# 24

# INTRODUCTION TO BAYESIAN ANALYSIS

**Chapter contributed by Bill Browne and Henrik Stryhn**

## 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 veterinary 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 and also 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 in 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 2 different approaches to statistical inference, which have different concepts and philosophical bases and will, in general, lead to different results. The rivalry between the 2 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 veterinary epidemiology such as risk assessment (*eg* Ranta *et al* (2005)) or comparison of diagnostic tests without a gold standard (*eg* Branscum *et al* (2005)), and to the analysis of multilevel data (*eg* Dohoo *et al* (2001)). 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. The analysis of virtually all complex models by the Bayesian approach is based on MCMC methods.

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* Gelman *et al* (2004) would be the proper starting point. Most Bayesian analyses require specialised software, and the standard choice is the (free) WinBUGS programme developed by the Medical Research Council Biostatistics Unit in Cambridge (http://www.mrc-bsu.cam.ac.uk/bugs/). 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 2.11).

### 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 $\theta$ the parameter (vector), and

- $L(Y|\theta)$ – the likelihood function,
- $f(\theta)$ – the prior distribution for $\theta$,
- $f(\theta|Y)$ – the posterior distribution for $\theta$ 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 $\theta$. Thus, the posterior distribution for $\theta$ 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 **credibility** 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.

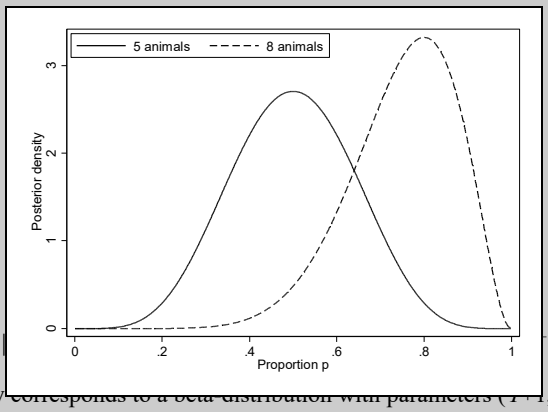
A common choice of prior (in particular among less-devoted Bayesian researchers) is a **non-informative** (flat, vague or diffuse) prior, which gives minimal preference to any particular values for $\theta$. As an extreme case, if we take $p(\theta) \equiv 1$ in Eq 24.1, the posterior distribution is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

**Example 24.1 Bayesian analysis of proportions**

Assume that we test 10 animals for a disease with highly variable prevalence. In one scenario, 5 of the animals tested positive; in another, 8 animals 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 animals out of 10 are the probabilities of the binomial (10, *P*) distribution. Therefore, if we observe *Y* positive animals, the posterior distribution has density:

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




$$f(P|$$ $$-Y$$

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 credibility 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 axis). 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 (MCMC) 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…}) 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 ($X_t$):

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

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, distr($X_t$)$\rightarrow\pi$ as time runs, where $\pi$ 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.

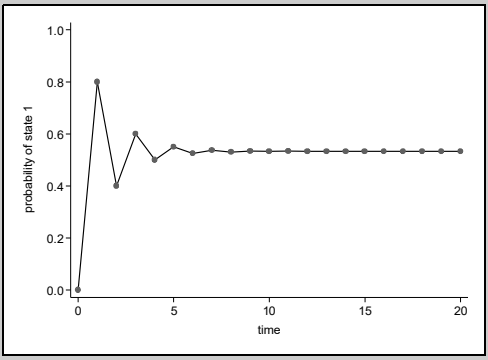



**Fig. 24.2 Convergence of a Markov chain to its limiting probability distribution**

### 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 $\pi$, 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 $\pi$, 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 $\pi$. 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($\pi$); 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 $\pi$ as the limiting distribution, when we haven't even got an analytical form for $\pi$! But this turns out to be possible for many multidimensional statistical models where $\pi$ 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 the example 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 ($\beta_0$), the slope ($\beta_1$), and the residual variance ($\sigma^2$), 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 instead sample in turn 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 not to be updated; for

example, if we update $\beta_0$ in the first step then the new value generated will be used in the subsequent steps to update $\beta_1$ and $\sigma^2$. 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 cows clustered in herds,

