24

INTRODUCTION TO BAYESIAN ANALYSIS

Chapter contributed by Henrik Stryhn and William Browne

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand the basic differences between Bayesian and classical (likelihood-based or frequentist) statistical approaches.

2. Understand how to fit standard regression models with non-informative priors and Markov chain Monte Carlo (MCMC) estimation.

3. Assess whether a chain produced by an MCMC procedure appears to be well-suited for sampling from the posterior distribution (and hence MCMC inference).

4. Use a Bayesian hierarchical model for analysing clustered data and extend this modelling to incorporate more complex data structures.

5. Understand how other modelling extensions such as missing data, measurement errors, and imperfect tests can be fitted using MCMC.

6. Understand how others have used the Bayesian framework and MCMC to combine existing data and expert opinions with new data using informative prior distributions.

24.1 Introduction

The previous 4 chapters have all looked at the problem of clustering (lack of independence among observations) in a dataset. We have seen how clustering is common to many datasets that we deal with in epidemiology. There are many methodological approaches to dealing with clustering, and in this chapter we introduce a completely different approach to statistics and associated methods that are useful in the mixed model setting as well as for both the simpler non-clustered datasets and other more complex structures.

This chapter will first describe the alternative Bayesian statistics paradigm and contrast it with the classical or 'frequentist' statistics that all other work in this book has so far relied upon. We will next describe the associated Markov chain Monte Carlo methods that are generally used to fit complex Bayesian models. We will then revisit examples from the earlier chapters and show what differences the Bayesian approach leads to before moving on to the mixed models described in the previous 4 chapters. We will finish the chapter by discussing other possible model elaborations such as more complex clustering structures, missing data, and measurement error that can be easily incorporated into the Bayesian framework and some mention of the incorporation of expert opinion into statistical analysis.

24.2 Bayesian analysis

Little known outside statistical science, there exist (at least) 2 different approaches for statistical inference, which have different concepts and philosophical bases and will, in general, lead to different results. The rivalry between the classical and Bayesian schools has persisted over decades, with neither emerging as the clear winner. Many statisticians cling to the middle ground believing that each of the 2 approaches has its weaknesses and strengths which make each of them attractive in particular situations. However, many (introductory) statistics courses are taught within the non-Bayesian (classical, likelihood-based, frequentist) framework with no reference to the Bayesian view.

Bayesian analysis has gained in popularity in recent years and has, for example, been applied to complex problems in epidemiology such as risk assessment (*eg* Aven and Eidesen (2007)) or evaluation of diagnostic tests without a gold standard (*eg* Dendukuri *et al* (2009)), and to the analysis of multilevel data (*eg* Gelman and Hill (2006); Goldstein *et al* (2002a)). The scope of practical Bayesian inference has been increased widely by the invention and recent advances of a simulation-based tool for statistical inference: **Markov chain Monte Carlo** (**MCMC**) estimation (*eg* Gilks *et al* (1996)). The analysis of most complex models by the Bayesian approach is based on MCMC methods, one alternative being the recently developed INLA (integrated nested Laplace approximation) method (Rue *et al*, 2009).

We hope the reader will bear with us for the inevitable inadequacy of a one-chapter introduction to a full, new statistical approach. Our aim can only be to give little more than a superficial impression of the ideas and steps involved in a Bayesian analysis. Recent textbooks on applied Bayesian analysis in the health and biological sciences (*eg* Christensen *et al* (2010); Gelman *et al* (2004)) would be a good starting point for a more detailed discussion. Most Bayesian analyses require specialised software, and one common choice is the (free) WinBUGS programme developed by the Medical Research Council Biostatistics Unit in Cambridge (http://www.mrc-bsu.cam.ac.uk/bugs/), or subsequent developments hereof. BUGS is short for

Bayesian analysis using Gibbs sampling, which is a particular type of MCMC analysis. The analyses of this section were, however, carried out using the MLwiN software (version 2.24).

24.2.1 Bayesian paradigm

Bayesian methodology owes its name to the fundamental role that **Bayes' theorem** (see Eq 24.1) plays in it. In Bayesian reasoning, uncertainty is attributed to the parameters while the sampled data is regarded as a fixed quantity once collected. This means that all parameters are modelled by distributions. Before any data are obtained, the knowledge about the parameters of a problem is expressed in the **prior distribution** of the parameters. Given actual data, the prior distribution and the data are combined to generate the **posterior distribution** of the parameters. The posterior distribution summarises our knowledge about the parameters after observing the data. The major differences between classical and Bayesian inference are outlined in Table 24.1, and will be detailed in the sections that follow.

Table 24.1 Bayesian versus classical approaches to statistics

Concept	Classical approach	Bayesian approach
Parameter	Fixed (unknown) constant	Distribution of possible values
Prior information on parameters	None	Prior distribution
Base of inference	Likelihood function	Posterior distribution
Parameter point estimate	Estimate (eg maximum likelihood estimate (MLE))	Statistic from posterior distribution eg mean, median or mode
Parameter interval estimate	Confidence interval	Bayesian credible interval
Hypothesis testing/ Model comparison	Test (eg LRT)/criterion (eg AIC)	Bayes' factors/criterion (eg DIC)

Let us briefly indicate the way the prior and the data are merged, and denote by Y the data, by θ the parameter (vector), and

- $L(Y|\theta)$ – the likelihood function,
- $f(\theta)$ – the prior distribution for θ,
- $f(\theta|Y)$ – the posterior distribution for θ after observing data Y

where the $f(\cdot)$s are either probability functions (discrete data) or probability densities (continuous data). With these definitions, Bayes' theorem states that:

$$f(\theta|Y) = const(Y) * L(Y|\theta) * f(\theta) \qquad \textit{Eq 24.1}$$

where $const(Y)$ is a constant depending on Y but not on θ. Thus, the posterior distribution for θ is essentially constructed by multiplying together the likelihood and the prior, and is a sort of compromise between the 2. In complex models, the constant depending on Y in Eq 24.1 is virtually impossible to calculate. This means that the posterior distribution cannot be calculated analytically, and therefore alternative methods need to be used. This intractability of the posterior distribution for all but the simplest problems meant that up until the early 1990s, Bayesian statistics was more of a theoretical subject than an applied one. The increase of computer speed and memory capacity and the introduction of simulation-based methods such as MCMC have had a great impact on Bayesian analysis and its use in real-world problems.

24.2.2 Statistical analysis using the posterior distribution

Even if it might seem awkward to discuss the posterior distribution before the prior distribution, let us see a simple example of Bayesian analysis (Example 24.1) before turning to the discussion of how to choose the prior distribution. The net result of a Bayesian analysis is a **distribution**, and the analysis might, therefore, be conveniently summarised by a graph (Fig. 24.1). Point estimates and confidence intervals are not truly Bayesian in spirit, but values such as the mean, median, or mode, and intervals comprising a certain probability mass of the posterior (sometimes called **probability** or **credible intervals**) might be calculated from the posterior distribution. Both the posterior mean and median are commonly used as point values, as they can be easily calculated directly from MCMC methods. The (joint) posterior mode is also used and is evaluated by finding the parameter point estimates simulated via MCMC that have generated the largest value of the posterior distribution, and hence it is also called a maximum *a posteriori* (MAP) estimate. In the classical framework, the maximum likelihood estimate (MLE) is the maximum of the likelihood function, and so for non-informative priors (as discussed next) the mode should agree with the MLE.

24.2.3 Choice of prior distributions

Generally, it can be said that the strength and weakness of Bayesian methods lie in the prior distributions. In highly multidimensional and complex problems, it is possible to incorporate model structure by means of prior distributions; such an approach has been fruitful, for example, in image analysis. The posterior of one analysis can also be taken as the prior for a subsequent study, thereby enabling successive updates of the collected and available information, as we will discuss later. On the other hand, the choice of prior distributions might seem open to a certain arbitrariness, even if subjectivity in the prior does not contradict the Bayesian paradigm. In the past, priors have often been chosen in a particular form allowing for explicit calculation of the posterior (**conjugate priors**) but, with access to MCMC methods, these have somewhat decreased in importance though are often still used.

Let us revisit Example 24.1 to explain how conjugate priors were part of the modelling. First, a binomial likelihood for the unknown proportion was combined with a uniform prior to create a beta posterior distribution. Then we showed that this beta posterior distribution can itself be combined as a prior distribution with further (binomial) data to again produce a beta posterior distribution. A **conjugate** prior distribution by definition is a prior which, when combined with a specific likelihood, produces a posterior of the same form as the prior. In this case, the beta distribution is the conjugate prior for the proportion/probability parameter in a binomial distribution. Also, the uniform prior initially used is equivalent to a beta-distribution with parameters (1,1), which explains why a beta posterior resulted when it was used as a prior.

Other conjugate prior distributions include the normal distribution for the mean of a normal likelihood, the gamma distribution for the precision (1/variance) of a normal likelihood and again the gamma distribution for the mean of a Poisson likelihood. **Note** A conjugate prior distribution determines only the type of distribution, not its specific parameters or characteristics such as the mean and variance.

