Bayesian Inference and Computational Methods

Contents

1 Bayesian and likelihood-based inference

1.1 Introduction to Bayesian inference

Given

1. a *statistical model* for the generation of the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ parametrised by an unknown $\theta \in \Theta$—which we wish to estimate—and hence a *likelihood function* $L(\theta; \boldsymbol{y})$,
2. a *prior (probability) distribution* for θ, represented by its density or probability function $\pi(\theta)$ and expressing our (usually subjective) *a priori* beliefs about the probability of θ taking different values,

we use Bayes' Theorem to calculate the *posterior distribution* of θ *given* the observed data $\boldsymbol{y}$. This has *density* or *probability function*

$$\begin{aligned} \pi(\theta \,|\, \boldsymbol{y}) &= \frac{\mathbf{P}(\theta \text{ and } \boldsymbol{y})}{\mathbf{P}(\boldsymbol{y})} \\ &= k(\boldsymbol{y})\pi(\theta)L(\theta; \boldsymbol{y}) \end{aligned} \tag{1}$$

where the normalising constant $k(\boldsymbol{y})$ (which may depend on the observed data $\boldsymbol{y}$, but not on θ) is such that

$$\int_{\theta' \in \Theta} \pi(\theta' \,|\, \boldsymbol{y})\, d\theta' = 1, \tag{2}$$

i.e.

$$k(\boldsymbol{y}) = \frac{1}{\int_{\theta' \in \Theta} \pi(\theta') L(\theta'; \boldsymbol{y})\, d\theta'}.$$

(Note that in the case where θ is discrete the integrals above are replaced by sums.)

In general we shall simply write expressions such as (1) as

$$\pi(\theta \,|\, \boldsymbol{y}) \propto \pi(\theta) L(\theta; \boldsymbol{y}). \tag{3}$$

since the normalising constant may always be calculated by the use of (2).

The *posterior distribution* $\pi(\theta \,|\, \boldsymbol{y})$ expresses our new belief about the probability of θ taking different values once the data $\boldsymbol{y}$ have been observed. It summarises completely the outcome of our inference about θ.

However, if we wished, for example, to supply an *point estimate* for θ, we might use the *mean*, the *median*, or the *mode* of this posterior distribution. Similarly if we wished an *interval estimate* for θ, we might calculate, for example, a 95% *Bayesian credible interval* as a region within which, under the distribution $\pi(\theta \,|\, \boldsymbol{y})$, the parameter θ lay with probability 0.95. This interval could be forced to be unique by, for example, the requirement that it be *equal-tailed*.

Example 1.1 *Bayesian estimation of binomial proportion p.*

A geneticist wishes to estimate the proportion of the population carrying a certain gene. They collect DNA from a random sample of 20 individuals, of whom 5 are found to carry the gene. Carry out an investigation of p using Bayesian techniques.

The first thing we need to do is construct the model for the generation of the data, and hence determine the likelihood $L(p)$. The model here is simply the distribution of the number Y of gene-carriers in the sample, and, since the sample is taken from

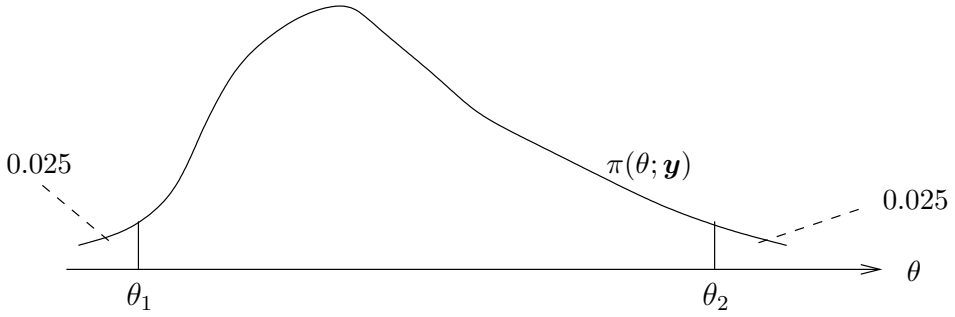

Figure 1: Determination of 95% equal-tailed Bayesian credible interval.

a large population, it is reasonable to assume that this is binomial with parameters 20 and p. Hence the likelihood $L(p)$ is given by

$$L(p) = P(Y = 5 \,|\, p) = \binom{20}{5} p^5 (1-p)^{15}.$$

Now, we need to specify a *prior density* for p. Let us suppose that the geneticist has little information on the value of p. To reflect this they choose for their prior density, $\pi(p) = 1$, that is a uniform distribution on the interval $(0, 1)$. We will later discuss the extent to which a uniform prior can be considered to represent *prior ignorance*.

We can now identify the *posterior density* of p as

$$\pi(p \,|\, y) = \frac{\pi(p)L(p)}{\int_0^1 \pi(p')L(p')dp'}$$

Since the denominator above is simply the reciprocal of the normalising constant we have, more simply,

$$\begin{aligned} \pi(p \,|\, y) &\propto \pi(p)L(p) \\ &= \binom{20}{5} p^5 (1-p)^{15} \\ &\propto p^5 (1-p)^{15}. \end{aligned}$$

This functional dependence on p identifies $\pi(p \,|\, y)$ as corresponding to a Beta$(6, 16)$ distribution.

Self-study exercise: Review the properties of the Beta distribution.

We can therefore identify the posterior mean of p as $3/11$, and the mode as $1/4$.

Further comment. Note that the choice of a uniform prior distribution for p means that the mode of the posterior distribution is just the value of p which maximises $L(p)$, i.e. is the conventional maximum likelihood estimate of p.

Example 1.2 *Inference for the Exp(λ) distribution.*

Suppose that the *lifetime* of a particular type of component is believed to be have an Exp(λ) distribution, where λ is unknown. In order to estimate λ, we select a random sample of 20 components and measure their lifetimes $t_1, t_2, t_3, ..., t_{20}$. Carry out a Bayesian analysis of these data to estimate λ.

We consider first the choice of *prior density* $\pi(\lambda)$ for λ. Again let us assume that we have little prior knowledge about the value of λ. Therefore we select a prior density

that 'supports' a broad range of values of λ. One possibility would be to use an $\text{Exp}(\alpha)$ distribution where α is small in some sense. Suppose we set $\alpha = 0.1$.

Now suppose that we take the observations and note that $\sum_{i=1}^{20} t_i = 10.0$ time units. Since (we assume) the observed lifetimes are independent of each other, the likelihood function $L(\lambda)$ is given by the joint density

$$\begin{aligned} L(\lambda; \boldsymbol{t}) &= \prod_{i=1}^{20} \lambda e^{-\lambda t_i} \\ &= \lambda^{20} e^{-\lambda \sum_{i=1}^{20} t_i}. \end{aligned}$$

Thus the *posterior density* is given by

$$\begin{aligned} \pi(\lambda \,|\, \boldsymbol{t}) &\propto \pi(\lambda) L(\lambda; \boldsymbol{t}) \\ &= \lambda^{20} e^{-\lambda(\alpha + \sum_{i=1}^{20} t_i)}. \end{aligned}$$

We can see that the posterior density of λ must correspond to a $\Gamma(21, 10.1)$ distribution. In particular the mean of the posterior distribution is given by

$$\mathbf{E}(\lambda \,|\, \boldsymbol{t}) = \frac{21}{\alpha + \sum_{i=1}^{20} t_i} = 2.08,$$

while its variance is given by

$$\text{var}(\lambda \,|\, \boldsymbol{t}) = \frac{21}{(\alpha + \sum_{i=1}^{20} t_i)^2} = 0.21.$$

Further comments. It is worth studying the effect of varying α on the posterior distribution. Note in particular that the choice of a small value of α places most of the 'weight' of the inference on the data. Note also that the conventional maximum likelihood estimate of λ is given by $\hat{\lambda} = 20/\sum_{i=1}^{20} t_i = 2.00$.

Example 1.3 *Censored observations*

Consider the same problem as that of Example 1.2 but suppose now that the data, rather than giving the precise times of failure of all components, only record the times of failure up to $t = 0.5$, at which time the experiment ceases. Suppose that 14 components (which we label $i = 1, \ldots, 14$) fail within this period with $\sum_{i=1}^{14} t_i = 2.2$, and the remaining 6 components are operational at $t = 0.5$. To construct the likelihood for this case we note that $\mathbf{P}_\lambda(T > 0.5) = e^{-0.5\lambda}$, and insert this factor into the likelihood function for each component whose lifetime exceeds 0.5. This yields a likelihood

$$\begin{aligned} L(\lambda) &= \lambda^{14} e^{-\lambda(3.0 + \sum_{i=1}^{14} t_i)} \\ &= \lambda^{14} e^{-5.2\lambda}. \end{aligned}$$

If we again use an $\text{Exp}(\alpha)$ distribution with $\alpha = 0.1$ as the prior distribution for λ, then the posterior distribution of λ is given by the density

$$\begin{aligned} \pi(\lambda \,|\, \boldsymbol{t}) &\propto \pi(\lambda) L(\lambda; \boldsymbol{t}) \\ &= \lambda^{14} e^{-\lambda(\alpha + 5.2)}, \end{aligned}$$

i.e. the posterior distribution is $\Gamma(15, 5.3)$. In this case the posterior mean and variance of λ are 2.83 and 0.53 respectively.

Notes

1. In comparison to more traditional approaches, the Bayesian approach to inference has the advantages of coherence and consistency. In particular, once the model for the data and the prior distribution for the unknown parameter θ are specified, there is no controversy about the probability theory leading to the posterior distribution.

2. Bayesian inferences are easily updated: let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be two vectors of observations obtained sequentially and independently of each other. Let $L(\theta; \boldsymbol{y}_1)$ and $L(\theta; \boldsymbol{y}_2)$ be the respective likelihood functions for θ. It follows from the above independence that the likelihood function, given both vectors of observations is

$$L(\theta; \boldsymbol{y}_1, \boldsymbol{y}_2) = L(\theta; \boldsymbol{y}_1) L(\theta; \boldsymbol{y}_2).$$

   If $\pi(\theta)$ denotes the prior density for θ, then the posterior density for θ, given the first set of observations $\boldsymbol{y}_1$, is given by

$$\pi(\theta \,|\, \boldsymbol{y}_1) \propto \pi(\theta) L(\theta; \boldsymbol{y}_1);$$

   the posterior density for θ, given the both sets of observations $\boldsymbol{y}_1$ $\boldsymbol{y}_2$, is given by

$$\begin{aligned} \pi(\theta \,|\, \boldsymbol{y}_1, \boldsymbol{y}_2) &\propto \pi(\theta) L(\theta; \boldsymbol{y}_1, \boldsymbol{y}_2) \\ &= \pi(\theta) L(\theta; \boldsymbol{y}_1) L(\theta; \boldsymbol{y}_2) \\ &\propto \pi(\theta \,|\, \boldsymbol{y}_1) L(\theta; \boldsymbol{y}_2). \end{aligned}$$

   Thus, given the second set of observations, the posterior distribution for θ is updated by treating the earlier posterior distribution as a prior for the new data.

3. The disadvantage of the Bayesian approach is the *subjectivity* involved in the choice of a prior distribution. However, all statistical inference involves some element of subjectivity, and it may be argued that the Bayesian approach simply places this 'up-front'.

1.2 Likelihood-based inference

Given a *statistical model* for the generation of the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ parametrised by an unknown $\theta \in \Theta$, *likelihood-based inference* uses only the *likelihood function* $L(\theta; \boldsymbol{y})$, to make inference about θ.

Note that, for the purposes of inference, the observations $\boldsymbol{y}$ are fixed and $L(\theta; \boldsymbol{y})$ is therefore to be regarded as a function of θ (we may sometimes write simply $L(\theta)$ for $L(\theta; \boldsymbol{y})$).

The importance of the likelihood is summarised by the *likelihood principle*, which (in essence) says the following.

> Suppose that two experimental outcomes $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ define likelihoods $L(\theta; \boldsymbol{y}_1)$ and $L(\theta; \boldsymbol{y}_2)$ that are proportional to each other, that is, for some $C > 0$, we have $L(\theta; \boldsymbol{y}_2) = CL(\theta; \boldsymbol{y}_2)$, for all θ. Then the conclusions about θ drawn from $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ should be identical.

The likelihood principle says that the all the information that the data give you about the parameters is embodied in the likelihood. A fuller account of the likelihood principle—and how it is a logical consequence of two other principles (the *conditionality* and *sufficiency* principles)—can be found in texts such as that by Casella and Berger. Further support for the likelihood as the most appropriate expression of information regarding the plausibility of different values of θ can be seen from the Neyman-Pearson Lemma in classical hypothesis testing.

Thus, in any problem of statistical estimation or inference it is a good idea to try to write down the likelihood function for the data. This requires the use the rules of probability theory in order to work out the probability or probability density of the observations given the parameter θ. Depending on whether the observations $\boldsymbol{y}$ are discrete, continuous, precisely measured, or censored (known only to lie in certain intervals) constructing likelihoods will typically require use of probability mass or density functions and cumulative distribution functions.

We now consider inference based on the use of the likelihood function $L(\theta; \boldsymbol{y})$ alone. Define also the log-likelihood function $l(\theta) = l(\theta; \boldsymbol{y}) = \log L(\theta; \boldsymbol{y})$

Point estimation. Choose $\hat{\theta} = \hat{\theta}(\boldsymbol{y})$ to maximise $L(\theta; \boldsymbol{y})$ (or equivalently $l(\theta; \boldsymbol{y})$). This is the *maximum likelihood estimate* of θ.

Interval estimation. We consider interval estimation for a single parameter θ. It is convenient (see below) to define the *deviance*

$$D(\theta; \boldsymbol{y}) = -2[l(\theta; \boldsymbol{y}) - l(\hat{\theta}; \boldsymbol{y})] \tag{4}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ defined above. Then a *likelihood interval* for θ is an interval of the form

$$I = \{\theta \colon D(\theta; \boldsymbol{y}) \leq d\}$$

for some appropriately chosen d. This is typically an interval $[\theta_l, \theta_u]$ (where θ_l and θ_u depend on $\boldsymbol{y}$) and is the interval within which the log-likelihood $l(\theta; \boldsymbol{y})$ is no more than $d/2$ below its maximum value.

