SMSTC (2018/19)

Statistics

Lecture 20: MCMC Methods 5: Advanced Topics

Valentin Popov
School of Mathematics and Statistics
University of St Andrews

www.smstc.ac.uk

Contents

20.1 Missing data

In practice, when analysing statistical data, problems can arise as a result of missing or unobserved data. For example, in the context of credit-scoring, suppose that we express the probability of an individual repaying their loan as a function of different covariates (such as age, sex, salary). For a given individual, if a covariate value is missing, the corresponding probability is unknown.

In general a data augmentation (or auxiliary variable) approach can be applied in the following two situations:

20–1. When there is missing data. In theory the distribution of the observed data can be obtained by integrating out the missing data. However, typically this is difficult (or impossible) to do analytically.

20–2. The likelihood of the data is intractable (or difficult to calculate), but, conditional on a collection of unobserved data, the likelihood becomes tractable (or simpler and quicker to calculate).

To implement an auxiliary variable approach, we adopt the following technique, by essentially extending the standard Bayes' Theorem. Let $\boldsymbol{x}_{obs}$ denote the observed data and $\boldsymbol{x}_{mis}$ the unobserved (or missing) data. We treat the missing data (or auxiliary variables) as additional parameters to be estimated, and form the joint posterior distribution over both these auxiliary variables and model parameters $\boldsymbol{\theta}$. Bayes' Theorem states that,

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}_{mis}|\boldsymbol{x}_{obs}) \propto f(\boldsymbol{x}_{mis}, \boldsymbol{x}_{obs}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

We use a standard MCMC algorithm to sample from the joint posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{x}_{mis}|\boldsymbol{x}_{obs})$. However, we are interested in the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x}_{obs})$. This is simply the marginal posterior distribution,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_{obs}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{x}_{mis}|\boldsymbol{x}_{obs}) d\boldsymbol{x}_{mis}.$$

Thus, we simply obtain a sample from the joint posterior distribution and use the realisations of $\boldsymbol{\theta}$ to obtain posterior summary statistics of interest.

20.1.1 Example: Genetic Linkage

We illustrate the missing data approach with a simple example relating to the genetic linkage of 197 animals each allocated to one of four categories:

$$\boldsymbol{Y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

with cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

Though it is possible to sample the posterior distribution of θ directly by the Metropolis-Hastings algorithm we use a data augmentation approach that allows us to simplify the likelihood and use the Gibbs sampler. Specifically, we augment the observed data $\boldsymbol{Y}$ by dividing the first cell into two with respective cell probabilities $1/2$ and $\theta/4$, giving an augmented data set $\boldsymbol{X} = (x_1, x_2, x_3, x_4, x_5)$ where $x_1 + x_2 = y_1$, $x_3 = y_2$, etc. Now specifying the prior,

$$\theta \sim U[0, 1],$$

we have

$$\pi(\theta|\boldsymbol{Y}) \propto (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}$$

whereas

$$\pi(\theta|\boldsymbol{X}) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}.$$

In this case, writing $\boldsymbol{X} = (\boldsymbol{Y}, Z)$ for "missing" data Z (i.e., $x_2 = z$, and $x_1 = y_1 - z$) we have

$$\theta | Z, \boldsymbol{Y} \sim Beta(Z + Y_4 + 1, Y_2 + Y_3 + 1)$$

and

$$Z | \theta, \boldsymbol{Y} \sim Bin\left(125, \frac{\theta}{2+\theta}\right).$$

Thus, we can update the parameter θ and auxiliary variables Z using the Gibbs sampler, with both posterior conditional distributions of standard form.

20.1.2 Comments

20–1. In the situation where the introduction of auxiliary variables simplifies the likelihood (as in the above genetic linkage example) there is typically a computational trade-off between the simplification of the likelihood and the additional computational expense of updating the auxiliary variables. The trade-off will be problem dependent. Clearly, if the likelihood is analytically intractable without the use of auxiliary variables, the trade-off is clear!

20–2. Consider the credit-scoring problem, where the probability of an individual repaying their loan is a function of different covariates. Let $\boldsymbol{y}$ denote the binary data of whether each individual repaid the loan; $\boldsymbol{x}_{obs}$ the observed covariate values and $\boldsymbol{x}_{mis}$ the missing covariate values. We can express the likelihood in the form,

$$f(\boldsymbol{y}, \boldsymbol{x}_{mis}, \boldsymbol{x}_{obs} | \boldsymbol{\theta}) = f(\boldsymbol{y} | \boldsymbol{\theta}, \boldsymbol{x}_{obs}, \boldsymbol{x}_{mis}) f(\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} | \boldsymbol{\theta}).$$

Note that we have an additional term $f(\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} | \boldsymbol{\theta})$, which can be interpreted as the underlying model (or prior) specified for the covariate values. Thus, to emphasise, in such problems, we need to specify an underlying model for the covariate values.

