MAS451: Principles of Statistics

Part 2: Bayesian inference, stochastic simulation and MCMC

Module description:

The central concern of modern statistical inference is estimation. The techniques used depend crucially on the inferential paradigm to which one is philosophically committed. Classical inference has been predominantly concerned with the study of estimators satisfying constraints which lead to tractable analytic solutions. In addition, inference was usually based on asymptotic properties of these estimators. The advent of modern computer technology has led to the development of new techniques which allow the relaxation of constraints and further, facilitate an understanding of their exact properties.

This module begins with a re-examination of classical estimation procedures before moving on to more modern computationally intensive techniques. For example, students will be introduced to Gibbs sampling, a simulation approach to Bayesian statistical inference which is at the cutting edge of current research. These modern techniques open up new opportunities for practical data analysis and this is reflected in the practical component of this module.

Course text:

D. Gamerman: *Markov Chain Monte Carlo* (Chapman and Hall, 1997).

WWW page:

`http://www.staff.ncl.ac.uk/d.j.wilkinson/teaching/mas451/`

Contents

Chapter 1

Bayesian inference

1.1 Introduction

Part 1 of the module started with the analysis of frequentist techniques for estimation. Many people now question the repeated sampling framework within which such techniques are grounded, and in any case, as soon as you move to more complex models, the frequentist machinery becomes unmanageable.

Next the theory of maximum likelihood was examined. This framework copes well with more complex models, particularly with the aid of a computer. However, the foundations of likelihood theory are not particularly compelling (to some), they provide no mechanism for incorporating prior information into the model (particularly problematic for models which are unidentifiable or weakly identifiable), and do not give full probabilistic information about parameters.

All of the problems associated with likelihood theory are addressed by the Bayesian approach to statistical inference. Until recently, however, Bayesian computation for complex models was prohibitively complex. Fortunately, if one is prepared to abandon an analytic approach to computation, and use stochastic simulation techniques, then Bayesian inference may now be used to analyse almost any statistical problem, of any complexity, and the analysis is more informative than any that could be carried out using frequentist or likelihood based techniques. That is not to say that Bayesian computation based on stochastic simulation is easy, or without many problems and difficulties — this is most certainly not the case. However, for many researchers, the power of modern Bayesian machinery is very appealing, and provides their favoured framework for statistical inference.

In this part of the module we will look briefly at why analytic approaches to Bayesian inference in complex models are generally intractable, before moving on to stochastic simulation and its application to Bayesian inference. Stochastic simulation is a very powerful technique with many applications outside Bayesian inference. It would therefore be inappropriate to introduce and develop all of the techniques of stochastic simulation in that context. Therefore, all of the simulation techniques will be developed in a general context, and only then applied to problems in Bayesian inference.

1.2 Bayesian inference

For two events E and F, the *conditional probability* of E given F is

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(E \cap F)}{\mathrm{P}(F)}.$$

From this we get the simplest version of Bayes theorem:

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(F|E)\ \mathrm{P}(E)}{\mathrm{P}(F)}.$$

The Theorem of Total Probability states that for an event F, and a partition $E_1, E_2, \ldots, E_n$ (one and only one will occur), we have

$$\begin{aligned}\mathrm{P}(F) &= \mathrm{P}(F|E_1)\ \mathrm{P}(E_1) + \mathrm{P}(F|E_2)\ \mathrm{P}(E_2) + \cdots + \mathrm{P}(F|E_n)\ \mathrm{P}(E_n) \\ &= \sum_{i=1}^{n} \mathrm{P}(F|E_i)\ \mathrm{P}(E_i),\end{aligned}$$

from which we get the more commonly used version of Bayes theorem:

$$\mathrm{P}(E_i|F) = \frac{\mathrm{P}(F|E_i)\ \mathrm{P}(E_i)}{\sum_{i=1}^{n} \mathrm{P}(F|E_i)\ \mathrm{P}(E_i)}.$$

Bayes theorem is useful because it tells us how to turn probabilities around. Often we are able to understand the probability of some *outcome* (F), conditional on various possible *hypotheses* (E_i). We can then compute probabilities of the form $\mathrm{P}(F|E_i)$. However, when we actually *observe* some outcome, we are interested in the probabilities of the hypotheses *conditional* on the outcome, $\mathrm{P}(E_i|F)$. Bayes theorem tells us how to compute these, but the answer also depends on the prior probabilities for the hypotheses, $\mathrm{P}(E_i)$, and hence to use Bayes theorem, these too must be specified. Thus Bayes theorem provides us with a coherent way of updating our prior beliefs about the hypotheses $\mathrm{P}(E_i)$ to $\mathrm{P}(E_i|F)$, our posterior beliefs based on the occurrence of F.

This is how it all works for purely discrete problems, but some adaptation is required before it can be used with continuous or mixed problems. For continuous problems, the hypotheses are represented by a continuous parameter, usually denoted by Θ, and the outcomes, or data by X. These may be scalars or vectors. We specify our prior beliefs about Θ in the form of a probability distribution with density $\pi(\theta)$. The data conditional on the hypotheses is modelled by the conditional density, $f(x|\theta)$. If this is regarded as a function of θ rather than x, it is known as the *likelihood*, and denoted $L(\theta;x)$. It is important to bear in mind that this is *not* a density for θ — it doesn't even integrate to one!

Once the prior and likelihood have been specified, the full joint density over all parameters and data is entirely determined:

$$f_{\Theta,X}(\theta,x) = \pi(\theta) f(x|\theta) = \pi(\theta) L(\theta;x).$$

Once we have this joint distribution, we can marginalise or condition as required. For example, we can obtain the marginal distribution for X by integrating over Θ:

$$f_X(x) = \int_\Theta f_{\Theta,X}(\theta,x)\, d\theta = \int_\Theta L(\theta;x)\pi(\theta)\, d\theta.$$

This is the equivalent of the Theorem of Total Probability for continuous variables. We can also condition on the data:

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{f_{\Theta,X}(\theta,x)}{f_X(x)} \\ &= \frac{\pi(\theta)L(\theta;x)}{\int_\Theta L(\theta;x)\pi(\theta)\,d\theta}. \end{aligned}$$

$f_{\Theta|X}(\theta|x)$ is the conditional density for Θ given X. It is known as the *posterior* density, and is usually denoted $\pi(\theta|x)$, leading to the continuous version of Bayes theorem:

$$\boxed{\pi(\theta|x) = \frac{\pi(\theta)L(\theta;x)}{\int_\Theta L(\theta;x)\pi(\theta)\,d\theta}}$$

Now, the denominator is not a function of θ, so we can in fact write this as

$$\boxed{\pi(\theta|x) \propto \pi(\theta)L(\theta;x)}$$

where the constant of proportionality is chosen to ensure that the density integrates to one. So, *the posterior is proportional to the prior times the likelihood*.

1.3 Bayesian computation

In principle, the previous section covers everything we need to know about Bayesian inference — the posterior is nothing more (or less) than a conditional distribution for the parameters given the data. In practice however, this may not be entirely trivial to work with.

The first problem one encounters is choosing the constant of proportionality so that the density integrates to one. If the density is non-standard (as is usually the case for non-trivial problems), then the problem reduces to integrating the product of the likelihood and the prior (known as the *kernel* of the posterior) over the support of Θ. If the support is infinite in extent, and/or multi-dimensional, then this is a highly non-trivial numerical problem.

Even if we have the constant of integration, if the parameter space is multi-dimensional, we will want know what the marginal distribution of each component looks like. For each component, we have a very difficult numerical integration problem.

1.3.1 Normal with unknown mean and variance

Consider the case where we have a collection of observations, X_i, which we believe to be iid Normal with unknown mean and precision (the reciprocal of variance). We write

$$X_i|\mu,\tau \sim N(\mu, 1/\tau).$$

The likelihood for a single observation is

$$L(\mu,\tau;x_i) = f(x_i|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}}\exp\left\{-\frac{\tau}{2}(x_i-\mu)^2\right\}$$

and so for n independent observations, $x = (x_1, \ldots, x_n)'$ is

$$\begin{aligned} L(\mu, \tau; x) = f(x|\mu, \tau) &= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}(x_i - \mu)^2\right\} \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}\left[(n-1)s^2 + n(\bar{x} - \mu)^2\right]\right\} \\ &\propto \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}\left[(n-1)s^2 + n(\bar{x} - \mu)^2\right]\right\} \end{aligned}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

For a Bayesian analysis, we need also to specify prior distributions for the parameters, μ and τ. There is a conjugate analysis for this problem based on the specifications:

$$\begin{aligned} \tau &\sim Gamma(a, b) \\ \mu|\tau &\sim N\left(c, \frac{1}{d\tau}\right). \end{aligned}$$

However, this specification is rather unsatisfactory — μ and τ are not independent, and in many cases our prior beliefs for μ and τ will separate into independent specifications. For example, we may prefer to specify independent priors for the parameters:

$$\begin{aligned} \tau &\sim Gamma(a, b) \\ \mu &\sim N\left(c, \frac{1}{d}\right). \end{aligned}$$

However, this specification is no longer conjugate, making analytic analysis intractable. Let us see why:

$$\begin{aligned} \pi(\mu) &= \sqrt{\frac{d}{2\pi}} \exp\left\{-\frac{d}{2}(\mu - c)^2\right\} \\ &\propto \exp\left\{-\frac{d}{2}(\mu - c)^2\right\} \end{aligned}$$

and

$$\begin{aligned} \pi(\tau) &= \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\} \\ &\propto \tau^{a-1} \exp\{-b\tau\} \end{aligned}$$

so

$$\pi(\mu, \tau) \propto \tau^{a-1} \exp\left\{-\frac{d}{2}(\mu - c)^2 - b\tau\right\}$$

giving

$$\pi(\mu,\tau|x) \propto \tau^{a-1}\exp\left\{-\frac{d}{2}(\mu-c)^2-b\tau\right\}\times\tau^{\frac{n}{2}}\exp\left\{-\frac{\tau}{2}\left[(n-1)s^2+n(\bar{x}-\mu)^2\right]\right\}$$
$$=\tau^{a+\frac{n}{2}-1}\exp\left\{-\frac{\tau}{2}\left[(n-1)s^2+n(\bar{x}-\mu)^2\right]-\frac{d}{2}(\mu-c)^2-b\tau\right\}.$$

The posterior density for μ and τ certainly won't factorise (μ and τ are not independent *a posteriori*), and will not even separate into the form of the conditional Normal-Gamma conjugate form mentioned earlier.

So, we have the kernel of the posterior for μ and τ, but it is not in a standard form. We can gain some idea of the likely values of (μ,τ) by plotting the bivariate surface (the integration constant isn't necessary for that), but we cannot work out the posterior mean or variance, or the forms of the marginal posterior distributions for μ or τ, since we cannot integrate out the other variable. We need a way of understanding posterior densities which does not rely on being able to analytically integrate the posterior density.