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

We have added 2 extra sets of parameters, the cluster effects $u_j$ and their variance $\sigma^2_u$, 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; thus, we only need to include an additional prior for $\sigma^2_u$ 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 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 ($\beta$) 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 last 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 $\beta$-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 & 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 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 $\theta$ 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 $\theta_t$ at iteration $t$. If we draw (simulate) the value $\theta^*$ 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 *Nocardia* 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 MCMC, we then firstly need to specify starting values for all unknown parameters. It seems

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

The table below presents results of the standard MCMC (in MLwiN) fitting of a logistic regression model to the *Nocardia* 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 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | mean | SD | 2.5% | 50% | 97.5% | mean | SD | 2.5% | 50% | 97.5% |
| dcpct | 0.023 | 0.008 | 0.008 | 0.023 | 0.039 | 0.023 | 0.008 | 0.008 | 0.023 | 0.039 |
| dneo | 3.015 | 0.744 | 1.608 | 3.005 | 4.538 | 2.977 | 0.738 | 1.633 | 2.943 | 4.537 |
| dclox | -1.279 | 0.613 | -2.562 | -1.264 | -0.117 | -1.315 | 0.608 | -2.536 | -1.304 | -0.151 |
| dbarn_2 | -1.574 | 0.691 | -3.012 | -1.540 | -0.313 | -1.492 | 0.681 | -2.924 | -1.460 | -0.247 |
| dbarn_3 | -0.273 | 1.260 | -2.788 | -0.287 | 2.187 | -0.214 | 1.234 | -2.634 | -0.222 | 2.231 |
| constant | -2.663 | 0.849 | -4.382 | -2.626 | -1.235 | -2.689 | 0.901 | -4.562 | -2.657 | -1.020 |

The effects of the various risk factors here are similar but slightly larger in magnitude than in Example 16.2; the 95% credible intervals from the posterior distribution are also wider than the 95% confidence intervals. We see some change in the estimates between 5,000 and 100,000 iterations, in particular for the first -dbarn- coefficient (whose distribution moved upwards) and the constant (whose distribution widened), suggesting that we needed the longer run length to get accurate estimates. **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 slight left skewness could be suspected in the distributions for the constant and -dbarn_2- parameters). 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 -dbarn_3- 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 -dneo- illustrates this situation.

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 (2009) 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 & 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 low values at around 3,700 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 = Nocardia

