Marine and Petroleum Geology 26 (2009) 525–535

Contents lists available at ScienceDirect

Marine and Petroleum Geology

journal homepage: www.elsevier.com/locate/marpetgeo

Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems

Kerry Gallagher [a,*], Karl Charvin [b,1], Soren Nielsen [c], Malcolm Sambridge [d], John Stephenson [e]

[a] Géosciences Rennes, Université de Rennes 1, Batiment 15, Campus de Beaulieu, Rennes 35042, France
[b] Dept. of Earth Sciences and Engineering, Imperial College London, London SW7 2AS, UK
[c] Dept. of Earth Sciences, University of Aarhus, Aarhus, Denmark
[d] Research School of Earth Sciences, Australian National University, Canberra, ACT 0200, Australia
[e] BP, Sunbury on Thames, Middlesex TW16 7LN, UK

ARTICLE INFO

Article history:
Received 12 September 2007
Received in revised form
11 February 2008
Accepted 2 January 2009
Available online 15 January 2009

Keywords:
Markov chain Monte Carlo
Inversion
Optimisation

ABSTRACT

We present an overview of Markov chain Monte Carlo, a sampling method for model inference and uncertainty quantification. We focus on the Bayesian approach to MCMC, which allows us to estimate the posterior distribution of model parameters, without needing to know the normalising constant in Bayes' theorem. Given an estimate of the posterior, we can then determine representative models (such as the expected model, and the maximum posterior probability model), the probability distributions for individual parameters, and the uncertainty about the predictions from these models. We also consider variable dimensional problems in which the number of model parameters is unknown and needs to be inferred. Such problems can be addressed with reversible jump (RJ) MCMC. This leads us to model choice, where we may want to discriminate between models or theories of differing complexity. For problems where the models are hierarchical (e.g. similar structure but with a different number of parameters), the Bayesian approach naturally selects the simpler models. More complex problems require an estimate of the normalising constant in Bayes' theorem (also known as the evidence) and this is difficult to do reliably for high dimensional problems. We illustrate the applications of RJMCMC with 3 examples from our earlier working involving modelling distributions of geochronological age data, inference of sea-level and sediment supply histories from 2D stratigraphic cross-sections, and identification of spatially discontinuous thermal histories from a suite of apatite fission track samples distributed in 3D.

1. Introduction

In basin modelling, a variety of strategies are used to constrain the evolution of a particular frontier region, basin or prospect, often with an emphasis on the maturation, generation and primary migration aspects of the kitchen, and/or the secondary migration, trapping and preservation aspects of individual reservoirs and traps. In general, strategies can be split into forward modelling and inverse modelling, in both of which the model is a mathematical representation of the underlying physical and chemical processes operating in the geological environment. In the former, the model parameters are specified a priori, and manually adjusted to achieve an acceptable fit to a suite of observations or calibration data. Forward modelling is also used when considering "what if?" scenarios to assess the sensitivity of model predictions to a range of values for one of more parameters. Inverse modelling attempts to infer the unknown model parameters from the data in a more direct way. This is achieved through a formal approach involving some type of guided search in the model space encompassing the range of acceptable values for each unknown parameter. In doing this, inverse methods generally require the solution of the forward problem or exploit the sensitivity of the forward problem to the unknown model parameters (i.e. we need to calculate derivatives, as in least squares solutions). The application of formal inversion methods in basin studies has been relatively limited, and primarily applied to thermal history and maturity modelling (Lerche et al., 1984; Lerche, 1990, 1991; Gallagher and Sambridge, 1992; Nielsen, 1995, 1996; Gallagher, 1998; Gallagher and Morrow, 1998). These applications generally used classical methods involving derivatives of the data fit (with respect to the model parameters) to improve the model parameter estimates, and some form of least squares approach, implicit in which is the assumption of Gaussian statistics

* Corresponding author.
E-mail address: kerry.gallagher@univ-rennes1.fr (K. Gallagher).
[1] Present address: Chevron Energy Technology Company, Chevron House, Hill of Rubislaw, Aberdeen AB15 6XL.

0264-8172/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.marpetgeo.2009.01.003

in terms of fitting the calibration data. Such methods were oriented towards finding a best model, usually defined in terms of the best fit to calibration data (but sometimes also including some constraints on the nature of the overall model, such as smoothness). We refer to this philosophy as optimisation in that the aim is usually to find one (best) model. Inversion is better defined in terms of characterising the whole model space, and clearly if this is done appropriately, a best model can be extracted. In addition, this also provides an ensemble, or distribution, of models and allows us to readily appraise the quality of the model, i.e. quantify uncertainty in terms of model resolution and predictive sensitivity.