In fact, there is nothing particularly special about the fact that the density represents a Bayesian posterior. Given any complex non-standard probability distribution, we need ways to understand it, to calculate its moments, to compute its conditional and marginal distributions and their moments. Stochastic simulation is one possible solution.

Chapter 2

Stochastic Simulation

2.1 Introduction

The rationale for stochastic simulation can be summarised very easily: to understand a statistical model simulate many realisations from it and study them.

2.1.1 Normal with unknown mean and variance

If we can simulate lots of realisations from the posterior distribution for μ and τ, we can look at histograms of the μ values, to get an idea of the marginal for μ. We can also look at the sample mean and variance of the μ values to find out the posterior mean and variance of the marginal for μ. In this respect, the simulation is a way of doing the integration by a chance, or *Monte Carlo* method.

2.1.2 Monte Carlo integration

Suppose we have a random variable X, with PDF, $f(x)$, and we wish to evaluate $\mathrm{E}(g(X))$ for some function $g(\cdot)$. We know that

$$\mathrm{E}(g(X)) = \int_X g(x) f(x)\, dx,$$

and so the problem is one of integration. However, if we can *simulate* realisations $x_1, \ldots, x_n$ of X, then we may approximate the integral by

$$\boxed{\mathrm{E}(g(X)) \simeq \frac{1}{n} \sum_{i=1}^{n} g(x_i).}$$

In fact, even if we can't simulate realisations of X, but can simulate realisations $y_1, \ldots, y_n$ of Y (a random variable with the same support as X), which has PDF $h(\cdot)$, then

$$\begin{aligned} \mathrm{E}(g(X)) &= \int_X g(x) f(x)\, dx \\ &= \int_X \frac{g(x) f(x)}{h(x)} h(x)\, dx \end{aligned}$$

and so $\mathrm{E}(g(X))$ may be approximated by

$$\mathrm{E}(g(X)) \simeq \frac{1}{n}\sum_{i=1}^{n}\frac{g(y_i)f(y_i)}{h(y_i)}.$$

This procedure is known as *importance sampling*, and can be very useful when there is reasonable agreement between $f(\cdot)$ and $h(\cdot)$. However, before we can do anything, we need a way of simulating random quantities from some standard distributions.

2.2 Uniform random numbers

Most stochastic simulation begins with a uniform random number generator. Most algorithms require the generation of independent observations uniformly over the interval $[0,1)$. Once we have $U \sim U[0,1)$, we can use it in order to simulate random quantities from any distribution we like. So, how do we do it?

Typically, a number theoretic method is used to generate a random integer from 0 to $2^N - 1$ (often, $N = 16$, 32 or 64). This can be divided by 2^N to give a number uniform on $[0,1)$. *Linear congruential generators* are often used for this purpose.

The algorithm begins with a *seed* x_0 then generates new values according to the (deterministic) rule

$$\boxed{x_{n+1} = (a x_n + b \mod 2^N)}$$

for carefully chosen a and b. If a "good" choice of a, b and N are used, this deterministic sequence of numbers will have a cycle length of 2^N and give every appearance of being random.

The NAG Fortran library uses

$$N = 59,\ b = 0,\ a = 13^{13}.$$

Most computer programming languages have a built-in function for returning a pseudo-random integer, or a pseudo-random $U[0,1)$ number. We will not worry too much about this, but rather take it as our starting point, and look at how these can be used to simulate from more exotic distributions.

2.3 Transformation methods

Suppose that we wish to simulate realisations of a random variable X, with PDF $f(x)$. If we also know the probability distribution function $F(x)$, and its inverse $F^{-1}(\cdot)$, we can simulate a realisation of X using a single $U \sim U[0,1)$ as follows. Put

$$\boxed{\tilde{X} = F^{-1}(U).}$$

Then, $\tilde{X}$ has distribution $F(\cdot)$, and hence has the same distribution as X. This follows as

$$\begin{aligned}
F_{\tilde{X}}(x) &= \mathrm{P}\left(\tilde{X} \le x\right) \\
&= \mathrm{P}\left(F^{-1}(U) \le x\right) \\
&= \mathrm{P}(U \le F(x)) \\
&= F_U(F(x)) \\
&= F(x). && (\text{as } F_U(u) = u)
\end{aligned}$$

2.3.1 Uniform random variates

Given $U \sim U[0,1)$, we can simulate $V \sim U[a,b)$ in the obvious way, that is

$$V = a + (b-a)U.$$

We can justify this as V has CDF

$$F(v) = \frac{v-a}{b-a}, \qquad a \leq v \leq b$$

and hence inverse

$$F^{-1}(u) = a + (b-a)u.$$

2.3.2 Exponential random variates

Consider $X \sim Exp(\lambda)$. This has density $f(x)$ and distribution $F(x)$, where

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0$$
$$F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0$$

and so

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

So, to simulate a realisation of X, simulate u from $U[0,1)$, and then put

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

Then x is a simulated value of X. Also, note that if $U \sim U[0,1)$, then $1-U \sim U(0,1]$, and so we can just put

$$x = -\frac{1}{\lambda}\log u$$

to obtain our exponential variates.

2.3.3 Scaling

It is worth noting the scaling issues in the above example, as these become more important for distributions which are more difficult to simulate from. If $U \sim U[0,1)$, then $Y = -\log U \sim Exp(1)$. The parameter, λ, of an exponential distribution is a *scale parameter* because we can obtain exponential variates with other parameters from a variate with a unit parameter by a simple linear scaling. That is, if

$$Y \sim Exp(1)$$

then

$$X = \frac{1}{\lambda}Y \sim Exp(\lambda).$$

In general, we can spot *location* and *scale* parameters in a distribution as follows. If Y has PDF $f(y)$ and CDF $F(y)$, and $X = aY + b$, then X has CDF

$$\begin{aligned} F_X(x) &= \mathrm{P}(X \leq x) \\ &= \mathrm{P}(aY + b \leq x) \\ &= \mathrm{P}\left(Y \leq \frac{x-b}{a}\right) \\ &= F\left(\frac{x-b}{a}\right) \end{aligned}$$

and PDF

$$f_X(x) = \frac{1}{a} f\left(\frac{x-b}{a}\right).$$

Parameters like a are scale parameters, and parameters like b are location parameters.

2.3.4 Gamma random variates

One way to simulate $X \sim Gamma(n, \lambda)$ random variates for integer n is to use the fact that if

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

and the Y_i are independent, then

$$X = \sum_{i=1}^{n} Y_i \sim Gamma(n, \lambda).$$

So, just simulate n exponential random variates and add them up. In particular, note that $Exp(\lambda) = Gamma(1, \lambda)$, and the independent sum $Gamma(n_1, \lambda) + Gamma(n_2, \lambda) = Gamma(n_1 + n_2, \lambda)$.

The first parameter of the gamma distribution (here n), is known as the *shape* parameter, and the second (here λ) is known as the *scale* parameter. The fact that the second parameter is a scale parameter is important, because many gamma generation algorithms will only generate gamma variables with a particular shape but unit scale. For example, in R, SPlus, and LISP-STAT, the gamma PDF, CDF and random variate generation functions only allow specification of a shape parameter — you must re-scale these appropriately yourself. We can do this easily because the $Gamma(\alpha, \beta)$ PDF is

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$

and so the $Gamma(\alpha, 1)$ PDF is

$$f_Y(y) = \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}, \quad y > 0.$$

We can see that if $Y \sim Gamma(\alpha, 1)$, then $X = Y/\beta \sim Gamma(\alpha, \beta)$ because

$$f_X(x) = \beta f_Y(\beta x).$$

Consequently, the CDFs must be related by

$$F_X(x) = F_Y(\beta x).$$

Techniques for efficiently generating gamma variates with arbitrary shape parameter is usually based on rejection techniques (to be covered later). Note however, that for shape parameters which are an integer multiple of 0.5, use can be made of the fact that $\chi^2_n = Gamma(n/2, 1/2)$. So, if you have a technique for generating χ^2 quantities, these can be used for generating gamma variates whose shape parameter is an integer multiple of 1/2.

2.3.5 Normal random variates

Note that all we need is a technique for simulating $Z \sim N(0,1)$ random variables. Then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. Also note that standard normal random variables can be used to generate χ^2 random variables. If $Z_i \sim N(0,1)$, and the Z_i are independent, then

$$C = \sum_{i=1}^{n} Z_i^2$$

has a χ^2_n distribution.

CLT based method

One simple way to generate Normal random variables is to make use of the Central Limit Theorem. Consider

$$Z = \sum_{i=1}^{12} U_i - 6$$

where $U_i \sim U[0,1)$. Clearly $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$, and by the central limit theorem, Z is approximately Normal. However, this method isn't exact. For example, Z only has support on $[-6,6]$, and is poorly behaved in the extreme tails. However, $\mathrm{P}(|Z| > 6) \simeq 2 \times 10^{-9}$, and so this method is good enough for many purposes.

Box-Muller method

A more efficient (and "exact") method for generating Normal random variates is the following. Simulate

$$\Theta \sim U[0, 2\pi)$$
$$R^2 \sim Exp(1/2)$$

independently. Then

$$X = R\cos\Theta$$
$$Y = R\sin\Theta$$

are two independent standard Normal random variables. It is easier to show this the other way around. That is, if X and Y are independent standard Normal random quantities, and they are

regarded as the Cartesian coordinates of a 2d random variable, then the polar coordinates of the variable are square-rooted exponential and uniform.

Suppose that $X, Y \sim N(0,1)$ and are independent. Then

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp\{-(x^2+y^2)/2\}.$$

Put

$$X = R\cos\Theta \quad \text{and} \quad Y = R\sin\Theta.$$

Then,

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= f_{X,Y}(x,y)\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| \\ &= \frac{1}{2\pi}e^{-r^2/2}\begin{vmatrix}\cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta\end{vmatrix} \\ &= \frac{1}{2\pi}\times re^{-r^2/2}. \end{aligned}$$

So, Θ and R are independent, $\Theta \sim U[0,2\pi)$ and $f_R(r) = re^{-r^2/2}$. It is then easy to show that $R^2 \sim Exp(1/2)$.

2.3.6 Mixtures

Suppose we can simulate from $f_X(x)$ and $f_{Y|X}(y|x)$, but that we want to simulate from the marginal for Y, $f_Y(y)$. First simulate x from $f_X(x)$, and then simulate $\tilde{Y}$ from $f_{Y|X}(y|x)$. Then $\tilde{Y}$ has the same distribution as Y:

$$\begin{aligned} F_{\tilde{Y}}(y) &= \mathrm{P}\left(\tilde{Y} \le y\right) \\ &= \int_X \mathrm{P}\left(\tilde{Y} \le y | X = x\right) f_X(x)\,dx \\ &= \int_X \int_{-\infty}^{y} f_{Y|X}(z|x)\,dz\, f_X(x)\,dx \\ &= \int_X \int_{-\infty}^{y} f_{X,Y}(x,z)\,dz\,dx \\ &= \int_{-\infty}^{y} dz \int_X f_{X,Y}(x,z)\,dx \\ &= \int_{-\infty}^{y} f_Y(z)\,dz \\ &= F_Y(y). \end{aligned}$$

Example

Consider again the conjugate prior for the Normal model with unknown mean and variance. We specify $\pi(\tau)$ and $\pi(\mu|\tau)$, thus determining

$$\pi(\tau,\mu) = \pi(\tau)\pi(\mu|\tau).$$

One way we can simulate from the marginal for μ, $\pi(\mu)$, is to first simulate τ from $\pi(\tau)$, and then use this to simulate μ from $\pi(\mu|\tau)$. This will be a realisation from the marginal $\pi(\mu)$. We could then calculate the mean and variance of a sample, look at a histogram of the values *etc.* in order to get an idea of the shape of the marginal.