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. 24.1

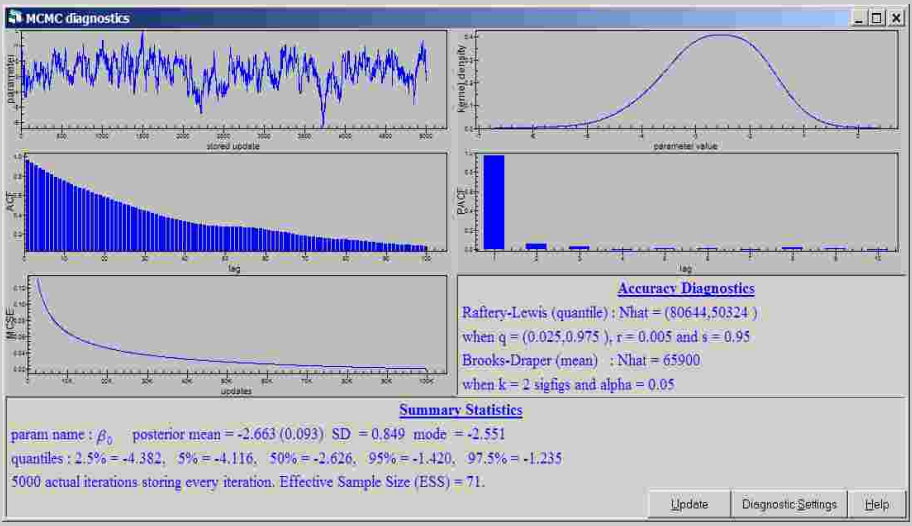



**Fig. 24.3 MCMC diagnostics for logistic regression intercept after 5,000 iterations**

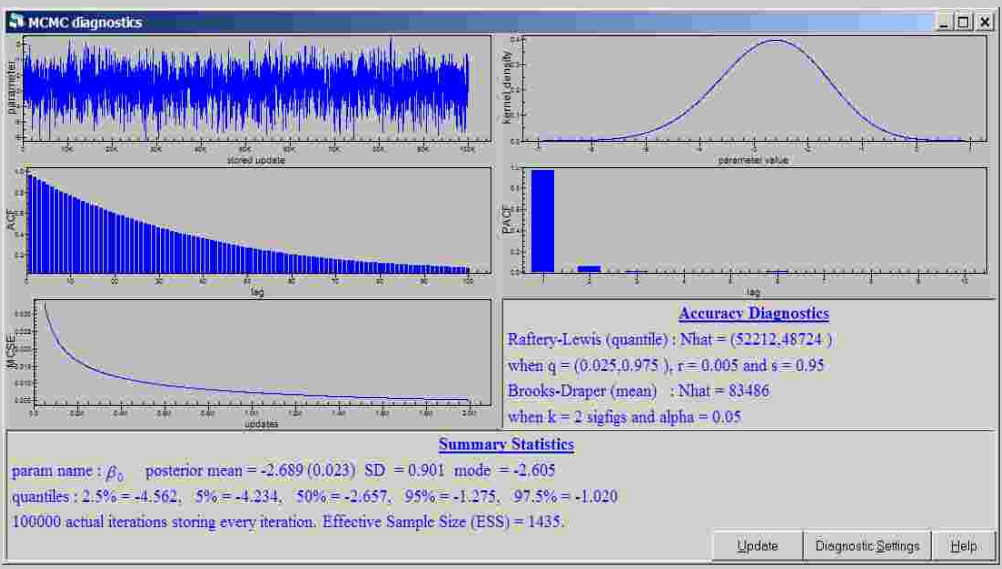



**Fig. 24.4 CMC Diagnostics for logistic regression intercept after 100,000 iterations**

(*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** (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 type 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 0.95, and even chain values around 30 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 & Lewis, (1992) 763-773) 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 80,000 and 50,000 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 100,000 iterations appears sufficient to estimate with 2 correct significant digits with probability at least 95% (=1-0.05).

Most of the **summary statistics** in the final panel we already used for Example 24.3. 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 1435 samples, thus reflecting the rather 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 2-level somatic cell count model of Chapter 21 (starting from Example 21.2). 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 the figure below, we consider the **variance partition coefficient** (*VPC*); this term was introduced by Goldstein *et al* (2002) as an alternative to the *ICC* (Section 21.2.1) used generally in this text. Recall that the *VPC* (or *ICC*) in a 2-level linear mixed model is defined as the proportion of variation present at the cluster level, and computed by the formula:

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

When using REML estimation in Chapter 21, we obtained a point estimate for the *VPC* by simply substituting the point estimates for the 2 variances in the above formula. 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 see that the posterior mean estimate (0.091) is close to the value obtained by simply plugging the variance posterior means into the *VPC* formula (0.158/(0.158+1.559)=0.092). We also see that the posterior distribution for this parameter is skewed, and can get a 95% credible interval of (0.052,0.149).