A common choice of prior (in particular among less-devoted Bayesian researchers) is a **non-informative** (flat, vague, or diffuse) prior, which gives minimal preference to any particular values for θ. As an extreme case, if we take $p(\theta) \equiv 1$ in Eq 24.1, the posterior distribution is just

Example 24.1 Bayesian analysis of proportions

Assume that we test 10 individuals for a disease or condition with a highly variable prevalence. In one scenario, 5 of the individuals tested positive; in another, 8 individuals tested positive. What information have we obtained about the disease prevalence in these 2 scenarios?

Recall that all Bayesian analyses involve a *prior* distribution, in this case for the disease prevalence P. Assume (somewhat unrealistically) that we had no particular prior information (due to the high variability of the disease) so that *a priori* all values of P would seem equally likely. Then we could choose a uniform distribution on (0,1) as our prior; this is an example of an non-informative prior (Section 24.2.3). The probability density of the uniform distribution is constant (1). The likelihood function for observing the number of positive individuals out of 10 are the probabilities of the binomial (10, P) distribution. Therefore, if we observe Y positive individuals, the posterior distribution has density:

$$f(P|Y) = const(Y) * P^{Y}(1-P)^{10-Y} * 1 = const(Y) P^{Y}(1-P)^{10-Y}$$

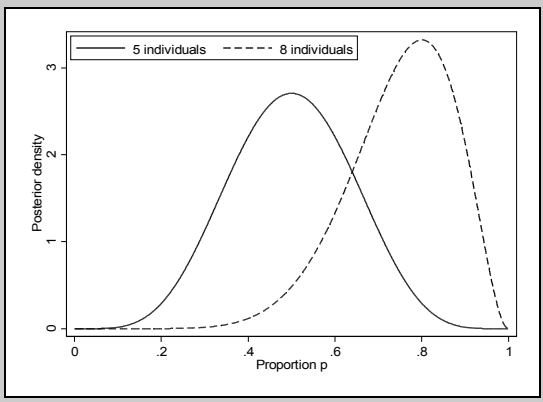

Fig. 24.1 Posterior distributions after 5 and 8 out of 10 individuals tested positive

This probability density corresponds to a beta-distribution with parameters (Y+1,10-Y+1). The constant, *const(Y)*, can be determined from Bayes' formula, but after having identified the posterior as a beta distribution, we get the constant from its density (it equals $(10+1)*\binom{10}{Y}$). Corresponding to observed values of Y=5 and Y=8, respectively, Fig. 24.1 shows beta distributions with parameters (6,6) and (9,3).

If we wanted to summarise our knowledge about P, we could use the mean, median, or mode of the distribution; for the 2 beta-distributions, they equal (0.5,0.5,0.5) and (0.75,0.764,0.8), respectively. These values can be compared with the usual estimates P=0.5 and P=0.8; the agreement of the mode and maximum likelihood estimate is no coincidence. If we wanted to summarise our knowledge about P into a 95% interval, we could choose the interval with endpoints equal to the 2.5 and 97.5 percentiles of the distribution; for the 2 beta-distributions they are (0.234,0.766) and (0.482,0.940). These intervals might be compared with the (exact) binomial confidence intervals of (0.187,0.813) and (0.444,0.975). The confidence intervals are wider than the credible intervals.

If instead we consider the 2 observations to be successive trials, then we could use the beta(6,6) distribution obtained from the first scenario as a prior for the second scenario. We then have:

$$f(P|Y_2) = const(Y_2) * P^{Y_2}(1-P)^{10-Y_2} * P^{6}(1-P)^{6} = const(Y_2) P^{Y_2+6}(1-P)^{16-Y_2}$$

With an outcome of the second trial of Y_2=8, this corresponds to a beta(14,8) distribution. We would get the same posterior if we had swapped the order of the 2 scenarios or indeed if we had considered all the data to be one dataset with 13 positive tests out of 20. This shows how Bayesian methods can be used in real time examples where data appear sequentially.

the likelihood function. So, for example, maximising the posterior (MAP estimate) yields exactly the maximum likelihood estimate. Therefore, we would by and large expect Bayesian inference with non-informative priors to be similar to likelihood-based inference. To take $p(\theta)$ constant is not always possible, but an alternative for a parameter (which can take any value) is a normal distribution with zero mean and a very large variance, effectively making values in a large interval around zero equally probable. As a technical note, it is sometimes possible to use an **improper** prior distribution, which is not a real probability distribution because it does not satisfy the condition of a finite probability of the entire sample space. The main example of an improper distribution is a constant value on an unbounded space (*eg* the constant 1 on the entire real line). Such a uniform prior can be thought of as a limiting case of normal distributions with very large variances. Despite the improper prior distribution, the posterior distribution may be perfectly well-defined, and therefore this type of uniform distribution is a popular choice for a non-informative prior. For a variance parameter, where values below zero are impossible, a standard non-informative distribution is a gamma distribution for the inverse of the variance with parameters that ensure the distribution to be concentrated close to zero (equivalent to very large variances).

24.3 Markov chain Monte Carlo estimation

Note This section uses a notation somewhat inconsistent with the rest of the book in order to stay reasonably in line with the usual notation in the field. In particular, X_1, X_2,... are not predictor variables.

Markov chains

A Markov chain (named after the Russian mathematician AA Markov) is a process (or sequence) $(X_0, X_1, X_2, \ldots)$ of random variables which satisfies the Markov property (below). The variables take values in a state space which can be either finite (*eg* $\{0,1\}$), discrete (*eg* $\{0,1,2,3\ldots\}$) or continuous (*eg* an interval, possibly infinite). The value of X_0 is the initial state of the chain, and the steps of the chain often correspond to evolution over time. The **Markov property** is a strong assumption about the probability distribution of the process (X_t):

$$\text{distribution of } (X_{t+1}, X_{t+2}, \ldots) \text{ given } (X_0, X_1, \ldots, X_t) = \text{distribution of } (X_{t+1}, X_{t+2}, \ldots) \text{ given only } (X_t) \qquad \textit{Eq 24.2}$$

In words, the future (of the process) depends on the past only through its present state. Thus, the chain has a 'short memory'. Some examples of Markov chains are processes describing games, population sizes, and queues. For example, Markov models for population size assume that the development of a population after any given time point depends only on the population's size at that time, and hence can be described solely in terms of birth, death, and migration rates. Examples of non-Markov processes are periodic phenomena and growth curves which do not have such 'short memory'. Our interest here is in **homogeneous** chains in which development does not change over time. For such chains the Markov condition (Eq 24.2) implies that whenever the chain has reached state x, it evolves from there as if it was restarted with $X_0=x$. The importance of homogeneous chains is that under some further, technical conditions they converge to limiting distributions as time runs. That is, $\text{distr}(X_t)\rightarrow\pi$ as time runs, where π is the limiting (or **stationary**) distribution (and in this case not the number 3.1415926…). This implies, for example, that $p(X_t=x)\rightarrow\pi(x)$. Example 24.2 illustrates the convergence of a simple Markov chain.

Example 24.2 Convergence of a homogeneous Markov chain

The simplest example of a homogeneous Markov chain has state space {0,1}. The states 0 and 1 could, for example, correspond to disease states (healthy/sick) or system states (busy/idle). The transitions from one state to the next are governed by a transition matrix

$$P=\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

where $p_{00}+p_{01}=1$ and $p_{10}+p_{11}=1$. For example, from state 0 the process continues to state 1 with probability p_{01} (and stays in state 0 with probability p_{00}). This chain has a stationary distribution whenever all probabilities are non-zero, and $\pi(1)= p_{01}/(p_{01}+p_{10})$. Fig. 24.2 shows the convergence of $p(X_t=1)$ from the initial state $X_0=0$ in a model with $p_{01}=0.8$ and $p_{10}=0.7$; the limiting probability of 0.5333 is reached very quickly.

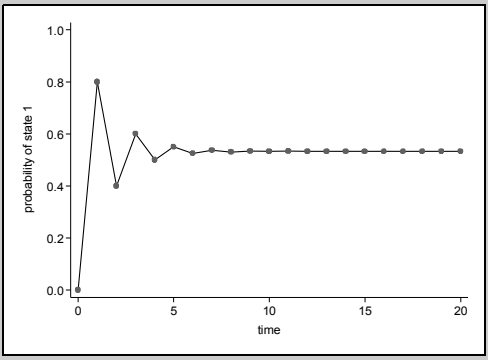

Fig. 24.2 Convergence of a Markov chain to its limiting probability distribution

24.3.1 Introduction to Markov chain Monte Carlo

The idea of MCMC estimation is simple, yet surprising. Suppose we were interested in a particular distribution π, but that quantities from this distribution were difficult to calculate because its analytical form is unknown (the distribution we have in mind is a posterior distribution from a complicated model). Suppose furthermore, that we were able to devise a Markov chain (X_t) such that distr$(X_t)\rightarrow\pi$. Then, in order to calculate statistics from π, we could run our Markov chain for a long time, for example, up to time step T (where T is large), to make the distribution of all X_t for $t \geq T$ a good approximation to π. Then in order to calculate, for example, the mean of the distribution, we could simply average over a sample of observations from the chain after time T. In a formula this would appear as:

$$\mathrm{E}(\pi)\approx \frac{1}{n}\sum_{t=T+1}^{s=T+n} X_t \qquad \textit{Eq 24.3}$$

Note that our sample from (X_t) is nothing like an independent sample (it is n successive values from a Markov chain which will be correlated). Despite the correlation, we can still use the formula to estimate E(π); however, our precision will be less than if we had an independent sample, and very much so if there is strong correlation in the chain. This precision will increase

as we run the chain for longer, and we can calculate a quantity called the Monte Carlo standard error (MCSE) which describes the uncertainty due to the simulation nature of the method. The MCSE is a function of the parameter's actual uncertainty, the correlation in the chain, and the length of the estimation sample from the chain (n). For uncorrelated chains, the MCSE is proportional to $1/\sqrt{n}$.