2.4 Rejection sampling

2.4.1 Uniform rejection method

Suppose we want to simulate from $f(x)$ with support on $[a,b]$, and that $f(x) \leq m$, $\forall x \in [a,b]$. Then simulate

$$X \sim U[a,b) \quad \text{and} \quad Y \sim U[0,m).$$

Simulate x and y from these distributions, and accept x as a simulated value from $f(x)$ if $y < f(x)$, otherwise *reject* and try again.

Why does this work? Intuitively, we can see that it will work because it has the effect of scattering points uniformly over the region bounded by the PDF and the x-axis. More formally, call the acceptance region A, and the accepted value $\tilde{X}$.

$$\begin{aligned}
F_{\tilde{X}}(x) &= \mathrm{P}\left(\tilde{X} \leq x\right) \\
&= \mathrm{P}(X \leq x | (X,Y) \in A) \\
&= \frac{\mathrm{P}((X \leq x) \cap ((X,Y) \in A))}{\mathrm{P}((X,Y) \in A)} \\
&= \frac{\displaystyle\int_a^b \mathrm{P}((X \leq x) \cap ((X,Y) \in A) | X = z) \times \frac{1}{b-a}\, dz}{\displaystyle\int_a^b \mathrm{P}((X,Y) \in A | X = z) \times \frac{1}{b-a}\, dz} \\
&= \frac{\dfrac{1}{b-a}\displaystyle\int_a^x \mathrm{P}((X,Y) \in A | X = z)\, dz}{\dfrac{1}{b-a}\displaystyle\int_a^b \mathrm{P}((X,Y) \in A | X = z)\, dz} \\
&= \frac{\displaystyle\int_a^x \frac{f(z)}{m} dz}{\displaystyle\int_a^b \frac{f(z)}{m} dz} \\
&= \int_a^x f(z)\, dz \\
&= F(x).
\end{aligned}$$

So, in summary, we simulate a value x uniformly from the support of X, and accept this value with probability $f(x)/m$, otherwise we reject and try again. Obviously the efficiency of this method depends on the overall proportion of candidate points that are accepted. The actual acceptance

probability for this method is

$$\begin{aligned}
\mathrm{P}(\text{Accept}) &= \mathrm{P}((X,Y) \in A) \\
&= \int_a^b \mathrm{P}((X,Y) \in A | X = x) \times \frac{1}{b-a} dx \\
&= \int_a^b \frac{f(x)}{m} \times \frac{1}{b-a} dx \\
&= \frac{1}{m(b-a)} \int_a^b f(x) dx \\
&= \frac{1}{m(b-a)}.
\end{aligned}$$

If this acceptance probability is very low, the procedure will be very inefficient, and a better procedure should be sought — the *envelope method* is one possibility.

2.4.2 Envelope method

Once we have established that scattering points uniformly over the region bounded by the density and the x-axis generates x-values with the required distribution, we can extend it to distributions with infinite support, and make it more efficient, by choosing our *enveloping* region more carefully.

Suppose that we wish to simulate X with PDF $f(\cdot)$, but that we can already simulate values of Y (with the same support as X), which has PDF $g(\cdot)$. Suppose further that there exists some constant a such that

$$f(x) \leq a g(x), \quad \forall x.$$

That is, a is an upper bound for $f(x)/g(x)$.

Consider the following algorithm. Draw $Y = y$ from $g(\cdot)$, and then $U = u \sim U[0, a g(y)]$. Accept y as a simulated value of X if $u < f(y)$, otherwise reject and try again. This works because it distributes points uniformly over a region covering $f(x)$, and then only keeps points in the required

region (under $f(x)$):

$$\begin{aligned}
\mathrm{P}\left(\tilde{X} \leq x\right) &= \mathrm{P}(Y \leq x | U \leq f(Y)) \\
&= \frac{\mathrm{P}([Y \leq x] \cap [U \leq f(Y)])}{\mathrm{P}(U \leq f(Y))} \\
&= \frac{\int_{-\infty}^{\infty} \mathrm{P}([Y \leq x] \cap [U \leq f(Y)] | Y = y)\, g(y) dy}{\int_{-\infty}^{\infty} \mathrm{P}(U \leq f(Y) | Y = y)\, g(y) dy} \\
&= \frac{\int_{-\infty}^{x} \mathrm{P}(U \leq f(Y) | Y = y)\, g(y) dy}{\int_{-\infty}^{\infty} \mathrm{P}(U \leq f(Y) | Y = y)\, g(y) dy} \\
&= \frac{\int_{-\infty}^{x} \frac{f(y)}{a g(y)} g(y) dy}{\int_{-\infty}^{\infty} \frac{f(y)}{a g(y)} g(y) dy} \\
&= \frac{\int_{-\infty}^{x} \frac{f(y)}{a} dy}{\int_{-\infty}^{\infty} \frac{f(y)}{a} dy} \\
&= \int_{-\infty}^{x} f(y)\, dy \\
&= F(x).
\end{aligned}$$

To summarise, simulate $Y = y$ from $g(\cdot)$, and accept this as X with probability $f(y)/[a g(y)]$, otherwise reject and try again.

Obviously, this method will work well if the overall acceptance rate is high, but not otherwise. What is the overall acceptance probability? We have

$$\begin{aligned}
\mathrm{P}(U < f(Y)) &= \int_{-\infty}^{\infty} \mathrm{P}(U < f(Y) | Y = y)\, g(y) dy \\
&= \int_{-\infty}^{\infty} \frac{f(y)}{a g(y)} g(y) dy \\
&= \int_{-\infty}^{\infty} \frac{f(y)}{a} dy \\
&= \frac{1}{a}.
\end{aligned}$$

Consequently, we want a to be as small as possible. Generally speaking, if $a > 50$, the envelope is not adequate — too many points will be rejected, so a better envelope needs to be found. If this isn't practical, then an entirely new approach is required — MCMC is a possibility (more on this later).

2.4.3 Bayes theorem by the rejection method

Often in Bayesian inference we will understand the prior well, and be able to simulate from it efficiently. However, we want to simulate from the posterior, and this is often more difficult.

One possibility for simulating from the posterior is to use the envelope rejection method with the prior as envelope. Put

$$\pi^*(\theta) = \pi(\theta)L(\theta;x).$$

So, $\pi^*(\theta) = k\pi(\theta|x)$ for some k, and so $\pi^*(\theta)$ is the *kernel* of the posterior for θ given x. Clearly

$$\pi^*(\theta) \leq \pi(\theta)L_{max}, \qquad \forall\theta$$

where L_{max} is the maximum value of $L(\theta;x)$ over θ for given x. So we can use the envelope method. Simulate θ from $\pi(\theta)$, and accept it with probability

$$\frac{\pi^*(\theta)}{L_{max}\pi(\theta)} = \frac{\pi(\theta)L(\theta;x)}{L_{max}\pi(\theta)} = \frac{L(\theta;x)}{L_{max}}.$$

otherwise reject and try again. This is intuitively reasonable: the posterior only has support where the prior has support, and you keep more of your simulated prior values where the likelihood is high.

This technique is very elegant and appealing, but sadly only works for low dimensional problems where there is reasonable agreement between the prior and the likelihood, otherwise the rejection rate is too high. In fact

$$\begin{aligned}
\mathrm{P}(\text{Accept}) &= \int_\Theta \frac{L(\theta;x)}{L_{max}}\pi(\theta)d\theta \\
&= \int_\Theta \frac{\pi^*(\theta)}{L_{max}}d\theta \\
&= \int_\Theta \frac{k\pi(\theta|x)}{L_{max}}d\theta \\
&= \frac{k}{L_{max}}.
\end{aligned}$$

We are usually using simulation methods because the constant of integration is unknown, so we may not be able to calculate the acceptance probability exactly.

Chapter 3

Markov Chains

The set $\{\theta^{(t)}|t=0,1,2,\dots\}$ is a *discrete time stochastic process*. The *state space* S is such that $\theta^{(t)} \in S,\ \forall t$ and may be discrete or continuous.

A (first order) *Markov chain* is a stochastic process with the property that the future states are independent of the past states given the present state. Formally, for $A \subseteq S$,

$$\mathrm{P}\left(\theta^{(n+1)} \in A|\theta^{(n)}=x,\theta^{(n-1)}=x_{n-1},\dots,\theta^{(0)}=x_0\right) = \mathrm{P}\left(\theta^{(n+1)} \in A|\theta^{(n)}=x\right), \qquad \forall x,x_{n-1},\dots,x_0 \in S.$$

The past states provide no information about the future state if the present state is known. The behaviour of the chain is therefore determined by $\mathrm{P}\left(\theta^{(n+1)} \in A|\theta^{(n)}=x\right)$. In general this depends on n, A and x. However, if there is no n dependence, so that

$$\mathrm{P}\left(\theta^{(n+1)} \in A|\theta^{(n)}=x\right) = \mathrm{P}(x,A), \quad \forall n,$$

then the Markov chain is said to be *homogeneous*, and the *transition kernel*, $\mathrm{P}(x,A)$ determines the behaviour of the chain. Note that $\forall x \in S,\ \mathrm{P}(x,\cdot)$ is a probability measure over S.

3.1 Discrete chains

3.1.1 Notation

When dealing with discrete state spaces, it is easier to write

$$\mathrm{P}(x,\{y\}) = \mathrm{P}(x,y) = \mathrm{P}\left(\theta^{(n+1)}=y|\theta^{(n)}=x\right).$$

In the case of a finite discrete state space, $S=\{x_1,\dots,x_r\}$, we can write $P(\cdot,\cdot)$ as a matrix

$$P = \begin{pmatrix} \mathrm{P}(x_1,x_1) & \cdots & \mathrm{P}(x_1,x_r) \\ \vdots & \ddots & \vdots \\ \mathrm{P}(x_r,x_1) & \cdots & \mathrm{P}(x_r,x_r) \end{pmatrix}.$$

Note that the elements are all non-negative and that the rows must sum to one. Such matrices are known as *stochastic matrices*. The product of 2 stochastic matrices is another stochastic matrix, and there is always at least one row eigenvalue equal to one (see the tutorial exercises!).