For example, suppose that we have a categorical covariate $\boldsymbol{z} = (z_1, \ldots, z_n)$, where some of the z_i values are missing, taking possible values $1, 2, \ldots, k$. An obvious model would be to specify,

$$z_i \sim MN(1, \boldsymbol{p}),$$

independently for each $i = 1, \ldots, n$, where MN denotes the Multinomial distribution with probabilities $\boldsymbol{p} = (p_1, \ldots, p_k)$, such that the probability that $z_i = j$ is simply equal to p_j. Note that we treat the probabilities $\boldsymbol{p}$ as parameters to be estimated within the Bayesian analysis (so we also need to specify a prior on the parameters $\boldsymbol{p}$ and update them within the MCMC algorithm).

Alternatively, for continuous covariates, one possible model would be to specify,

$$z_i \sim N(\mu, \sigma^2),$$

where μ and σ^2 are again parameters to be estimated, and will require priors. Clearly, the underlying model specified on the covariate values will be problem dependent.

20.1.3 Example: Rats

WinBUGS[a] is able to deal with missing values. For example, see the example considered in Section 18.3, and the extension to the case where some of the weights of the rats are unobserved, as described within WinBUGS (`Help` → `Examples Vol I` → `Rats: Normal hierarchical model`). Essentially, the missing data values are replaced with "`NA`" in the appropriate input file.

[a] and OpenBUGS and JAGS

20.2 Prediction

We make a small note that the issue of predicting future events is trivial given the previous missing data approach. In particular, we have that future events are simply treated as missing data at the end of the given time-series, where the actual states of the missing data values are of interest. Since the missing data are treated as parameters and updated within each step of the MCMC algorithm, the posterior distribution of these parameters can easily be extracted from the given MCMC simulation.

20.3 Random effects

Random effects within the Bayesian framework can once more be dealt with in a straightforward manner. We considered one example already, in Section 18.3.2. Here is another. Consider the credit-scoring example, where we may specify,

$$\text{logit } p_i = \mu + \boldsymbol{\beta}^T \boldsymbol{Y}_i,$$

where $\boldsymbol{\beta}^t$ denotes the set of regression coefficients and $\boldsymbol{X}_i$ the set of covariate values for individual i. Then, the parameters to be estimated are μ and $\boldsymbol{\beta}$. However, there may be additional covariates that are not considered that influence the probability of not repaying the loan (p_i). One possible approach is to consider a random effects model to model the additional (individual) variability not explained by the covariates. For example, we may specify,

$$\text{logit } p_i = \mu + \boldsymbol{\beta}^T \boldsymbol{Y}_i + \epsilon_i$$

where,

$$\epsilon_i \sim N(0, \sigma^2).$$

Thus, the random effects allow for additional variability not explained by the covariates.

In the presence of random effects, we consider the analogous approach to the problem where we had missing covariate values. Essentially we treat the nuisance parameters, $\boldsymbol{\epsilon}$, as auxiliary variables, or parameters, to be updated within the MCMC algorithm and integrated out. Then, for notational purposes, letting $\boldsymbol{\theta} = \{\mu, \boldsymbol{\beta}, \sigma^2\}$, the joint posterior distribution can be expressed in the form,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\epsilon}, \boldsymbol{\theta}) f(\boldsymbol{\epsilon}|\boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where $\boldsymbol{x}$ denotes the observed data (binary outcome of loan repayment and corresponding covariate values for each individual in the study). The likelihood term can once more be easily evaluated when conditioning on the parameters $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\epsilon}$ imputed within the MCMC algorithm. Similarly, the underlying model for the random effects, $f(\boldsymbol{\epsilon}|\boldsymbol{\theta})$, (defined simply by the Normal distribution), can be evaluated. Thus, we can obtain a sample from this joint posterior distribution and simply marginalise to obtain the posterior distribution of interest.

We note that within a Bayesian framework, the difference between a fixed effect model and a random effects model is simply the prior specification. In particular a random effects model can be regarded as simply a hierarchical prior (see Section 16.5.3).

20.4 Model selection

Often, there may be a number of different models that can be fitted to the data. For example, in a variable selection context, the parameter of interest (such as probability of repaying a loan or survival rate) may be written as a function of different covariates. The estimates of the parameters may be dependent on the model specified (in terms of the covariates present in the model). Alternatively, the underlying model itself may be of interest, providing information regarding the covariates that influence the given system under study.

In the Bayesian framework the information criteria described above no longer have a fixed value, but a distribution as the parameters, $\boldsymbol{\theta}$, has a distribution. An information criterion (the Deviance Information Criterion) has been suggested that takes into account that $\boldsymbol{\theta}$ has a distribution, rather than a fixed value. This approach is aimed to be analogous to the AIC statistics and considers a trade-off between the complexity of the model and the corresponding fit of the model to the data. A very different approach, and as we shall see a natural extension to the construction of the posterior distribution of interest using Bayes' Theorem, is to consider the model itself to be a parameter within the analysis. This allows us to calculate posterior model probabilities and quantitatively discriminate between competing models, in an easily interpretable manner. We discuss each of these methods in turn.