## 24.5 Extensions of Bayesian and MCMC modelling

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

**Example 24.5 Bayesian MCMC analysis of somatic cell count data**
data = scc40_2level

Two MCMC analyses were carried out using the 2-level somatic cell count dataset (the full dataset was not used to avoid the complications of repeated measures correlation structures). One analysis used Gibbs sampling (the recommended method for linear mixed models), the other used Metropolis-Hastings sampling (for fixed parameters). In theory, both procedures are valid provided convergence of the chains. In the table below, we restate for convenience also the linear mixed model estimates from Example 21.2 (without centring the 2 continuous predictors).

| Method | Mixed model | | Bayesian mixed model and MCMC | | | |
|---|---|---|---|---|---|---|
| Option | REML estimation | | Gibbs Sampling | | Metropolis-Hastings | |
| Variable | β | SE | β* (SE #) | ESS | β* (SE #) | ESS |
| hsize (in 100s) | 0.408 | 0.377 | 0.404 (0.386) | 1.9k | 0.387 (0.383) | 0.2k |
| heifer | -0.737 | 0.055 | -0.736 (0.056) | 18.8k | -0.737 (0.055) | 12.3k |
| season=spring | 0.161 | 0.091 | 0.161 (0.091) | 16.5k | 0.160 (0.091) | 2.9k |
| season=summer | 0.002 | 0.086 | 0.001 (0.087) | 18.0k | 0.000 (0.087) | 3.0k |
| season=fall | 0.001 | 0.092 | 0.002 (0.092) | 18.6k | 0.001 (0.093) | 3.7k |
| dim | 0.277 | 0.050 | 0.278 (0.050) | 17.1k | 0.278 (0.050) | 7.0k |
| constant | 4.641 | 0.197 | 4.642 (0.202) | 2.5k | 4.654 (0.202) | 0.2k |
| herd variance | 0.149 | 0.044 | 0.158 (0.048) | 8.1k | 0.158 (0.048) | 38.9k |
| error variance | 1.557 | 0.048 | 1.559 (0.048) | 18.9k | 1.559 (0.048) | 93.5k |

*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 some of the fixed parameters and therefore, estimation was extended to 100,000 samples. Overall, the agreement between the 3 sets of estimates is very good. The only noteworthy disagreements are in the herd-level parameters. The Metropolis-Hastings estimate for -hsize- is somewhat off the other 2 estimates, but the chain for this parameter was extremely highly correlated and thus, the posterior distribution not estimated well. We can see that even though the actual Metropolis-Hastings runs are 5 times as long, the ESS for all fixed effects for these methods is less than for Gibbs sampling. Also the posterior distributions for -hsize- and the constant show slightly higher standard deviations than the SEs from REML estimation.

### 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 in estimating complex data structures. We follow in part the multiple membership multiple classification (MMMC) framework of models described by Browne *et al* (2001a) and borrow an example from this 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). Crossed classifications are often seen in for example genetics examples where