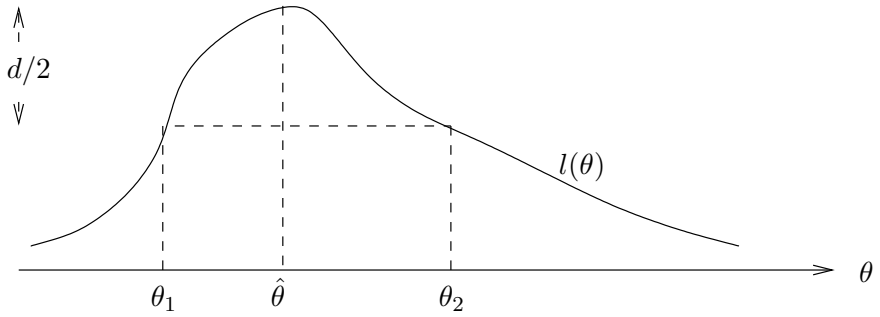

Figure 2: Determination of likelihood interval.

In order to associate a particular value of d with a particular confidence level, we can make use of some of the traditional frequentist theory for determining confidence intervals: for a given 'true' value of θ, it can be shown that, asymptotically and under mild regularity conditions,

$$D(\theta; \boldsymbol{y}) \sim \chi^2_1, \tag{5}$$

i.e. has a chi-squared distribution with 1 degree of freedom.
For any $\alpha \in (0, 1)$ let $\chi^2_1(\alpha)$ be the percentage point of the χ^2_1-distribution such that

$$\mathbf{P}(\chi^2_1 > \chi^2_1(\alpha)) = \alpha \tag{6}$$

(where χ^2_1 has a χ^2_1-distribution). Define the (random) interval

$$\begin{aligned}[\theta_l(\boldsymbol{y}),\, \theta_u(\boldsymbol{y})] &= \{\theta\colon D(\theta; \boldsymbol{y}) \le \chi^2_1(\alpha)\} \\ &= \{\theta\colon l(\hat{\theta}; \boldsymbol{y}) - l(\theta; \boldsymbol{y}) \le \chi^2_1(\alpha)/2\}.\end{aligned}$$

Then, for the given 'true' value of θ, it follows from (5) and (6) that

$$\begin{aligned}\mathbf{P}\left(\theta \in [\theta_l(\boldsymbol{y}),\, \theta_u(\boldsymbol{y})]\right) &= \mathbf{P}\left(D(\theta; \boldsymbol{y}) \le \chi^2_1(\alpha)\right) \\ &= 1 - \alpha.\end{aligned}$$

Thus $[\theta_l(\boldsymbol{y}),\, \theta_u(\boldsymbol{y})]$ is, asymptotically, a $100(1-\alpha)$ confidence interval for θ.
For example, for a 95% confidence interval, we take the deviance $d = \chi^2_1(0.05) \approx 3.84$, so that the likelihood interval consists of those values of θ for which $l(\theta; \boldsymbol{y})$ is no more than 1.92 below $l(\hat{\theta}; \boldsymbol{y})$.

Example 1.4 Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ where $y_1, \ldots, y_n$ are independent identically distributed as $N(\mu, \sigma^2)$ and σ is assumed known. We wish to derive a likelihood interval for μ associated with 95% confidence.
We have

$$l(\mu; \boldsymbol{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + k = -\frac{n}{2\sigma^2}(\bar{y} - \mu)^2 + k',$$

where $\bar{y} = (\sum_{i=1}^n y_i)/n$ and k, k' are terms which do not depend on μ. (*Exercise: check both the identities above.*) Hence $l(\mu; \boldsymbol{y})$ is maximised at the maximum likelihood estimate $\hat{\mu} = \bar{y}$. Further the likelihood interval for μ is given by

$$\{\mu\colon\ -2(l(\mu; \boldsymbol{y}) - l(\hat{\mu}; \boldsymbol{y})) \le d\} = \left\{\mu\colon (\bar{y} - \mu)^2 \le d\frac{\sigma^2}{n}\right\} = \left(\bar{y} - z\frac{\sigma}{\sqrt{n}},\, \bar{y} + z\frac{\sigma}{\sqrt{n}}\right),$$

where $d = 3.84$ (see above) and where $z = d^{1/2} = 1.96$. Thus we see that in this case the likelihood interval agrees with the traditionally derived confidence interval. The theory of maximum likelihood estimation tells us that for other distributions a similar result holds asymptotically.

1.3 Bayesian inference: choice of prior distribution

1.3.1 Conjugate prior distributions

Suppose that we are given a *family* $\mathcal{F}$ of distributions for the data (e.g. binomial, Poisson, exponential, normal) parametrised by some (possibly vector) θ to be estimated.

Definition. The *class* $\mathcal{P}$ of distributions is **conjugate** for $\mathcal{F}$, if, for any *prior distribution* in the class $\mathcal{P}$, the *posterior distribution* also belongs to the class $\mathcal{P}$.

A *conjugate class* $\mathcal{P}$ for $\mathcal{F}$ is usually sufficiently broad as to allow a wide choice of possible prior distributions. The advantages of using it are those of analytical simplicity and of ease of updating the *posterior distribution* given additional data (see the second note at the end of Section 1.1).

Conjugate classes $\mathcal{P}$ exist for all the main families $\mathcal{F}$ of distributions for data. We give some examples now and further ones subsequently.

Example 1.5 *Bayesian estimation of binomial proportion p.*

We generalise Example 1.1. Suppose that, in n independent trials, each of which has a probability of *success* p, a total of y *successes* are observed. Then, given the data, the *likelihood function* for p is

$$L(p; y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Suppose that we choose as *prior distribution* for p a $\beta(a, b)$ distribution, $a > 0$, $b > 0$, i.e. density

$$\pi(p) \propto p^{a-1}(1-p)^{b-1} \qquad a > 0, \quad b > 0.$$

(Note that in particular the choice $a = b = 1$ corresponds to a $\mathrm{U}(0, 1)$ prior distribution for p.) Then the *posterior distribution* for p has density

$$\begin{aligned} \pi(p \,|\, y) &\propto \pi(p) L(p; y) \\ &\propto p^{a+y-1}(1-p)^{b+n-y-1} \end{aligned}$$

and hence is a $\beta(a + y, b + n - y)$ distribution. It follows that the class of *beta* distributions is *conjugate* for the *binomial* family.

Note that, provided a and b are small, the *posterior distribution* depends mostly on the data. The *mean* of the *posterior distribution* is $(a+y)/(a+b+n)$ and its *mode* is $(a+y-1)/(a+b+n-2)$. In the case of a *uniform prior distribution* ($\beta(1, 1)$) the *mode* of the *posterior distribution* corresponds to the maximum likelihood estimate—as would be expected.

Now suppose that a further m independent observations are obtained, of which z are found to be *successes*. We may use the previously found *posterior distribution* as the *prior distribution* before the incorporation of the new data, after which the *new posterior distribution* is found to be $\beta(a+y+z, b+n+m-(y+z))$—the same as would have resulted had we obtained all the data in one go.

Example 1.6 *Bayesian estimation of Poisson parameter* λ.

Suppose that we have n *independent* (nonnegative integer) *observations* $\boldsymbol{y} = (y_1, \ldots, y_n)$ from a *Poisson* distribution with unknown parameter (*mean*) $\lambda > 0$ to be estimated. [For example, y_i might be the *number of insurance claims* in week i, if it was considered that such claims arose as a homogeneous Poisson process.] The *probability function* for any single observation y is

$$p_\lambda(y) = e^{-\lambda}\frac{\lambda^y}{y!}$$

and so, given the data, the *likelihood function* for λ is

$$\begin{aligned} L(\lambda; \boldsymbol{y}) &\propto \prod_{i=1}^{n} e^{-\lambda}\lambda^{y_i} \\ &= e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i}. \end{aligned}$$

Suppose that we choose as *prior distribution* for λ a $\Gamma(a, b)$ distribution, $a > 0$, $b > 0$, i.e. density

$$\pi(\lambda) \propto \lambda^{a-1}e^{-b\lambda} \qquad a > 0, \quad b > 0.$$

(Note that in particular the choice $a = 1$ corresponds to an $\mathrm{Exp}(b)$ prior distribution for λ.) Then the *posterior distribution* for λ has density

$$\begin{aligned} \pi(\lambda \,|\, \boldsymbol{y}) &\propto \pi(\lambda)L(\lambda; \boldsymbol{y}) \\ &\propto \lambda^{a-1+\sum_{i=1}^n y_i}e^{-(b+n)\lambda} \end{aligned}$$

and hence is a $\Gamma(a + \sum_{i=1}^n y_i, b + n)$ distribution. It follows that the class of *gamma* distributions is *conjugate* for the *Poisson* family.

Note again that, provided a and b are small, the *posterior distribution* depends mostly on the data. The *mean* of the *posterior distribution* is $(a + \sum_{i=1}^n y_i)/(b + n)$ and its *mode* is $(a + \sum_{i=1}^n y_i - 1)/(b + n)$ (*exercise*).

Now suppose that a further m independent observations $y_{n+1}, \ldots, y_{n+m}$ are obtained. We may again use the previously found *posterior distribution* as the *prior distribution* before the incorporation of the new data, after which the *new posterior distribution* is found to be $\Gamma(a + \sum_{i=1}^{n+m} y_i, b + n + m)$—again the same as would have resulted had we obtained all the data in one go.

1.3.2 Noninformative prior distributions

When we have little prior knowledge about the likely values of the unknown parameter θ, we seek to model its initial distribution by the use of a **noninformative**, or **flat**, *prior*. Then, in a sense, the *posterior distribution* depends mostly on the observed data.

In Example 1.5 (Bayesian estimation of binomial proportion p), the $\beta(a, b)$ distribution with small values of a and b corresponds to the use of a (relatively) *noninformative prior distribution*—in particular, as we have already observed, the choice $a = b = 1$ corresponds to a *uniform prior*. Similarly, in Example 1.6 (Bayesian estimation of Poisson parameter λ), the $\Gamma(a, b)$ distribution with small values of a and b again corresponds to the use of a *noninformative prior distribution*. We give a further example, which allows us to introduce the idea of an *improper prior distribution*.

Example 1.7 *Bayesian estimation for the mean μ of the normal distribution with known variance σ^2.*

Suppose that we have n independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ from a *normal* distribution $\mathrm{N}\left(\mu, \sigma^2\right)$ with *known* variance $\sigma^2 > 0$ and *unknown* mean μ to be estimated. [For example, the y_i might be *temperature* measurements, or (for actuaries) the sizes of *insurance claims*.] The *probability density function* for any single observation y is

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

and so, given the data, the *likelihood function* for μ is

$$\begin{aligned} L(\mu; \boldsymbol{y}) &\propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right), \qquad \textit{(exercise!)}, \end{aligned}$$

where $\bar{y} = (\sum_{i=1}^{n} y_i)/n$ is the *sample mean*.

Suppose now that we choose as *prior distribution* for μ a $\mathrm{N}\left(\mu_0, \sigma_0^2\right)$ distribution, with density

$$\pi(\mu) \propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

Some (tedious) algebra (*exercise*) shows that the *posterior distribution* for μ is

$$\mathrm{N}\left(\frac{\mu_0\sigma_0^{-2} + n\bar{y}\sigma^{-2}}{\sigma_0^{-2} + n\sigma^{-2}}, (\sigma_0^{-2} + n\sigma^{-2})^{-1}\right).$$

It follows again that the class of *normal* distributions is *conjugate* for the *normal* family with σ^2 known. However, in the case where little is known about μ before the data $\boldsymbol{y}$ are observed, it is tempting to take σ_0^2 as large as possible. As $\sigma_0^2 \to \infty$, the *posterior distribution* tends (regardless of μ_0) to

$$\mathrm{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

The density of this *posterior distribution* is proportional to the *likelihood function* above. It is the same as would have resulted from taking the *prior distribution* for μ to be *uniformly* distributed ("*flat*") on the entire real line. However such a "distribution" cannot be *normalised* so as to correspond to a total probability of 1. It is an example of an *improper prior distribution*, which is nevertheless very useful in practice.

Jeffreys' prior distribution *Jeffreys' prior distribution* is a particular choice of (relatively) *noninformative prior*, and is given by the *prior density*

$$\pi(\theta) \propto [J(\theta)]^{1/2},$$

where $J(\theta)$ is the *Fisher information* for θ given by

$$J(\theta) = \mathbf{E}_\theta\left[\left(\frac{d \log L(\theta; \boldsymbol{y})}{d\theta}\right)^2\right] = -\mathbf{E}_\theta\left[\frac{d^2 \log L(\theta; \boldsymbol{y})}{d\theta^2}\right].$$

[This has the important property of being *invariant* under transformation of the parameter θ, i.e. under any transformation $\phi = \phi(\theta)$ the distribution induced by $\pi(\theta)$ on ϕ is just the Jeffreys' prior for ϕ.]

Example 1.8 *Jeffreys' prior for estimation of binomial proportion p.*
We return to Example 1.5, and again suppose that, in n independent trials, each of which has a probability of *success* p, a total of y *successes* are observed. Then, given the data, the *likelihood function* for p is

$$L(p; y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Hence, the *log-likelihood function* $l(p; \boldsymbol{y}) = \log L(p; \boldsymbol{y})$ is given by

$$l(p; y) = y \log p + (n-y) \log(1-p) + k$$

where k does not depend on p. Thus

$$-\frac{d^2 l(p; \boldsymbol{y})}{dp^2} = \frac{y}{p^2} + \frac{n-y}{(1-p)^2}$$

and so

$$\begin{aligned} -\mathbf{E}_p \left(\frac{d^2 l(p; \boldsymbol{y})}{dp^2} \right) &= \frac{n}{p} + \frac{n}{1-p} \\ &= \frac{n}{p(1-p)}. \end{aligned}$$

It follows that the *Jeffreys' prior distribution* for p is given by the density

$$\pi(p) \propto p^{-1/2} (1-p)^{-1/2},$$

i.e. is a $\beta(1/2, 1/2)$ distribution.

1.3.3 Prior elicitation

Suppose that we have some beliefs about the probabilities that the unknown parameter θ will fall within given regions. Then we may reasonably choose a *prior distribution* for θ (for example, from within a *conjugate class*) so as to reflect these beliefs.

Example 1.9 *Example 1.1 again.* Suppose now that the geneticist, prior to the collection of the data, has reason to believe that the probability that the binomial proportion p is less than 0.1 is (at most) 0.1, and similarly that the probability that the binomial proportion p is greater than 0.5 is (at most) 0.1. In this case a reasonable choice of *prior distribution* $\pi(p)$ for p would be a $\beta(a, b)$ distribution such that

$$\int_0^{0.1} \pi(p)\, dp = 0.1 \quad \text{and} \quad \int_{0.5}^1 \pi(p)\, dp = 0.1,$$

i.e. a and b are such that the 0.1-quantile of $\beta(a, b)$ is 0.1 and that 0.9-quantile of $\beta(a, b)$ is 0.5. This is given by taking $a = 2.20$ and $b = 5.50$. (*Check!*)
After the observation of 5 instances of the gene in a random sample of 20 individuals, the posterior distribution for p is $\beta(7.20, 20.50)$.