20.4.1 Bayesian information criteria

The AIC (and BIC, etc.) are defined "classically", where the likelihood is evaluated at the MLE of the parameters. An alternative information criterion, is the Deviance Information Criterion (DIC; Spiegelhalter *et al*, 2002). This discriminates between models by considering both the fit and complexity of the model, but was developed for use within the Bayesian framework. The fit of model m is evaluated in terms of the deviance, while the complexity of the model is defined in terms of the "effective number of parameters", denoted by $p_D(m)$. If we let $\bar{\boldsymbol{\theta}} = \mathbb{E}_\pi(\boldsymbol{\theta}|\text{data})$, i.e., $\bar{\boldsymbol{\theta}}$ denote the posterior mean of the parameters $\boldsymbol{\theta}$, then we can express this effective number of parameters for model m in the form,

$$p_D(m) = -2\mathbb{E}_\pi(\log f_m(\boldsymbol{x}|\boldsymbol{\theta})) + 2\log f_m(\boldsymbol{x}|\bar{\boldsymbol{\theta}}),$$

where the expectation is once more with respect to the posterior distribution. Thus, the effective number of parameters is defined to be the difference between the posterior expectation of the deviance and the deviance of the posterior mean of the parameters.

The DIC for model m can then be expressed in the form,

$$DIC_m \quad = \quad -2\mathbb{E}_\pi(\log f_m(\boldsymbol{x}|\boldsymbol{\theta})) + p_D(m),$$

where, once more, the expectation is with respect to the posterior distribution. The DIC can easily be calculated within any MCMC procedure with practically no additional computational expense and can be calculated within WinBUGS, OpenBUGS and JAGS. However, we need to individually fit each possible model to the data, which may be infeasible if there are a larger number of possible models. Also, the DIC does not provide a quantitative comparison between competing models that is readily interpretable. Despite having been around for over 10 years now, DIC is still somewhat controversial (see Discussion in Spiegelhalter *et al*, 2002, and Spiegelhalter *et al*, 2014).

Example: Rats

We return to the rats example. We let $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$. We consider 4 different models for the mean:

$$\begin{aligned}
\text{Model } m_1\text{:} \qquad \mu_{ij} &= \alpha; \\
\text{Model } m_2\text{:} \qquad \mu_{ij} &= \alpha + \beta_1 z_{1t}; \\
\text{Model } m_3\text{:} \qquad \mu_{ij} &= \alpha + \beta_2 z_{2t}; \\
\text{Model } m_4\text{:} \qquad \mu_{ij} &= \alpha + \beta_1 z_{1t} + \beta_2 z_{2t}.
\end{aligned}$$

where $\epsilon_t \sim N(0, \sigma^2)$ and z_{1t} denotes the centred time at time t (i.e. raw time minus mean time) and z_{2t} denotes the centred squared time (i.e. squared time minus mean squared time). Thus, these models correspond to the constant model (model m_1), linear regression on time (model m_2), linear regression on the time squared (model m_3) and quadratic regression on time (model m_4).

Note that in such polynomial regression model m_3 is often not considered (i.e., which is essentially a quadratic regression without a linear component). For each parameter, if it is present in the model we specify independent priors,

$$\alpha \sim N(0, 10^5); \qquad \beta_1 \sim N(0, 10^5); \qquad \beta_2 \sim N(0, 10^5); \qquad \sigma^2 \sim \Gamma^{-1}(0.001, 0.001).$$

We initially run the saturated model (i.e., model m_4) for 30,000 iterations with a conservative burn-in of 1,000 iterations and obtain the following summary statistics:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	245.6	2.024	0.01129	241.6	245.6	249.5	1001	29000
beta	6.184	0.12	8.358E-4	5.924	6.184	6.445	1001	29000
beta2	-0.03398	0.0159	8.669E-5	-0.06471	-0.0314	-0.002674	1001	29000
sigma2	15.95	0.9412	0.005797	14.24	15.9	17.92	1001	29000

The parameter names correspond to `alpha` $= \alpha$; `beta` $= \beta_1$; `beta2` $= \beta_2$; $\sigma 2 = \sigma^2$. Independent replications provided essentially identical results and the convergence diagnostics did not suggest any lack of convergence. By looking at these posterior estimates it once more appears that there is a strong linear relationship with a 95% symmetric credible interval of $(5.924, 6.445)$ for β_1, which clearly does not contain the value of 0. The 95% symmetric credible interval for β_2 is $(-0.065, -0.003)$. Although this does not contain the value of 0, it is very close, so that it is unclear whether we would prefer model m_4 to model m_2. We need to perform a more rigorous model discrimination technique - see Exercise 20-2 for performing model selection via the DIC statistic.

As we said before, when applying the DIC, we need to fit each possible model of interest to the data, which may be infeasible if there are a large number of possible models to be considered; in addition, the DIC does not provide a quantitative comparison between competing models that is readily interpretable. Within the general Bayesian framework, there is a more natural way of dealing with the issue of model discrimination via posterior model probabilities, and a simple extension to Bayes Theorem to take into account additional model uncertainty. Essentially, this alternative approach considers the model itself as a parameter, and allows us to quantitatively compare competing models, and to incorporate any information concerning the relative *a priori* probabilities of the different models. We concentrate on this latter approach.