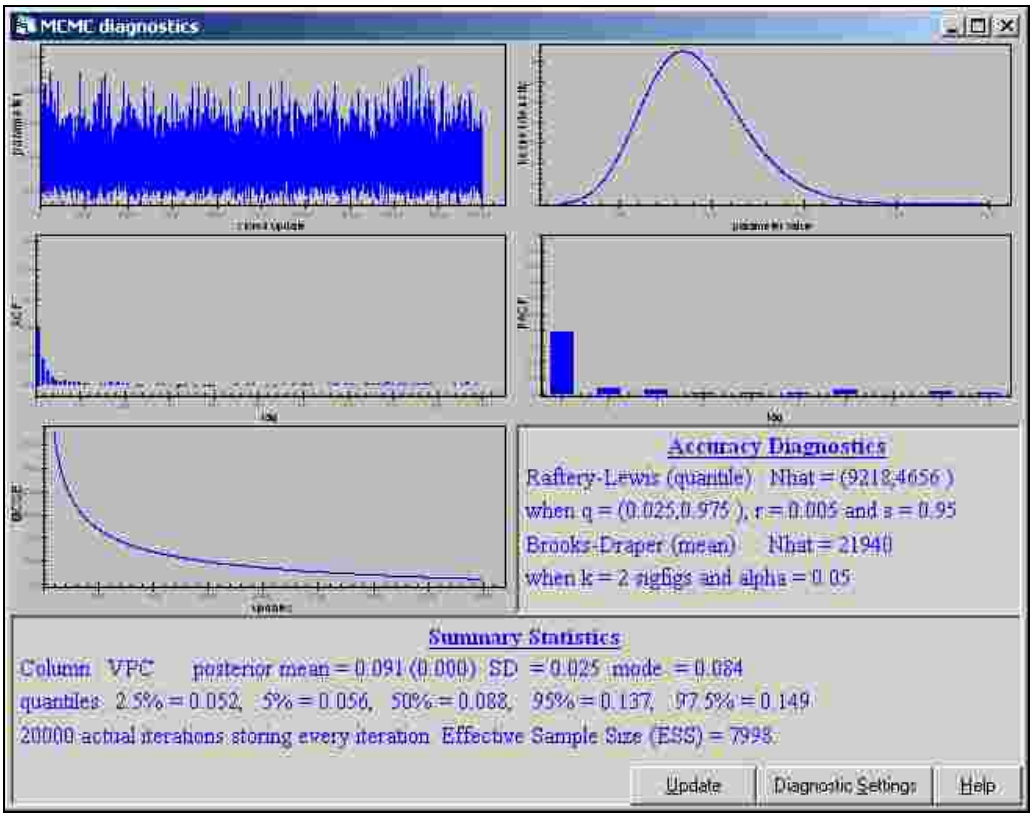



**Fig. 24.5 MLwiN diagnostic plot for variance partitioning coefficient parameter from Example 24.5**

we may have details of the sire and dam of each offspring animal. A cross-classified linear mixed model for a continuous measure (such as growth or yield) 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 blocked 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, cows may be bought and sold over time and hence (historically) belong to several herds 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 (2009), Chapter 15), 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. There is, however, very

**Example 24.6 *Salmonella* in Danish chicken**

Browne *et al* (2001a) examined a dataset kindly 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 are 2 separate levels of clustering to consider in the modelling. First, the production hierarchy in which the production flocks are nested within chicken houses (of which there are 725), which again are nested within farms (304). Second, the breeding hierarchy, in which there are 200 breeding (parent) flocks (in Denmark at that time) which produce the eggs that create the production flocks. The precise proportions of chickens that come from each parent flock (up to 6) to make up the production flock are known.

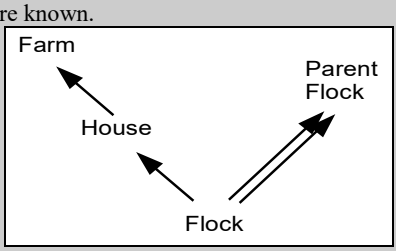



**Fig. 24.6 Classification diagram for Danish chicken salmonella**

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 are hatched. The model for flock $i$ can be written as follows:

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

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 Variable | Adaptive rejection Estimate* (SE#) | Metropolis-Hastings 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 gave a more significant increase in *Salmonella* than hatcheries 2 and 4. We also see that there are large effects from the parent flocks used and the farm on which the chickens are housed, but smaller effects for houses within farms.

little literature on using MCMC for missing data examples in veterinary epidemiology.

### 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 ($\sigma^2_m$) 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 veterinary epidemiology when we consider diagnostic tests, as effectively 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 tests 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 disease. Examples discussed include the association between sub-clinical paratuberculosis infection and fertility in sheep and goats, critical control points for *Salmonella* cross-contamination of pig carcasses in 2 Greek slaughterhouses, and factors associated with the serological prevalence of *Salmonella enterica* in Greek finishing swine herds.

### 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 & 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 '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: also 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* (2009)).

We illustrate this short discussion of Bayesian latent class models by reanalysing the ISA 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 Davis, California, 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 in veterinary science. Green *et al* (2009) considered the use of what is known as a community of prior distributions that represent a spectrum of clinical prior beliefs that incorporate scepticism, enthusiasm and uncertainty of veterinarians to the effectiveness of a specified mastitis control plan. They then combined these differing prior opinions with the observed results of the interventions prescribed in the control plan to discover via the posterior distribution how the various possible veterinary opinions would change in light of the data. They included projected financial benefits into the plan and found that a severely sceptical prior would result in a posterior belief of greater than 0.5 that the financial return would be less than £5 per cow. An enthusiast would conversely have a posterior belief of greater than 0.5 that the financial return would be greater than £20 per cow based on the data. They also considered the impact of increasing the size of the dataset on the posterior and found that with more evidence the sceptics become more convinced of the efficacy of the plan.