1.4 Bayesian inference: reporting conclusions

Having derived the *posterior distribution* of a parameter θ there are several ways in which we can express the results. For single parameters, a plot of the *posterior density* is very informative and shows clearly the range of values consistent with your posterior beliefs. We can also quote quantities such as the *posterior mean* or the *posterior variance* of θ. Indeed any summary of a distribution can be used.

A natural analogue of the frequentist confidence interval for a parameter is the **Bayesian credible interval**. For example, suppose that, given data $\boldsymbol{y}$ you derive the *posterior density* of θ as $\pi(\theta \,|\, \boldsymbol{y})$. Then a 95% credible interval (a, b) is any interval whose posterior probability of containing θ is 0.95. Often we might quote an equal-tailed interval (obtained by selecting the 97.5% and 2.5% critical points of $\pi(\theta \,|\, \boldsymbol{y})$), or a minimum-width interval (obtained by thresholding $\pi(\theta \,|\, \boldsymbol{y})$) assuming this can be calculated.

Depending on the circumstances, we may be interested in the posterior probability of a parameter being greater than or less than some threshold. For example, this might be the case where the experiment has been done for the purpose of quality control.

1.4.1 Predictive distributions

Usually when we carry out a Bayesian analysis to obtain $\pi(\theta \,|\, \boldsymbol{y})$ our interest lies in predicting some other quantity z of practical importance, whose distribution is determined by θ. (For example, in the case of Example 1.2, our interest may be in the lifetime of the next component that we select.) Having obtained $\pi(\theta \,|\, \boldsymbol{y})$, what we really want to do is determine the distribution of z given $\boldsymbol{y}$, given by the density (or probability function) $f(z \,|\, \boldsymbol{y})$. This distribution, known as the **predictive distribution** of z, is given as a *mixture distribution* over the possible values of θ. This has *density* (or *probability function*)

$$f(z \,|\, \boldsymbol{y}) = \int f(z \,|\, \theta)\pi(\theta \,|\, \boldsymbol{y})\, d\theta$$

where $f(z \,|\, \theta)$ denotes the density (or probability function) of z given θ. In this section we give some examples of calculating predictive distributions and highlight some of the mixture distributions that arise in standard problems.

Example 1.10 Following change in regulations, students are suppose to bring their own calculators to examinations. However, invariably a number forget and invigilators bring a small number of calculators to exams for these individuals. Suppose that in the first exam after the rule change an invigilator finds that 2 students out of a class of 30 have forgotten their calculators. The next week she has to invigilate an exam with 25 students. How many calculators should she bring in order to be 95% certain that she will have enough? Assume she is a Bayesian who takes a pessimistic view of the organisational skills of students.

A Bayesian solution. First of all, she needs to assume some statistical model for the number of students that forget to bring a calculator to an exam. A natural assumption is that this number follows a $\text{Bin}(n, p)$ distribution where n denotes the number of students taking the exam, and that p is the same for all exams.

Next she needs to identify a *prior distribution* for p, representing her beliefs before having seen the data $\boldsymbol{y}$ from the first exam, and hence a *posterior distribution* for

p given this data. Being pessimistic she assumes a $\mathrm{U}(0,1)$ *prior distribution* for p. Now, as in earlier examples, she immediately calculates that the *posterior density* $\pi(p\,|\,\boldsymbol{y})$ for p given the data is that corresponding to the $\beta(3,29)$ distribution.

Finally she must calculate the *predictive distribution* of Z, the number of students who forget their calculators in the next exam. Given p, we have $Z \sim \mathrm{Bin}(25,p)$ and so the *probability function* of Z given p is

$$f(z\,|\,p) = \binom{25}{z} p^z (1-p)^{25-z}, \qquad 0 \le z \le 25. \tag{7}$$

Thus the *predictive probability function* of Z given the data $\boldsymbol{y}$ is

$$\begin{aligned} f(z\,|\,\boldsymbol{y}) &= \int_0^1 f(z\,|\,p)\pi(p\,|\,\boldsymbol{y})\,dp \qquad (8)\\ &= \int_0^1 \binom{25}{z} p^z (1-p)^{25-z} \frac{\Gamma(32)}{\Gamma(3)\Gamma(29)} p^2(1-p)^{28}\,dp \\ &= \frac{\Gamma(32)}{\Gamma(3)\Gamma(29)} \binom{25}{z} \int_0^1 p^{z+2}(1-p)^{53-z}\,dp \\ &= \frac{\Gamma(32)}{\Gamma(3)\Gamma(29)\Gamma(57)} \binom{25}{z} \Gamma(3+z)\Gamma(54-z) \\ &= \frac{31!25!}{2!28!56!}\,\frac{(2+z)!(53-z)!}{z!(25-z)!}, \qquad 0 \le z \le 25. \end{aligned}$$

[Note that the various multiplicative constants in the above expressions, which are independent of z, may be omitted, provided we replace "=" by "$\propto$" and are prepared to calculate the correct normalising constant at the end of the calculation—in this case no work is saved by doing this. Note also that, for any integer $k \ge 1$, we have $\Gamma(k) = (k-1)!$.]

Now by examining the associated cumulative distribution function, we find that $\mathbf{P}(Z \le 5\,|\,\boldsymbol{y}) = 0.931$ while $\mathbf{P}(Z \le 6\,|\,\boldsymbol{y}) = 0.966$. Therefore she should bring 6 calculators to be 95% certain of having enough.

As an alternative to analysis, note that the expression on the right side of (8) is the *expectation* of the binomial probability $f(z\,|\,p)$ of z given p with respect to the *posterior distribution* of p. Suppose, therefore, that we can simulate a large number n of values of p from its posterior density $\pi(p\,|\,\boldsymbol{y})$. For each z, we may calculate $f(z\,|\,p)$ (as given by (7)) for the simulated values of p; the sample mean of these values is then an estimate of $f(z\,|\,\boldsymbol{y})$. For example, appropriate **R** code is

```
p = rbeta(10000,3,29)                          # 10000 simulations of p
f = rep(0,26)                                  # initialise prob fun f
for(z in 0:25) f[z+1]=mean(dbinom(z,25,p))     # estimate f
sum(f)                                         # check
f                                              # display f
cumsum(fsim)                                   # display cum probs
```

Exercises. Use **R** to calculate the *exact* values $f(z\,|\,\boldsymbol{y})$, $0 \le z \le 25$ of the predictive probability function (check also that the probabilities sum to 1). Use also simulation to estimate $f(z\,|\,\boldsymbol{y})$ as above, and compare your results. What would happen if a more informative prior were used, giving more weight to smaller values of p?

Example 1.11 Consider the component experiment of Example 1.2 above in which the lifetime of a component was $\mathrm{Exp}(\lambda)$ where the *posterior distribution* of λ, given

the observed data $\boldsymbol{t}$, was $\Gamma(\alpha, \beta)$ (with *density* $\pi(\lambda \,|\, \boldsymbol{t})$) for suitably chosen parameters α and β. Let us suppose that this component is used in a space vehicle which has to perform a flight of duration 1 day? If the component fails during the flight then it is replaced immediately from a pool of identical components whose lifetimes are all independent of each other. How many components in total are needed to ensure that the vehicle completes the flight with at least 90% certainty?

To solve this we need to consider the *predictive distribution*, again given the data, of the number Z of components which fail during a 1-day flight. It follows from the above assumption that component lifetimes are independent identically distributed $\text{Exp}(\lambda)$ that, *given* λ, component failures occur as a Poisson process with rate λ and so, again given λ, the distribution of Z is $\text{Pois}(\lambda)$ with *probability function*

$$f(z \,|\, \lambda) = \frac{e^{-\lambda}\lambda^z}{z!}, \qquad z = 0, 1, 2, \ldots. \tag{9}$$

Therefore the predictive probability function $f(z \,|\, \boldsymbol{t})$ of Z given the observed data $\boldsymbol{t}$ is given by the expectation of $f(z \,|\, \lambda)$ with respect to the above $\Gamma(\alpha, \beta)$ posterior distribution of λ, i.e. by

$$\begin{aligned}
f(z \,|\, \boldsymbol{t}) &= \int_0^\infty f(z \,|\, \lambda)\pi(\lambda \,|\, \boldsymbol{t})\, d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^z}{z!}\frac{1}{\Gamma(\alpha)}\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}\, d\lambda, \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)z!}\int_0^\infty \lambda^{z+\alpha-1} e^{-(1+\beta)\lambda}\, d\lambda, \\
&= \frac{\beta^\alpha \Gamma(z+\alpha)}{\Gamma(\alpha)z!(1+\beta)^{z+\alpha}} \\
&= \frac{\Gamma(z+\alpha)}{\Gamma(\alpha)z!}p^\alpha(1-p)^z, \qquad\qquad z = 0, 1, 2, \ldots.
\end{aligned}$$

where $p = \beta/(1+\beta)$. Thus the predictive distribution of Z is a *(shifted origin) negative binomial* distribution with parameters α and p (when α is an integer this distribution is the sum of α independent copies of a geometric distribution with starting value shifted from 1 to 0). By considering the cumulative distribution function for the case $\alpha = 21$, $\beta = 10.1$ of Example 1.2 we note that, under this distribution, $\mathbf{P}(Z \leq 3) = 0.83$, while $\mathbf{P}(Z \leq 4) = 0.93$. It follows that 5 components are required to give at last 90% certainty of completing the 1-day flight successfully.

Again an alternative is to simulate a large number n of values of λ from its posterior distribution (with density $\pi(\lambda \,|\, \boldsymbol{t})$). For each $z = 0, 1, 2, \ldots$, we may calculate $f(z \,|\, \lambda)$ (as given by (9)) for the simulated values of λ; the sample mean of these values is then an estimate of $f(z \,|\, \boldsymbol{t})$. *This is left as an exercise using* ***R****.*

These examples of calculating predictive distributions show how mixture distributions naturally arise. In the cases considered so far, there has been only a single parameter and the integrals have been analytically tractable. More generally, Bayesian inference and prediction can require calculation of integrals that may be multidimensional (in the case of more complex models), or may fail to be analytically tractable. One of the barriers to widespread implementation of Bayesian ideas in the past was the complexity of the integrations that naturally arose. As we see later in the course, this difficulty has been overcome to a major extent through the use of stochastic integration techniques, coupled with modern computer power. This allows such integrals to be estimated numerically.

1.5 Multiparameter inference

In many practical situations there will be more than a single unknown parameter to estimate. The Bayesian approach can again be applied except that our one-dimensional integrals considered before now become multivariate integrals. We illustrate the approach in the case of inference for the normal distribution, and may consider other situations later.

1.5.1 Joint inference for the normal mean and variance

Suppose that we observe a random sample $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ from a population whose distribution we believe to be $\mathrm{N}(\mu, \phi)$ where the mean μ and variance ϕ are unknown. Let us suppose that we propose that the *joint prior distribution* of (μ, ϕ) should be the *product* (corresponding to *independence*) of a noninformative, improper, *uniform prior distribution* for μ, and a noninformative, improper, *prior distribution* for ϕ whose *density* is proportional to ϕ^{-1}. (This prior distribution for ϕ is equivalent to the assumption of a uniform prior distribution for $\log \phi$—*exercise!*) Then the *joint prior distribution* for these two parameters has *density*

$$\pi(\mu, \phi) \propto \frac{1}{\phi}, \qquad -\infty < \mu < \infty, \quad 0 < \phi < \infty. \tag{10}$$

To carry out a Bayesian analysis we mirror the procedure in the one-dimensional case. The *likelihood function* is given by

$$\begin{aligned} L(\mu, \phi; \boldsymbol{y})) &\propto \phi^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\phi}\right) \\ &\propto \phi^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2}{2\phi}\right) \\ &\propto \phi^{-n/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\phi}\right), \end{aligned}$$

where, as usual, $\bar{y}$ is the *sample mean* and s^2 is the *sample variance*. Thus the *joint posterior density* $\pi(\mu, \phi \,|\, \boldsymbol{y})$ can be written as

$$\pi(\mu, \phi \,|\, \boldsymbol{y}) \propto \phi^{-(n+2)/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\phi}\right)$$

If we wish to make inference on μ and ϕ separately then we can do this from the *marginal posterior densities* of these parameters. This requires us to integrate the above *posterior density* with respect to ϕ and μ respectively. When we do this we find (see, for example, Gelman *et al*) that

$$\pi(\mu \,|\, \boldsymbol{y}) \propto \left(1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}\right)^{-n/2}.$$

Comparing this with the form of the density of the t-distribution (again see, for example, Gelman *et al*) we find that, *under the posterior distribution*,

$$\frac{\mu - \bar{y}}{s/n^{1/2}} \sim t_{n-1},$$

i.e. has a t-distribution with $n-1$ degrees of freedom. Since μ is a linear function of this quantity, then we can make posterior inferences about it quite easily.

To get the *posterior density* of ϕ we have to integrate the *joint posterior density* with respect to μ. We find that (once more see, for example, Gelman *et al*)

$$\pi(\phi \,|\, \boldsymbol{y}) \propto \psi^{\frac{n+1}{2}} e^{-\frac{1}{2}\psi},$$

where $\psi = \frac{(n-1)s^2}{\phi}$. The posterior density of ψ itself is proportional to $\pi(\phi \,|\, \boldsymbol{y})$ multiplied by the modulus of $\frac{d\phi}{d\psi}$. Therefore we have

$$\pi(\psi \,|\, \boldsymbol{y}) \propto \psi^{\frac{n-3}{2}} e^{-\frac{1}{2}\psi},$$

from which we see that $(n-1)s^2/\phi \sim \chi^2_{n-1}$.

For this choice of *prior distribution* for (μ, ϕ) we see that the *posterior credible intervals* we would calculate for either parameter would correspond exactly with the corresponding confidence intervals calculated using classical methods. Priors which lead to posterior inferences that "match" the results of classical analyses are sometimes called *probability matching priors*.

The above example is a case where things can be tackled analytically. The posterior densities for both parameters can be identified and their properties are well known, and are tabulated in statistical tables. However, in many other situations in Bayesian analyses the resulting integrals will not be so simple. We will then be forced to resort to computational methods of working out posterior densities, or to investigate their properties by simulating directly from them. Simulation methods are at the heart of many developments in modern statistics.

2 Introduction to simulation techniques

2.1 Introduction

The use of **simulation** is to replace difficult or impossible analytical probability calculations with numerical estimation based on either extended or repeated realisations of appropriate probability models.

In this section we consider (generally) simulation from single probability distributions on (possibly subsets) of the real line $\mathbb{R}$.