Other statistics as well as the mean might be computed from the limiting distribution. The initial part of the chain, $X_0,\ldots,X_T$, is called the **burn-in** period and the parameter values associated with the burn-in are discarded before summary measures are calculated (as shown in Eq 24.3).

Apparently the flaw of this idea is the necessity to construct a Markov chain with π as the limiting distribution, when we haven't even got an analytical form for π. However, this turns out to be possible for many multidimensional statistical models where π is known only up to a proportionality constant (such as $const(Y)$ in Eq 24.1). To construct a Markov chain, one needs to specify its transition mechanism (in Example 24.2 above, the transition matrix P), whereas the starting value is of minor importance. There are 2 major, general techniques for doing this: **Gibbs sampling** and **Metropolis-Hastings sampling** (technically, Gibbs sampling is a special case of Metropolis-Hastings sampling but usually is considered to be a separate method). One major practical complication involved in MCMC estimation is the length of the burn-in period, in order to make estimation from Eq 24.3 valid. Constructed Markov chains might converge rapidly or very slowly to their limiting distribution, sometimes so slowly that the chain is useless for estimation purposes. Therefore, it is crucial to have tools for monitoring the convergence and the required length of burn-in periods. The MCMC software will provide some diagnostics tools for monitoring. In the next 2 sections we will provide a brief explanation of how Gibbs and Metropolis-Hastings sampling works. Gibbs sampling can be easily applied to normal response models, whereas Metropolis-Hastings sampling can be applied more generally but might result in highly correlated and very slowly converging chains.

24.3.2 Gibbs sampling for linear and linear mixed models

The Gibbs sampling algorithm for a regression model is based on the conjugate distributions for the mean and variance parameters in a normal likelihood/model (Section 24.2.3). Let us first consider a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Here we have 3 unknown parameters: the intercept (β_0), the slope (β_1), and the residual variance (σ^2), which in a Bayesian regression all need prior distributions. We will generally choose conjugate priors, namely normal priors for the intercept and slope and an inverse gamma prior for the variance (equivalently a gamma prior for the precision). It is actually possible in this setting to derive the posterior distribution (a normal-inverse gamma distribution), however we will illustrate how we would implement the Gibbs sampling algorithm for this problem.

The full posterior distribution is $f(\beta_0, \beta_1, \sigma^2|Y)$, but in Gibbs sampling instead of sampling directly from this multivariate distribution, we sample from the series of conditional posterior distributions,

$$f(\beta_0|Y, \beta_1, \sigma^2), f(\beta_1|Y, \beta_0, \sigma^2), \text{ and } f(\sigma^2|Y, \beta_0, \beta_1)$$

In each sampling step, we use the current values for the parameters when updating other

parameters; for example, if we update β_0 in the first step, then the new value generated will be used in the subsequent steps to update β_1 and σ^2. It can be shown that sampling from these 3 distributions in turn produces (dependent) chains from the posterior distribution, and when conjugate priors are used, then the forms of the 3 conditional posterior distributions are known distributions that can easily be simulated from (2 normals and an inverse gamma). To run the Gibbs sampling algorithm requires choosing starting values for the 3 unknown parameters and then performing a burn-in as described earlier, until the chains have moved away from the starting values and are sampling from the posterior distribution.

The beauty of MCMC algorithms is that because they consist of a series of steps to update individual parameters, it is easy to fit expanded models by including additional steps and modifying existing steps. Let us expand the above model by including random effects, say corresponding to measures on individuals clustered in groups:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij}, \quad u_j \sim N(0, \sigma_u^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

We have added 2 extra sets of parameters, the cluster effects u_j and their variance σ_u^2 , and so we now have 2 additional steps to the algorithm. By expressing the cluster effects as random, we have given them a prior distribution (normal); thus, we only need to include an additional prior for σ_u^2 which we would normally give a conjugate inverse gamma prior. The existing steps will also be modified as the cluster effects need to be conditioned on. Our Gibbs sampling algorithm therefore simulates from the following distributions in turn:

$$f(\beta_0|Y, u, \beta_1, \sigma^2), f(\beta_1|Y, u, \beta_0, \sigma^2), f(u_j|Y, \beta_0, \beta_1, \sigma_u^2, \sigma^2), j = 1, \ldots, J,$$
$$f(\sigma_u^2|u_j) \text{ and } f(\sigma^2|Y, u, \beta_0, \beta_1)$$

Here we see that there is actually one step for each cluster effect (as we loop over *j*), but these are all of the same form, and there is one step for the variance. You will also note that some steps are not conditioning on all the other variables, for example the cluster variance only conditions on the cluster effects. This is because some of the variables are conditionally independent—effectively here the variance only appears in the prior distribution of the random effects and so is conditionally independent of all other parameters. All of the above distributions are either normal or inverse gamma distributions, and so are easily simulated from. As an additional simplification, we would often combine the intercept and slope into a vector (β) and update them together as this vector will have a multivariate normal conditional posterior distribution.

24.3.3 Gibbs and Metropolis-Hastings sampling for non-normal models

In the previous section, we showed how the Gibbs sampling algorithm works by constructing the conditional posterior distributions for each group of parameters and taking simulated draws from each distribution in turn. Let us consider here a different model, namely the logistic regression model for binary responses (Chapter 16):

$$p(Y_i = 1) = p_i, \quad logit(p_i) = \beta_0 + \beta_1 X_i$$

To convert this model to a Bayesian framework, we should choose priors for the unknown β-parameters. As these parameters can take values on the whole real line, a common choice is a normal prior distributions with mean 0 and a small precision (*ie* a large variance).

The conditional posterior distributions for a similar development of the Gibbs sampler as above (*eg* $f(\beta_0|Y, \beta_1)$) in this case don't equate to standard statistical distributions and so are too difficult to simulate from directly. There is a technique built on rejection sampling known as adaptive rejection (AR) sampling (Gilks and Wild, 1992), which can be used for certain non-standard distributions to circumvent the problem; the WinBUGS software has the option to use this technique in logistic regression models.

The other technique commonly used and implemented in both MLwiN and WinBUGS (as the default method in many models, including logistic regression) is Metropolis-Hastings sampling. In Metropolis-Hastings sampling we do not simulate from the conditional posterior distribution, but instead simulate from a **proposal** distribution. The simulated parameter is then either accepted or rejected, and the accept/reject rule ensures that the technique is equivalent to sampling from the correct conditional posterior distribution. Metropolis-Hastings is different from AR sampling in the way it deals with a rejected proposed value: in Metropolis-Hastings sampling, the parameter value from the last iteration is carried over, whereas for AR sampling the procedure is rerun until a value is accepted. The proposal distribution can be of almost any form, provided that all feasible parameter values can be reached in a finite number of iterations and that the proposal distribution doesn't force oscillating behaviour in the chain (known as aperiodicity).

Let's indicate how Metropolis-Hastings sampling works for a general parameter θ and its posterior distribution $p(\theta|Y)$. The proposal distribution may depend on the current value of the chain; let $q(\theta|\theta_t)$ be the proposal distribution given the current value θ_t at iteration t. If we draw (simulate) the value θ^* from $q(\theta|\theta_t)$ at iteration $(t+1)$, we accept this new value with probability

$$\alpha(\theta^*, \theta_t) = \min\left(1, \frac{p(\theta^*|Y)\,q(\theta_t|\theta^*)}{p(\theta_t|Y)\,q(\theta^*|\theta_t)}\right) \qquad \textit{Eq 24.4}$$

In practice, this means that we draw another random number from a uniform distribution on (0,1) to decide whether to accept the proposal or not: if this random number exceeds $\alpha(\theta^*,\theta_t)$, the proposal is not accepted and the chain stays put (*ie* $\theta_{t+1}=\theta_t$). The acceptance probability involves 2 ratios, the ratio of the posteriors for the proposed and current variables and the Hastings ratio, which is the ratio of probabilities of the proposed move against its reverse and accounts for non-symmetric proposals. One of the most common Metropolis-Hastings samplers is the random walk Metropolis algorithm where we use a normal proposal distribution centred around the current value and with a fixed variance. This proposal is symmetric and so the Hastings ratio in the above is not required (as it always takes value 1).