Suppose that at time n, we have

$$\begin{aligned}
\mathrm{P}\left(\theta^{(n)} = x_1\right) &= \pi^{(n)}(x_1) \\
\mathrm{P}\left(\theta^{(n)} = x_2\right) &= \pi^{(n)}(x_2) \\
\vdots \qquad &\qquad \vdots \\
\mathrm{P}\left(\theta^{(n)} = x_r\right) &= \pi^{(n)}(x_r).
\end{aligned}$$

We can write this as an r-dimensional row vector

$$\pi^{(n)} = (\pi^{(n)}(x_1), \pi^{(n)}(x_2), \ldots, \pi^{(n)}(x_r)).$$

What is the probability distribution at time $n+1$? Using the Theorem of Total Probability, we have

$$\mathrm{P}\left(\theta^{(n+1)} = x_1\right) = P(x_1,x_1)\pi^{(n)}(x_1) + P(x_2,x_1)\pi^{(n)}(x_2) + \cdots + P(x_r,x_1)\pi^{(n)}(x_r),$$

and similarly for $\mathrm{P}\left(\theta^{(n+1)} = x_2\right)$, $\mathrm{P}\left(\theta^{(n+1)} = x_3\right)$, *etc.* We can write this in matrix form as

$$(\pi^{(n+1)}(x_1), \pi^{(n+1)}(x_2), \ldots, \pi^{(n+1)}(x_r)) = (\pi^{(n)}(x_1), \pi^{(n)}(x_2), \ldots, \pi^{(n)}(x_r)) \begin{pmatrix} \mathrm{P}(x_1,x_1) & \cdots & \mathrm{P}(x_1,x_r) \\ \vdots & \ddots & \vdots \\ \mathrm{P}(x_r,x_1) & \cdots & \mathrm{P}(x_r,x_r) \end{pmatrix}$$

or equivalently

$$\pi^{(n+1)} = \pi^{(n)} P.$$

So,

$$\begin{aligned}
\pi^{(1)} &= \pi^{(0)} P \\
\pi^{(2)} &= \pi^{(1)} P = \pi^{(0)} PP = \pi^{(0)} P^2 \\
\pi^{(3)} &= \pi^{(2)} P = \pi^{(0)} P^2 P = \pi^{(0)} P^3 \\
\vdots \quad &= \quad \vdots \\
\pi^{(n)} &= \pi^{(0)} P^n.
\end{aligned}$$

That is, the initial distribution $\pi^{(0)}$, together with the transition matrix P, determine the probability distribution for the state at all future times.

3.1.2 Stationary distributions

A distribution π is said to be a *stationary distribution* of the homogeneous Markov chain governed by the transition matrix P if

$$\boxed{\pi = \pi P.}$$

Note that π is a row eigenvector of the transition matrix, with corresponding eigenvalue equal to one. It is also a fixed point of the linear map induced by P. The stationary distribution is so-called because if at some time n, we have $\pi^{(n)} = \pi$, then $\pi^{(n+1)} = \pi^{(n)} P = \pi P = \pi$, and similarly

$\pi^{(n+k)} = \pi,\ \forall k \geq 0$. That is, if a chain has a stationary distribution, it retains that distribution for all future time. Note that

$$\begin{aligned} \pi = \pi P &\iff \pi - \pi P = 0 \\ &\iff \pi(I - P) = 0 \end{aligned}$$

where I is the $r \times r$ identity matrix. Hence the stationary distribution of the chain may be found by solving

$$\pi(I - P) = 0.$$

3.1.3 Convergence

Convergence of Markov chains is a rather technical topic, which we don't have time to examine in detail here. This short section presents a very informal explanation of why Markov chains often do converge to their stationary distribution, and how the rate of convergence can be understood.

Let π be a (row) eigenvector of P with corresponding eigenvalue λ. Then

$$\pi P = \lambda \pi.$$

Also $\pi P^n = \lambda^n \pi$. It is easy to show that for stochastic P we must have $|\lambda| \leq 1$ (see exercises!). We also know that at least one eigenvector is equal to one (the corresponding eigenvector is a stationary distribution). Let

$$(\pi_1, \lambda_1),\ (\pi_2, \lambda_2), \ldots, (\pi_r, \lambda_r)$$

be the full eigen-decomposition of P, with $|\lambda_i|$ in decreasing order, so that $\lambda_1 = 1$, and π_1 is a (re-scaled) stationary distribution. Now, for *any* initial distribution $\pi^{(0)}$, we may write

$$\pi^{(0)} = a_1\pi_1 + a_2\pi_2 + \cdots + a_r\pi_r$$

uniquely, for appropriate choice of a_i, as the eigenvectors of P form a basis (this isn't always true, but making this assumption keeps the maths simple!). Then

$$\begin{aligned} \pi^{(n)} &= \pi^{(0)} P^n \\ &= (a_1\pi_1 + a_2\pi_2 + \cdots + a_r\pi_r)P^n \\ &= a_1\pi_1 P^n + a_2\pi_2 P^n + \cdots + a_r\pi_r P^n \\ &= a_1\lambda_1^n\pi_1 + a_2\lambda_2^n\pi_2 + \cdots + a_r\lambda_r^n\pi_r \\ &\to a_1\pi_1, \qquad \text{as } n \to \infty, \end{aligned}$$

provided that $|\lambda_2| < 1$. The rate of convergence is governed by the second eigenvalue, λ_2. Provided $|\lambda_2| < 1$, the chain eventually converges to the stationary distribution, irrespective of the initial distribution. If there is more than one unit eigenvalue, then there is an infinite family of stationary distributions, and convergence to any particular distribution is not guaranteed.

3.1.4 Reversible chains

If $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(N)}$ is a Markov chain, then the reversed sequence of states, $\theta^{(N)}, \theta^{(N-1)}, \ldots, \theta^{(0)}$ is also a Markov chain. To see this, consider the conditional distribution of the current state given the future:

$$\begin{aligned}
&\mathrm{P}\left(\theta^{(n)} = y | \theta^{(n+1)} = x_{n+1}, \ldots, \theta^{(N)} = x_N\right) \\
&\qquad = \frac{\mathrm{P}\left(\theta^{(n+1)} = x_{n+1}, \ldots, \theta^{(N)} = x_N | \theta^{(n)} = y\right)\ \mathrm{P}\left(\theta^{(n)} = y\right)}{\mathrm{P}\left(\theta^{(n+1)} = x_{n+1}, \ldots, \theta^{(N)} = x_N\right)} \\
&= \frac{\mathrm{P}\left(\theta^{(n+1)} = x_{n+1} | \theta^{(n)} = y\right) \cdot \cdots\cdots\ \mathrm{P}\left(\theta^{(N)} = x_N | \theta^{(N-1)} = x_{N-1}\right)\ \mathrm{P}\left(\theta^{(n)} = y\right)}{\mathrm{P}\left(\theta^{(n+1)} = x_{n+1}\right)\ \mathrm{P}\left(\theta^{(n+2)} = x_{n+2} | \theta^{(n+1)} = x_{n+1}\right) \cdot \cdots\cdots\ \mathrm{P}\left(\theta^{(N)} = x_N | \theta^{(N-1)} = x_{N-1}\right)} \\
&\qquad = \frac{\mathrm{P}\left(\theta^{(n+1)} = x_{n+1} | \theta^{(n)} = y\right)\ \mathrm{P}\left(\theta^{(n)} = y\right)}{\mathrm{P}\left(\theta^{(n+1)} = x_{n+1}\right)} \\
&\qquad = \mathrm{P}\left(\theta^{(n)} = y | \theta^{(n+1)} = x_{n+1}\right).
\end{aligned}$$

This is exactly the condition required for the reversed sequence of states to be Markovian.

Now let $P_n^*(x,y)$ be the transition kernel for the reversed chain. Then

$$\begin{aligned}
P_n^*(x,y) &= \mathrm{P}\left(\theta^{(n)} = y | \theta^{(n+1)} = x\right) \\
&= \frac{\mathrm{P}\left(\theta^{(n+1)} = x | \theta^{(n)} = y\right)\ \mathrm{P}\left(\theta^{(n)} = y\right)}{\mathrm{P}\left(\theta^{(n+1)} = x\right)} \qquad \text{(Bayes theorem)} \\
&= \frac{\mathrm{P}(y,x)\,\pi^{(n)}(y)}{\pi^{(n+1)}(x)}.
\end{aligned}$$

Therefore in general, the reversed chain is not homogeneous. However, if the chain has reached its stationary distribution, then

$$P^*(x,y) = \frac{\mathrm{P}(y,x)\,\pi(y)}{\pi(x)},$$

and so the reversed chain is homogeneous, and has a transition matrix which may be determined from the transition matrix for the forward chain (and its stationary distribution).

If

$$P^*(x,y) = P(x,y), \qquad \forall x,y$$

then the chain is said to be *reversible*, and we have the *detailed balance equations*:

$$\boxed{\pi(x)P(x,y) = \pi(y)P(y,x), \quad \forall x,y.} \qquad (*)$$

If we have a chain with transition kernel $P(x,y)$ and a distribution $\pi(\cdot)$ satisfying (*), then it follows that the chain is reversible with stationary distribution $\pi(\cdot)$. The chain also has other nice properties (such as positive recurrence) which make it comparatively well-behaved. We can see that $\pi(\cdot)$ must be a stationary distribution by summing both sides over x. Once we know this, reversibility follows immediately.

Consider for a moment the converse problem to the one we have been considering. That is, given a stationary distribution $\pi(\cdot)$, can we find a transition kernel $P(\cdot,\cdot)$ such that (*) is satisfied? That is, can we construct a reversible Markov chain which has $\pi(\cdot)$ as its stationary distribution? The answer is "yes", and we can do it in many different ways, but we shall return to this later, as it is essentially what MCMC is all about.

3.2 Continuous state space Markov chains

Here we are still working with discrete time, but we are allowing the state space S of the Markov chain to be continuous (*eg.* $S \subseteq \mathbf{R}$).

Example — AR(1)

Consider the AR(1) model

$$Z_t = \alpha Z_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2).$$

It is clear that the conditional distribution of Z_t given $Z_{t-1} = z_{t-1}$ is just

$$Z_t | (Z_{t-1} = z_{t-1}) \sim N(\alpha z_{t-1}, \sigma^2),$$

and that it does not depend on any other previous time points. Thus, the AR(1) is a Markov chain and its state space is the real numbers, so it is a continuous state space Markov chain. Note however that other classical time series models such as MA(1) and ARMA(1,1) are *not* Markov chains. The AR(2) is a *second order* Markov chain, but we will not be studying these.

3.2.1 Transition kernels

Again, for a homogeneous chain, we can define

$$\mathrm{P}(x, A) = \mathrm{P}\left(\theta^{(n+1)} \in A | \theta^{(n)} = x\right).$$

For continuous state spaces we always have $P(x, \{y\}) = 0$, so instead we define $P(x,y)$ by

$$\begin{aligned} P(x,y) &= \mathrm{P}\left(\theta^{(n+1)} \leq y | \theta^{(n)} = x\right) \\ &= \mathrm{P}\left(\theta^{(1)} \leq y | \theta^{(0)} = x\right), \ \forall x, y \in S, \end{aligned}$$

the conditional cumulative distribution function. This is the distributional form of the transition kernel for continuous state space Markov chains, but we can also define the corresponding conditional density

$$p(x,y) = \frac{\partial}{\partial y} P(x,y), \quad x, y, \in S.$$

We can use this to define the density form of the *transition kernel* of the chain. This can also be used more conveniently for vector Markov chains, where the state space is multidimensional (say $S \subseteq \mathbf{R}^n$).