Why is this useful? Suppose that we are able to simulate a *large* number n of *independent* realisations $\boldsymbol{y} = (y_1, \ldots, y_n)$ (i.e. a random sample of size n) from a distribution with density (or probability) function f on $\mathbb{R}$. Then, for any function g on $\mathbb{R}$, the *(Weak) Law of Large Numbers* gives the estimate

$$\mathbf{E}_f g := \int_{\mathbb{R}} g(y) f(y)\, dy \approx \frac{1}{n} \sum_{i=1}^{n} g(y_i). \tag{11}$$

Here $\mathbf{E}_f g$ is just a compact notation for $\mathbf{E}g(Y)$ where Y is a random variable with distribution given by the density (or probability) function f. Further, it is generally the case, from the *Central Limit Theorem*, that the error in (11) is proportional to $n^{1/2}$, a result which allows us to estimate how large n needs to be for a given level of accuracy.

Hence to estimate, for example, the *mean* of this distribution, or equivalently $\mathbf{E}Y$, we would take $g(y) = y$. To estimate the *second moment (about 0)*, or equivalently $\mathbf{E}Y^2$, we would take $g(y) = y^2$.

To estimate $\mathbf{P}(Y \leq y_0)$ for some given y_0, we would take g to be the indicator function given by

$$g(y) = \begin{cases} 1, & y \leq y_0 \\ 0, & y > y_0. \end{cases}$$

Thus we need techniques for simulating *independent* realisations from given distributions. Usually this is done with a computer package which is able to simulate from a small number of *given* distributions. The challenge is then to use this capability to simulate from any required *target* distribution.

The standard situation is that the package is only able to simulate from the $\mathrm{U}(0,1)$ distribution. However, **R** is able to simulate directly from a wide range of standard distributions, in each case for all values of their parameters.

In practice computer packages generate sequences of *pseudorandom* random numbers. In the case of the $\mathrm{U}(0,1)$ distribution, any such sequence $(u_1, u_2, u_3, \ldots)$ is in fact a *deterministic* function of its initial *seed* u_0, but behaves for all practical applications *as if* it were a sequence of independent realisations from the $\mathrm{U}(0,1)$ distribution.

It is therefore *extremely important* to be able to have confidence in the quality of the package's "random" number generator. In the case of a well-established and tested package such as **R**, there is generally no need for any concern.

In the rest of this section we consider various methods of simulating from given **target** distributions.

2.2 Inverse transform method

Define the *generalised inverse* of a distribution function F by

$$F^{-1}(u) = \min\{y\colon F(y) \geq u\}, \qquad u \in (0,1). \tag{12}$$

Note that F^{-1} is essentially the *quantile function* of the distribution given by F, and that in the case where F is *continuous* and *strictly increasing*, F^{-1} is just the usual inverse of F.

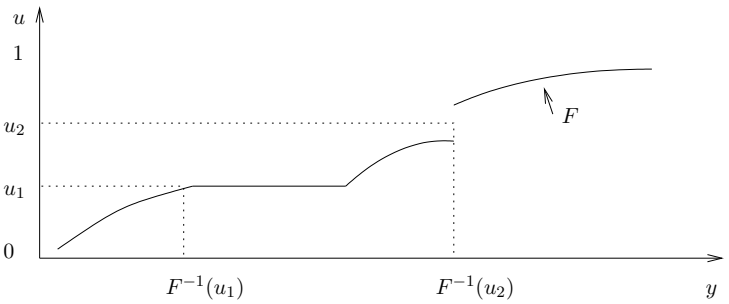

Figure 3: Definition of the generalised inverse function.

Then, since F is increasing, for any $u \in (0,1)$ and any $y \in \mathbb{R}$,

$$F^{-1}(u) \leq y \qquad \Longleftrightarrow \qquad u \leq F(y). \tag{13}$$

Result 2.1 Let F be any given **target** distribution function, and let U be a random variable such that $U \sim \mathrm{U}(0,1)$. Then the random variable $Y := F^{-1}(U)$ has distribution function F.

Proof. For any $y \in \mathbb{R}$, from (13),

$$\mathbf{P}(Y \leq y) = \mathbf{P}(F^{-1}(u) \leq y) = \mathbf{P}(U \leq F(y)) = F(y)$$

(where the last equality above follows since $U \sim \mathrm{U}(0,1)$).

Hence we can, in principle, simulate any distribution on $\mathbb{R}$ by this method.

Limitation: F^{-1} may be too difficult to routinely calculate for some distributions, e.g. the normal.

Example 2.1 *Simulation of the* $\mathrm{Exp}(\lambda)$ *distribution.* Here we have $F(y) = 1 - e^{-\lambda y}$ and so

$$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u), \qquad u \in (0,1).$$

Hence, if $U \sim \mathrm{U}(0,1)$, then $Y := -\frac{1}{\lambda}\ln(1-U)$ has an $\mathrm{Exp}(\lambda)$ distribution. (Note that here $Y' := -\frac{1}{\lambda}\ln U$ also has an $\mathrm{Exp}(\lambda)$ distribution (*why?*).)

Example 2.2 *Simulation of the Bernoulli distribution with parameter* p*, i.e.* $\mathrm{Bin}(1,p)$. This distribution has an atom of probability of size $1-p$ at 0 and an atom of probability of size p at 1. Hence

$$F^{-1}(u) = \begin{cases} 0, & 0 < u \leq 1-p \\ 1, & 1-p < u < 1. \end{cases}$$

Hence, if $U \sim \mathrm{U}(0,1)$, we may simulate the required Bernoulli random variable Y by taking $Y = 0$ if $U \leq 1-p$ and $Y = 1$ otherwise. (Note that it is easy to see this directly (*why?*).)

Note that the method of Example 2.2 is easily seen to apply to the simulation of a random variable Y with any given discrete distribution: divide the interval $(0, 1)$ into intervals with lengths equal to the probabilities associated with the distribution, generate $U \sim \mathrm{U}(0, 1)$ and choose Y according to the interval within which U falls.

2.3 Rejection sampling

This is mainly useful in simulation from continuous distributions, in the case where it is not easy to simulate directly from the *target* distribution, and we will consider the continuous case.

Result 2.2 Suppose that

1. we wish to simulate a realisation of a continuous random variable Y with a given **target** distribution on $\mathbb{R}$ whose *density* is proportional to some function f;
2. we have available a method for simulating from a continuous **envelope** distribution on $\mathbb{R}$ with *density* proportional to some function g and such that, for some constant $M < \infty$,

$$\frac{f(y)}{Mg(y)} \leq 1, \qquad \text{for all } y. \tag{14}$$

Consider the following **algorithm**:

1. Simulate a realisation z of a random variable Z with the given **envelope** distribution (*density* proportional to g);
2. set $Y = z$, i.e. **accept** z, *independently* with probability $\dfrac{f(z)}{Mg(z)}$; otherwise **reject** z, go back to step 1, and perform as many *independent* **repetitions** as are necessary to obtain an *acceptance*.

The Y has the required **target** distribution.

Proof. Suppose that the *envelope* distribution has exact density kg for some (possibly unknown) constant k. From the above construction, based on *independent repetitions*, we have, for any y,

$$\begin{aligned}
\mathbf{P}(Y \leq y) &= \mathbf{P}(Z \leq y \,|\, Z \text{ is accepted}) \\
&= \frac{\mathbf{P}(Z \leq y \text{ and } Z \text{ is accepted})}{\mathbf{P}(Z \text{ is accepted})} \\
&= \int_{-\infty}^{y} \frac{f(z)}{Mg(z)} kg(z)\, dz \Big/ \int_{-\infty}^{\infty} \frac{f(z)}{Mg(z)} kg(z)\, dz \\
&= \int_{-\infty}^{y} f(z)\, dz \Big/ \int_{-\infty}^{\infty} f(z)\, dz\,.
\end{aligned}$$

But this implies that Y has density proportional to f as required.

Notes.

1. At each attempt, i.e. each repetition, in the above procedure, it is necessary to perform two simulations, one to generate z, and one to decide whether to *accept* z. (For the latter we may simulate $U \sim \mathrm{U}(0,1)$ and accept if and only if $U \leq f(z)/Mg(z)$.)
2. The condition (14) ensures that, conditional on the value z of Z obtained in step 1 of the above algorithm, the *acceptance probability* never exceeds one. Clearly, given the density g of the *envelope* distribution, *acceptance probabilities* are maximised by choosing M as small as possible, i.e.

$$M = \sup_y \frac{f(y)}{g(y)}$$

   and further the *unconditional acceptance probability* is maximised by choosing g so that the ratio $f(y)/g(y)$ varies as little as is reasonably possible. (In particular g may not have tails which are essentially lighter than those of f.) The *limitation* of *rejection sampling* is that, for some *target* distributions, it may be difficult to find an *envelope* distribution satisfying these conditions, and from which it is possible to simulate easily.
3. Clearly, in order to obtain a large sample of *independent* realisations form the *target* distribution, our practical procedure is to obtain a, necessarily even larger sample of *independent* realisations form the *envelope* distribution and to keep just those which are accepted.

Example 2.3 Suppose that we wish to simulate from the $\beta(2,2)$ distribution, i.e. with *density proportional* to the function f defined by

$$f(y) = \begin{cases} y(1-y), & y \in [0,1] \\ 0, & \text{otherwise.} \end{cases}$$

Then we could take

$$g(y) = \begin{cases} 1, & y \in [0,1] \\ 0, & \text{otherwise.} \end{cases}$$

Since $f(y)/g(y) \leq 1/4$ for all y (clearly we may restrict y to the interval $[0,1]$) we could then take $M = 1/4$, so that in this case our *rejection sampling algorithm* becomes

1. Simulate a realisation z of a random variable $Z \sim \mathrm{U}(0,1)$ (*density* g);
2. *accept* z, *independently* with probability $4z(1-z)$;
   otherwise *reject* z, go back to the first step, and perform as many *independent repetitions* as are necessary to obtain an *acceptance*.

Some ***R*** *code:*

```
z = runif(10000)            # sample from U(0,1) distribution (density g)
u = runif(10000)
accept = u < 4*z*(1-z)      # logical vector of acceptance decisions
sample = z[accept]          # sample (mean size 6667) with density f

plot(qbeta(ppoints(sample),2,2), sort(sample))     # Q-Q plot to check
```

Example 2.4 Suppose that we wish to simulate a random variable Y with a *target* $\mathrm{N}(0,1)$ distribution.

Since this distribution is symmetric about 0, it is convenient to use rejection sampling to simulate $|Y|$; the random variable Y itself is then obtained by multiplying $|Y|$ by an *independent* random variable which takes the values 1 and -1 with equal probabilities $1/2$.

The random variable $|Y|$ has density proportional to the function f on $\mathbb{R}_+ = [0, \infty)$ defined by

$$f(y) = e^{-y^2/2}.$$

We take as *envelope* the Exp(1) distribution. This has density g on $\mathbb{R}_+$ given by

$$g(y) = e^{-y}$$

and so we make take

$$\begin{aligned} M &= \sup_{y \in \mathbb{R}_+} \frac{f(y)}{g(y)} \\ &= \sup_{y \in \mathbb{R}_+} e^{y - y^2/2} \\ &= e^{1/2}. \end{aligned}$$

Thus our *rejection sampling algorithm* to simulate Y becomes

1. Simulate a realisation z of a random variable $Z \sim \mathrm{Exp}(1)$ (using, e.g. the *inverse transform method*, i.e. $Z = -\ln U$ where $U \sim \mathrm{U}(0,1)$);
2. *accept* z, *independently* with probability $e^{-1/2+z-z^2/2} = e^{-(z-1)^2/2}$; otherwise *reject* z, go back to the first step, and perform as many *independent repetitions* as are necessary to obtain an *acceptance*.
3. Finally the accepted random variable is multiplied by an *independent* random variable which takes the values 1 and -1 with equal probabilities $1/2$.

As usual we generate a large sample of independent realisations of Y (i.e. from the $\mathrm{N}(0,1)$ distribution) by generating an even larger sample of independent realisations of Z and keeping only those which are accepted.

Note also that the *unconditional* acceptance probability here is given by

$$\int_0^\infty e^{-(z-1)^2/2} \, e^{-z} \, dz = \left(\frac{\pi}{2e}\right)^{1/2} = 0.7602$$

*Some **R** code:*

```
z = -log(runif(10000))            # sample from Exp(1) distribution
u = runif(10000)
accept = u < exp(-(z-1)^2/2)      # logical vector of acceptance decisions
sample = z[accept]                # sample (mean size 7602) from dist of |Y|
flip = runif(10000) < 0.5         # logical vector of sign change decisions
sample[flip] = - sample[flip]     # change signs

qqnorm(sample)                    # Q-Q plot to check
```

2.4 Composition method

Result 2.3 Suppose that the distribution function F of the **target** distribution has a representation of the form

$$F(y) = \sum_{i=1}^{n} p_i F_i(y), \tag{15}$$

where, for each i, F_i is a distribution function (for a distribution from which we can easily simulate), and where each $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$.
Consider the following **algorithm**:

1. Choose i with probability p_i.
2. Conditional on the choice i, simulate the random variable Y to have distribution function F_i.

Then the unconditional distribution of Y is given by the **target** distribution function F.

Proof. We have, for any y,

$$\begin{aligned}
\mathbf{P}(Y \le y) &= \sum_{i=1}^{n} \mathbf{P}(i \text{ chosen}) \mathbf{P}(Y \le y \,|\, i \text{ chosen}) \\
&= \sum_{i=1}^{n} p_i F_i(y) \\
&= F(y), \qquad \text{by (15)}.
\end{aligned}$$

Example 2.5 Suppose that a random variable Y with probability 0.9 is sampled from a $\mathrm{N}\left(6, 1^2\right)$ distribution and with probability 0.1 is sampled from a $\mathrm{N}\left(8, 4^2\right)$ distribution (a typical 'contaminated' distribution). Suppose further that we wish to simulate 10000 independent realisations of Y. Then the following ***R*** *code* would be sufficient.

```
y1 = rnorm(10000,6,1)            # sample from N(6,1)
y2 = rnorm(10000,8,4)            # sample from N(8,16)
choice = runif(10000) < 0.9      # logical vector to indicate choice
y = y1*choice + y2*(!choice)     # R converts choice to vector of 1's and 0's !

qqnorm(y)                        # a normal Q-Q plot is instructive
```

2.5 Techniques for particular distributions

1. *Exponential with mean* λ^{-1}: $\mathrm{Exp}(\lambda)$.

   Set $Y = -\ln(U)$ where $U \sim \mathrm{U}(0,1)$.