Informative prior distributions were also used extensively by Jewell (2009); Jewell *et al* (2009) in his work on using MCMC to predict the progress of infectious disease epidemics in livestock. In his thesis Jewell considered modelling both a potential avian influenza epidemic in the UK and the small UK foot-and-mouth epidemic in 2007. He used data from the 2001 foot-

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

In continuation of Example 5.12, we show estimates from a Bayesian analysis of the conditional independence latent class model for 3 tests and 3 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 informative priors for the specificity parameters for IFAT and VI tests. Based on previous studies on these tests for detection of ISA virus in the same population (see Nerette *et al* (2008) for details), informative beta prior distributions were constructed with the aid of BetaBuster software. The *Sp* of IFAT was given a beta(128.43,6.31) prior distribution, corresponding to the specification of a mode at 0.96 and a 5% percentile at 0.92. Similarly, a beta(458.21,1) distribution was used for the *Sp* of VI, for a mode at 1 and 1% percentile at 0.99. Both analyses were carried out using WinBUGS version 1.4.2 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 | | | IFAT | | PCR | | VI | |
|---|---|---|---|---|---|---|---|---|---|
| | Low | Med | High | Se | Sp | Se | Sp | Se | Sp |
| Estimate | 0.024 | 0.296 | 0.828 | 0.693 | 0.980 | 0.994 | 0.726 | 0.963 | 0.980 |
| Lower CrI | 0.004 | 0.212 | 0.735 | 0.599 | 0.952 | 0.968 | 0.657 | 0.886 | 0.974 |
| Upper CrI | 0.068 | 0.390 | 0.910 | 0.778 | 0.994 | 1.000 | 0.792 | 0.996 | 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 endpoints equal to 1.000 are still strictly less than 1 but listed as 1.000 after rounding off to 3 decimals).

Medians (with 95% CrI) for Model (B) with informative priors for IFAT and VI specificity:

| Model (B) | Prevalence | | | IFAT | | PCR | | VI | |
|---|---|---|---|---|---|---|---|---|---|
| | Low | Med | High | Se | Sp | Se | Sp | Se | Sp |
| Estimate | 0.026 | 0.297 | 0.828 | 0.691 | 0.971 | 0.994 | 0.728 | 0.963 | 0.999 |
| Lower CrI | 0.006 | 0.213 | 0.736 | 0.597 | 0.949 | 0.968 | 0.659 | 0.885 | 0.994 |
| Upper CrI | 0.070 | 0.393 | 0.911 | 0.775 | 0.986 | 1.000 | 0.794 | 0.996 | 1.000 |

In comparison with model (A) it is seen that the *Sp* for IFAT dropped slightly, and the *Sp* for VI increased slightly. These changes were expected as the prior distributions were centred slightly below and above, respectively, the posterior distributions based on the data alone. In addition, the CrI for these parameters were shrunk, due to the added information. All other parameters were virtually unaffected by the informative priors for the 2 *Sp* parameters.

and-mouth epidemic to give strong prior distributions for parameters in the model of the 2007 epidemic. The area of epidemic modelling is an exciting and important one for veterinary epidemiologists (see also Chapter 27), and Bayesian statistical modelling is likely to play a vital role here.