20.4.2 Posterior model probabilities

Within the Bayesian framework we can consider the model itself to be a discrete parameter to be estimated. The set of possible values that this parameter can take is simply the set of possible models. By applying Bayes' theorem, we can form the joint posterior distribution over both the model and the parameters. Formally, we have,

$$\pi(\boldsymbol{\theta}_m, m|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m)p(m),$$

where $f(\boldsymbol{x}|\boldsymbol{\theta}_m, m)$ denotes the likelihood of the data given model m, and the corresponding parameter values; $p(\boldsymbol{\theta}_m|m)$ the prior for the parameters in model m; and $p(m)$ the prior probability for model m. Thus, since we are treating the model as a parameter to be estimated, we need to specify the corresponding prior probability mass function for each possible model. Note that we use a subscript on the parameters $\boldsymbol{\theta}_m$ to denote the set of parameters in model m. Typically a parameter may be present in more than one model, for example, a regression coefficient corresponding to a given covariate. In general, we can specify a different set of prior distributions for the parameters in each of the different models.

Of particular interest within this model uncertainty framework is the posterior marginal distribution of the different possible models. In other words, the updated support for each possible

model, following the data being observed. Suppose that the set of plausible models is denoted by $\boldsymbol{m} = (m_1, \cdots, m_k)$. The corresponding posterior model probability for model m_i is given by,

$$\pi(m_i|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|m_i)p(m_i)}{\sum_{i=1}^{k} f(\boldsymbol{x}|m_i)p(m_i)}, \tag{20.1}$$

where,

$$f(\boldsymbol{x}|m_i) = \int f(\boldsymbol{x}|m_i, \boldsymbol{\theta}_{m_i})p(\boldsymbol{\theta}_{m_i}|m_i)d\boldsymbol{\theta}. \tag{20.2}$$

This posterior (marginal) distribution of the models quantitatively discriminates between different models by calculating the corresponding posterior model probability of each model. A related (and equivalent) statistic relating to the posterior model probabilities of different models are their corresponding Bayes Factors.

Bayes Factors

Bayes factors can be used to compare two opposing models, given the data observed. Suppose that we are comparing two models, labelled m_1 and m_2. Then, the Bayes factor of m_1 against m_2 can be expressed as,

$$B_{12} = \frac{\pi(m_1|\boldsymbol{x})/\pi(m_2|\boldsymbol{x})}{p(m_1)/p(m_2)},$$

where $p(\cdot)$ denotes the prior for the given model; and $\pi(\cdot|\text{data})$ the posterior probability of the model, given the data. Thus the Bayes factor can be interpreted as the ratio of the posterior odds to its prior odds of model m_1 to model m_2. Using Bayes theorem, (given in equation (20.1)) we can also express the Bayes factor in the form,

$$B_{12} \quad = \quad \frac{f(\boldsymbol{x}|m_1)}{f(\boldsymbol{x}|m_2)},$$

where $f(\boldsymbol{x}|m)$ denotes the likelihood of the data under model m.
Kass and Raftery (1995) give the following "rule of thumb" guide for interpreting Bayes Factors:

Bayes Factor	Interpretation
< 3	Not worth mentioning
$3 - 20$	Positive evidence of model m_1 to m_2
$20 - 150$	Strong evidence of model m_1 to m_2
> 150	Very strong evidence of model m_1 to m_2

20.4.3 Model averaging

Bayesian model-averaging obtains an estimate of a parameter, based on all plausible models, taking into account both parameter and model uncertainty. Thus, model-averaging can overcome the problem associated with different plausible models giving very different parameter estimates. Formally, the model-averaged posterior distribution of some parameter vector $\boldsymbol{\theta}$ common to all models m_i, $i = 1, \cdots, K$ is given by,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{K} \pi(\boldsymbol{\theta}|\boldsymbol{x}, m_i)\pi(m_i|\boldsymbol{x}).$$

Essentially, the model-averaging approach obtains a single parameter estimate based on all plausible models by weighting each according to their corresponding posterior model probability.

Comments

20–1. Parameters should only be averaged over different models when the interpretation of the parameter in the different models is the same. For example, suppose that we specify the survival rate as a function of a number of covariates, where there is prior uncertainty relating to the covariates present in the model. The survival rate has the same interpretation within each possible model (although the set of covariates that it is a function of may vary), and hence a model-averaged estimate of the survival rate takes into account both parameter and model uncertainty, with respect to covariate dependence. Now consider the regression coefficient of one of the possible covariates. The interpretation of this coefficient is different, dependent on whether the covariate is present in the model or not. If the covariate is present in the model, then the regression coefficient describes how the survival rate is related to the given covariate; however, if the covariate is NOT present in the model, then there is no relationship between the survival rate and the covariate value (the regression coefficient is identically equal to zero). Thus, it does not make sense to provide a model-averaged estimate of the regression coefficient over all models; although it does make sense to provide a model-averaged estimate of the regression coefficient, conditional on the covariate being present in the model.