   (*Inverse transform* method with $1-U$ replaced by U.)

   R: use `rexp`.

2. *Bernoulli with parameter* p: $\mathrm{Bin}(1,p)$.

   Generate $U \sim \mathrm{U}(0,1)$ and set

   $$Y = \begin{cases} 0, & U \le 1-p \\ 1, & \text{otherwise.} \end{cases}$$

   (This is just the implementation of the discrete *inverse transform* method.)

   R: use `rbinom`.

3. *Binomial with parameters* n *and* p: $\mathrm{Bin}(n,p)$.

   If n is not too large $Y \sim \mathrm{Bin}(n,p)$ may be conveniently simulated as the sum of n *independent* Bernoulli random variables each with parameter p.

   Alternatively, we may use the discrete *inverse transform* method: generate $U \sim \mathrm{U}(0,1)$ and set $Y = F^{-1}(U)$ where F^{-1} is the inverse of the distribution function F of the $\mathrm{Bin}(n,p)$ distribution, i.e. is the *quantile* function. This reduces to dividing $(0,1)$ into intervals according to the required binomial probabilities, and choosing Y according to the interval in which U falls.

   R: use `rbinom`.

4. *Geometric with parameter* p: $\mathrm{Geo}(p)$.

   Here we require $\mathbf{P}(Y = k) = (1-p)^{k-1}p$, $k = 1,2,3,\ldots$. One possible algorithm is:

   (a) Generate $Z = \mathrm{Exp}(\lambda)$ where $\lambda = -\ln(1-p)$;
   (b) set $Y = 1 + [Z]$, where $[Z]$ denotes the integer part of Z.

   Proof. We have, for integer $k \ge 1$,

   $$\mathbf{P}(Y \ge k) = \mathbf{P}(Z \ge k-1) = e^{-\lambda(k-1)} = (1-p)^{k-1},$$

   which implies that Y has the required distribution (*why?*).

   R: use `rgeom`. However, note that this simulates from the geometric distribution with 'origin' at 0 (rather than 1), i.e. the distribution with probability function $p(k) = (1-p)^k p$, $k = 0,1,2,\ldots$. Hence for the more usual $\mathrm{Geo}(p)$ distribution we need to add 1 to each of the realisations produced by `rgeom`.

5. *Poisson with parameter (mean)* λ: $\mathrm{Pois}(\lambda)$.

   $Y \sim \mathrm{Pois}(\lambda)$ can be simulated as the number of events by time λ in a Poisson process with rate 1: thus, for a sequence $U_1, U_2, \ldots$ of independent $\mathrm{U}(0,1)$ random variables, we may take

   $$\begin{aligned} Y &= \max\{k\colon\ -\sum_{i=1}^{k} \ln U_i \le \lambda\} \\ &= \max\{k\colon\ \prod_{i=1}^{k} U_i \ge e^{-\lambda}\}. \end{aligned}$$

To implement this, we need to form the successive products $\prod_{i=1}^{k} U_i$, until we first obtain N such that $\prod_{i=1}^{N} U_i > \lambda$, and then take $Y = N - 1$.

Here is a simple (but not very efficient) **R** function to generate a random sample of a Pois(λ) random variable:

```
spois = function(n, lambda){
  expl = exp( - lambda)
  z = numeric(n)                        #create vector to hold result
  for(i in 1:n) {                       #loop of size size
      m = 0                             #initialize Poisson count
      prod = 1                          #initialize product
      while(prod > expl) {              #loop to create count
          m = m + 1
          prod = prod * runif(1)
      }                                 #end of while loop
      z[i] = m - 1                      #define value of realisation i
  }                                     #end of for loop
  return(z)                             #return value of function
}                                       #end of function definition
```

Simulations from the Pois(λ) distribution are also readily obtained via the discrete *inverse transform* method.

In practice it is simpler to use the built-in **R** function `rpois`, which allows us to simulate a sample of arbitrary size.

6. *Normal with mean μ and standard deviation σ*: $\mathrm{N}(\mu, \sigma)$.

   Note first that if $Z \sim \mathrm{N}(0,1)$, and $Y = \mu + \sigma Z$, then $Y \sim \mathrm{N}(\mu, \sigma)$, so that it is sufficient to consider simulation from the $\mathrm{N}(0,1)$ distribution.

   One possibility is to use *rejection sampling* as in Example 2.4.

   Another possibility is the *Box-Müller* method for simulating a pair of independent $\mathrm{N}(0,1)$ random variables:

   (a) Generate independent $\mathrm{U}(0,1)$ random variables U_1, U_2.
   (b) Set $\Theta = 2\pi U_1$ and $R = (-2 \ln U_2)^{1/2}$.
   (c) Then $X := R\cos\Theta$ and $Y := R\sin\Theta$ are independent $\mathrm{N}(0,1)$ random variables.

   Proof. (*Outline.*) Clearly Θ and R are independent, $\Theta \sim \mathrm{U}(0, 2\pi)$, and (by the inverse transform method) R has distribution function F given by $F(r) = 1 - e^{-r^2/2}$. Consideration of the transformation $(R, \Theta) \longrightarrow (X, Y)$ (polar to Cartesian coordinates) now gives that X and Y are independent $\mathrm{N}(0,1)$ random variables. (*Exercise!*)

3 Markov chain Monte Carlo

3.1 Introduction to Markov chains

3.1.1 Definitions and examples

A *discrete time stochastic (i.e. random) process* $\{X_n\}_{n\geq 0}$ taking values in a *discrete state space* S (whose **states** are typically labelled by the integers or some subset of the integers) is a **Markov chain** if and only if

$$\mathbf{P}(X_{n+1} = j \,|\, X_n = i,\, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbf{P}(X_{n+1} = j \,|\, X_n = i)$$

for all $n \geq 0$ and for all $j, i, i_{n-1}, \ldots, i_1 \in S$.

Thus, *given* the evolution of the process $\{X_n\}_{n\geq 0}$ up to any "current" time n, the probabilistic description of its behaviour at time $n+1$ (and, by induction, the probabilistic description of all its subsequent behaviour) depends only on the current state $\{X_n = i\}$, and not on the previous history of the process. This is the **Markov property**.

Additionally the *Markov chain* $\{X_n\}_{n\geq 0}$ is **time homogeneous**, if, for all $i, j \in S$, there is some probability p_{ij} such that

$$\mathbf{P}(X_{n+1} = j \,|\, X_n = i) = p_{ij}$$

independently of n.

The matrix $P = (p_{ij})_{i,j\in S}$ is then referred to at the **transition matrix** of the Markov chain.

Note that necessarily

$$p_{ij} \geq 0 \;\text{ for all } i, j \in S, \qquad \sum_{j\in S} p_{ij} = 1 \;\text{ for all } i \in S,$$

i.e. P is a **stochastic matrix**.

For a *time homogeneous Markov chain* $\{X_n\}_{n\geq 0}$, given its evolution up to any "current" time n, the probabilistic description of its behaviour at time $n+1$ (and, by extension, the probabilistic description of all its subsequent behaviour) depends only on the current state $\{X_n = i\}$, and not on the previous history of the process *nor* on the time n itself.

Except where explicitly stated otherwise, we shall assume that *all Markov chains are time homogeneous.*

Example 3.1 *Device state.* Suppose that a device can be in one of three states: 1 = working properly; 2 = working badly; 3 = broken. Its states on successive days might form a *Markov chain* with *transition matrix*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{3}{4} & \frac{1}{4} \\ 1 & 0 & 0 \end{pmatrix}.$$

Thus, given the states of the device up to any day n, its state on day $n+1$ depends (statistically) only on its state on day n and not on its state on earlier days. Note that if the device is broken on one day, it has been replaced and is working properly the next day.

Interesting questions:

1. if the device is in state 1 on day 0, what is the probability it is in state 1 on day n?
2. if the device is in state 1 on day 0, what is the probability it remains in state 1 every day up to and including day n?
3. what is the average time between successive replacements of the device (note that this is here the reciprocal of the long-run frequency of time spent in state 3)?
4. what are the long-run proportions of time spent in each of the three states?

Note: the last two quantities above are in principle random variables, though we shall see later that, for this example, they are (almost surely) constant.

Example 3.2 *Device with absorbing state.* Consider again Example 3.1, but suppose instead that when the device is broken it is not repaired. Then the process of successive states remains a *Markov chain* with new *transition matrix*

$$P' = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus once the device enters state 3 it remains within that state forever, and can never again reach either state 1 or state 2. The state 3 is referred to as **absorbing**.

Example 3.3 *Random walk on the integers.* Let a be a integer constant and let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed integer-valued random variables. Define the process $\{X_n\}_{n\geq 0}$ by

$$X_0 = a, \qquad X_n = a + \sum_{i=1}^{n} \xi_i, \quad n \geq 1.$$

Then $\{X_n\}_{n\geq 0}$ is a **random walk**.

To see that $\{X_n\}_{n\geq 0}$ is a *Markov chain*, observe that, for any n and any set of states $i_0, \ldots, i_{n-1}, i, j$

$$\begin{aligned} \mathbf{P}(X_{n+1} = j \,|\, X_n = i,\, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) &= \mathbf{P}(\xi_{n+1} = j - i) \qquad \text{(by independence)} \\ &= \mathbf{P}(X_{n+1} = j \,|\, X_n = i). \end{aligned}$$

The *random walk* $\{X_n\}_{n\geq 0}$ is **simple** if the random variables ξ_i can only take the values $+1$ or -1, in which case there is some probability p between 0 and 1 such that

$$\begin{aligned} \mathbf{P}(\xi_i = 1) &= p \\ \mathbf{P}(\xi_i = -1) &= q, \end{aligned}$$

where $q = 1 - p$.

Interesting questions: perhaps the most interesting question for the *simple random walk* is to determine the probability, starting at $X_0 = a$, where $a > 0$, that the process $\{X_n\}$ ever hits the state 0 (at some random future time). This is the *probability of ruin.*

Example 3.4 *Random walk with reflecting barrier(s).* Consider again the *simple random walk* $\{X_n\}_{n\geq 0}$ defined above, but with *state space* S restricted to the *non-negative integers*, i.e. $S = \{0, 1, 2, \dots\}$. Suppose further that the behaviour of the random walk is *modified* so that whenever the process is in state 0, its next step is to remain in state 0 with probability q or to go to state 1 with probability p (where p and q are as defined above), *independently of all previous moves*. Then it is again easy to see that this modified process is a *Markov chain*. The state 0 is referred to as a **reflecting barrier** for the chain. Markov chains of this sort arise frequently in queueing theory and in the theory of population processes.

A further possible modification is to have an additional (downwards) *reflecting barrier* at some state $a > 0$, so that the state space becomes $S = \{0, 1, 2, \dots\ a\}$

Interesting question: in either case calculate the long-term proportion of time spent in each state.

3.1.2 Simple probability calculations

Suppose that $\{X_n\}_{n\geq 0}$ is a *time homogeneous Markov chain* with transition matrix $P = (p_{ij})_{i,j\in S}$.

Further, let $\boldsymbol{\mu} = (\mu_i)_{i\in S}$ be the *distribution* of the *initial random variable* X_0, i.e. $\mu_i = \mathbf{P}(X_0 = i)$.

Then the probabilistic description of the entire process $\{X_n\}_{n\geq 0}$ is determined by $\boldsymbol{\mu}$ and P. We may think of $\boldsymbol{\mu}$ as representing (the probability of) the *initial state* of the process, and P as representing its subsequent *dynamics*.

In particular we have, for any sequence of states $i_0, \dots, i_n$,

$$\mathbf{P}(X_0 = i_0,\, X_1 = i_1, \dots, X_n = i_n) = \mu_{i_0} p_{i_0 i_1} \dots p_{i_{n-1} i_n}, \tag{16}$$

and so more general probabilities can also be calculated by summing the probabilities of elementary events of this form.

For any $n \geq 0$, define the **n-step transition probabilities**

$$p_{ij}^{(n)} = \mathbf{P}(X_{m+n} = j \,|\, X_m = i), \qquad i, j \in S.$$

These are *independent* of m by *time homogeneity*. Note also that $p_{ij}^{(0)} = \delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Then, from (16), writing i for i_0, j for i_n, summing probabilities over all intermediate states $i_1, \dots, i_{n-1}$, and finally conditioning on $X_0 = i$ (or equivalently taking $\mathbf{P}(X_0 = i) = 1$), we have

$$p_{ij}^{(n)} = (P^n)_{ij}, \qquad i, j \in S,$$

i.e. the *n-step transition probabilities $p_{ij}^{(n)}$, are given by (the components of) the matrix P^n*. This is also necessarily a *stochastic matrix*. Further it should also be possible to see the above result directly.

Similarly, from (16), writing j for i_n, and summing probabilities over all intermediate states $i_0, i_1, \dots, i_{n-1}$, we have

$$\mathbf{P}(X_n = j) = (\boldsymbol{\mu} P^n)_j, \qquad j \in S,$$

i.e. the *distribution* of X_n is given by the vector $\boldsymbol{\mu} P^n$.

Finally we have the **Chapman-Kolmogorov equations**: for any $m > 0$, $n > 0$, and $i, j \in S$,

$$\begin{aligned} p_{ij}^{(m+n)} &= \mathbf{P}(X_{m+n} = j \,|\, X_0 = i) \\ &= \sum_{k \in S} \mathbf{P}(X_n = k,\, X_{m+n} = j \,|\, X_0 = i) \\ &= \sum_{k \in S} \mathbf{P}(X_n = k \,|\, X_0 = i) \mathbf{P}(X_{m+n} = j \,|\, X_0 = i, X_n = k) \\ &= \sum_{k \in S} \mathbf{P}(X_n = k \,|\, X_0 = i) \mathbf{P}(X_{m+n} = j \,|\, X_n = k) \qquad \text{(by the Markov property)} \\ &= \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)} \end{aligned}$$

and the **Chapman-Kolmogorov inequalities**: for any $m > 0$, $n > 0$, and $i, j, k \in S$,

$$p_{ij}^{(m+n)} \geq p_{ik}^{(m)} p_{kj}^{(n)},$$

which follow immediately from the *Chapman-Kolmogorov equations*.