**Example 24.8 Hierarchical centring of somatic cell count data**
data = scc40_2level

Hierarchical centring simply means rewriting a random effects model so that the random effects are centred around any cluster level predictors in the model. So for the somatic cell data we write the model as

$$Y_{ij}=\beta_2 X_{2ij}+\beta_3 X_{3ij}+\ldots+\beta_6 X_{6ij}+u_j^*+e_{ij},\ \ u_j^*\sim N(\beta_0+\beta_1 X_{1j},\sigma_u^2),\ \ e_{ij}\sim N(0,\sigma^2)$$

Here we have centred the random effect in herd $j$ around the intercept $\beta_0$ and the herd size effect $\beta_1 X_{1j}$. The centred random effects $u_j^*$ are not the same as the original uncentred random effects $u_j$; 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 and the fixed effects. In the table below we compare results for the centred and uncentred (from Example 24.5) parameterisations.

| Parameterisation | Uncentred | | Centred | |
|---|---|---|---|---|
| Variable | Estimate* (SE#) | ESS | Estimate* (SE#) | ESS |
| hsize (in 100s) | 0.404 (0.386) | 1.9k | 0.407 (0.385) | 17.1k |
| heifer | -0.736 (0.056) | 18.8k | -0.736 (0.055) | 18.9k |
| season = spring | 0.161 (0.091) | 16.5k | 0.160 (0.091) | 17.6k |
| season = summer | 0.001 (0.087) | 18.0k | 0.002 (0.087) | 18.3k |
| season = fall | 0.002 (0.092) | 18.6k | 0.000 (0.093) | 18.4k |
| dim | 0.278 (0.050) | 17.1k | 0.278 (0.050) | 18.2k |
| constant | 4.642 (0.202) | 2.5k | 4.642 (0.201) | 17.5k |
| herd variance | 0.158 (0.048) | 8.1k | 0.157 (0.048) | 8.8k |
| error variance | 1.559 (0.048) | 18.9k | 1.560 (0.048) | 19.8k |

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

In the table above, we see good agreement between parameterisations and how the hierarchical centring has improved performance in terms of ESS for the intercept and herd size parameters. The other parameters are little changed in terms of ESS.

### 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 mastitis incidence in dairy cattle. We will (Example 24.8) illustrate here one such technique, **hierarchical centring**, to refit the model in Example 24.5.

## REFERENCES

Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modelling. Prev Vet Med. 2005; 68: 145-63.

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

Browne W, Goldstein H, Rasbash J. Multiple membership multiple classification (MMMC) models. Statistical modelling. 2001a; 1: 103-24.

Browne W, 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. MCMC Estimation in MlwiN. 2009 available from http://www.bristol.ac.uk.cmm.

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 Royal Statl Soc A. 2009; 172: 579-98.

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

Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island Prev Vet Med. 2001; 50: 127-44.

Gelman A, Carlin J, Stern H, Rubin D. Bayesian Data Analysis, 2nd Ed. Chapman and Hall; London. 2004.

Gilks W, Wild P. Adaptive rejection sampling for Gibbs sampling J R Stat Soc C. 1992; 41: 337-48.

Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models Understanding Statistics. 2002; 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;91(2-4):209-17.

Jewell C. Real-time Interference and Risk-Prediction for notifiable diseases in Animals. University of Warwick; Warwick. 2009.

Jewell C, Kypraios T, Christley R, Roberts G. A novel approach to real-time risk prediction for emerging infectious diseases: a case study in avian influenza (H5N1) Pre Vet Med. 2009; 91: 19-28.

Johnson W, Gastwirth J. 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:

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: 155-62.

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

Nerette P, Stryhn H, Dohoo I, Hammell K. Using pseudogold standards and latent class analysis in combination to evaluate the accuracy of three diagnostic tests. Prev Vet Med. 2008; 85: 207-25.

Raftery A, Lewis S. How many iterations in the Gibbs sampler? In J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (eds.) Bayesian Statistics 4, 763-773. Oxford: Oxford University Press; 1992.

Ranta J, Tuominen P, Maijala R. Estimation of true Salmonella prevalence jointly in cattle herd and animal populations using Bayesian hierarchical modeling Risk Anal. 2005; 25: 23-37.

Tinline RR, MacInnes CD. Ecogeographic patterns of rabies in southern Ontario based on time series analysis. J Wildl Dis. 2004; 40(2): 212-21.

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