20–2. Model-averaging should always be performed with some thought (i.e., common sense). For example consider the (extreme but highly demonstrative) case where there are only two plausible models, each having (approximately) equal posterior support. Suppose that the corresponding posterior distribution of the parameter θ, common to both models has a bi-modal distribution, with each mode corresponding to each possible model. Using the posterior mean of the parameter, conditional on the model gives a good representation of the location of the parameter within the individual models. However, taking the posterior model-averaged mean of θ will typically not be a good summary description of the posterior distribution of θ. A better description would be the marginal density or at least the HPDI credible interval if a summary statistic is required. In this example, it is the use of the summary statistic that is at fault, rather than the idea of model-averaging. This is nothing new, since the problem can occur within any Bayesian analysis when any parameter has a non-unimodal distribution. However, multi-modal posterior distributions of parameters are more common within model-averaging, since it is possible that the posterior distribution of the parameter may differ between the models with reasonable posterior support.

20.4.4 Estimating posterior model probabilities

We can write down an expression for the posterior model probabilities, in the form of an integral using (20.1) and (20.2). However, the integration over the parameter space in (20.1) is, in general, analytically intractable. There are a number of different methods that have been proposed for estimating the posterior model probabilities. There are two main approaches: (i) Monte Carlo estimates; and (ii) MCMC-type estimates. (We briefly point to some other alternatives at the end of this subsection.)

(i) Monte Carlo estimates

Monte Carlo estimates involve obtaining an estimate of equation (20.2), for the different possible models, which can then be used to combine with the prior model probability to obtain an estimate of the posterior model probabilities, given in (20.1). We can express (20.2) in the form,

$$f(\boldsymbol{x}|m_i) = \mathbb{E}_p[f_{m_i}(\boldsymbol{x}|\boldsymbol{\theta})],$$

i.e., the expectation of the likelihood with respect to the prior distribution $p(\boldsymbol{\theta})$. (Note that we drop the subscript m_i notation for the set of parameters $\boldsymbol{\theta}$ for notational convenience since it is

clear from the context that the set of parameters refers to those for model m_i.) We can use the Monte Carlo estimate,

$$\widehat{f}(\boldsymbol{x}|m_i) = \frac{1}{K}\sum_{j=1}^{K} f(\boldsymbol{x}|\boldsymbol{\theta}_j).$$

Then, we can estimate the posterior probability of model m_i as,

$$\pi(m_i|\boldsymbol{x}) \propto p(m_i)\widehat{f}(\boldsymbol{x}|m_i).$$

We repeat this process for each model $m = m_1, \ldots, m_k$ and renormalise the estimates obtained for each model to obtain an estimate of the corresponding posterior model probabilities.

However, this estimate of the likelihood is generally inefficient and very unstable. An alternative Monte Carlo estimate for the likelihood can be derived using importance sampling, where we use a sample from an alternative distribution and then reweight the estimate. An obvious choice for the importance sampling distribution is the posterior distribution, which we can sample from using the MCMC algorithm, to obtain the parameters $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$ from the posterior distribution. The estimate of the likelihood in (20.2) simplifies to the harmonic mean, i.e.

$$\widehat{f}(\boldsymbol{x}|m_i) = \left[\frac{1}{K}\sum_{j=1}^{K}\frac{1}{f(\boldsymbol{x}|\boldsymbol{\theta}_j)}\right]^{-1}.$$

However, this estimate is very sensitive to small likelihood values and can once again be very unstable. This estimate actually has an infinite asymptotic variance. There are various other (typically more complex) Markov chain estimation procedures and for further details of these estimation procedures see, for example, Kass and Raftery (1995) and Gamerman (1997) Ch. 7.2.

The Monte Carlo estimates are the easiest to program and conceptualise. However, they are often very inefficient and do not always converge within a feasible number of iterations. In addition, each individual model needs to be considered in turn, which may be infeasible when there are a large number of models. We consider an alternative, more general, approach, which allows us to feasibly calculate the posterior model probabilities of all possible models.

(ii) Reversible Jump MCMC

We have a posterior distribution (over parameter and model space) defined up to proportionality:

$$\pi(\boldsymbol{\theta}_m, m|\text{ data}) \propto L(\text{data }|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m)p(m).$$

If we can sample from this posterior distribution then we are able to obtain posterior estimates of summary statistics of interest. We use an MCMC-type approach. However, different models, generally have a different number of parameters, and so moving between models can involve a change in the dimension. As a result we cannot use the Metropolis-Hastings algorithm (see Lecture 17), since this is only defined for moves that do not alter the dimension of the Markov chain, i.e. for moves that do not alter the number of parameters of the Markov chain. Thus we need to consider an alternative algorithm.