Over the last few years, a class of sampling-based methods for inversion/parameter estimation developed in statistics has begun to be applied to Earth Science problems. These methods come under the general name of Markov chain Monte Carlo (MCMC) and are superficially similar to the well known Monte Carlo approach. A useful introduction to MCMC methods is given by Gilks et al. (1996). The approach is based on a random walk to produce a sequence of models from the model space. The Markov chain aspect is that a new model is generated conditional on the previous one. However, this process is independent of how the previous model was arrived at. So, the basis of MCMC is that we choose a model from some a priori specified range or distribution of the model parameters, and then propose a new model, conditional on the current model. The fundamental aspect of MCMC is how we decide whether or not to replace the current model with the proposed model or not. A second important consideration for the efficiency of MCMC is how we generate the proposed model. In this paper, we give a brief overview of MCMC, and focus particularly on the Bayesian version, which is formulated in terms of probability distributions. We will start with a basic discussion of Bayes' theorem and describe the basic MCMC approach in this framework. We will go on to discuss briefly how MCMC can be generalised to problems of variable dimension, and finally will show 3 examples of the approach applied to geological problems.

2. Bayes' theorem

Bayes' theorem (Bayes, 1763) is well known, and some basic introductions are given in Lee (2004) and Sivia and Skilling (2006) while more detailed/advanced treatments can be found in Tarantola (2005) and Bernardo and Smith (1994). An interesting tutorial on the philosophical differences between classical and Bayesian statistics is given by Scales and Snieder (1997) and Malinverno and Parker (2005) discuss the different approaches when considering uncertainty.

Mathematically, Bayes' theorem is expressed as

$$p(\boldsymbol{m}|\boldsymbol{d})p(\boldsymbol{d}) = p(\boldsymbol{d}|\boldsymbol{m})p(\boldsymbol{m}) \tag{1}$$

where p is probability, $\boldsymbol{m}$ and $\boldsymbol{d}$ are model parameter and data vectors respectively, and $p(a|b)$ means the probability of a given b, or a conditional on b. At this stage, we will drop the term $p(\boldsymbol{d})$, and rewrite Bayes' rule in terms of proportionality relationship,

$$p(\boldsymbol{m}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{m})p(\boldsymbol{m}) \tag{2}$$

Beginning with the last term on the right, $p(\boldsymbol{m})$ is the prior probability distribution on the model parameters, i.e. what we think we know before collecting data (e.g. Scales and Tenorio, 2001; Curtis and Wood, 2004). For example, if we were interested in heat flow, we know that typical values are between 30 and 120 m Wm^{-2}. Then we could choose $p(\boldsymbol{m})$ to be a uniform distribution in this range, or a normal distribution with an appropriate mean and standard deviation.

The term $p(\boldsymbol{d}|\boldsymbol{m})$ is the data likelihood, and can be regarded as a measure of how well we fit the observed data. This is the probability that we would obtain the data (or observations) $\boldsymbol{d}$, given the underlying model parameters, $\boldsymbol{m}$. A common form of likelihood (as used in least squares) is a Normal or Gaussian distribution, i.e.

$$p(\boldsymbol{d}|\boldsymbol{m}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{V}|^{N/2}} e^{-\frac{1}{2}((\boldsymbol{d}-g(\boldsymbol{m}))^t|\boldsymbol{V}|^{-1}(\boldsymbol{d}-g(\boldsymbol{m})))} \tag{3}$$

where $g(\boldsymbol{m})$ is the prediction from model parameters $\boldsymbol{m}$, $\boldsymbol{V}$ is the data covariance matrix, N is the number of data and t is the vector transpose. In terms of optimisation, we may want to find the model that maximises this probability (known as the maximum likelihood method). This is equivalent to minimising the argument of the exponential, which is same as the sum of squares misfit function used in conventional least squares.

The term $p(\boldsymbol{m}|\boldsymbol{d})$ is the posterior probability distribution on the model parameters, or the probability of a particular set of model parameters, $\boldsymbol{m}$, given the observations or data, $\boldsymbol{d}$. Again, if we want to choose a 'best' model, then we could take the model which has the highest posterior probability. Depending on the nature of the prior function, this may or may not be the same as the model that maximises the data likelihood.