Example

If we write our AR(1) in the form

$$\theta^{(t+1)} = \alpha\theta^{(t)} + \varepsilon_t, \qquad \varepsilon_t \sim N(0,\sigma^2)$$

then

$$\theta^{(t+1)}|(\theta^{(t)} = x) \sim N(\alpha x, \sigma^2),$$

and so the density form of the transition kernel is just

$$p(x,y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{y-\alpha x}{\sigma}\right)^2 \right\}.$$

3.2.2 Stationarity and reversibility

Let the state at time n, $\theta^{(n)}$ be represented by a probability density function, $\pi^{(n)}(x),\ x \in S$. By the continuous version of the Theorem of Total Probability, we have

$$\pi^{(n+1)}(y) = \int_S p(x,y)\pi^{(n)}(x)\,dx. \qquad (\dagger)$$

We see from (†) that a stationary distribution must satisfy

$$\boxed{\pi(y) = \int_S p(x,y)\pi(x)\,dx,}$$

which is the continuous version of the discrete matrix equation $\pi = \pi P$.

Again, we can use Bayes theorem to get the transition density for the reversed chain

$$p_n^*(x,y) = \frac{p(y,x)\pi^{(n)}(y)}{\pi^{(n+1)}(x)},$$

which homogenises in the stationary limit to give

$$p^*(x,y) = \frac{p(y,x)\pi(y)}{\pi(x)}.$$

So, if the chain is reversible, we have the continuous form of the detailed balance equations

$$\boxed{\pi(x)p(x,y) = \pi(y)p(y,x), \quad \forall x,y \in S.} \qquad (\ddagger)$$

Again, any chain satisfying (‡) is reversible with stationary distribution $\pi(\cdot)$. We can see that detailed balance implies stationarity of $\pi(\cdot)$ by integrating both sides of (‡) with respect to x. Once we know that $\pi(\cdot)$ is the stationary distribution, reversibility follows immediately.

Example — AR(1)

We know that linear combinations of Normal random variables are Normal, so we expect the stationary distribution of our example AR(1) to be Normal. At convergence, successive distributions are the same. In particular, the first and second moments at successive time points remain constant. First, $\mathrm{E}\left(\theta^{(n+1)}\right) = \mathrm{E}\left(\theta^{(n)}\right)$, and so

$$\begin{aligned}
\mathrm{E}\left(\theta^{(n)}\right) &= \mathrm{E}\left(\theta^{(n+1)}\right) \\
&= \mathrm{E}\left(\alpha\theta^{(n)} + \varepsilon_n\right) \\
&= \alpha\,\mathrm{E}\left(\theta^{(n)}\right) \\
\Rightarrow \mathrm{E}\left(\theta^{(n)}\right) &= 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{Var}\left(\theta^{(n)}\right) &= \mathrm{Var}\left(\theta^{(n+1)}\right) \\
&= \mathrm{Var}\left(\alpha\theta^{(n)} + \varepsilon_n\right) \\
&= \alpha^2\mathrm{Var}\left(\theta^{(n)}\right) + \sigma^2 \\
\Rightarrow \mathrm{Var}\left(\theta^{(n)}\right) &= \frac{\sigma^2}{1-\alpha^2}.
\end{aligned}$$

So, we think the stationary distribution is Normal with mean zero and variance $\sigma^2/(1-\alpha^2)$. That is, we think the stationary density is

$$\begin{aligned}
\pi(x) &= \frac{1}{\sqrt{\frac{2\pi\sigma^2}{1-\alpha^2}}} \exp\left\{-\frac{1}{2}\frac{x^2}{\frac{\sigma^2}{1-\alpha^2}}\right\} \\
&= \sqrt{\frac{1-\alpha^2}{2\pi\sigma^2}} \exp\left\{-\frac{x^2(1-\alpha^2)}{2\sigma^2}\right\}.
\end{aligned}$$

Since we know the transition density for this chain, we can see if this density satisfies detailed balance:

$$\begin{aligned}
\pi(x)p(x,y) &= \sqrt{\frac{1-\alpha^2}{2\pi\sigma^2}} \exp\left\{-\frac{x^2(1-\alpha^2)}{2\sigma^2}\right\} \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\alpha x}{\sigma}\right)^2\right\} \\
&= \frac{\sqrt{1-\alpha^2}}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}[x^2 - 2\alpha xy + y^2]\right\}
\end{aligned}$$

after a little algebra. But this expression is exactly symmetric in x and y, and so

$$\pi(x)p(x,y) = \pi(y)p(y,x).$$

So we see that $\pi(\cdot)$ does satisfy detailed balance, and so the AR(1) is a *reversible* Markov chain and *does* have stationary distribution $\pi(\cdot)$.

3.3 Simulation

3.3.1 Simulating Markov chains

Markov chain simulation is easy provided that we can simulate from the initial distribution, $\pi^{(0)}(x)$, and from the conditional distribution represented by the transition kernel, $p(x,y)$.

First we simulate $\theta^{(0)}$ from $\pi^{(0)}(\cdot)$, using one of the techniques discussed in Chapter 2. We can then simulate $\theta^{(1)}$ from $p(\theta^{(0)},\cdot)$, as this is just a density. In general, once we have simulated a realisation of $\theta^{(n)}$, we can simulate $\theta^{(n+1)}$ from $p(\theta^{(n)},\cdot)$, using one of the standard techniques from Chapter 2.

Example — AR(1)

Let us start our AR(1) off at $\theta^{(0)} = 0$, so we don't need to simulate anything for the initial value. Next we want to simulate $\theta^{(1)}$ from $p(\theta^{(0)},\cdot) = p(0,\cdot)$, that is, we simulate $\theta^{(1)}$ from $N(0,\sigma^2)$. Next we simulate $\theta^{(2)}$ from $p(\theta^{(1)},\cdot)$, that is, we simulate $\theta^{(2)}$ from $N(\alpha\theta^{(1)},\sigma^2)$. In general, having simulated $\theta^{(n)}$, we simulate $\theta^{(n+1)}$ from $N(\alpha\theta^{(n)},\sigma^2)$.

3.3.2 "Burn-in" and the stationary distribution

As n gets large, the distribution of $\theta^{(n)}$ tends to the distribution with density $\pi(\cdot)$, the stationary distribution of the chain. All values sampled after convergence has been reached are draws from $\pi(\cdot)$. There is a "burn-in" period before convergence is reached, so if interest is in $\pi(\cdot)$, these values should be discarded before analysis takes place.

3.3.3 Analysis

If we are interested in an integral

$$\int_S g(x)\pi(x)dx = \mathrm{E}_\pi(g(\Theta)),$$

then if $\theta^{(1)},\theta^{(2)},\ldots,\theta^{(n)}$ are draws from $\pi(\cdot)$, this integral may be approximated by

$$\mathrm{E}_\pi(g(\Theta)) \simeq \frac{1}{n}\sum_{i=1}^{n} g(\theta^{(i)}).$$

However, *draws from a Markov chain are not independent*, so the variance of the sample mean cannot be computed in the usual way.

Suppose $\theta^{(i)} \sim \pi(\cdot)$, $i = 1,2,\ldots$. Then

$$\mathrm{E}_\pi(\Theta) \simeq \frac{1}{n}\sum_{i=1}^{n} \theta^{(i)} = \bar{\theta}_n.$$

Let $\mathrm{Var}\,(\Theta) = \mathrm{Var}\left(\theta^{(i)}\right) = \nu^2$. Then if the θ_i were independent, we would have

$$\mathrm{Var}\left(\bar{\theta}_n\right) = \frac{\nu^2}{n}.$$

However, if the $\theta^{(i)}$ are *not* independent (*eg.* sampled from a non-trivial Markov chain), then

$$\boxed{\text{Var}\left(\bar{\theta}_n\right) \neq \frac{\nu^2}{n}.}$$

Example — AR(1)

$$\text{Var}\left(\theta^{(i)}\right) = \frac{\sigma^2}{1-\alpha^2} = \nu^2$$

and

$$\gamma(k) = \text{Cov}\left(\theta^{(i)}, \theta^{(i+k)}\right) = \nu^2 \alpha^k,$$

so

$$\begin{aligned}\text{Var}\left(\bar{\theta}_n\right) &= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n}\theta^{(i)}\right) \\ &= \frac{\nu^2}{n^2}\left(n + \sum_{i=1}^{n-1} 2(n-i)\alpha^i\right).\end{aligned}$$

We can get Maple to sum this up for us as follows:

```
> S:=sum(2*(n-i)*alphaˆi,i=1..(n-1));
> V:=((sigmaˆ2)/(nˆ2*(1-alphaˆ2)))*(n+s);
> simplify(expand(V));
```

which gives

$$\begin{aligned}\text{Var}\left(\bar{\theta}_n\right) &= \frac{1}{n^2}\frac{\sigma^2}{1-\alpha^2}\frac{n + 2\alpha^{n+1} - 2\alpha - n\alpha^2}{(1-\alpha)^2} \\ &= \frac{1}{n}\frac{\sigma^2}{1-\alpha^2}\left[\frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^n)}{n(1-\alpha)^2}\right].\end{aligned}$$

To a first approximation, we have

$$\text{Var}\left(\bar{\theta}_n\right) \simeq \frac{1}{n}\frac{\sigma^2}{1-\alpha^2}\frac{1+\alpha}{1-\alpha},$$

and so the "correction factor" for the naive calculation based on an assumption of independence is $(1+\alpha)/(1-\alpha)$. For α close to one, this can be very large. *eg.* for $\alpha = 0.95$, $(1+\alpha)/(1-\alpha) = 39$, and so the variance of the sample mean is actually around 40 times bigger than calculations based on assumptions of independence would suggest. Similarly, confidence intervals should be around 6 times wider than calculations based on independence would suggest.

We can actually use this analysis in order to analyse other Markov chains. If the Markov chain is reasonably well approximated by an AR(1) (and many are), then we can estimate the variance of our sample estimates by the AR(1) variance estimate. For an AR(1), α is just the lag 1 autocorrelation of the chain ($\text{Corr}(\theta^{(i)}, \theta^{(i+1)}) = \alpha$), and so we can estimate the α of any simulated chain by the sample autocorrelation at lag 1. We can then use this to compute or correct sample variance estimates based on sample means of chain values.