The reversible jump (RJ) MCMC algorithm (Green, 1995) is an extension of the Metropolis-Hastings algorithm, which allows us to construct a Markov chain with stationary distribution equal to the joint posterior distribution of both models and parameters, $\pi(m, \boldsymbol{\theta}_m|\boldsymbol{x})$. Note that we are able to traverse the posterior model space, in terms of the models and corresponding parameters, in a single Markov chain. Within each iteration of the Markov chain, the algorithm involves two steps:

1. Update the parameters, $\boldsymbol{\theta}_m$, conditional on the model using the Metropolis-Hastings algorithm; and

2. Update the model, m, conditional on the current parameter values using the RJMCMC algorithm described below.

The posterior model probabilities can be simply estimated as the proportion of time that the constructed Markov chain is in any given model, effectively integrating out over the parameters in the model, $\boldsymbol{\theta}_m$ (as required for equation (20.2)). Note that the standard MCMC issues apply such as using an appropriate burn-in for the Markov chain, considering prior sensitivity analyses etc. We now describe the reversible jump step for updating the model within the Markov chain in more detail.

RJMCMC algorithm

The RJ step of the algorithm involves two steps:

20–1. Proposing to move to a different model with some given parameter values;

20–2. Accepting this proposed move with some probability.

Thus, the general structure of the RJ step is the same as for the Metropolis-Hastings algorithm. We begin with the first step. Suppose that at iteration k the Markov chain is in model m with parameters $(\boldsymbol{\theta}, m)_k$ and that we propose to move to state $(\boldsymbol{\theta}', m')$. To do this, we define a bijective function g, such that $(\boldsymbol{\theta}', \boldsymbol{u}') = g(\boldsymbol{\theta}, \boldsymbol{u})$, where $\boldsymbol{u}$ and $\boldsymbol{u}'$ are sets of random variables with density function $q(\boldsymbol{u})$ and $q'(\boldsymbol{u}')$, respectively. Note that by definition of the bijective function $|\boldsymbol{\theta} \cup \boldsymbol{u}| = |\boldsymbol{\theta}' \cup \boldsymbol{u}'|$. We accept this proposed move with probability $\min(1, A)$, where,

$$A = \frac{\pi(\boldsymbol{\theta}', m'|\boldsymbol{x})P(m|m')q'(\boldsymbol{u}')}{\pi(\theta, m|\boldsymbol{x})P(m'|m)q(\boldsymbol{u})} \left| \frac{\partial(\boldsymbol{\theta}', \boldsymbol{u}')}{\partial(\boldsymbol{\theta}, \boldsymbol{u})} \right|,$$

where $\boldsymbol{x}$ denotes the data; and $P(m'|m)$ the probability of proposing to move from model m to model m' and the last term is the absolute value of the determinant of the Jacobian matrix, necessary because we are changing variables from $(\boldsymbol{\theta}, \boldsymbol{u})$ to $(\boldsymbol{\theta}', \boldsymbol{u}')$. If the move is accepted, we set $(\boldsymbol{\theta}, m)_{k+1} = (\boldsymbol{\theta}', m')$; else we set $(\boldsymbol{\theta}, m)_{k+1} = (\boldsymbol{\theta}, m)_k$. Note that the formula given is not as complicated as it appears!

Example 1: Nested models

Suppose that a response variable may be linearly regressed on an explanatory variable. We specify,

$$Y_i \overset{iid}{\sim} N(\mu_i, \sigma^2).$$

There are two possible models for μ_i:

$$\begin{array}{lrcl} \text{Model } m_1\text{:} & \mu_i & = & \beta_0; \\ \text{Model } m_2\text{:} & \mu_i & = & \beta_0 + \beta_1 x_i. \end{array}$$

The parameters β_0 and β_1 (for model m_2) are to be estimated and x_i denotes some explanatory variable. For each step of the MCMC algorithm we firstly update each parameter, conditional on the model; then, we update the model, conditional on the current parameter values of the Markov chain, using the RJ algorithm. We only describe the updating of the model here and consider each step of the RJ step in turn:

Step 1: Propose new model

Suppose that at iteration k, the Markov chain is in state $(\boldsymbol{\theta}, m)_k$, where $\boldsymbol{\theta} = (\beta_0)$ and $m = m_1$. Then, since there are only two possible models we always propose to move to the alternative

model. We propose to move to model $m' = m_2$, with parameter values $\boldsymbol{\theta}' = (\beta_0', \beta_1')$. We set,

$$\begin{aligned} \beta_0' &= \beta_0 \\ \beta_1' &= u, \end{aligned}$$

where $u \sim q(u)$, i.e. we simulate a value of u from some (arbitrary) proposal distribution q which has the same (or larger) support as β_1.

Step 2: Accept/reject step

We accept the proposed model move to state $(\boldsymbol{\theta}', m')$ and set $(\boldsymbol{\theta}, m)_{k+1} = (\boldsymbol{\theta}', m')$ with probability $\min(1, A)$, where,

$$A = \frac{\pi(\boldsymbol{\theta}', m'|\boldsymbol{x})P(m|m')}{\pi(\boldsymbol{\theta}, m|\boldsymbol{x})P(m'|m)q(u)} \left|\frac{\partial(\beta_0', \beta_1')}{\partial(\beta_0, u)}\right|,$$

where $P(m'|m)$ denotes the probability of proposing to move to model m', given the current state of the chain is m, and vice versa (which in this case is always equal to one); and the final expression denotes the Jacobian and is given by,

$$\begin{aligned} \left|\frac{\partial(\beta_0', \beta_1')}{\partial(\beta_0, u)}\right| &= \begin{vmatrix} \frac{\partial\beta_0'}{\partial\beta_0} & \frac{\partial\beta_1'}{\partial\beta_0} \\ \frac{\partial\beta_0'}{\partial u} & \frac{\partial\beta_1'}{\partial u} \end{vmatrix} \\ &= \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1. \end{aligned}$$

Else, if we reject the proposed move, we set, $(\boldsymbol{\theta}, m)_{k+1} = (\boldsymbol{\theta}, m)_k$.