Example 3.5 Consider again Example 3.1 (*device state*). The matrices of the n-step transition probabilities, for $n = 1, 2, \ldots, 6$, are given (to 3 sig. figs.) by

$$P = \begin{pmatrix} 0.500 & 0.250 & 0.250 \\ 0.000 & 0.750 & 0.250 \\ 1.000 & 0.000 & 0.000 \end{pmatrix} P^2 = \begin{pmatrix} 0.500 & 0.312 & 0.188 \\ 0.250 & 0.562 & 0.188 \\ 0.500 & 0.250 & 0.250 \end{pmatrix} P^3 = \begin{pmatrix} 0.438 & 0.359 & 0.203 \\ 0.312 & 0.484 & 0.203 \\ 0.500 & 0.312 & 0.188 \end{pmatrix}$$

$$P^4 = \begin{pmatrix} 0.422 & 0.379 & 0.199 \\ 0.359 & 0.441 & 0.199 \\ 0.438 & 0.359 & 0.203 \end{pmatrix} P^5 = \begin{pmatrix} 0.410 & 0.390 & 0.200 \\ 0.379 & 0.421 & 0.200 \\ 0.422 & 0.379 & 0.199 \end{pmatrix} P^6 = \begin{pmatrix} 0.405 & 0.395 & 0.200 \\ 0.390 & 0.410 & 0.200 \\ 0.410 & 0.390 & 0.200 \end{pmatrix}$$

Note the rapidity with which, for any fixed j, the dependence on i of $p_{ij}^{(n)}$ declines as n increases—the chain tends to quickly "forget its initial state".

3.2 Stationary distributions

Throughout this and the succeeding sections $\{X_n\}_{n \geq 0}$ continues to be a *time homogeneous Markov chain* with *state space* S and *transition matrix* $P = (p_{ij})_{i,j \in S}$.

We further assume throughout that the *Markov chain* $\{X_n\}_{n \geq 0}$ is **irreducible**. This is defined by the requirement that all states of the chain **intercommunicate**, i.e. that for every $i, j \in S$, there is some $n \geq 0$ such that $p_{ij}^{(n)} > 0$, so that it is possible to reach every state of the chain from every other state of the chain in some number of steps with *strictly positive* probability. Equivalently, the entire state space S of the chain is said to form a *single (closed) class* in this case. *Irreducible* chains feature in many applications.

Finally we assume—mainly for simplicity—that the *Markov chain* $\{X_n\}_{n \geq 0}$ is **aperiodic**. [A chain is **periodic** if there is some $d > 1$ such that, starting in any given state, the chain can only return to that state at multiples of the time d; otherwise it is **aperiodic**. *Periodicity* is just a nuisance, both for theory and practice. It is easy to see how to deal with it once the *aperiodic* case is well-understood.]

Recall also that a vector $\boldsymbol{\pi} = (\pi_i)_{i \in S}$ on S is a **(probability) distribution** if and only if $\pi_i \geq 0$ for all i and $\sum_{i \in S} \pi_i = 1$.

Definition. A *(probability) distribution* $\boldsymbol{\pi}$ on S is **stationary** for the Markov chain $\{X_n\}_{n\geq 0}$ if and only if

$$\boldsymbol{\pi}P = \boldsymbol{\pi} \qquad \text{or equivalently} \qquad \sum_{i\in S} \pi_i p_{ij} = \pi_j, \quad j \in S. \tag{17}$$

We then have also, for all $n \geq 1$,

$$\boldsymbol{\pi}P^n = \boldsymbol{\pi} \qquad \text{or equivalently} \qquad \sum_{i\in S} \pi_i p_{ij}^{(n)} = \pi_j, \quad j \in S. \tag{18}$$

i.e. if the *distribution* of X_0 is $\boldsymbol{\pi}$, then, for all $n \geq 1$, the *distribution* of X_n is *also* $\boldsymbol{\pi}$.

A **stationary distribution** is also known as an **equilibrium distribution** for the chain. An informal interpretation (which will be made formal in Section 3.5) is that it gives the *long-term proportion of time* spent in each of the states.

Important results. (See Section 3.4 for more details.)

- The Markov chain $\{X_n\}_{n\geq 0}$ *may or may not* have a stationary distribution;
- (under our assumption that the chain is *irreducible*) when the chain *does have a stationary distribution* this is *unique*;
- when S is *finite*, the chain *always* has a stationary distribution.

Note also that if a *distribution* $\boldsymbol{\pi}$ is *stationary*, then $\pi_j > 0$ for all $j \in S$. This follows since there is clearly *some* i such that $\pi_i > 0$ (since $\sum_{i\in S} \pi_i = 1$), and then, since the chain is assumed to be *irreducible*, for any other $j \in S$, there is some n such that $p_{ij}^{(n)} > 0$, and so, from (18), $\pi_j > 0$.

To *find* the *stationary distribution* (given the *transition matrix* P) we may either *solve* the equations (17), or simply *guess* the answer (which is often possible) and *verify* that it *satisfies* the equations (17).

In Section 3.3 below we consider a further possibility, which only works for certain *transition matrices* P.

Example 3.6 *No-claims discount scheme.* Consider the transition matrix

$$P = \begin{pmatrix} p & 1-p & 0 \\ p & 0 & 1-p \\ p & 0 & 1-p \end{pmatrix},$$

which might be appropriate to modelling the levels of an individual in successive years in a simple no-claims discount scheme.

The equations (17) for the *stationary distribution* $\boldsymbol{\pi}$ here become

$$\begin{aligned} p\pi_1 + p\pi_2 \quad + p\pi_3 \quad &= \pi_1 \\ (1-p)\pi_1 \qquad\qquad\quad &= \pi_2 \\ (1-p)\pi_2 + (1-p)\pi_3 \quad &= \pi_3 \end{aligned}$$

While this looks like 3 equations in 3 unknowns, *any given one of these equations is implied by the remainder* (as is always the case with the equations (17)—add them

up to see this), and we must add the *further requirement* that $\boldsymbol{\pi}$ be a *distribution*, i.e.

$$\pi_1 + \pi_2 + \pi_3 = 1.$$

We thus obtain that the *Markov chain* corresponding to the *transition matrix* P has the unique *stationary distribution*

$$\boldsymbol{\pi} = (p,\, p(1-p),\, (1-p)^2).$$

Example 3.7 *Simple random walk.* For the simple random walk we can check directly from the equations (17) that there is *no* stationary distribution (*exercise, if you like—you will need to remember that the components of a stationary distribution sum to* 1). However, suppose the contrary, i.e. that a, necessarily unique, stationary distribution π does exist. Then by the *(spatial) translation invariance* of the random walk, π_i is constant for all $i \in S$ and also $\sum_{i\in S} \pi_i = 1$, which is impossible since S is infinite. Indeed this argument shows that *any* random walk fails to have a stationary distribution.

3.3 Detailed balance

We continue to assume that the *time-homogeneous Markov chain* $\{X_n\}_{n\geq 0}$ is *irreducible* and *aperiodic*.

Implicit in the ideas of the Section 3.2 is that we wish to solve the following problems:

1. determine *whether* there is a *stationary distribution* for the Markov chain $\{X_n\}_{n\geq 0}$; (this always exists when S is *finite*);
2. find it.

Both these problems may always be solved, in principle, by looking for a *distribution* $\boldsymbol{\pi}$ ($\pi_i \geq 0$ for all i and $\sum_{i\in S} \pi_i = 1$) which satisfies the equations (17).

The solution of the equations (17) can be relatively difficult. However, *sometimes* there is an *alternative* approach to the solution of both the above problems.

We shall say that the transition matrix P possesses the **detailed balance** property if there is a *strictly positive* vector $\boldsymbol{\pi}$ on S satisfying the **detailed balance equations**

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \text{for all } i, j \in S. \tag{19}$$

We now have the following result.

Theorem 3.1 Suppose that the *transition matrix* P possesses the *detailed balance* property, i.e. there exists a strictly positive vector $\boldsymbol{\pi}$ on S satisfying the equations (19).

(a) If $\sum_{i\in S} \pi_i < \infty$ (which is always the case when S is finite), then we can then choose (*normalise*) $\boldsymbol{\pi}$ so that $\sum_{i\in S} \pi_i = 1$, and $\boldsymbol{\pi}$ is then the, *necessarily unique, stationary distribution* for the chain.

(b) If $\sum_{i\in S} \pi_i = \infty$, then there is *no stationary distribution* for the chain.

Proof. We prove (a) only. This follows from the observation that if a *distribution* $\boldsymbol{\pi}$ satisfies the *detailed balance equations* (19), then, for each fixed j,

$$\begin{aligned}\sum_{i\in S} \pi_i p_{ij} &= \pi_j \sum_{i\in S} p_{ji} \\ &= \pi_j.\end{aligned}$$

and so, by (17), $\boldsymbol{\pi}$ is *stationary*.

Only some *transition matrices* possess the *detailed balance property*—in (19) there are typically far more equations than unknowns. When the *transition matrix* P *does* possess *detailed balance*, the equations (19) *are much simpler to solve* than the usual equations (17) for the *stationary distribution*. However, when the *transition matrix* P *does not* possess *detailed balance*, we can conclude *nothing* by the approach of this section (not even whether a stationary distribution exists), and must revert to the approach of the previous section.

One case where the *transition matrix* P *always* possesses *detailed balance* is where it is possible to *order* the states of S so that one-step transitions are possible between and only between *neighbouring* states. The *equations* (19) may then be solved recursively.

A *Markov chain* for which the *detailed balance equations* have a solution $\boldsymbol{\pi}$ with $\sum_{i\in S}\pi_i < \infty$, which may then be *normalised* to the stationary distribution, and which is started with this *stationary distribution*, is often referred to as **reversible**.

Example 3.8 *No-claims discount scheme.* Consider again the Example 3.6 with the *transition matrix*

$$P = \begin{pmatrix} p & 1-p & 0 \\ p & 0 & 1-p \\ p & 0 & 1-p \end{pmatrix}.$$

The *detailed balance equations* (19) are here

$$\begin{aligned} \pi_1(1-p) &= \pi_2 p \\ 0 &= \pi_3 p \\ \pi_2(1-p) &= 0 \end{aligned}$$

and it is clear that we *cannot* find a probability distribution $\boldsymbol{\pi}$ satisfying these equations. Nevertheless we have already shown that the *distribution*

$$\boldsymbol{\pi} = (p,\, p(1-p),\, (1-p)^2)$$

is *stationary* for this chain.

Example 3.9 *n-state system with cyclic symmetry.* We shall take $n = 3$; the argument for any $n > 3$ is the same. Consider a 3-state *Markov chain* with *transition matrix*

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix}.$$

where $0 < p < 1$ and $p + q = 1$. The *detailed balance equations* (19) are here

$$\begin{aligned} \pi_1 p &= \pi_2 q \\ \pi_2 p &= \pi_3 q \\ \pi_3 p &= \pi_1 q, \end{aligned}$$

which have a solution if and only if $p = q = 1/2$, giving the *stationary distribution* $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$ in this case. If $p \neq q$, the *detailed balance equations* do *not* have a solution, and the equations (17) must be solved instead to find the *stationary distribution* (which is still $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$).

Example 3.10 *Simple random walk with reflecting barrier at* 0. In Example 3.4 we introduced the *simple random walk* on the *nonnegative integers* modified so that the state 0 is a *reflecting barrier*. This has *transition matrix* $P = (p_{ij})$ where

$$\begin{aligned} p_{i,i+1} &= p, \qquad i \geq 0 \\ p_{i,i-1} &= q, \qquad i \geq 1 \\ p_{ij} &= 0, \qquad \text{for all other pairs } i, j, \end{aligned}$$

where, as usual, $p + q = 1$. The (nontrivial) *detailed balance equations* are here

$$\pi_i p = \pi_{i+1} q, \qquad i \geq 0,$$

which may be solved recursively to give

$$\pi_i = \pi_0 \left(\frac{p}{q}\right)^i, \qquad i \geq 0.$$

Hence we have $\sum_{i\geq 0} \pi_i < \infty$ if and only if $p < q$, i.e. $p < 1/2$. In this case for $\sum_{i\geq 0} \pi_i = 1$ we require $\pi_0 = 1 - p/q$. We deduce that a *stationary distribution* exists if and only if $p < 1/2$ and is then given by $\boldsymbol{\pi}$ where

$$\pi_i = \left(1 - \frac{p}{q}\right)\left(\frac{p}{q}\right)^i, \qquad i \geq 0.$$

Example 3.11 *Markov chain Monte Carlo.* Suppose that we have a **target** distribution $\boldsymbol{\pi}$ on a *state space* S such that $\pi_i > 0$ for all $i \in S$. We wish to construct a *Markov chain* on S for which $\boldsymbol{\pi}$ is the *stationary distribution*.

Let $Q = (q_{ij})_{i,j\in S}$ be *any stochastic matrix* which would itself define an *irreducible* chain; thus, for each i, the vector of probabilities $(q_{ij})_{j\in S}$ forms a probability distribution—the **proposal** distribution. Define the *Markov chain* $\{X_n\}_{n\geq 0}$ with *transition matrix* $P = (p_{ij})_{i,j\in S}$ given by, for all i,

$$\begin{aligned} p_{ij} &= \min\left(q_{ij}, \frac{\pi_j q_{ji}}{\pi_i}\right), \qquad j \neq i, \\ p_{ii} &= 1 - \sum_{j\neq i} p_{ij}. \end{aligned}$$

Then, for any i, j with $j \neq i$, we have

$$\begin{aligned} \pi_i p_{ij} &= \min(\pi_i q_{ji}, \pi_j q_{ji}) \\ &= \pi_j p_{ji}, \end{aligned}$$

and so the *distribution* $\boldsymbol{\pi}$ here satisfies the *detailed balance equations* for the chain $\{X_n\}_{n\geq 0}$, and so is the *stationary distribution* of this chain.

3.4 Stationary and limiting distributions

Recall that the Markov chain $\{X_n\}_{n\geq 0}$ is assumed to be *irreducible* (S is a *single closed class*), and that it is also assumed to be *aperiodic*.

The following result summarises the connection between *stationary* and *limiting distributions* for the chain.

Theorem 3.2 *Either*

(a) there is a *unique stationary distribution* $\boldsymbol{\pi}$ for the chain $\{X_n\}_{n\geq 0}$, and then

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j \qquad \text{for all } i,j \in S;$$

in this case the chain is said to be **ergodic**; *or*

(b) there is *no stationary distribution* for the chain, and then

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0 \qquad \text{for all } i,j \in S.$$

In the case where S is finite the case (a) always holds, i.e. when S is *finite*, an *irreducible aperiodic* chain is always *ergodic*.

Proof. We give a outline proof for the case where S is *finite*, i.e. we show that here the result (a) follows.

