule after observing a sample of size n with corresponding Bayes risk

$$\rho^*(f(\theta \,|\, x)) = \frac{(\alpha + \sum_{i=1}^n x_i)(\alpha + \sum_{i=1}^n x_i + 1)}{(\beta + n)^3}. \tag{4.16}$$

We now consider the risk of the sampling procedure. From (4.4) the Bayes decision function is (4.15) viewed as a random variable, that is

$$\delta^* = \frac{\alpha + \sum_{i=1}^n X_i + 1}{\beta + n}$$

From (4.5), the risk of the sampling procedure, ρ_n^*, is the expected value of (4.16) when viewed as a random variable,

$$\begin{aligned} \rho_n^* &= E\left\{ \frac{(\alpha + \sum_{i=1}^n X_i)(\alpha + \sum_{i=1}^n X_i + 1)}{(\beta + n)^3} \right\} \\ &= \frac{\alpha(\alpha+1) + (2\alpha+1)E(\sum_{i=1}^n X_i) + E\{(\sum_{i=1}^n X_i)^2\}}{(\beta+n)^3} \end{aligned} \tag{4.17}$$

[2] See, for example, Question Sheet Three Exercise 1.(b).

Now, we utilise the tower property of expectations[3] *to find* $E(\sum_{i=1}^n X_i)$ *and* $E\{(\sum_{i=1}^n X_i)^2\}$. *We have that*

$$\begin{aligned} E\left(\sum_{i=1}^n X_i\right) &= E\left\{E\left(\sum_{i=1}^n X_i \,\middle|\, \theta\right)\right\} \\ &= E\left\{\sum_{i=1}^n E(X_i \,|\, \theta)\right\} \\ &= \sum_{i=1}^n E(\theta) \;=\; \frac{n\alpha}{\beta}; \end{aligned} \tag{4.18}$$

$$\begin{aligned} E\left\{\left(\sum_{i=1}^n X_i\right)^2\right\} &= E\left[E\left\{\left(\sum_{i=1}^n X_i\right)^2 \,\middle|\, \theta\right\}\right] \\ &= E\left\{Var\left(\sum_{i=1}^n X_i \,\middle|\, \theta\right) + E^2\left(\sum_{i=1}^n X_i \,\middle|\, \theta\right)\right\} \\ &= E(n\theta + n^2\theta^2) \;=\; \frac{n\alpha}{\beta} + \frac{n^2\alpha(\alpha+1)}{\beta^2}. \end{aligned} \tag{4.19}$$

Notice that we have exploited the independence of the X_i *given* θ *and that* $X_i \,|\, \theta \sim Po(\theta)$ *with* $\theta \sim Gamma(\alpha, \beta)$. *A slightly quicker, though less general, approach is to note that* $\sum_{i=1}^n X_i \,|\, \theta \sim Po(n\theta)$. *Substituting (4.18) and (4.19) into (4.17) gives*

$$\begin{aligned} \rho_n^* &= \frac{1}{(\beta+n)^3}\left\{\alpha(\alpha+1) + (2\alpha+1)\frac{n\alpha}{\beta} + \frac{n\alpha}{\beta} + \frac{n^2\alpha(\alpha+1)}{\beta^2}\right\} \\ &= \frac{\alpha(\alpha+1)}{\beta^2(\beta+n)^3}\left\{\beta^2 + 2\beta n + n^2\right\} \\ &= \frac{\alpha(\alpha+1)}{\beta^2(\beta+n)}. \end{aligned}$$

Notice that when $n = 0$, $\rho_{n=0}^* = \frac{\alpha(\alpha+1)}{\beta^3} = \rho^*(f(\theta))$, *the Bayes risk of the immediate decision. As* n *increases then* ρ_n^* *decreases.*

[3] See Question Sheet One Exercise 5.