For the reverse move, where we propose to move from state $(\boldsymbol{\theta}', m')$, to state $(\boldsymbol{\theta}, m)$, we remove the parameter β_1' (or equivalently set the value equal to zero). This move is then accepted with probability $\min(1, A^{-1})$, where A is given above.

Example 2: Non-nested models

We extend the above linear regression to multiple regression where there are 2 possible explanatory variables, and hence a total of four possible models:

$$\begin{aligned} \text{Model } m_1\text{:} \quad \mu_i &= \beta_0; \\ \text{Model } m_2\text{:} \quad \mu_i &= \beta_0 + \beta_1 x_{1i} \\ \text{Model } m_3\text{:} \quad \mu_i &= \beta_0 + \beta_2 x_{2i}; \\ \text{Model } m_4\text{:} \quad \mu_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}; \end{aligned}$$

where x_{1i} and x_{2i} are two different explanatory variables.

We consider moving between each of the four possible models with equal probability, i.e., $\mathbb{P}(m_i|m_j) = \frac{1}{3}$ for $i \neq j$. Note that if we propose to move between models 2 and 3, then this involves simultaneously adding and removing a parameter within the same model move. To demonstrate the model move between non-nested models, suppose that at iteration k of the Markov chain the state of the chain is $(\boldsymbol{\theta}, m)_k$, where $m = m_2$ and $\boldsymbol{\theta} = \{\beta_0, \beta_1\}$. Then, we propose to move each each other model with equal probability, and we randomly choose to move to model $m' = m_3$, with parameters $\boldsymbol{\theta}' = \{\beta_0', \beta_2'\}$. This is a non-nested model move.

We initially define $\boldsymbol{u} = \{u\}$ and $\boldsymbol{u}' = \{u'\}$. We then define the bijective function g, such that,

$$\begin{aligned} \beta_0' &= \beta_0 \\ \beta_2' &= u \\ u' &= \beta_1. \end{aligned}$$

Finally, we need to define the corresponding proposal distributions, denoted by q and q', for u and u'. For example, we may assume that these are both independent Normal distributions, each with mean 0 and standard deviation σ and σ', respectively. This once more completely defines the reversible jump move. Once again, the bijective function g is simply the identity function, so that the corresponding Jacobian is equal to unity. Then, the corresponding acceptance probability for moving between models $m = m_2$ and $m' = m_3$ is simply, $\min(1, A)$, where A is given by,

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\theta}', m'|\boldsymbol{x})P(m|m')q'(u')}{\pi(\theta, m|\boldsymbol{x})P(m'|m)q(u)} \left|\frac{\partial(\boldsymbol{\theta}', \boldsymbol{u}')}{\partial(\boldsymbol{\theta}, \boldsymbol{u})}\right| \\ &= \frac{\pi(\boldsymbol{\theta}', m'|\boldsymbol{x})q'(u')}{\pi(\theta, m|\boldsymbol{x})q(u)}. \end{aligned}$$

Example - Rats

We implement a RJMCMC algorithm in WinBUGS for the rats example assuming polynomial regression up to a quadratic term which we rephrase as there being 2 covariates - time and time squared. We specify,

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2).$$

There are 4 possible models.

$$\begin{aligned} &\text{Model } m_1: \quad \mu_{ij} = \alpha; \\ &\text{Model } m_2: \quad \mu_{ij} = \alpha + \beta_1 z_{1j}; \\ &\text{Model } m_3: \quad \mu_{ij} = \alpha + \beta_2 z_{2j}; \\ &\text{Model } m_4: \quad \mu_{ij} = \alpha + \beta_1 z_{1j} + \beta_2 z_{2j}, \end{aligned}$$

where z_{1j} denotes the *normalised* time at time t and and z_{2j} *normalised* squared time at time t. We also normalise the observed (response) weights (due to the implementation of RJMCMC in WinBUGS).

For this particular example we are able to implement the RJMCMC within WinBUGS (see code "ratsrj.odc" on the SMSTC website). We specify the prior specification,

$$\alpha \sim N(0, \sigma_1^2); \qquad \beta_1 \sim N(0, \sigma_1^2); \qquad \beta_2 \sim N(0, \sigma_1^2);$$

where σ_1^2 is to be specified. Table 20.1 provides the corresponding posterior model probabilities for each model and marginal posterior probability of the regression parameters being present for different prior specifications on the regression coefficients.