We end this brief introduction into construction of Markov chains for MCMC estimation by noting that despite all the methods described being (theoretically) 'correct', their utility for a specific model may be very different. In addition to the ease with which the chains can be simulated, the chains may also not take the same time to reach the target distribution, and may have different degrees of correlation (it is desirable to have as little correlation in the chains as possible). This raises the need for diagnostics to assess the utility of the MCMC estimates, one of the topics of the next section. Generally speaking, Metropolis-Hastings samples are easy to generate but may lead to more correlated chains, partly due to the fact that rejected proposals result in the chain not moving. Also, different algorithms may be combined for different parameters—another feature of MCMC which makes the set of MCMC techniques a very flexible framework for fitting statistical models.

24.4 Statistical analysis based on MCMC estimation

In the previous section we described in detail the algorithms that can be used within MCMC estimation. In this section we will begin by looking at how we perform an actual Bayesian analysis. Here we will answer questions such as, how long do we run our MCMC sampler for?; and how do we summarise our estimates?

24.4.1 MCMC in practice: logistic regression

In Example 24.3, we consider a logistic regression model fitted to the low birth weight dataset in Example 16.2. To translate the logistic regression model to a Bayesian framework, we have included uniform (improper) priors for all the fixed coefficients. To fit a statistical model using

Example 24.3 Fitting a logistic regression model using MCMC in MLwiN
data = bw5k

The table below presents results of the standard MCMC (in MLwiN) fitting of a logistic regression model to the low birth weight dataset. To the left, we show results after 5,000 iterations following a burn-in of 500 iterations; to the right, results after a longer run of 100,000 iterations.

Estimation	After 5,000 iterations					After 100,000 iterations				
Predictor	mean	SD	2.5%	50%	97.5%	mean	SD	2.5%	50%	97.5%
smk	0.520	0.175	0.161	0.524	0.838	0.520	0.184	0.149	0.521	0.871
white	-0.302	0.175	-0.659	-0.297	0.047	-0.321	0.182	-0.677	-0.320	0.038
frace=hisp	-0.407	0.193	-0.792	-0.407	-0.032	-0.435	0.203	-0.841	-0.434	-0.043
frace=black	0.217	0.181	-0.156	0.219	0.559	0.193	0.194	-0.193	0.194	0.575
previs	-0.056	0.015	-0.089	-0.055	-0.027	-0.059	0.014	-0.087	-0.059	-0.031
constant	-1.734	0.217	-2.150	-1.743	-1.257	-1.680	0.238	-2.143	-1.683	-1.203

The posterior means and standard deviations (for 100,000 iterations) are very close to the ML estimates and standard errors in Example 16.2. We see some change in the estimates between 5,000 and 100,000 iterations, in particular for the 3 parameters related to race and for the constant, and all posterior distributions widened, suggesting that we needed the longer run length to get accurate estimates and probability intervals. **Note** As the estimation procedure involves simulation, the actual values are always subject to random noise. The posterior means and medians are close because all distributions are fairly symmetrical (a result of the large dataset). For roughly symmetrical distributions, it makes no big difference whether one reports the posterior mean or median. The Bayesian approach does not provide P-values for tests of the individual coefficients, but one may assess their 'significance' (this term has no well-defined meaning in Bayesian statistics) by the location of the value 0 in the posterior distribution. If the distribution includes a substantial range of values both below and above zero, one could say that there is 'no evidence' against the value being zero (but it could also be negative or positive), see the parameter for -frace-=black for an example. If, on the other hand, the distribution is well above zero, and the 95% credible interval does not include zero, one could say there is evidence that the parameter is greater than zero; the parameter for -smk- illustrates this situation. As an intermediate case, the coefficient for -white- is negative but the upper bound of the credible interval slips just above zero. In classical statistics this would indicate a P-value just above 0.05, and we could choose to interpret it in the same way after Bayesian estimation, though without formal justification.

MCMC, we then first need to specify starting values for all unknown parameters. It seems natural to use the estimates from 'classical' estimation (as is done in the MLwiN software). In this case, the estimates from Example 16.2. MLwiN uses a Metropolis-Hastings algorithm for a logistic regression model, and so we also need to decide on proposal distributions for each parameter. Here, MLwiN uses scaled-up standard errors from the classical methods, and an adapting method that tunes the proposal variances to get a desired acceptance rate (*ie* the rate of Metropolis-Hastings proposals accepted) for each parameter (see Browne (2012) for more details).

As in Example 24.3, we could assess the robustness of our MCMC results to the settings of the estimation (such as the starting values, length of burn-in period, and run length) by comparing results from different scenarios. In practice, this is cumbersome and difficult to do in a systematic way, and it also provides little insight into potential problems with the chains. Instead we largely rely on **MCMC diagnostics**, a set of descriptive tools and statistics based on the actual chain for each parameter obtained in a single run. These diagnostics should allow us to detect major flaws with the chains (and therefore, with the estimates derived from them) and guide us to a suitable run length. The diagnostics offered by different software packages vary to some extent; we'll focus on the most common features as well as a few useful special features of MLwiN. **Note** There is one set of diagnostics for each parameter, and the behaviour of the chains will usually differ substantially between parameters.

Before presenting the diagnostics, let's recap the key issues to consider when running an MCMC estimation algorithm. First, we need to be sure that the start of the chain we are using for our inference has converged to the desired posterior distribution. To this end, we may need to adjust the burn-in length to throw away more iterations that may occur prior to convergence. In this example, we started from the classical (maximum likelihood) estimates which should be very close to the mode of the posterior, and hence convergence should be almost instantaneous and not an issue. In more complex models which are difficult to fit using classical methods, we cannot use 'good' starting values, and so ensuring the algorithm has burned in is important. The standard diagnostic procedure is to use multiple chains from spread out starting values to ensure that not only has the algorithm converged, but that the chains converge to the same place, and hence that the posterior is unimodal (*ie* has only one peak). The WinBUGS software offers the user the opportunity to run multiple chains and compute the modified Gelman-Rubin convergence diagnostic (Brooks and Gelman, 1998). If the diagnostic doesn't appear to converge, then by inspection of the chains we may diagnose multimodality. In this situation, increasing the run length will not help matters although, in most other cases, increasing run length should result in eventual convergence and more accurate estimates. Fortunately, in most modelling situations covered in this book, posterior multimodality would be very unusual.

The second consideration with regard to run length is that, after convergence, we should run long enough to give accurate estimates. Given the autocorrelated nature of the chains produced, the desirable run length will depend on required parameter accuracy and the magnitude of the autocorrelation: the larger the autocorrelation, the less information in the contained sample of the chain, and the larger sample size required. Example 24.4 displays the autocorrelation as well as other MCMC diagnostics for some of the chains behind the results in Example 24.3.

The diagnostic displays in Example 24.4 contain 7 panels that we will consider in turn. The **trace plot** in the upper left panel shows the whole MCMC chain that has been run. In Fig. 24.3, we can see that the chain wanders fairly slowly around the posterior and, for example, only explores very high values in a few stints (*eg* around 4,600 iterations). Fig. 24.4 is a much better looking chain where the bulk of the posterior is explored in every small subsection of the chain.

Example 24.4 MCMC diagnostics in MLwiN
data = bw5k

Figs. 24.3 and 24.4 show MCMC diagnostics for the constant (intercept) parameter of the logistic regression model of Example 24.3 after 5,000 iterations and 100,000 iterations, respectively.

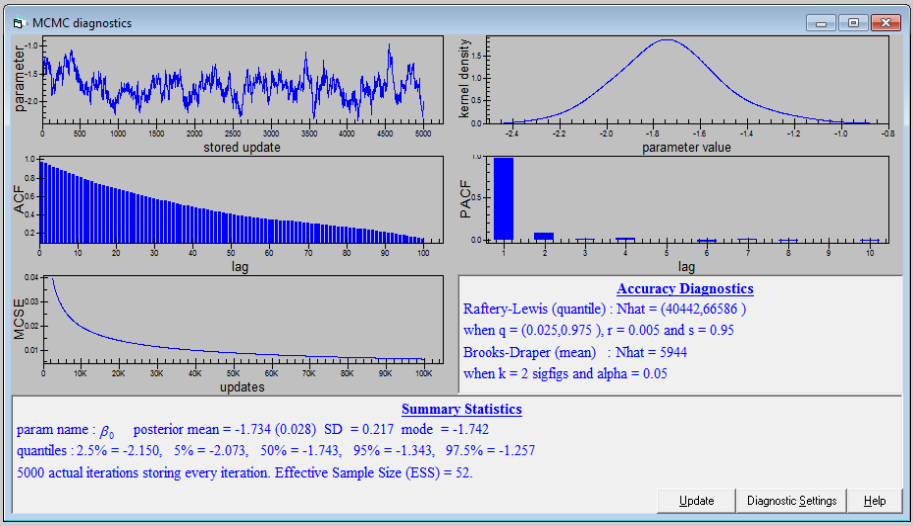

Fig. 24.3 MCMC diagnostics for logistic regression intercept after 5,000 iterations