Chapter 4

Markov Chain Monte Carlo

4.1 Introduction

In Chapter 2, we discussed the fundamentals of stochastic simulation techniques. In Chapter 3, we looked at Markov chains, their properties and how to simulate them. In this chapter we will be concerned with combining the two in order to provide a flexible framework for simulating from complex distributions based on simulated values from carefully constructed Markov chains. The techniques are generally referred to as *Markov Chain Monte Carlo* techniques, and are often abbreviated to MCMC. There has been an explosion in the use of MCMC in statistics over recent years, primarily because of their application in Bayesian inference. However, MCMC has application in other areas of statistics too, and so the theory will be presented in a general context before being applied to the problem of simulating from Bayesian posterior distributions.

There are many different MCMC techniques, but we only have time to look briefly at two of the most fundamental. The first is the *Gibbs sampler*, which was at the forefront of the recent MCMC revolution, and the second is generally known as *Metropolis-Hastings* sampling. In fact, MCMC schemes based on the combination of these two fundamental techniques are still at the forefront of MCMC research.

4.2 The Gibbs sampler

4.2.1 Introduction

The Gibbs sampler is a way of simulating from multivariate distributions based only on the ability to simulate from conditional distributions. In particular, it is appropriate when sampling from marginal distributions is not convenient or possible.

Example

Reconsider the problem from Chapter 1 of Bayesian inference for the mean and variance of a normally distributed random sample. In particular, consider the non-conjugate approach based on independent prior distributions for the mean and variance. We had

$$\begin{aligned} X_i|\mu,\tau &\sim N(\mu, 1/\tau) \text{ independently, } i = 1,\ldots,n \\ \tau &\sim Gamma(a,b) \\ \mu &\sim N(c, 1/d). \end{aligned}$$

We used these to derive the joint posterior distribution for μ and τ based on a sample x of size n in terms of the sufficient statistics

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The posterior took the form

$$\pi(\mu,\tau|x) \propto \tau^{a+\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2}\left[(n-1)s^2 + n(\bar{x}-\mu)^2\right] - \frac{d}{2}(\mu-c)^2 - b\tau\right\}.$$

As explained previously, this distribution is not in a standard form. However, whilst clearly not conjugate, this problem is often referred to as *semi-conjugate*, because the two *full conditional* distributions $\pi(\mu|\tau,x)$ and $\pi(\tau|\mu,x)$ *are* of standard form, and further, *are* of the same form as the independent prior specifications, that is, $\tau|\mu,x$ is gamma distributed, and $\mu|\tau,x$ is normally distributed. In fact,

$$\tau|\mu,x \sim Gamma\left(a+\frac{n}{2}, b+\frac{1}{2}\left[(n-1)s^2 + n(\bar{x}-\mu)^2\right]\right)$$
$$\mu|\tau,x \sim N\left(\frac{cd+n\tau\bar{x}}{n\tau+d}, \frac{1}{n\tau+d}\right).$$

So, providing that we can simulate normal and gamma quantities, we can simulate from the full conditionals. How can we simulate from the joint density or the marginals?

4.2.2 Sampling from bivariate densities

Consider a bivariate density $\pi(x,y)$. We have

$$\pi(x,y) = \pi(x)\pi(y|x)$$

so we can simulate from $\pi(x,y)$ by first simulating $X = x$ from $\pi(x)$, and then simulating $Y = y$ from $\pi(y|x)$. On the other hand, if we can simulate from the marginal for y, we can write

$$\pi(x,y) = \pi(y)\pi(x|y)$$

and simulate $Y = y$ from $\pi(y)$ and then $X = x$ from $\pi(x|y)$. Either way we need to be able to simulate from one of the marginals!

Let's just suppose we can, that is, we have an $X = x$ from $\pi(x)$. Given this, we can now simulate a $Y = y$ from $\pi(y|x)$ to give a pair of points (x,y) from the bivariate density. However, in that case the y value must be from the marginal $\pi(y)$, and so we can simulate an $X' = x'$ from $\pi(x'|y)$ to give a new pair of points (x',y) also from the joint density. But now x' is from the marginal $\pi(x)$, and so we can keep going. This alternate sampling from conditional distributions defines a bivariate Markov chain, and we have just given an intuitive explanation for why $\pi(x,y)$ is its stationary distribution. The transition kernel for this bivariate Markov chain is

$$p((x,y),(x',y')) = \pi(x',y'|x,y) = \pi(x'|x,y)\pi(y'|x',x,y) = \pi(x'|y)\pi(y'|x').$$

4.2.3 The Gibbs sampler

Suppose the density of interest is $\pi(\theta)$, where $\theta = (\theta_1, \ldots, \theta_d)'$, and that the full conditionals

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d) = \pi(\theta_i|\theta_{-i}) = \pi_i(\theta_i), \qquad i = 1, \ldots, d$$

are available for simulating from. The Gibbs sampler follows the following algorithm:

1. Initialise the iteration counter to $j = 1$. Initialise the state of the chain to $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})'$.
2. Obtain a new value $\theta^{(j)}$ from $\theta^{(j-1)}$ by successive generation of values
$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\theta_d^{(j)} &\sim \pi(\theta_d|\theta_1^{(j)}, \ldots, \theta_{d-1}^{(j)})
\end{aligned}$$
3. Change counter j to $j+1$, and return to step 2.

This clearly defines a homogeneous Markov chain, as each simulated value depends only on the previous simulated value, and not on any other previous values or the iteration counter j. However, we need to show that $\pi(\theta)$ is a stationary distribution of this chain. The transition kernel of the chain is

$$p(\theta, \phi) = \prod_{i=1}^{d} \pi(\phi_i|\phi_1, \ldots, \phi_{i-1}, \theta_{i+1}, \ldots, \theta_d).$$

Therefore, we just need to check that $\pi(\theta)$ is the stationary distribution of the chain with this transition kernel. Unfortunately, the traditional *fixed-sweep* Gibbs sampler just described is *not* reversible, and so we cannot check stationarity by checking for detailed balance (as detailed balance fails). We need to do a direct check of the stationarity of $\pi(\theta)$, that is, we need to check that

$$\pi(\phi) = \int_S p(\theta, \phi)\pi(\theta)\, d\theta.$$

For the bivariate case, we have

$$\begin{aligned}
\int_S p(\theta, \phi)\pi(\theta)\, d\theta &= \int_S \pi(\phi_1|\theta_2)\pi(\phi_2|\phi_1)\pi(\theta_1, \theta_2)\, d\theta_1 d\theta_2 \\
&= \pi(\phi_2|\phi_1) \int_{S_1} \int_{S_2} \pi(\phi_1|\theta_2)\pi(\theta_1, \theta_2)\, d\theta_1 d\theta_2 \\
&= \pi(\phi_2|\phi_1) \int_{S_2} \pi(\phi_1|\theta_2)\, d\theta_2 \int_{S_1} \pi(\theta_1, \theta_2)\, d\theta_1 \\
&= \pi(\phi_2|\phi_1) \int_{S_2} \pi(\phi_1|\theta_2)\pi(\theta_2)\, d\theta_2 \\
&= \pi(\phi_2|\phi_1)\pi(\phi_1) \\
&= \pi(\phi_1, \phi_2) \\
&= \pi(\phi).
\end{aligned}$$

The general case is similar. So, $\pi(\theta)$ is a stationary distribution of this chain. Discussions of uniqueness and convergence are beyond the scope of this course. In particular, these issues are complicated somewhat by the fact that the sampler described is not reversible.

4.2.4 Reversible Gibbs samplers

Whilst the fixed-sweep Gibbs sampler itself is not reversible, each *component update* is, and hence there are many variations on the fixed-sweep Gibbs sampler which *are* reversible, and hence do satisfy detailed balance. Let us start by looking at why each component update is reversible.

Suppose we wish to update component i, that is, we update θ by replacing θ_i with ϕ_i drawn from $\pi(\phi_i|\theta_{-i})$. All other components will remain unchanged. The transition kernel for this update is

$$p(\theta,\phi) = \pi(\phi_i|\theta_{-i})I(\theta_{-i} = \phi_{-i})$$

where

$$I(E) = \begin{cases} 1 & \text{if } E \text{ is true,} \\ 0 & \text{if } E \text{ is false.} \end{cases}$$

Note that the density is zero for any transition changing the other components. Now we may check detailed balance:

$$\begin{aligned} \pi(\theta)p(\theta,\phi) &= \pi(\theta)\pi(\phi_i|\theta_{-i})I(\theta_{-i} = \phi_{-i}) \\ &= \pi(\theta_{-i})\pi(\theta_i|\theta_{-i})\pi(\phi_i|\theta_{-i})I(\theta_{-i} = \phi_{-i}) \\ &= \pi(\phi_{-i})\pi(\theta_i|\phi_{-i})\pi(\phi_i|\phi_{-i})I(\theta_{-i} = \phi_{-i}) && \text{(as } \theta_{-i} = \phi_{-i}\text{)} \\ &= \pi(\phi)\pi(\theta_i|\phi_{-i})I(\theta_{-i} = \phi_{-i}) \\ &= \pi(\phi)p(\phi,\theta). \end{aligned}$$

Therefore detailed balance is satisfied, and hence the update is reversible with stationary distribution $\pi(\cdot)$.

If this particular update is reversible, and preserves the equilibrium distribution of the chain, why bother updating any other component? The reason is that the chain defined by a single update is *reducible*, and hence will not converge to the stationary distribution from an arbitrary starting point. In order to ensure *irreducibility* of the chain, we need to make sure that we update each component sufficiently often. As we have seen, one way to do this is to update each component in a fixed order. The drawback of this method is that *reversibility* is lost when we do this.

An alternative to the *fixed-sweep* strategy is to pick a component at random at each stage, and update that. This gives a reversible chain with the required stationary distribution, and is known as the *random scan* Gibbs sampler.

An even simpler way to restore the reversibility of the chain is to first scan through the components in fixed order, and then scan backwards through the components. This does define a reversible Gibbs sampler. We can check that it works in the bivariate case as follows. The algorithm starts with (θ_1,θ_2) and then generates (ϕ_1,ϕ_2) as follows:

$$\begin{aligned} \theta_1' &\sim \pi(\theta_1'|\theta_2) \\ \phi_2 &\sim \pi(\phi_2|\theta_1') \\ \phi_1 &\sim \pi(\phi_1|\phi_2). \end{aligned}$$

Here θ_1' is an auxiliary variable that we are not interested in *per se*, and which needs to be integrated out of the problem. The full transition kernel is

$$p(\theta,(\theta_1',\phi)) = \pi(\theta_1'|\theta_2)\pi(\phi_2|\theta_1')\pi(\phi_1|\phi_2),$$

and integrating out the auxiliary variable gives

$$\begin{aligned} p(\theta,\phi) &= \int \pi(\theta_1'|\theta_2)\pi(\phi_2|\theta_1')\pi(\phi_1|\phi_2)\,d\theta_1' \\ &= \pi(\phi_1|\phi_2)\int \pi(\theta_1'|\theta_2)\pi(\phi_2|\theta_1')\,d\theta_1'. \end{aligned}$$

We can now check for detailed balance:

$$\begin{aligned} \pi(\theta)p(\theta,\phi) &= \pi(\theta)\pi(\phi_1|\phi_2)\int \pi(\theta_1'|\theta_2)\pi(\phi_2|\theta_1')\,d\theta_1' \\ &= \pi(\theta_2)\pi(\theta_1|\theta_2)\pi(\phi_1|\phi_2)\int \pi(\theta_1'|\theta_2)\pi(\phi_2|\theta_1')\,d\theta_1' \\ &= \pi(\theta_1|\theta_2)\pi(\phi_1|\phi_2)\int \pi(\theta_2)\pi(\theta_1'|\theta_2)\pi(\phi_2|\theta_1')\,d\theta_1' \\ &= \pi(\theta_1|\theta_2)\pi(\phi_1|\phi_2)\int \pi(\theta_1',\theta_2)\pi(\phi_2|\theta_1')\,d\theta_1' \\ &= \pi(\theta_1|\theta_2)\pi(\phi_1|\phi_2)\int \pi(\theta_1')\pi(\theta_2|\theta_1')\pi(\phi_2|\theta_1')\,d\theta_1', \end{aligned}$$

and, as this is symmetric in θ and ϕ, we must have

$$\pi(\theta)p(\theta,\phi) = \pi(\phi)p(\phi,\theta).$$

This chain is therefore reversible with stationary distribution $\pi(\cdot)$.