The form of Bayes' theorem given in equation (2) can be expressed in words as the likelihood updates (or maps) what we thought about the model parameters before we had any data (the prior) into what can additionally be learnt once we have acquired some data (the posterior). Clearly if the prior and the posterior distributions are found to be the same, this implies we have learnt nothing abut the model from the data that we didn't know already. This could mean we have managed to specify the prior information to characterise perfectly the model space. Alternatively, it may mean that the model is not particularly sensitive to the data we have (i.e. the likelihood is flat in the region of the model space defined by the prior). As shown in Fig. 1, the Bayesian approach can be thought of in terms of or updating information or information retrieval (e.g. Tarantola and Valette, 1982). To characterise the model space, clearly we want to be able to determine the posterior distribution. In most real problems, this generally involves analytically intractable integration, and so we need to use numerical methods.

3. Markov chain Monte Carlo

As described earlier the basic algorithm for MCMC involves a current and a proposed model, and the critical step is deciding whether to replace the current model with the proposed model. If we keep the number of unknown model parameters constant, and now refer to the current model parameter vector as $\boldsymbol{m}$ and the proposed model parameter vector as $\boldsymbol{m}'$, then the criterion can be formulated as

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{m}')p(\boldsymbol{d}|\boldsymbol{m}')q(\boldsymbol{m}|\boldsymbol{m}')}{p(\boldsymbol{m})p(\boldsymbol{d}|\boldsymbol{m})q(\boldsymbol{m}'|\boldsymbol{m})}\right\} \tag{4}$$

where the first 2 terms on both the upper and lower lines are the prior and likelihood for the proposed and current models, respectively. The term $q(\boldsymbol{m}'|\boldsymbol{m})$ is the proposal function, and is used to propose a new model parameter vector, $\boldsymbol{m}'$, given the current model parameter vector, $\boldsymbol{m}$. A requirement of the algorithm is that we can return from the proposed model to the current model (this is known as reversibility), so a proposal function occurs on both the upper and lower lines. Having determined a value for α, (which is always between 0 and 1), we generate a random number (u) from a uniform distribution (also between 0 and 1). If u is less than α, we

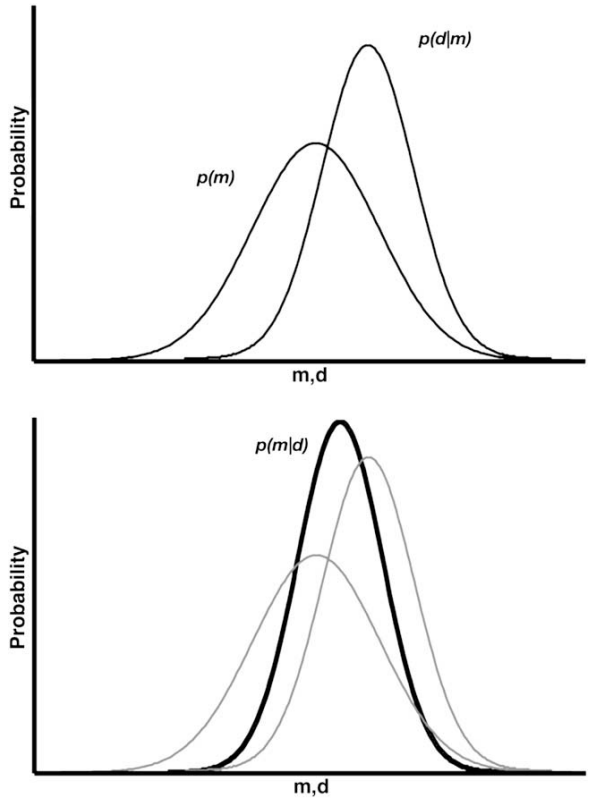

Fig. 1. An illustration of the relationship between the prior, $p(\boldsymbol{m})$, likelihood, $p(\boldsymbol{d}|\boldsymbol{m})$, and posterior, $p(\boldsymbol{m}|\boldsymbol{d})$, in Bayes' theorem. In this illustration the maximum probability for the prior and the posterior correspond to different models. The likelihood can be thought of as a way of updating the prior information to the posterior, so that hopefully we learn more about the model from the data.

accept the proposed model. If not, we reject the proposed model, and stay at the current model. This represents one iteration and the whole process is repeated many times. Ferrero and Gallagher (2002) used MCMC sampling procedure in the context of modelling heat flow histories in sedimentary basins. They give an example of a basic algorithm for MCMC and we will not repeat that here.