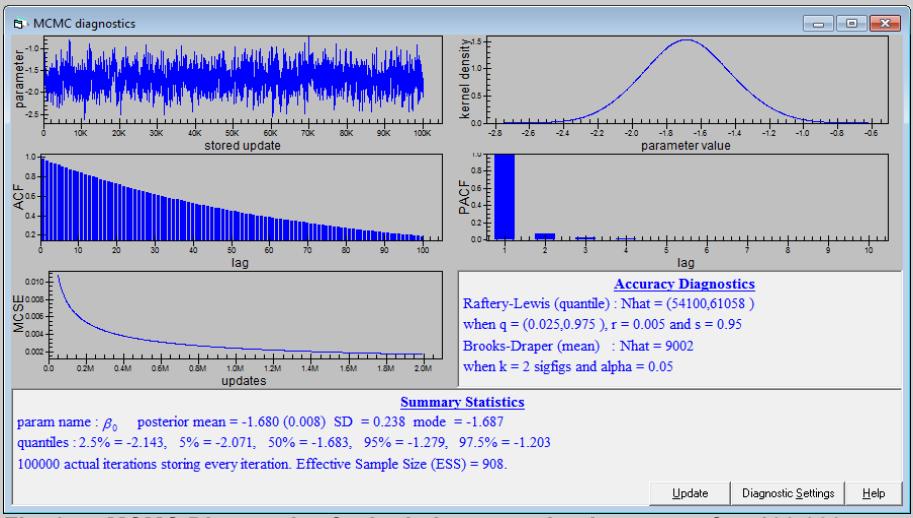

Fig. 24.4 MCMC Diagnostics for logistic regression intercept after 100,000 iterations

(*continued on next page*)

Example 24.4 (*continued*)

As the 2 figures depict different segments of the same chain, the similarity between them is no surprise. The trace plot in the upper left corner looks much more dense for the longer chain, simply because of the larger number of observations; also, the autocorrelation function (ACF) in the left middle panel is more smooth for the longer chain. The accuracy diagnostics in the lower right panel are different but of similar magnitude in the 2 figures. The suggested run lengths based on these diagnostics ('Nhat' in the listing) are greater than 5,000 and less than 100,000; that is, in the first instance the diagnostics suggest running for longer than 5,000 iterations, and in the second instance they indicate that extension beyond the actual 100,000 iterations is not necessary. See the text for explanation of the individual plots and statistics.

The upper right hand panel contains a **kernel density plot** of the posterior distribution, which is a kind of smoothed histogram and is in fact the desired summary of the posterior distribution. In both Figs 24.3 and 24.4, the plot looks roughly symmetric and bell-shaped, although Fig. 24.3 appears slightly less symmetric with a flatter peak, presumably due to not enough iterations being performed.

The next 2 panels contain the **autocorrelation function** (ACF) and partial autocorrelation functions (PACF) (these functions were introduced in Section 14.11). The ACF shows the correlation between each iteration and one that is lagged by a specified number; in particular, the ACF value at lag 1 is the estimated correlation between X_t and X_{t+1} across the chain. Ideally, the ACF values should be zero for independence but the 'ski-ramp' appearance we see is symptomatic of a poorly (or slowly) mixing chain, where 'mixing' refers to the ability of the chain to traverse all parts of the distribution. The first order autocorrelation (*ie* at lag 1) is around 0.98, and even chain values around 40 iterations apart have a correlation of 0.5. The PACF is useful mostly to confirm that the chains are truly Markovian and the behaviour we see —a large peak at lag 1 followed by virtually zero values for other lags—confirms this.

The third row panels contain **accuracy diagnostics**. The left panel shows a graph estimating the Monte Carlo standard error (Section 24.3.1) of the posterior mean estimate for various potential iterations. The MCSE is an indication of the precision of the estimated posterior mean and this panel allows users to calculate how long to run for a desired MCSE. The other diagnostics are the Raftery-Lewis (Raftery and Lewis, 1992) and Brooks-Draper diagnostics, which both aim to give a suggested run length to the user. The Raftery-Lewis diagnostic is based on estimating a particular quantile (or percentile) of the distribution with specified accuracy; Fig. 24.3 gives estimated required run lengths ('Nhat') of about 40,000 and 67,000 iterations for estimation of the 2.5% and 97.5% quantiles within 0.005 (with 95% probability). In poorly behaved chains, one sometimes encounters the paradoxical situation that increasing the run length leads to further increased required run lengths, but in Fig. 24.4, the required run lengths are well below the actual run length, so we have satisfied the diagnostic. The Brooks-Draper diagnostic instead looks at estimating the posterior mean to a given accuracy; we see that only about 10,000 iterations would appear sufficient to estimate the intercept mean with 2 correct significant digits (*ie* -1.7) with probability at least 95% (=1-0.05). Note the difference between significant digits and decimal places—it would take many more iterations to get 2 decimal places (~3 significant digits) correct for the intercept.

Most of the **summary statistics** in the final panel we have already used for Example 24.3. The mode is computed as the peak of the kernel smoothed density curve. The **effective sample size** (ESS) diagnostic provides an indirect measure of the correlation in the chains. It is defined as:

$$\text{ESS} = n/\kappa, \quad \text{with} \quad \kappa = 1 + 2\sum_{i=1}^{\infty} \rho(i), \qquad \textit{Eq 24.5}$$

where n is the number of iterations run, and $\rho(i)$ is the estimated autocorrelation for lag i. For practical calculation, the sum is approximated by stopping when a value of i is reached where $\rho(i)<0.1$. A basic interpretation of the ESS is as the number of independent samples that contains equivalent information to the dependent sample from the Markov chain. In Fig. 24.4, the 100,000 actual iterations corresponded to an ESS of only 908 samples, thus reflecting the large autocorrelation in the chain.

24.4.2 MCMC in practice: linear mixed model

Our first example was for a non-normal response model, which required us to use Metropolis sampling and hence run the chains for longer. We further illustrate the use of MCMC techniques for random effects models by the 3-level random intercept model for blood pressure of Chapter 23 (Example 23.5). Inclusion of error correlation structure is certainly possible in Bayesian analysis, but we prefer the simpler model here to avoid technical discussions of prior distributions on matrices. All prior distributions were taken as non-informative using the default values of the MLwiN software: the fixed effects parameters were modelled by uniform priors, and the 2 variances were given inverse gamma priors. Details of the estimates obtained using both Gibbs sampling and Metropolis-Hastings sampling are given in Example 24.5 to illustrate the differences between the methods.

One aspect of MCMC sampling that is really a major advantage of all simulation-based techniques, is the ability to derive posterior distributions and hence, also point and interval estimates for other derived quantities in a model. In Fig. 24.5, we consider the **variance partition coefficient** (*VPC*) introduced by Goldstein *et al* (2002b) that we have also referred to informally as the proportion of variance explained at different levels in previous chapters. Recall that a *VPC* in certain models, *eg* the 2-level random intercept model, can also be interpreted as an *ICC* (Sections 21.2.1 and 22.2.3). In the 3-level model for blood pressure, the *VPC* for subjects (patients) is the proportion of variance at the patient level, computed by the formula:

$$VPC = \sigma_p^2 / (\sigma_c^2 + \sigma_p^2 + \sigma^2) \qquad \textit{Eq 24.6}$$

When using REML estimation in previous chapters, we obtained a point estimate for the *VPC* by simply substituting the point estimates for the variances into such formulae. As MCMC is a simulation-based method, we can go one step further and employ the above formula at each iteration of the chain, thereby producing an entire new chain for the *VPC* variable. Fig. 24.5 shows the diagnostics for the *VPC* variable based on the Gibbs sampling method in Example 24.5. We note that the posterior mean estimate (0.461) is almost identical to the value obtained by directly plugging the variance posterior means into the *VPC* formula: 35.55/(4.95+35.55+36.46)=0.462. We also see that the posterior distribution for this parameter is symmetrical (**Note** VPC values close to 0 or 1 tend to have a skewed distribution), and we can get a 95% credible interval of (0.388,0.535).

24.5 Extensions of Bayesian and MCMC modelling

The examples in the previous section demonstrated that good agreement between likelihood-

Example 24.5 Bayesian MCMC analysis of blood pressure data
data = bp

Two MCMC analyses were carried out for a 3-level random intercept model for the blood pressure data. One analysis used Gibbs sampling (the recommended method for linear mixed models), the other used (univariate) Metropolis-Hastings sampling (for fixed parameters). In theory, both procedures are valid provided convergence of the chains. In the table below, we also show the linear mixed model estimates (similar to those presented in Example 23.5 for a model with autocorrelated errors).