First, since the chain is *irreducible* and *aperiodic* a *coupling* argument, discussed in the lectures, shows that there is some vector $\boldsymbol{\pi}$ such that, for all $i,j \in S$,

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j. \tag{20}$$

independently of i. (This implies in particular that the chain "eventually forgets its initial state".) Since, for any i and for all n, $\sum_{j\in S} p_{ij}^{(n)} = 1$, letting $n \to \infty$ and using (20), we obtain

$$\sum_{j\in S} \pi_j = 1.$$

Further, for any $n \geq 1$ and for any $i,j \in S$, we have, by the *Chapman-Kolmogorov equations* (or just see it directly)

$$p_{ij}^{(n+1)} = \sum_{k\in S} p_{ik}^{(n)} p_{kj}.$$

Again letting $n \to \infty$ and using (20), we obtain, for all j,

$$\pi_j = \sum_{k\in S} \pi_k p_{kj}.$$

Hence $\boldsymbol{\pi}$ is a *stationary distribution*.

To prove *uniqueness*, suppose that $\boldsymbol{\pi}'$ is *any stationary distribution* for the chain. Then, letting $n \to \infty$ in equation (18), we obtain, for all j,

$$\begin{aligned}\pi'_j &= \sum_{i\in S} \pi'_i \pi_j \\ &= \pi_j \sum_{i\in S} \pi'_i \\ &= \pi_j\end{aligned}$$

since $\sum_{i\in S} \pi'_i = 1$.

Note. When the *Markov chain* is *ergodic*, i.e. the case (a) of Theorem 3.2, it follows that all states of the chain are *positive persistent* (or *positive recurrent*). In the case (b) it turns out that either all states are *transient* or all states are *null persistent*.

Example 3.12 *No-claims discount scheme.* In the earlier Example 3.8 it is clear that the Markov chain is *aperiodic* and well as being *irreducible*, and we conclude that, for all $i, j \in S$, $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$, where $\boldsymbol{\pi} = (p,\, p(1-p),\, (1-p)^2)$.

3.5 The ergodic theorem for Markov chains

We suppose now that the chain $\{X_n\}_{n\geq 0}$ is **ergodic**, i.e that it is *irreducible* and *aperiodic*, and there exists a (necessarily unique) *stationary distribution* $\boldsymbol{\pi}$.

From Theorem 3.2, for all $i, j \in S$ we have

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \tag{21}$$

and also $\pi_j > 0$ for all j. It follows from the *positive persistence* (*positive recurrence*) of S in this case that, with probability 1, all states of the chain are visited *infinitely often*. We might expect from (21) that, for each j, π_j would correspond to the *long-term proportion of time spent in state* j. That this is true is a special case of the **Ergodic Theorem** below.

Theorem 3.3 (Ergodic theorem for Markov chains) Suppose that $\{X_n\}_{n\geq 0}$ is *ergodic* with stationary distribution $\boldsymbol{\pi}$. Then, for any function f on the state space S such that $\sum_{i\in S} \pi_i |f(i)| < \infty$ (i.e. f has a *finite expectation* with respect to the stationary distribution $\boldsymbol{\pi}$),

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(X_n) = \sum_{i\in S} \pi_i f(i) \qquad \text{with probability 1.} \tag{22}$$

For the special case referred to above, fix any $j \in S$ and define the function f_j on S by

$$f_j(i) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Then the left side of equation (22) is precisely the *long-term proportion of time spent in state* j, while the right side is simply π_j, so that the *ergodic theorem* here does indeed assert that, with probability 1, the *long-term proportion of time spent in state* j is equal to π_j.

Proof. We give a brief outline of the proof of the *ergodic theorem*. It is sufficient to prove it for every function f_j of the form defined above—the result for more general functions f (with finite expectation with respect to π) is then easily deduced.

The **strong law of large numbers** is a key result of basic probability theory which asserts that, with probability 1, the *average* of a sum of n *independent identically distributed random variables* converges, as $n \to \infty$, to their common *expectation*. It follows fairly simply from this that, for any $j \in S$, there is some constant π'_j such that, for any *initial distribution* of the chain,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_j(X_n) = \pi'_j \qquad \text{with probability 1.} \tag{23}$$

Taking the initial distribution to be $\boldsymbol{\pi}$, and taking expectations in (23), it follows that $\pi'_j = \pi_j$ as required.

3.6 MCMC: Introduction

3.6.1 Objective

Often we wish to calculate functions of a distribution $\boldsymbol{\pi}$, as given by its *probability function* or *density function* $\pi(\cdot)$, and defined on some space S. For example, $\boldsymbol{\pi}$ might be a *Bayesian posterior distribution*, and we might wish to calculate its *mean*, its *variance*, or the *probability* it takes a particular value or lies within a given range.
In general such quantities can be expressed as the *expectation*

$$\mathbf{E}_{\boldsymbol{\pi}} g = \begin{cases} \int_{x \in S} \pi(x) g(x)\, dx & \text{if } S \text{ is continuous} \\ \sum_{x \in S} \pi(x) g(x) & \text{if } S \text{ is discrete} \end{cases}$$

of some *function* g on S with respect to $\boldsymbol{\pi}$.

Examples

1. For the *mean* μ of $\boldsymbol{\pi}$, we take $g(x) = x$.
2. For the *variance* of $\boldsymbol{\pi}$, we could take $g(x) = (x - \mu)^2$, where the mean μ has been previously calculated or estimated.
3. For the *probability* assigned by $\boldsymbol{\pi}$ to the *interval* $[a, b]$, we take $g(x) = 1$ if $a \leq x \leq b$ and $g(x) = 0$ otherwise.
4. For the *probability* assigned by $\boldsymbol{\pi}$ to any particular *value* a we take $g(x) = 1$ if $x = a$ and $g(x) = 0$ otherwise (this is really a special case of the previous example).

Suppose now that $\pi(\cdot)$ is only known up to a *multiplicative constant*—the exact value of the *normalising constant* being too difficult to calculate. This is often the case in applied probability models with complex constraints and in Bayesian statistics (where $\boldsymbol{\pi}$ is the *posterior distribution* of the parameters to be estimated.)
We regard $\boldsymbol{\pi}$ as a **target distribution**, and construct an **irreducible Markov chain** $\{X_n\}_{n \geq 0}$ with **transition matrix** or **kernel** $P = (p(x, y))_{x,y \in S}$ for which $\boldsymbol{\pi}$ is the **stationary** or **invariant** distribution.
Then, by the **ergodic theorem**, for any function g on S such that $\mathbf{E}_{\boldsymbol{\pi}} g$ is finite, and for any time k

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=k+1}^{k+N} g(X_n) = \mathbf{E}_{\boldsymbol{\pi}} g \tag{24}$$

Hence we can *estimate* $\mathbf{E}_{\boldsymbol{\pi}} g$ by *simulating* the chain for a sufficiently long period of time to obtain a good approximation to the left side, and so also the right side, of (24). Note that we typically neglect an initial segment of the chain of length k—see below.
Recall also that, when the chain is *aperiodic*, the *distribution* of X_n converges to $\boldsymbol{\pi}$. This remains true, in a suitably time-averaged sense, in the *periodic* case.
Note that for a *given distribution* $\boldsymbol{\pi}$, there is a huge choice of *Markov chains* which have $\boldsymbol{\pi}$ as their *stationary distribution*.
Notation: as previously we will use a common notation for the discrete and continuous cases. In particular, in the *continuous* case $p(x, \cdot)$ represents the *density* of the distribution of X_{n+1} conditional on $X_n = x$, while in the *discrete* case $p(x, \cdot)$ represents the *probability (mass) function* of the distribution of X_{n+1} conditional on $X_n = x$ (so that here $P = (p(x, y))_{x,y \in S}$ is just the usual transition matrix of the chain).

3.6.2 Notes

1. **Burn-in.** It is desirable to choose k *sufficiently large* that the *distribution* of X_k, i.e. of the *chain* at time k, has converged close to its *stationary distribution* $\boldsymbol{\pi}$. This greatly increases the *accuracy* of *approximations* based on (24). There is a huge literature on this, but in summary approaches to the choice of k are:

   (a) *Theoretical.* If possible the chain $\{X_n\}_{n\geq 0}$ (i.e. the *kernel* of transition probabilities P) should be chosen so that the convergence of the distribution of X_n to $\boldsymbol{\pi}$ is *geometrically* (*exponentially*) fast—ideally *uniformly* over all starting points. Then k can be taken to be small.

   (b) *Diagnostic.* In particular appropriate *plots* (e.g. simple plots of X_n, or $g(X_n)$, against n) may be used to assess whether, by some time k, the distribution of X_n has settled sufficiently close to $\boldsymbol{\pi}$ for all $n \geq k$.

   The length of the **burn-in** period depends crucially on the choice of chain. However, in most situations the length N of the simulation required for reasonably accurate estimation is such that the burn-in period k need not exceed more than 1–2% of the total length $k + N$ of the simulation.

2. **Efficiency.** It is important that the chain should **mix** reasonably rapidly, and in particular that it should not get stuck in particular states for lengthy periods of time (because of low *acceptance probabilities*—see below). The ideal situation is when, for all x, $p(x, \cdot)$ is reasonably close to the *target distribution* $\boldsymbol{\pi}$, so that successive observations of the chain are close to being independent identically distributed. (Of course if this could be achieved exactly there would be no need to think in terms of MCMC!)

   In *good* situations (e.g. when S is *finite*, and under *stationarity*)

   $$\frac{1}{N}\sum_{n=k+1}^{k+N} g(X_n)$$

   is an, at least asymptotically, *unbiased* estimator of $\mathbf{E}_{\boldsymbol{\pi}} g$ with *standard error* proportional to $1/\sqrt{N}$. However, when the *mixing* is *poor*, the *constant of proportionality* will be much greater than in the case of i.i.d. sampling.

3. **Mixing kernels.** Frequently, in order to achieve *ergodicity* and *good mixing*, it is necessary to choose at each time step from one of a number of *kernels*, typically either in *rotation* or *randomly* (in the latter case we are really using a single more general kernel). An example occurs in the use of the *Gibbs sampler*—again see below. At each time step the *kernel* P must be such that $\boldsymbol{\pi}P = \boldsymbol{\pi}$.

4. **Auxiliary variables.** Sometimes it is desirable to run the Markov chain $\{X_n\}_{n\geq 0}$ on a state space S which has been enlarged by the introduction of an additional *auxiliary* variable z, say.

   For example, suppose that x represents the *parameter(s) of interest* in a *Bayesian* model. Suppose also that (a) the *prior distribution* of x is given by $\boldsymbol{\pi}_0$; (b) for any given values of the parameter(s) x, *unobserved* random variables z are generated in accordance with a density $f(x, z)$; (c) for given z, the *observed data* y are generated in accordance with a density $h(z, y)$.

Then the *posterior distribution* of x, given the observations y, is given by the probability function or density

$$\pi(x) = K\pi_0(x) \int_z f(x,z)h(z,y)\,dz$$

for some (usually incalculable) *normalising constant* K. The difficulties involved in the integration are typically such that it is simpler to observe that the *posterior joint distribution* of (x,z) is given by

$$\hat{\pi}(x,z) = \hat{K}\pi_0(x)f(x,z)h(z,y)$$

(where again $\hat{K}$ is the appropriate normalising constant). A *Markov chain* $(X_n, Z_n)_{n\geq 0}$ may then be constructed on the space S of all (x,z) with the above *joint distribution* of (x,z) as its *target*. Since what is of interest is the *marginal distribution* $\boldsymbol{\pi}$ of x, any quantity of the form $\mathbf{E}_{\boldsymbol{\pi}} g$, where g depends only on x, may be estimated via (24) as usual, i.e. by ignoring the observed values of Z_n.

3.7 MCMC: Algorithms

3.7.1 The Metropolis-Hastings algorithm

This is a method of constructing a Markov chain *kernel* $P = (p(x,y))_{x,y\in S}$ such that, for the given *target distribution* $\boldsymbol{\pi}$ (with *probability function* or *density* $\pi(\cdot)$) we have $\boldsymbol{\pi}P = \boldsymbol{\pi}$. The constructed *kernel* possesses the *detailed balance property*, indeed it is this that makes its construction possible.

The algorithm is completely defined by the specification, for each $x \in S$, of a **proposal distribution**, given by its *probability function* (or *density*) $q(x,\cdot)$. The *kernel*, i.e. effectively the *chain*, is then defined by the two-step procedure:

1. given the value x of the *chain* at the current time n, say, choose a **candidate** y according to the *proposal probabilities* $q(x,\cdot)$, i.e. choose y with *probability (density)* $q(x,y)$.
2. **accept** the *candidate* y as the next value of the *chain* (at time $n+1$) with the *acceptance probability*

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right); \tag{25}$$

otherwise (with probability $1-\alpha(x,y)$) the next value of the *chain* (at time $n+1$) is again taken to be x. Note that it is *crucial* that if the candidate is *not accepted*, the *chain nevertheless moves forward one time step*—while remaining in the same state.

Note also that if the *candidate* y is the *same* as the current state x, then it is accepted with probability 1, and the chain again moves forward one time step while remaining in the same state—again it is *crucial* that the *chain* does moves forward one time step.

Finally, note that in (25) it is indeed only necessary to know the *target* probability function or density $\pi(\cdot)$ up to a multiplicative constant, since the *acceptance probability* $\alpha(x,y)$ involves only the *ratio* $\pi(y)/\pi(x)$.

The *transition probabilities* of the resulting *Markov chain* are then given by, for $y \neq x$,

$$p(x,y) = \alpha(x,y)q(x,y) = \min\left(q(x,y), \frac{\pi(y)q(y,x)}{\pi(x)}\right),$$

and so, again for $y \neq x$,

$$\pi(x)p(x,y) = \min\left(\pi(x)q(x,y), \pi(y)q(y,x)\right) = \pi(y)p(y,x),$$

so that the *kernel* $P = (p(x,y))_{x,y\in S}$ does indeed possess the *detailed balance property*.

Additionally, in order for the result (24) to hold, we require the *proposal* to be such that the *chain* defined by the *kernel* P is *irreducible*. In some cases the property is obvious, while in others it requires careful checking.

Ideally the *proposal distribution* should be reasonably close to the *target distribution*, but in many problems the chain will instead have to move in fairly small steps.

3.7.2 Special cases of the Metropolis-Hastings algorithm

We mention some special cases, corresponding to particular choices of the *proposal distribution* $q(x,\cdot)$.

The Metropolis algorithm. In this case the *proposal distribution* is such that $q(x,y) = q(y,x)$ for all $x, y \in S$, so that the *acceptance probability* (25) becomes

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right). \tag{26}$$

Frequently we have $q(x,y) = q(|y-x|)$, i.e. the *proposal* is to move a distance from x which is chosen *independently* of the current state x. This is known as **random walk Metropolis**.