We have seen that there are ways of adapting the standard fixed-sweep Gibbs sampler in ways which ensure reversibility. However, reversibility is not a requirement of a useful algorithm — it simply makes it easier to determine the properties of the chain. In practice, the fixed-sweep Gibbs sampler often has as good or better convergence properties than its reversible cousins. Given that it is slightly easier to implement and debug, it is often simpler to stick with the fixed-sweep scheme than to implement a more exotic version of the sampler.

4.2.5 Simulation and analysis

Suppose that we are interested in a multivariate distribution $\pi(\theta)$ (which may be a Bayesian posterior distribution), and that we are able to simulate from the full conditional distributions of $\pi(\theta)$. Simulation from $\pi(\theta)$ is possible by first initialising the sampler somewhere in the support of θ, and then running the Gibbs sampler. The resulting chain should be monitored for convergence, and the "burn-in" period should be discarded for analysis. After convergence, the simulated values are all from $\pi(\theta)$. In particular, the values for a particular component will be simulated values from the marginal distribution of that component. A histogram of these values will give an idea of the "shape" of the marginal distribution, and summary statistics such as the mean and variance will be approximations to the mean and variance of the marginal distribution. The accuracy of the estimates can be gauged using the techniques from the end of Chapter 3.

Example

Returning to the case of the posterior distribution for the normal model with unknown mean and precision, we have already established that the full conditional distributions are

$$\tau|\mu,x \sim Gamma\left(a+\frac{n}{2}, b+\frac{1}{2}\left[(n-1)s^2+n(\bar{x}-\mu)^2\right]\right)$$
$$\mu|\tau,x \sim N\left(\frac{cd+n\tau\bar{x}}{n\tau+d}, \frac{1}{n\tau+d}\right).$$

We can initialise the sampler anywhere in the half-plane where the posterior (and prior) has support, but convergence will be quicker if the chain is not started in the tails of the distribution. One possibility is to start the sampler near the posterior mode, though this can make convergence more difficult to diagnose. A simple strategy which is often easy to implement for problems in Bayesian inference is to start off the sampler at a point simulated from the prior distribution, or even at the mean of the prior distribution. Here, the prior mean for (τ,μ) is $(a/b,c)$. Once initialised, the sampler proceeds with alternate simulations from the full conditional distributions. The first few (hundred?) values should be discarded, and the rest can give information about the posterior marginal distributions.

Of course, the Gibbs sampler tacitly assumes that we have some reasonably efficient mechanism for simulating from the full conditional distributions, and yet this isn't always the case. Fortunately, the Gibbs sampler can be combined with Metropolis-Hastings algorithms when the full conditionals are difficult to simulate from.

4.3 Metropolis-Hastings sampling

4.3.1 Introduction

Let us return to the problem we considered in Chapter 2: given a distribution, how can we simulate from it? If none of the techniques from Chapter 2 are (obviously) appropriate, what can we do? One possibility is to construct a reversible Markov chain which has the distribution of interest as its stationary distribution. Then, by simulating such a Markov chain we can obtain random variates from the distribution of interest. Obviously such a strategy will not result in a sequence of independent values from the distribution, but this does not necessarily matter. The question is then how to construct such a Markov chain? Obviously, it must be easier to simulate the Markov chain (using methods from Chapter 2) than to simulate from the stationary distribution itself, or nothing has been gained. There are actually many ways we can do this, but there is a general class of methods known as Metropolis-Hastings schemes, which are generally applicable and widely used.

4.3.2 Metropolis-Hastings algorithm

Suppose that $\pi(\theta)$ is the density of interest. Suppose further that we have some (arbitrary) transition kernel $q(\theta,\phi)$ (known as the *proposal distribution*) which is easy to simulate from, but does not (necessarily) have $\pi(\theta)$ as its stationary density. Consider the following algorithm:

1. Initialise the iteration counter to $j = 1$, and initialise the chain to $\theta^{(0)}$.
2. Generate a *proposed* value ϕ using the kernel $q(\theta^{(j-1)},\phi)$.
3. Evaluate the *acceptance probability* $\alpha(\theta^{(j-1)},\phi)$ of the proposed move, where
$$\alpha(\theta,\phi) = \min\left\{1, \frac{\pi(\phi)q(\phi,\theta)}{\pi(\theta)q(\theta,\phi)}\right\}.$$
4. Put $\theta^{(j)} = \phi$ with probability $\alpha(\theta^{(j-1)},\phi)$, and put $\theta^{(j)} = \theta^{(j-1)}$ otherwise.
5. Change the counter from j to $j+1$ and return to step 2.

In other words, at each stage, a new value is generated from the proposal distribution. This is either accepted, in which case the chain moves, or rejected, in which case the chain stays where it is. Whether or not the move is accepted or rejected depends on an acceptance probability which itself depends on the relationship between the density of interest and the proposal distribution. Note that the density of interest $\pi(\cdot)$ only enters into the acceptance probability as a ratio, and so the method can be used when the density of interest is only known up to a scaling constant.

The Markov chain defined in this way is reversible, and has stationary distribution $\pi(\cdot)$ irrespective of the choice of proposal distribution, $q(\cdot,\cdot)$. Let us see why. The transition kernel is clearly given by

$$p(\theta,\phi) = q(\theta,\phi)\alpha(\theta,\phi), \quad \text{if } \theta \neq \phi.$$

But there is also a finite probability that the chain will remain at θ. This is one minus the probability that the chain moves, and thus is given by

$$1 - \int q(\theta,\phi)\alpha(\theta,\phi)\,d\phi.$$

So, the transition kernel is part continuous and part discrete. We can easily write down the cumulative distribution form of the transition kernel as

$$P(\theta,\phi) = \int_{-\infty}^{\phi} q(\theta,\phi)\alpha(\theta,\phi)\,d\phi + I(\phi \geq \theta)\left[1 - \int q(\theta,\phi)\alpha(\theta,\phi)\,d\phi\right].$$

We then get the full density form of the kernel by differentiating with respect to ϕ as

$$p(\theta,\phi) = q(\theta,\phi)\alpha(\theta,\phi) + \delta(\theta-\phi)\left[1 - \int q(\theta,\phi)\alpha(\theta,\phi)\,d\phi\right],$$

where $\delta(\cdot)$ is the Dirac δ-function. Now we have the transition kernel we can check whether detailed balance is satisfied:

$$\begin{aligned}\pi(\theta)p(\theta,\phi) &= \pi(\theta)q(\theta,\phi)\min\left\{1,\frac{\pi(\phi)q(\phi,\theta)}{\pi(\theta)q(\theta,\phi)}\right\} \\ &\quad + \delta(\theta-\phi)\left[\pi(\theta) - \int \pi(\theta)q(\theta,\phi)\min\left\{1,\frac{\pi(\phi)q(\phi,\theta)}{\pi(\theta)q(\theta,\phi)}\right\} d\phi\right] \\ &= \min\{\pi(\theta)q(\theta,\phi),\pi(\phi)q(\phi,\theta)\} \\ &\quad + \delta(\theta-\phi)\left[\pi(\theta) - \int \min\{\pi(\theta)q(\theta,\phi),\pi(\phi)q(\phi,\theta)\}\, d\phi\right].\end{aligned}$$

The first term is clearly symmetric in θ and ϕ. Also, the second term must be symmetric in θ and ϕ, because it is only non-zero precisely when $\theta = \phi$. Consequently, detailed balance is satisfied, and the Metropolis-Hastings algorithm defines a reversible Markov chain with stationary distribution $\pi(\cdot)$, irrespective of the form of $q(\cdot,\cdot)$.

Complete freedom in the choice of the proposal distribution $q(\cdot,\cdot)$ leaves us wondering what kinds of choices might be good, or generally quite useful. Some commonly used special cases are discussed below.

4.3.3 Symmetric chains (Metropolis method)

The simplest case is the Metropolis sampler, which is based on the use of a symmetric proposal with $q(\theta,\phi) = q(\phi,\theta),\ \forall\theta,\phi$. We see then that the acceptance probability simplifies to

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

and hence does not involve the proposal density at all. Consequently proposed moves which will take the chain to a region of higher density are always accepted, while moves which take the chain to a region of lower density are accepted with probability proportional to the ratio of the two densities — moves which will take the chain to a region of very low density will be accepted with very low probability. Note that any proposal of the form $q(\theta,\phi) = f(|\theta-\phi|)$ is symmetric, where $f(\cdot)$ is an arbitrary density. In this case, the proposal will represent a symmetric displacement from the current value. This also motivates the following.

4.3.4 Random walk chains

In this case, the proposed value ϕ at stage j is $\phi = \theta^{(j-1)} + w_j$ where the w_j are iid random variables (completely independent of the state of the chain). Suppose that the w_j have density $f(\cdot)$, which is easy to simulate from. We can then simulate an *innovation*, w_j, and set the *candidate* point to $\phi = \theta^{(j-1)} + w_j$. The transition kernel is then $q(\theta,\phi) = f(\phi-\theta)$, and this can be used to compute the acceptance probability. Of course, if $f(\cdot)$ is symmetric about zero, then we have a symmetric chain, and the acceptance probability does not depend on $f(\cdot)$ at all.

So, suppose you decide to use a symmetric random walk chain with proposed mean zero innovations. There is still the question of how they should be distributed, and what variance they should have. A simple, easy to simulate from distribution is always a good idea, such as uniform

or normal (normal is generally better, but is a bit more expensive to simulate). So, what variance should we choose? The choice of variance will affect the acceptance probability, and hence the overall proportion of accepted moves. If the variance of the innovation is too low, then most proposed values will be accepted, but the chain will move very slowly around the space — the chain is said to be too "cold". On the other hand, if the variance of the innovation is too large, very few proposed values will be accepted, but when they are, they will often correspond to quite large moves — the chain is said to be too "hot". Experience suggests that an overall acceptance rate of around 30% is desirable, and so it is possible to "tune" the variance of the innovation distribution to get an acceptance rate of around this level.