Method	Mixed model		Bayesian mixed model and MCMC			
Option	REML estimation		Gibbs Sampling		Metropolis-Hastings	
Parameter	Estim.	SE	Estim.* (SE #)	ESS	Estim.* (SE #)	ESS
tx = Nifedipine	-1.243	0.975	-1.260 (0.983)	2.8k	-1.295 (0.980)	1.4k
tx = Atenolol	-3.001	0.965	-2.988 (0.974)	2.5k	-3.047 (0.957)	1.3k
visit = 4	-0.994	0.509	-1.000 (0.507)	19.6k	-0.990 (0.510)	6.9k
visit = 5	-2.643	0.516	-2.648 (0.515)	19.7k	-2.640 (0.516)	6.8k
visit = 6	-3.113	0.522	-3.115 (0.526)	19.9k	-3.114 (0.525)	7.0k
dbp1c	0.474	0.086	0.475 (0.087)	2.6k	0.473 (0.087)	3.8k
constant	94.69	0.89	94.62 (0.90)	1.7k	94.69 (0.92)	0.8k
centre variance	4.82	2.52	4.95 (2.88)	1.4k	5.01 (2.92)	6.4k
patient variance	35.02	3.97	35.55 (4.12)	6.3k	35.50 (4.10)	32.0k
error variance	36.34	1.82	36.46 (1.86)	10.3k	36.46 (1.83)	55.2k

*mean of posterior distribution; #standard deviation of posterior distribution; ESS=Effective Sample Size (k=1000s)

The Gibbs-sampled chain converged more rapidly and showed less correlation, so only 20,000 samples were used for estimation after a burn-in of 10,000 samples. The Metropolis-Hastings chain showed high correlation for all of the fixed parameters and therefore, estimation was extended to 100,000 samples. Overall, the agreement between the 3 sets of estimates and their standard errors (or posterior standard deviations) is good. The 2 upper level variances are estimated at slightly higher values by the Bayesian methods than REML estimation. The minor differences for treatment parameters could, with the fairly low effective sample sizes for the MCMC estimates, be due to sampling variability (the MCSEs were around 0.02). We can see that even though the actual Metropolis-Hastings runs are 5 times as long, the ESS for almost all fixed effects is less than for Gibbs sampling.

based and Bayesian estimation with non-informative priors can be achieved (without asserting this to always be the case). One additional advantage of the Bayesian approach is that the models can quite easily be extended to include, for example, non-normal random effects and further structure in the data. In this section, we will discuss several model extensions that can be handled using MCMC.

24.5.1 Cross-classified and multiple membership models

In Chapter 20, we introduced the concept of a cross-classified data structure and contrasted it with the hierarchical data structure predominantly encountered in the previous chapters. Here we describe another complex data structure and demonstrate how a Bayesian MCMC approach may help when dealing with complex data structures. We follow in part the multiple

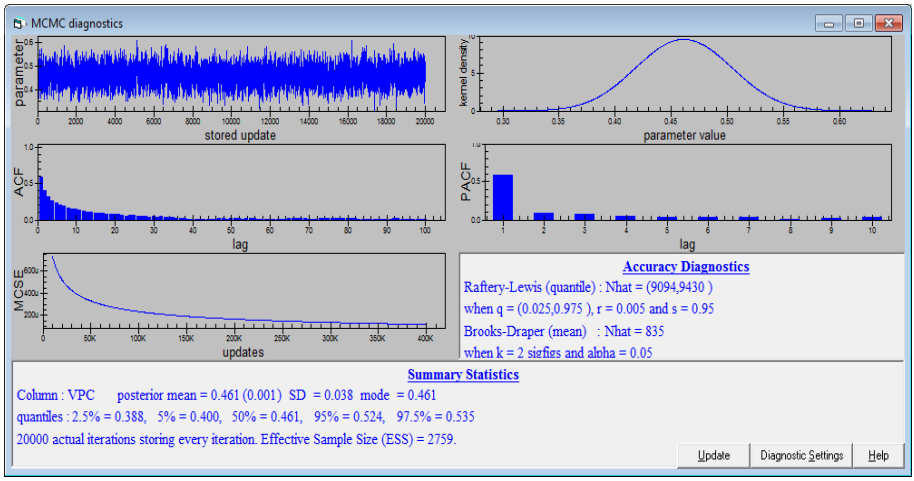

Fig. 24.5 MLwiN diagnostic plot for variance partitioning coefficient parameter from Example 24.5 and Eq 24.6

membership multiple classification (MMMC) framework of models described by Browne *et al* (2001a) and borrow an example from that paper.

Recall that a cross-classification exists when each observation (observational unit) can be included under 2 (or more) classifications that are not nested (hierarchical) within each other (Section 20.2). In addition to the health care examples discussed in Chapter 20, we could also imagine health or test performance measurements on children in secondary school, cross-classified by their primary and secondary schools. A cross-classified linear mixed model for a continuous measure would then take the form shown in Eq 21.9. In a Bayesian framework, we would typically use standard inverse gamma distributions as priors for the variance parameters. Cross-classified models can be more difficult to fit in some classical statistical algorithms that rely on the block-diagonal matrix structures, that exist in nested models, for speed. However, as MCMC algorithms consist of updating parameters in individual conditional steps, they are not affected in the same way by cross-classified structures.

The other model extension contained in MMMC models is the **multiple membership** model. Here, we remove the restriction of a one-to-one relationship between an observation and a classification unit. These structures are useful for accounting for changing group membership. For example, patients may change family doctors or children may change school over time and hence (historically) belong to several units, each of which might influence their current response. The natural way to model this is to give weightings to each clustering unit that influences the observation, with these weightings summing to 1. Such models induce a complicated correlation structure that is difficult to fit by classical procedures without relying on crude maximisation of the likelihood function (which may be numerically ineffective). We will show in Example 24.6 how to include multiple memberships (and cross-classifications) in an example from Denmark on *Salmonella* incidence in chickens.

To finish this section it should be noted that the MMMC modelling framework can also be used for modelling spatial effects (Browne (2012), Chapter 17), and that MCMC methods are particularly useful for spatial modelling (Chapter 26).

24.5.2 Missing data

We supplement our brief discussion of missing data in Section 15.5 by outlining the Bayesian approach to missing data. From an MCMC and Bayesian perspective, missing data are handled in a modelling approach where the missing data are treated as additional parameters in the model. For missing response variables, we already have a distribution for them and so they can be simulated as an extra step in the model. For missing predictor variables, an additional prior distribution is required for the missing values. The type of missing predictor variable will influence the form of the prior distribution and care has to be taken for example with categorical predictors to ensure that the prior distribution is given for the original categorical predictor, rather than the dummy variables that are actually fitted in the model. Bayesian approaches to missing data are dealt with in separate chapters in both Congdon (2007) and Gelman *et al* (2004) which give more details on how this is achieved. Bayesian MCMC methods have also proved to be particularly useful for complex modelling of informative missing data mechanisms (*eg* Carpenter *et al* (2002); Mason *et al* (2012)).

Example 24.6 *Salmonella* in Danish chickens

Browne *et al* (2001a) examined a dataset provided by Mariann Chriel, where the interest lies in the causes and sources of variability in S*almonella* outbreaks in poultry farms from 1995 to 1997. The observation level in this situation is a flock of chickens (for meat), and over the 3 years 10,127 flocks were observed. There were 2 separate levels of clustering to consider in the modelling. First, the production hierarchy in which the production flocks were nested within chicken houses (of which there were 725), which again were nested within farms (304). Second, the breeding hierarchy, in which there were 200 breeding (parent) flocks (in Denmark at that time) which produce the eggs that created the production flocks. The precise proportions of chickens that came from each parent flock (up to 6) to make up the production flock were known.

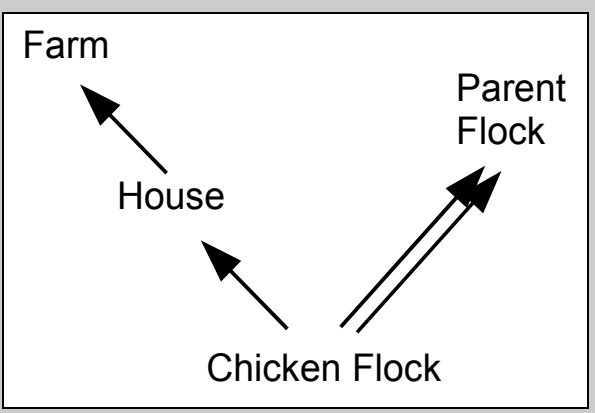

Fig. 24.6 Classification diagram for Danish chicken salmonella data

Our binary response variable indicates whether the flock had *Salmonella* isolated, and we also have 2 additional predictor variables, namely the year of the flock and the hatchery from which the flock were hatched. The model for flock i can be written as follows:

$$p_i = P(Y_i = 1), \text{ and } logit(p_i) = (X\beta)_i + \sum_{j \in \text{parent flock}(i)} w_{ij}^{(2)} u_j^{(2)} + u_{\text{house}(i)}^{(3)} + u_{\text{farm}(i)}^{(4)},$$
$$\text{with } u_j^{(2)} \sim N(0, \sigma^2_{u(2)}), u_h^{(3)} \sim N(0, \sigma^2_{u(3)}), u_f^{(4)} \sim N(0, \sigma^2_{u(4)}),$$

where $w_{ij}^{(2)}$ is the proportion of chickens in flock i originating from parent flock j, and independence of all random effects is still assumed. The associated classification diagram is shown in Fig. 24.6; here we use a double arrow to represent a multiple membership relationship.