The independence sampler. Here, for all $x, y \in S$, we have $q(x,y) = \bar{q}(y)$ for some function $\bar{q}$ on S, i.e. the *proposal probabilities* $q(x,\cdot)$ are *independent* of the current state x. The *acceptance probability* (25) then becomes

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)\bar{q}(x)}{\pi(x)\bar{q}(y)}\right).$$

Example 3.13 Suppose that the state space S is the set of all $n!$ *permutations* $\delta = (\delta_1, \ldots, \delta_n)$ of the integers 1 to n for some n sufficiently large that $n!$ is a big number. Suppose further that the *target distribution* is given by its *probability function* $\pi(\cdot)$ on S, typically known only up to some *multiplicative constant* which cannot be calculated on account of the huge size of S.

Consider the following *proposal* for the simulation of a *Markov chain* on S with stationary distribution given by $\pi(\cdot)$ by the use of the *Metropolis-Hastings algorithm*: whatever the current state δ of the chain (δ is some permutation of $\{1, \ldots, n\}$), two indices i and j are chosen at random from $(1, \ldots, n)$, all pairs being equally likely. The new *candidate state* is then given by swapping δ_i and δ_j. Thus the *proposal* probabilities are given by

$$q(\delta, \delta') = \begin{cases} \binom{n}{2}^{-1} & \text{if } \delta \text{ and } \delta' \text{ differ at exactly 2 places} \\ 0 & \text{otherwise.} \end{cases}$$

Since $q(\delta, \delta') = q(\delta', \delta)$ for all δ, δ' this is an instance of the *Metropolis algorithm*. The acceptance probability $\alpha(\delta, \delta')$ is given by (26) (with x and y replaced by δ and δ'). In the case where $\pi(\delta) > 0$ for all $\delta \in S$ it is clear the the resulting Markov chain is *irreducible* (since every permutation can be obtained from every other through some finite number of swaps). The *mixing* of the chain is relatively slow, so that many steps of the chain are required for accurate estimation of quantities of interest, but this is compensated for by the simplicity of implementation of the algorithm.

An interesting particular case arises when S is composed of a set A of equiprobable *allowed* states and the remaining set $S \setminus A$ of *forbidden* states, so that we may take

$$\pi(\delta) = \begin{cases} k^{-1} & \text{if } \delta \in A \\ 0 & \text{if } \delta \notin A, \end{cases}$$

where $k = |A|$ is typically unknown. For example, the set $\{1, \ldots, n\}$ might correspond to n *individuals* and the permutation $\delta = (\delta_1, \ldots, \delta_n\}$ might correspond to their *ranks* (according to some ordering). Under the target distribution $\boldsymbol{\pi}$ on S, all permutations (rankings) might be equally likely provided only that they satisfied some constraints of the form $\delta_i < \delta_j$ (individual i ranks ahead of individual j) for certain specified ordered pairs (i, j). Thus A would here be the set of permutations satisfying these constraints. For n of even moderate size (e.g. $n = 30$) such constraints might well ensure that the size of the set A of allowed permutations is too large and too difficult to calculate. Here the Markov chain, if started in A, remains within A thereafter, the acceptance probabilities being 1 for proposed swaps such that the new state also belongs to A, and 0 otherwise. It is necessary to ensure that the set A is such that the corresponding chain is *irreducible*, i.e. that every state within A can be reached from every other by a sequence of simple swaps while remaining within the set A (otherwise it will be necessary to use a more sophisticated *proposal*). A typical quantity of interest might be the probability, under the distribution on S given by $\pi(\cdot)$, of the event $\{\delta_i = j\}$ for some $i, j \in \{1, \ldots, n\}$. This would be estimated by the long-run proportion of those states of the chain for which this event occurred.

3.7.3 The Gibbs sampler

Frequently the space S (on which the *target distribution* $\boldsymbol{\pi}$ of interest is defined) is *multidimensional* (this is where MCMC is most useful). This happens in *Bayesian estimation* whenever there is more than one parameter to be estimated, and, for example, in *spatial processes*.

For simplicity, we shall assume that S is *two-dimensional* so that we may write a typical state $\boldsymbol{x} \in S$ as $\boldsymbol{x} = (x_1, x_2)$. In many cases, while the *target probabilities* (or *density*) $\pi(\cdot, \cdot)$ is known only up to a multiplicative constant, the one-dimensional *conditional probabilities* $\pi_{2|1}(x_2 \,|\, x_1)$ of x_2 given x_1 and $\pi_{1|2}(x_1 \,|\, x_2)$ of x_1 given x_2 can easily be calculated *exactly*. The **Gibbs sampler** is then a simulation of a *Markov chain* on S in which the two coordinates of $\boldsymbol{x}$ are updated *alternately* by the use of these two conditional probability (density) functions. Thus, if the current state of the chain is $\boldsymbol{x} = (x_1, x_2)$ and x_1 is to be kept fixed, then x_2 is updated in accordance with the conditional probability (density) function $\pi_{2|1}(\cdot \,|\, x_1)$ (to obtain a new state (x_1, x_2')) while if x_2 is to be kept fixed, then x_1 is updated in accordance with the conditional probability (density) function $\pi_{1|2}(\cdot \,|\, x_2)$ (to obtain a new state (x_1', x_2)). We thus obtain an instance of a Markov chain, which is not quite *time-*

homogeneous, but which has two *kernels* P_1 and P_2, say, which are used *alternately* at successive steps of the chain.

To show that the *target distribution* $\boldsymbol{\pi}$ is a stationary distribution of this chain, it is necessary to verify that $\boldsymbol{\pi}P_i = \boldsymbol{\pi}$ for both $i = 1$ and $i = 2$. For this it will be sufficient to verify the *detailed balance equations*

$$\pi(x_1, x_2)\pi_{2|1}(x_2' \,|\, x_1) = \pi(x_1, x_2')\pi_{2|1}(x_2 \,|\, x_1)$$
$$\pi(x_1, x_2)\pi_{1|2}(x_1' \,|\, x_2) = \pi(x_1', x_2)\pi_{1|2}(x_1 \,|\, x_2)$$

corresponding to each of the alternating steps of the chain. But these equations are immediate simply from the definition of conditional probability. Thus, provided we have *irreducibility* of the chain so that the *ergodic theorem* again holds (the extension to the *alternating chain* follows easily by considering the *homogeneous chain* $\{X_{2n}\}_{n\geq 0}$), expectations with respect to $\boldsymbol{\pi}$ may be estimated as usual by considering long-run frequencies and using (24).

Figure 4 shows the alternate updating procedure in the state space S.

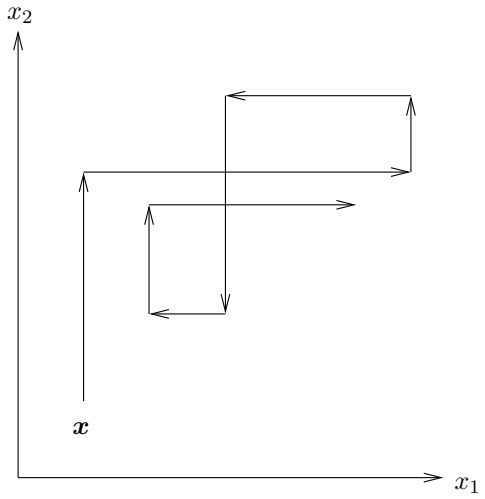

Figure 4: Two-dimensional state space: alternate updating of x_1 and x_2.

An alternative is, at each step, to make a random choice of the coordinate of S to be updated, resulting in a Markov chain which is now again *time-homogeneous*.

We observe that the *Gibbs sampler* may also be viewed as an instance of the *Metropolis-Hastings algorithm*, with the use of the *conditional distributions* as alternating *proposals* and with *acceptance probabilities* which always turn out to be equal to 1. *[Exercise!]*

Finally we remark that the above ideas extend naturally to a space S of any *finite* number of dimensions: at each step of the chain one coordinate of the state is updated, using the conditional distribution of that coordinate given the current values of the remaining coordinates. The coordinates may be chosen *cyclically* or *randomly*.

Example 3.14 *A spatial process.* Consider an $N \times N$ lattice in which each vertex (i, j) (where $1 \leq i \leq N$, $1 \leq j \leq N$) is in either state 0 or state 1. The state of the entire system is therefore $\boldsymbol{x}$ where $\boldsymbol{x} = (x_{ij})_{1\leq i\leq N, 1\leq j\leq N}$ and each x_{ij} is either 0 or 1 according to the state of vertex (i, j). Figure 5 shows a possible state $\boldsymbol{x}$ of the

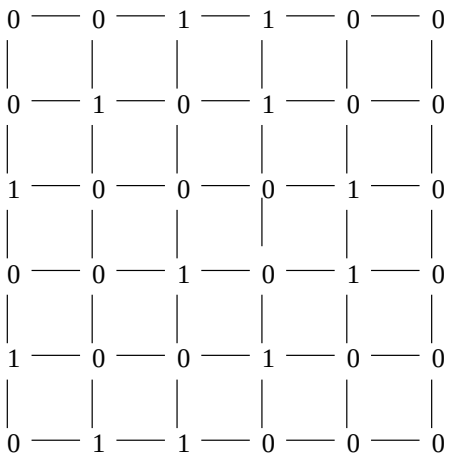

Figure 5: Spatial model: possible state $\boldsymbol{x}$ of 6×6 lattice

system. The probability distribution $\boldsymbol{\pi}$ on the space S of all such $\boldsymbol{x}$ is such that the *conditional probability* that vertex (i, j) is in state 1, *given* the states of all other vertices, is

$$\frac{\alpha\beta^{y_{ij}}}{1 + \alpha\beta^{y_{ij}}} \tag{27}$$

where

$$y_{ij} = x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1}$$

is the total number of 1s at *immediately neighbouring* vertices. Thus, for $\beta < 1$, the state 1 at any vertex is *less* likely to occur the *more* the number of 1s at neighbouring vertices (*repulsion*), while, for $\beta > 1$, the state 1 is *more* likely to occur the *more* the number of 1s at neighbouring vertices (*attraction*). (In the special case $\beta = 0$ (*extreme repulsion*), no vertex may be in state 1 if any of its neighbouring vertices is in state 1. Also the case $\beta = 1$ corresponds to *independent* states at each vertex.) *Repulsion* is natural in, for example, models of plant competition (where the probability of a plant at a given location decreases with the total number of plants at neighbouring locations), while *attraction* is natural in, for example, models of ferromagnetism (where the magnetic polarity at one location is likely to line up with that at neighbouring locations). There are many further examples of this model.

It turns out that the distribution $\boldsymbol{\pi}$ is given *uniquely* by

$$\pi(\boldsymbol{x}) = K\alpha^{n_1(\boldsymbol{x})}\beta^{n_2(\boldsymbol{x})} \tag{28}$$

where $n_1(\boldsymbol{x}) = \sum_i \sum_j x_{ij}$ is the total number of 1s, $n_2(\boldsymbol{x})$ is the total number of *pairs* of immediately neighbouring vertices for which *both* vertices in the pair are in state 1, and K is the appropriate *normalising constant*, which, for even moderately large N is too difficult to calculate. (*Exercise: verify that the distribution π given by* (28) *does indeed lead to the conditional distributions* (27).)

A Markov chain for which the *target* distribution $\boldsymbol{\pi}$ is *stationary* may be simulated by the use of the *Gibbs sampler*, using the *conditional distributions* (27) to update *one* component of $\boldsymbol{x}$ (i.e. the state x_{ij} at one vertex (i, j) of the lattice) at a time. The vertices may be taken in any order, or, at each step, a vertex might be chosen at random.

Typically we might be interested in the *probability* (under the distribution $\boldsymbol{\pi}$) of a

1 at some particular vertex (i, j). We would thus require $\mathbf{E}_{\boldsymbol{\pi}} g_{ij}$ where

$$g_{ij}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } x_{ij} = 1 \\ 0 & \text{if } x_{ij} = 0. \end{cases}$$

Under the *simulation* the required *probability* would thus be estimated by the *long-run proportion* of 1s at the vertex (i, j).

Note that when β is far from 1 and when N is even moderately large (e.g. $N = 10$), the length of simulation required for accurate estimation may be impossibly long. In this case matters may be improved by, at each step of the chain, simultaneously updating the state at several vertices, using the joint conditional distribution at these vertices, given the state at the remaining vertices.

3.8 MCMC: Assessment of uncertainty

Since the successive observations of the Markov chain are not independent identically distributed, assessments of uncertainty, e.g. standard errors, are more difficult to determine. We mention two possibilities.

1. **Blocking.** This idea applies to many instances of estimation where we have a long sequence of dependent observations, in which the degree of dependence decreases with increasing separation of the observations. Let

$$\bar{g}_N = \frac{1}{N} \sum_{n=k+1}^{k+N} g(X_n) \tag{29}$$

be the estimate of $\mathbf{E}_{\boldsymbol{\pi}} g$ based on observation of the segment of the chain from time $k+1$ to time $k+N$ (where k corresponds to the burn-in time). For good estimation we require N to be large. Suppose that it is sufficiently large that we may divide this segment of the chain into b nonoverlapping **blocks**, each of length M where both b and M are reasonably large. Let $\bar{g}_{M,i}$ denote the usual estimate of $\mathbf{E}_{\boldsymbol{\pi}} g$ based on the observations in the ith block, $1 \leq i \leq b$. Then, since M is reasonably large, the b estimates $\bar{g}_{M,i}$ may reasonably be treated as *independent identically distributed* observations, each with mean $\mathbf{E}_{\boldsymbol{\pi}} g$ (each is unbiased). Further the mean of these estimates is just $\bar{g}_N$. Hence, by the usual theory of estimation based on the use of the sample mean, since also b is reasonably large, $\bar{g}_N$ is an unbiased estimator of $\mathbf{E}_{\boldsymbol{\pi}} g$ with variance

$$\operatorname{var} \bar{g}_N \approx \frac{1}{b(b-1)} \sum_{i=1}^{b} (\bar{g}_{M,i} - \bar{g}_N)^2$$

and so the standard error of $\bar{g}_N$ (as an estimator of $\mathbf{E}_{\boldsymbol{\pi}} g$) is just $(\operatorname{var} \bar{g}_N)^{1/2}$.

2. **Regeneration.** Suppose for example that the state space S is discrete, so that, in the constructed Markov chain, which is ergodic, return to any *given fixed state* occurs infinitely often. Since the chain is Markov, its behaviour between any two successive returns is independent of that between all other pairs of successive returns, both with regard to the return times and with regard to the states visited. If the segment of the chain from time $k+1$ to time $k+N$ on which estimation of $\mathbf{E}_{\boldsymbol{\pi}} g$ is based starts and finishes at the given state, both N and $\sum_{n=k+1}^{k+N} g(X_n)$ may be expressed as sums of independent identically distributed random variables. This gives another approach to calculation of the standard error of the estimator $\bar{g}_N$ defined by (29).