4.3.5 Independence chains

In this case (reminiscent of the envelope rejection method and importance sampling), the proposed transition is formed independently of the previous position of the chain, and so $q(\theta,\phi) = f(\phi)$ for some density $f(\cdot)$. Here the acceptance probability becomes

$$\alpha(\theta,\phi) = \min\left\{1, \frac{\pi(\phi)}{\pi(\theta)} \Big/ \frac{f(\phi)}{f(\theta)}\right\},$$

and we see that the acceptance probability can be increased by making $f(\cdot)$ as similar to $\pi(\cdot)$ as possible (in this case, the higher the acceptance probability, the better).

Bayes Theorem via independence chains

In the context of Bayesian inference, just as with the envelope method, one possible choice for the proposal density is the prior density. The acceptance probability then becomes

$$\alpha(\theta,\phi) = \min\left\{1, \frac{L(\phi;x)}{L(\theta;x)}\right\},$$

and hence depends only on the likelihood ratio of the candidate point and the current value.

4.4 Hybrid methods

We have now seen how we can use the Gibbs sampler to sample from multivariate distributions provided that we can simulate from the full conditionals. We have also seen how we can use Metropolis-Hastings methods to sample from awkward distributions (perhaps full conditionals). If we wish, we can combine these in order to form hybrid Markov chains whose stationary distribution is a distribution of interest.

4.4.1 Componentwise transition

Given a multivariate distribution with full conditionals that are awkward to sample from directly, we can define a Metropolis-Hastings scheme for each full conditional, and apply them to each component in turn for each iteration. This is like the Gibbs sampler, but each component update is a Metropolis-Hastings update, rather than a direct simulation from the full conditional. This is in fact the original form of the Metropolis algorithm.

4.4.2 Metropolis within Gibbs

Given a multivariate distribution with full conditionals, some of which may be simulated from directly, and others which have Metropolis-Hastings updating schemes, the Metropolis within Gibbs algorithm goes through each in turn, and simulates directly from the full conditional, or carries out a Metropolis-Hastings update as necessary.

4.4.3 Blocking

The components of a Gibbs sampler, and those of Metropolis-Hastings chains, can be vectors (or matrices) as well as scalars. For many high-dimensional problems, it can be helpful to group related parameters together into blocks, and use multivariate simulation techniques to update those together if possible.

It is very useful to know of the existence of these hybrid methods for more complex problems, but the actual implementation of such schemes is beyond the scope of this course.

4.5 Summary and conclusions

The purpose of this part of the course was two-fold.

1. To develop an understanding of the advanced simulation techniques used in modern statistical analysis.

2. To show how these techniques can be used to carry out Bayesian inference for complex models where analytic analysis is intractable.

We will finish the course by looking briefly at an example of Bayesian inference for a problem which is a little more involved than any we have analysed previously.

Example

Consider the following simple *hierarchical* (or *one-way random effects*) model:

$$\begin{aligned} Y_{ij}|\theta_i,\tau &\sim N(\theta_i, 1/\tau), \quad \text{independently, } i = 1,\ldots,m,\ j = 1,\ldots n_i \\ \theta_i|\mu,\nu &\sim N(\mu, 1/\nu), \quad i = 1,\ldots,m. \end{aligned}$$

Such a model could be used to model a quality attribute of an industrial batch process, where m batches of items are produced, and batch i contains n_i items. Y_{ij} is the measurement made on the jth item in batch i. We assume that batch i has mean θ_i and that the measurements are normally distributed. We also assume that the θ_i are themselves normally distributed. Essentially, the model has the effect of inducing a correlation between items in a batch, due to the fact that we expect items within a batch to be more similar than items from different batches. Note that this generic scenario can be applied to a range of situations. For example, the items of interest could be schools, and the batches could represent LEAs. By measuring the performance of the schools within LEAs, inferences can be made about the quality of the LEAs themselves.

We will consider the most general (and quite typical) case where μ, τ and ν are all unknown. We wish to make inferences about these parameters, and the unknown θ_i. Thus there are $m+3$ parameters of interest in this model.

The specification of the model is completed with independent priors for μ, τ and ν:

$$\begin{aligned}\mu &\sim N(a, 1/b)\\ \tau &\sim Gamma(c,d)\\ \nu &\sim Gamma(e,f).\end{aligned}$$

In principle we have now completely specified the model, and can compute the posterior distribution. Of course, the posterior distribution is very high dimensional and, more importantly, not of a standard form. Here MCMC techniques can be used to describe the posterior distribution. The likelihood contribution for each observation y_{ij} is

$$L(\theta_i, \tau; y_{ij}) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}(y_{ij}-\theta_i)^2\right\}$$

and so the full likelihood is

$$\begin{aligned}L(\theta, \tau; y) &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} L(\theta_i, \tau; y_{ij})\\ &= \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{m}\left[(n_i-1)s_i^2 + n_i(y_{i\cdot}-\theta_i)^2\right]\right\}\end{aligned}$$

where

$$N = \sum_{i=1}^{m} n_i, \quad y_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij}-y_{i\cdot})^2.$$

The prior takes the form

$$\pi(\mu, \tau, \nu, \theta) = \pi(\mu)\pi(\tau)\pi(\nu)\pi(\theta|\mu,\nu)$$

where

$$\begin{aligned}\pi(\mu) &\propto \exp\left\{-\frac{b}{2}(\mu-a)^2\right\}\\ \pi(\tau) &\propto \tau^{c-1}\exp\{-d\tau\}\\ \pi(\nu) &\propto \nu^{e-1}\exp\{-f\nu\}\\ \pi(\theta_i|\mu,\nu) &= \sqrt{\frac{\nu}{2\pi}}\exp\left\{-\frac{\nu}{2}(\theta_i-\mu)^2\right\}\\ \Rightarrow \pi(\theta|\mu,\nu) &\propto \nu^{m/2}\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{m}(\theta_i-\mu)^2\right\}\end{aligned}$$

and therefore,

$$\pi(\mu, \tau, \nu, \theta) \propto \nu^{e+m/2-1}\tau^{c-1}\exp\left\{-\frac{1}{2}\left[2d\tau + 2f\nu + b(\mu-a)^2 + \nu\sum_{i=1}^{m}(\theta_i-\mu)^2\right]\right\}.$$

Now we have the likelihood and the prior, we can write down the posterior distribution:

$$\pi(\mu,\tau,\nu,\theta|y) \propto$$
$$\tau^{c+N/2-1}\nu^{e+m/2-1}\exp\left\{-\frac{1}{2}\left[2d\tau+2f\nu+b(\mu-a)^2+\sum_{i=1}^{m}\left(\nu(\theta_i-\mu)^2+\tau(n_i-1)s_i^2+\tau n_i(y_{i\cdot}-\theta_i)^2\right)\right]\right\}.$$

It is difficult to do anything analytic with this, so we will try and construct a Gibbs sampler in order to investigate it. This is fairly straightforward, and the full conditionals are as follows. The full conditional for μ is

$$\begin{aligned}\pi(\mu|\cdot) &\propto \exp\left\{-\frac{1}{2}\left[b(\mu-a)^2+\sum_{i=1}^{m}\nu(\theta_i-\mu)^2\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}(b+m\nu)\left(\mu-\frac{ba+m\nu\bar{\theta}}{b+m\nu}\right)^2\right\}, \quad \text{where } \bar{\theta}=\frac{1}{m}\sum_{i=1}^{m}\theta_i,\end{aligned}$$

that is

$$\boxed{\mu|\cdot} \sim \boxed{N\left(\frac{ba+n\nu\bar{\theta}}{b+m\nu},\frac{1}{b+m\nu}\right)}.$$

The conditional for τ is

$$\pi(\tau|\cdot) \propto \tau^{c+N/2-1}\exp\left\{-\tau\left[d+\frac{1}{2}\sum_{i=1}^{m}\left((n_i-1)s_i^2+n_i(y_{i\cdot}-\theta_i)^2\right)\right]\right\},$$

that is

$$\boxed{\tau|\cdot} \sim \boxed{Gamma\left(c+N/2, d+\frac{1}{2}\sum_{i=1}^{m}\left[(n_i-1)s_i^2+n_i(y_{i\cdot}-\theta_i)^2\right]\right)}.$$

The conditional for ν is

$$\pi(\nu|\cdot) \propto \nu^{e+m/2-1}\exp\left\{-\nu\left[f+\frac{1}{2}\sum_{i=1}^{m}(\theta_i-\mu)^2\right]\right\},$$

that is

$$\boxed{\nu|\cdot} \sim \boxed{Gamma\left(e+m/2, f+\frac{1}{2}\sum_{i=1}^{m}(\theta_i-\mu)^2\right)}.$$

The conditional for θ_i is

$$\begin{aligned}\pi(\theta_i|\cdot) &\propto \exp\left\{-\frac{1}{2}\left[\nu(\theta_i-\mu)^2+\tau n_i(y_{i\cdot}-\theta_i)^2\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}(\nu+n_i\tau)\left(\theta_i-\frac{\nu\mu+n_iy_{i\cdot}\tau}{\nu+n_i\tau}\right)^2\right\},\end{aligned}$$

that is

$$\boxed{\theta_i|\cdot} \sim \boxed{N\left(\frac{\nu\mu + n_i y_{i\cdot}\tau}{\nu + n_i\tau}, \frac{1}{\nu + n_i\tau}\right)}, \quad \boxed{i = 1, \ldots m}.$$

Therefore, we have a Gibbs sampler with $m+3$ components. We just have to specify the prior parameters a, b, c, d, e, f and compute the data summaries $m, n_i, N, y_{i\cdot}, s_i^2,\ i = 1, \ldots, m$. Then, we initialise the sampler by simulating from the prior, or by starting off each component at its prior mean. The sampler is then run to convergence, and samples from the stationary distribution are used to understand the marginals of the posterior distribution. This model is of sufficient complexity that assessing convergence of the sampler to its stationary distribution is a non-trivial task. At the very least, multiple large simulation runs are required, with different starting points, and the first portion (say, a third) of any run should be discarded as "burn-in".

We have seen that the algebra can get quite tricky as we build up more and more complex models. Ultimately, we want the computer to look after this for us, as well as the actual simulation. The computer program WinBUGS does exactly this. Using the software, the model prior and data is specified. The software then works out for itself how to sample from the full conditionals. With such ease of use, some control over how the sampling is carried out is lost, but usually this is a price worth paying. Using WinBUGS, models of considerable complexity and flexibility may be built. However, as the complexity of the model increases, problems with assessment of the convergence of the sampler increase. Again, we would ideally want the computer to take care of this for us. There are many software tools available for MCMC convergence diagnostics (such as CODA), but their use is far from automatic, and beyond the scope of this course.