(*continued on next page*)

Example 24.6 (*continued*)

Results of fitting this model using both Metropolis-Hastings sampling in MLwiN and adaptive rejection sampling in WinBUGS are given in the following table:

MCMC sampling	Adaptive rejection	Metropolis-Hastings
Parameter	**Estimate* (SE#)**	**Estimate* (SE#)**
constant	-2.330 (0.208)	-2.329 (0.216)
year=1996	-1.242 (0.164)	-1.238 (0.165)
year=1997	-1.163 (0.193)	-1.159 (0.194)
hatchery=2	-1.733 (0.255)	-1.730 (0.259)
hatchery=3	-0.200 (0.252)	-0.201 (0.247)
hatchery=4	-1.054 (0.380)	-1.056 (0.381)
parent flock variance $\sigma^2_{u(2)}$	0.890 (0.181)	0.884 (0.182)
house variance $\sigma^2_{u(3)}$	0.202 (0.113)	0.199 (0.112)
farm variance $\sigma^2_{u(4)}$	0.924 (0.193)	0.922 (0.203)

* mean of posterior distribution; # standard deviation of posterior distribution

Here we see good agreement between the 2 MCMC methods and the following substantive conclusions: that *Salmonella* was greater at the start of the study (1995) than in the 2 following years; that hatcheries 1 and 3 had substantially higher levels of *Salmonella* than hatcheries 2 and 4. We also see that there are large effects from the parent flocks used and from the farm on which the chickens are housed, but smaller effects for houses within farms.

24.5.3 Measurement errors and imperfect tests

Measurement error modelling was discussed in Chapter 12 and several classical approaches were mentioned there. In the Bayesian world, we would think of measurement error modelling as a missing data problem, as the true values are missing and we instead observe a value that contains errors. Browne *et al* (2001b) give an MCMC algorithm for adjusting for measurement errors in continuous predictors in a multilevel modelling situation. Their example model for a 2-level structure and a single continuous predictor (X) is given below in a simplified form (omitting the random slope for X):

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij}, \quad \text{with}$$
$$u_j \sim N(0, \sigma_u^2), \varepsilon_{ij} \sim N(0, \sigma^2) \text{ and } X^o_{ij} \sim N(X_{ij}, \sigma_m^2), X_{ij} \sim N(\theta, \phi^2)$$

Here the multilevel model is defined in terms of the true (unobserved) predictor values $X_{ij,}$ with a distribution given for the link between the observed values X^o_{ij} and the true predictor values, and a prior distribution for the latter. A simulation study showed that if the magnitude of measurement error (σ_m^2) is known, then the correct parameter estimates can be recovered. Congdon (2007) gives several other examples of the use of MCMC estimation for accounting for measurement errors.

When measurement errors occur in categorical variables we normally call them misclassifications. These misclassifications are commonly studied in epidemiology when we

consider diagnostic tests, as sensitivity and specificity are quantifiers of the proportions of the 2 forms of misclassification possible in a binary outcome variable. The aim of including misclassification in the modelling may be to estimate the diagnostic test characteristics (discussed in the next section) or to adjust a regression or mixed model for the imperfect test characteristics. McInturff *et al* (2004) reviewed the Bayesian methodology involved in a multiple logistic regression with misclassification and illustrated this with an example from human health with fairly strong priors for both misclassification rates. Kostoulas *et al* (2009) used MCMC methods to adjust estimates of the variance partition coefficient (*VPC*) when faced with an imperfect test for (animal) disease.

24.5.4 Latent class models for diagnostic test evaluation

In this section, we supplement the review of latent class models in Section 5.8 with a few comments on the Bayesian approach and add the Bayesian equivalent of the maximum likelihood analysis in Example 5.12. Bayesian methods for imperfect tests were introduced in the early 1990s when MCMC methods were still in their infancy (Johnson and Gastwirth, 1991; Joseph *et al*, 1995), and have since become the standard analytical approach within the field. As already mentioned in Chapter 5, the reason for the success of Bayesian methods lies primarily in their ability to both include prior information and tackle complex estimation problems. Test sensitivity and specificity are prime examples of parameters where one would often have access to substantial prior information from previous work within the same or a similar population or from the published literature in general. Unless one was indeed faced with a new and untested diagnostic procedure, a truly Bayesian approach would not use the uniform prior distribution (from Example 24.1) for sensitivity and specificity. It is customary to specify the prior as a beta distribution, and tools exist to determine its 2 parameters (a,b) from more intuitive characteristics of the distribution. The BetaBuster software is downloadable from the diagnostic tests from the Bayesian Epidemiologic Screening Techniques (BEST) website referenced at the end of this section, and allows specification by the mode and a percentile. An algebraic formula can give the values of (a,b) from the distribution's mean ($\mu=a/(a+b)$) and variance ($\sigma^2=\mu(1-\mu)/(a+b+1)$), but these are less intuitive to specify than the mode and a percentile. A restricted form of the beta distribution, determined only from its mode, minimum, and maximum (if these differ from 0 and 1, respectively) is often used in risk analysis, and in this context bears the name Pert distribution (*eg* Vose (2008)). Uniform prior distributions are sometimes justified by referring to the previously discussed 'correspondence' between frequentist and Bayesian analyses (with uniform priors), although devoted Bayesians will turn this around as an argument against a frequentist approach when prior information is available.

In the context of latent class models, the ability of Bayesian methods to tackle more complex estimation problems relaxes the requirement for separate 'populations' with different prevalence (Section 5.8.1), which is unnatural unless built into the sampling design, and allows for inclusion of conditional dependence between tests (Section 5.8.7). Three explanations can be offered of this increase in scope by MCMC estimation in a Bayesian framework; the most obvious one is that genuine prior distributions provide extra information on which the estimation can be based. More technically, the estimation avoids searching for the maximum of a potentially very difficult function to maximize (*eg* the likelihood function may be multimodal), and a non-uniform prior distribution usually exerts a smoothing of the target function (the posterior density) which simplifies the estimation. One word of caution: in Bayesian analysis it is required that model parameters are identifiable, so it is not true that any

model (extension) leads to a meaningful analysis by MCMC methods. Loosely stated, identifiability means that the likelihood function or posterior distribution contains the necessary information to determine the parameters of the model without ambiguity. We would usually expect identifiable 'frequentist' models to lead to identifiable parameters in a Bayesian posterior distribution based on the same likelihood, while the reverse is not true. Non-identifiability may be difficult to diagnose directly from the simulated Markov chains, and only recently has progress been made towards a better theoretical understanding of the necessary and sufficient conditions for identifiability (Jones *et al*, 2010).

We illustrate this short discussion of Bayesian latent class models by reanalysing the norovirus data of Example 5.12 with both uniform and informative prior distributions in Example 24.7. We also restate (from Chapter 5) the reference to the BEST website at University of California at Davis, which contains a wealth of information (papers and software) on Bayesian approaches to diagnostic testing problems which are beyond the scope of the present text; the web address is http://www.epi.ucdavis.edu/diagnostictests/.

24.5.5 Further examples of informative priors and expert opinion

In this section, we give a few extra examples of the use of informative priors. Green *et al* (2009) adapted ideas from human clinical practice (in part from Spiegelhalter *et al* (2004)) to a veterinary context in order to quantify how different (synthetic) data scenarios might impact the opinions about the effectiveness of a certain disease control program among veterinary clinicians. This was achieved through the use of a community of prior distributions that incorporated scepticism, enthusiasm, and uncertainty of the clinicians to the program. Comparisons of the prior and posterior distributions yielded information about how changes in opinion related to the prior beliefs held.

Informative prior distributions were also used extensively in work by Jewell *et al* (2009) involving MCMC methods to predict the progress of infectious disease epidemics, specifically an avian influenza epidemic, in livestock. Also transmission of human disease, such as influenza (Cauchemez *et al*, 2004) and gastroenteritis in hospitals (Clancy and O'Neill, 2008), has been studied by Bayesian MCMC methods. The area of epidemic modelling is an exciting and important one for epidemiologists (see also Chapter 27), and Bayesian statistical modelling is likely to play a vital role here.

24.5.6 Improving MCMC algorithms

In this chapter, we have shown how MCMC methods have revolutionised the estimation of Bayesian statistical models. We have also seen that the MCMC modelling framework is very flexible and that we can create lots of different MCMC algorithms for the same model. Browne *et al* (2009) show how, by changing the parameterisation of a model, we can improve the performance in terms of speed and chain autocorrelation, including an application of such techniques to a model for discontinuation of the use of contraception by women in Indonesia. We will (Example 24.8) illustrate here one such technique, **hierarchical centring**, to refit the model in Example 24.5.

Example 24.7 Bayesian latent class model estimation of *Se* and *Sp*
data = nv

In continuation of Example 5.12, we show estimates from a Bayesian analysis of the conditional independence latent class model for 3 tests and 2 populations (see Chapter 5 for details of the study, a full data listing and maximum likelihood estimates of the parameters). Two versions of the Bayesian model were run: (A) with all prior distributions taken as uniform on the interval (0,1), and (B) with weakly informative priors for all *Se* and *Sp* parameters. For the sole purpose of illustrating the use of informative priors, we developed a rudimentary prior 'belief' about these by studying the preceding literature cited in Fisman *et al* (2009). One particular (yet not uncommon) difficulty with the previously published studies was that they did not use the same version of the tests. Specifically, the commercial EIA kit used and real-time RT-PCR were not covered in the references, and our belief had to be based on the performance of similar tests.