Model	$\sigma_1^2 = 1000$	$\sigma_1^2 = 100$	$\sigma_1^2 = 10$	$\sigma_1^2 = 1$
Model m_1	0.000	0.000	0.000	0.000
Model m_2	0.997	0.992	0.976	0.927
Model m_3	0.000	0.000	0.000	0.000
Model m_4	0.003	0.008	0.024	0.073
β_1	1.000	1.000	1.000	1.000
β_2	0.003	0.008	0.024	0.073

Table 20.1: The posterior probability for each model and corresponding marginal posterior probability for β_1 and β_2 to be present in the model for different values of σ_1^2.

Thus, there appears to be very strong evidence for only a linear term in the model (i.e., very strong evidence for model m_2). This is contrary to the result for the DIC (see Exercises 20-2), although this is for a slightly different model specification in terms of the scale of the response used and transformation on the explanatory variables. There is no reason for the Bayes factor (or

posterior model probability) to rank the models in the same posterior order and it is relatively common for these different methods to provide different results. However, both analyses show the strong linear dependence of weight on time.

Comments

20–1. In practice several model updates may be performed within each iteration of the Markov chain. For example, for the covariate example, we may cycle through each individual covariate and propose to add/remove each covariate independently.

20–2. The RJ algorithm in the presence of model uncertainty has the advantage that the acceptance probability is easy to calculate for a given proposed move. In addition, irrespective of the number of possible models within an analysis, only a single chain needs to be run to obtain both the estimates of the posterior model probabilities and of the parameters within those models. Thus, this approach can be implemented even when consideration of all possible models individually is not feasible. Essentially time is spent exploring the models with reasonable posterior support given the observed data, and models not supported by the data (and priors) are not well explored.

20–3. The RJ algorithm described is able to calculate model-averaged estimates of the parameters within the Markov chain. They can simply be obtained as the estimates of the parameters, irrespective of the model that the chain is in (so long as they do contain that parameter), i.e., we "sum" out over the different models. Thus, the reversible jump procedure obtains parameter estimates under individual models, together with the posterior model probabilities and finally model-averaged parameter estimates, all within a single chain.

20–4. Care should be taken in specifying the priors on the parameters $p(\boldsymbol{\theta}|m)$ in the presence of model uncertainty, since these priors can have a significant impact on the corresponding posterior model probabilities (Lindley's paradox). A prior sensitivity analysis should always be performed for a number of reasonable (i.e. sensible) priors and compared.

20–5. The RJMCMC algorithm has been incorporated into WinBUGS, as an additional add-on package "jump". However, the RJMCMC algorithm is only possible for two situations, namely for variable selection problems (i.e., covariate analyses) in the presence of random effects and spline regressions. Alternatively, bespoke code needs to be written in order to implement the RJMCMC algorithm.

(iii) Other approaches

There is a diversity of other approaches available for estimating posterior model probabilities and related quantities. Some of these are relatively easily programmed in BUGS, but have limited application (e.g., only work for relatively simple model). An example is the "Gibbs Variable Selection" method (also called the Carlin and Chib method) described in O'Hara and Sillanpaa (2009, and references therein). See that paper for a review and discussion of some other methods.

20.5 Further reading

There are a plethora of Bayesian textbooks available that describe Bayesian statistics and the associated computationally intensive techniques, including MCMC. An excellent general introductory textbook that discusses the underlying ideas of Bayesian inference in an intelligible manner is "Bayesian Statistics: An Introduction" by Lee. Further in-depth concepts and ideas can be found in the additional textbooks "Bayesian Data Analysis" by Gelman, Carlin, Stern and Rubin; and "The Bayesian Choice" by Robert. Alternatively, for a Bayesian "encyclopaedia", "Kendall's Advanced Theory of Statistics Volume 2B: Bayesian Inference" is a good reference. In addition to

general textbooks on Bayesian inference (with many different examples), there are many subject-specific Bayesian textbooks.

Most books on Bayesian inference discuss the method of MCMC and its associated issues. Alternatively, there are others that focus more on MCMC itself. An excellent introductory book on MCMC is "Markov chain Monte Carlo" by Gamerman - even though it is now > 10 years old, it is typically the first book that I refer to for reference. Brooks (1998) also gives a very good and accessible overview of MCMC methods. Further in-depth issues are discussed in "Monte Carlo Statistical Methods" by Robert and Cassella and "Markov chain Monte Carlo in Practice" Gilks, Richardson and Spiegelhalter, which considers a number of applications.

An interesting-looking practical book I have not yet red is "Doing Bayesian Data Analysis: A tutorial with R, JAGS and Stan" by John K. Kruschke.

If you are interested in the philosophical and practical differences and similarities between Bayesian and classical statistics, then the classic book is "Comparative Statistical Inference" by Barnett, now in its 3rd edition.

Lastly, MCMC is not the only way to perform compuational Bayesian inference! One nice (but now a bit out of date) text book that empasizes the conceptual links between several alternative approaches is "Monte Carlo Strategies for Scientific Computing" by Jun Liu.

Once more this is not a comprehensive list of useful references, but simply a selection of textbooks that discuss the ideas and associated methodology of MCMC (and other methods) in greater depth, and may be useful for further reference.