All 3 tests appeared to have high *Sp* with no clear distinction between them; thus we took the same prior *Sp* for all 3 tests, assuming a mode of 0.95 and 95% certainty that the *Sp* would exceed 0.80. This specification yielded a beta(21.20,2.06) distribution in the BetaBuster software. The distribution is fairly wide and could be thought of as 'weakly informative'. Similarly we constructed weakly informative priors for the *Se* of the RT-PCR and EIA tests, from postulated modes at 0.8 and 0.7, and 95% lower bounds of 0.6 and 0.5, respectively. Finally, reports of low *Se* for EM tests existed and lead us to a mode of 0.35 and a 95% upper bound of 0.6. The analysis with either uniform or weakly informative priors were carried out using WinBUGS version 1.4 software with 5,000 burn-in samples and an estimation chain of 50,000 samples. The chains showed only little autocorrelation, and all MCMC diagnostics were satisfactory.

Median estimates (and 95% credible intervals (CrI)) for Model (A) with all priors uniform on (0,1):

Model (A)	Prevalence		PCR		EIA		EM	
	Low	High	Se	Sp	Se	Sp	Se	Sp
Estimate	0.220	0.686	0.960	0.866	0.832	0.953	0.173	0.993
Lower CrI	0.131	0.560	0.879	0.762	0.708	0.885	0.101	0.964
Upper CrI	0.334	0.794	0.998	0.961	0.939	0.994	0.268	1.000

The estimates generally agree well with the MLE (Example 5.12). No Bayesian estimates are on the boundary of the interval, and even those estimates close to the boundary have moved inwards. Credible intervals are available for all parameters (note that the upper CrI endpoint equal to 1.000 is still strictly less than 1 but listed as 1.000 after rounding off to 3 decimals).

Medians (with 95% CrI) for Model (B) with weakly informative priors:

Model (B)	Prevalence		PCR		EIA		EM	
	Low	High	Se	Sp	Se	Sp	Se	Sp
Estimate	0.239	0.710	0.935	0.899	0.796	0.958	0.182	0.986
Lower CrI	0.147	0.592	0.859	0.812	0.685	0.902	0.115	0.954
Upper CrI	0.353	0.810	0.982	0.970	0.894	0.991	0.267	0.998

In comparison with model (A) we see minor movements in the centres of the posterior distributions, always in the direction of the centre of the corresponding prior distribution. Most posterior distributions are slightly narrower with informative priors, but when the data and prior are not centred closely the posterior can also become wider (*eg* the *Sp* for EM). The main message of the results is perhaps that the 'weakly informative' priors have a real impact on the posterior distributions, and that it therefore is critically important that the prior distributions are scientifically well justified (which they weren't in the present exploration for illustrative purposes).

Example 24.8 Hierarchical centring of linear mixed model for blood pressure data
data = bp

Hierarchical centring simply means rewriting a random effects model so that the random effects are centred around any cluster level parameters or predictors in the model. The blood pressure model has 2 hierarchical levels, and by clustering at the top (centre) level we would write the model as

$$Y_{ijk} = \beta_1 X_{1ijk} + \ldots + \beta_6 X_{6ijk} + u_{jk} + v_k^* + e_{ijk}, \quad u_j \sim N(0, \sigma_u^2), \quad v_k^* \sim N(\beta_0, \sigma_v^2), \quad e_{ijk} \sim N(0, \sigma^2)$$

Here we moved the intercept β_0 from the fixed effects to the distribution of the centre random effects. Any predictors at the centre level would have been moved together with the intercept but the model did not contain any centre-level predictors; thus, everything else was left unchanged. The centred random effects, v_k^*, are not the same as the original uncentred random effects, v_k, however, by subtracting their mean we can easily move between parameterisations. The above centred parameterisation can be fitted using Gibbs sampling and will potentially give less correlated chains as there should be less correlation between the centred random effects (for centres) and the intercept. We would expect a greater impact by centring at the patient level because 3 fixed effect parameters reside at this level (2 treatment parameters and the slope for -dbp1c-). The modelling equation is rewritten in a similar way, with the relevant fixed effects moved into the mean of u_{jk}^*. In the table below, we show results for the 2 centred parameterisations which can be compared with the uncentred results in Example 24.5.

Parameterisation	Centred at centres		Centred at patients	
Parameter	Estimate* (SE#)	ESS	Estimate* (SE#)	ESS
tx = Nifedipine	-1.222 (0.970)	2.9k	-1.248 (0.978)	18.9k
tx = Atenolol	-2.999 (0.962)	2.7k	-3.009 (0.971)	18.9k
visit = 4	-0.995 (0.510)	19.3k	-0.990 (0.511)	19.7k
visit = 5	-2.644 (0.513)	19.2k	-2.639 (0.518)	19.4k
visit = 6	-3.117 (0.517)	19.4k	-3.116 (0.523)	19.5k
dbp1c	0.474 (0.086)	2.6k	0.475 (0.087)	13.8k
constant	94.66 (0.91)	4.0k	94.68 (0.91)	3.7k
centre variance	5.04 (2.88)	1.6k	4.96 (2.85)	1.5k
patient variance	35.50 (4.10)	6.4k	35.53 (4.05)	6.7k
error variance	36.45 (1.84)	11.5k	36 47(1.85)	11.0k

*mean of posterior distribution; #standard deviation of posterior distribution; ESS=Effective Sample Size (k=1000s)

The table shows good agreement between estimates of centred and uncentred (Example 24.5) Gibbs sampling. We also see that the top level centring only improved the estimation moderately for the intercept, whereas centring at the patient level improved estimation substantially for the 3 parameters at the patient level. It was expected that the patient-level centring would work better because it involved more fixed effects parameters and because a larger proportion of variance resided at that level. The other parameters, including all the variances, changed only little in terms of ESS because they were were not involved in the reparameterisation.

References

Aven T, Eidesen K. A predictive Bayesian approach to risk analysis in health care. BMC Med Res Methodol. 2007;7:38.

Brooks SP, Gelman A. Alternative methods for monitoring convergence of iterative simulations. J Comp Graph Statistics. 1998;7:434-55.

Browne WJ. Estimation in MLwiN (Version 2.25). Bristol: Centre for Multilevel Modelling, U. of Bristol; 2012.

Browne WJ, Goldstein H, Rasbash J. Multiple membership multiple classification (MMMC) models. Stat Modelling. 2001a;1:103-24.

Browne WJ, Goldstein H, Woodhouse G, Yang M. An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. Multilevel Modelling Newsletter. 2001b;13 (1):4-10.

Browne WJ, Steele F, Golalizadeh M, Green MJ. The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. J R Stat Soc A. 2009;172(3):579-98.

Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. Stat Med. 2002 Apr 30;21(8):1043-66.

Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boelle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Stat Med. 2004 Nov 30;23(22):3469-87.

Christensen R, Johnson WO, Branscum AJ, Hanson TE. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. Boca Raton: CRC Press; 2010.

Clancy, O'Neill P. Bayesian estimation of the basic reproduction number in stochastic epidemic models. Bayesian Analysis. 2008;3:737-58.

Congdon P. Bayesian Statistical Modelling, 2nd Ed. Chichester: Wiley; 2007.

Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat Med. 2009 Feb 1;28(3):441-61.

Fisman DN, Greer AL, Brouhanski G, Drews SJ. Of gastro and the gold standard: evaluation and policy implications of norovirus test performance for outbreak detection. J Transl Med. 2009;7:23.

Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis, 2nd Ed. London: Chapman and Hall; 2004.

Gelman A, Hill J. Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press; 2006.

Gilks W, Richardson S, Spiegelhalter D. Markov chain Monte Carlo in Practice. Boca Raton: Chapman & Hall / CRC; 1996.

Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. J R Stat Soc C.

1992;41:337-48.

Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. Stat Med. 2002a Nov 15;21(21):3291-315.

Goldstein H, Browne WJ, Rasbash J. Partitioning variation in multilevel models. Underst Statistics. 2002b;1:223-32.

Green MJ, Browne WJ, Green LE, Bradley AJ, Leach KA, Breen JE, *et al*. Bayesian analysis of a mastitis control plan to investigate the influence of veterinary prior beliefs on clinical interpretation. Prev Vet Med. 2009 Oct 1;91(2-4):209-17.

Jewell CP, Kypraios T, Christley RM, Roberts GO. A novel approach to real-time risk prediction for emerging infectious diseases: a case study in avian influenza (H5N1). Prev Vet Med. 2009.

Johnson WO, Gastwirth JL. Asymptotics for the Bayesian analysis of medical screening tests: application to AIDS data. J R Stat Soc B. 1991;53:427-39.

Jones G, Johnson WO, Hanson TE. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics. 2010;66:855-63.

Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141 (3):263-72.

Kostoulas P, Leontides L, Browne WJ, Gardner IA. Bayesian estimation of variance partition coefficients adjusted for imperfect test sensitivity and specificity. Prev Vet Med. 2009;89:155-162.

Mason A, Richardson S, Plewis I, Best N. Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. J Offic Stat. 2012;(pending).

McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. Stat Med. 2004 Apr 15;23(7):1095-109.

Raftery AE, Lewis SM. How many iterations in the Gibbs sampler? Bayesian Statistics. 1992;4.

Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J R Stat Soc B. 2009;71:319-92.

Spiegelhalter D, Abrams K, Myles J. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester: John Wiley & Sons; 2004.

Vose D. Risk Analysis: A Quantitative Guide, 3rd Ed: Wiley; 2008.