After an initial period sampling (known as the burn-in), the current model from each iteration is taken as representing a sample from the posterior distribution. The whole collection, or ensemble, of samples can then be used as a proxy for the multi-dimensional posterior distribution. Note that if we reject a proposed model, and stay at the current model, we add another copy of the current model to the ensemble. This is what we want to do as we are trying to sample each model in proportion to its posterior probability. The approach has a sound theoretical basis (Hastings, 1970; Green, 1995) although in practice the algorithm needs to be tuned in terms of the number of iterations, the length of the burn-in and subsequent number of iterations, and also the proposal function. While the proposal function can be chosen arbitrarily, the performance of the algorithm can be sensitive to the nature of the parameterisation (see Fig. 2 caption). Furthermore, there are different ways to propose the model parameters. For example, each parameter can be updated individually and independently, or a group of parameters can be updated in a block. The latter is useful if a particular set of parameters are known to be correlated for a given problem. The proposal function can be designed to incorporate correlation and improve the efficiency of the sampler. A variant of the basic algorithm known as delayed rejection (e.g. Green and Mira, 2001; Al-Awadhi et al., 2002), can be used to improve the acceptance rate by modifying the proposal function during an iteration. For example, if a relatively large change occurs in the model proposal, and we reject the proposed model, we can return to the current model and propose a smaller change which may then be more likely to be accepted.

Given that we can approximate the posterior distribution, we can then derive many useful quantities from this. Firstly, simply by plotting the distribution of samples we can visualize the joint distribution for various combinations of parameters. For example, if we consider the well known straight line regression problem, for a series of observations, y^j, made at a series of values x^j,

$$y^j = ax^j + b \tag{5}$$

Here the model parameters are the slope (a) and intercept (b), so $\boldsymbol{m} = \{m_a, m_b\}$. This problem is generally solved using least squares methods, assuming Gaussian statistics for the observations. In this case, we can calculate the posterior distribution exactly (e.g. Tarantola, 2005). In Fig. 3a, we show the true joint distribution on m_a and m_b, $p(m_a,m_b)$, overlain by the distribution of samples from MCMC, showing how the latter captures the form of the underlying probability distribution. If we are interested in the distribution on one of the parameters, a, for example, then formally we need to solve the following integral

$$p(a) = \int p(a,b)\mathrm{d}b \tag{6}$$

Here, for a given value of a, we need integrate out the variation in b. This is known as marginalizing and $p(a)$ is the marginal probability distribution of a. Using the MCMC samples, we just plot all value of a as a histogram (Fig. 3b), as the sampling effectively deals with the integration.

Also, it is straightforward to calculate estimates of the expected value for any parameter, simply by averaging over all the samples accepted for that parameter, i.e.

$$E(m_i) = \frac{1}{N_a}\sum_{j=1}^{N_a} m_i^j \tag{7}$$

where N_a is the number of samples accepted (post-burn-in) for model parameter m_i.

Perhaps less obviously, we may be more interested in the predictions from the whole range of models, rather than the values of the model parameters themselves. For example, in reservoir production history matching, where we need to estimate a permeability model from the production history (e.g. Ballester and Carter, 2007), we are usually concerned in predicting the future production than the details of the permeability model itself. To estimate the expected value of the predictions ($\boldsymbol{d}^{\mathrm{pred}}$), as above, we need to integrate over all models, given the observed data ($\boldsymbol{d}$),

$$E\left(\boldsymbol{d}^{\mathrm{pred}}|\boldsymbol{d}\right) = \int \boldsymbol{d}^{\mathrm{pred}}(\boldsymbol{m})p(\boldsymbol{m}|\boldsymbol{d})\mathrm{d}\boldsymbol{m} \tag{8}$$

The MCMC estimate of this is similar to equation (7), i.e.

$$E\left(\boldsymbol{d}^{\mathrm{pred}}|\boldsymbol{d}\right) = \frac{1}{N_a}\sum_{j=1}^{N_a} \boldsymbol{d}^{\mathrm{pred}}(\boldsymbol{m}_j) \tag{9}$$

as the values of $\boldsymbol{m}$ are sampled according to $p(\boldsymbol{m}|\boldsymbol{d})$, and the predictions, $\boldsymbol{d}^{\mathrm{pred}}$, are a function of each model parameter vector.

Finally, it is straightforward to quantify the uncertainty on the predictions. For example, suppose we use N_a post-burn-in samples to make N_a predictions for a given variable. Then the 95% credible interval can be determined by ranking the predictions in increasing

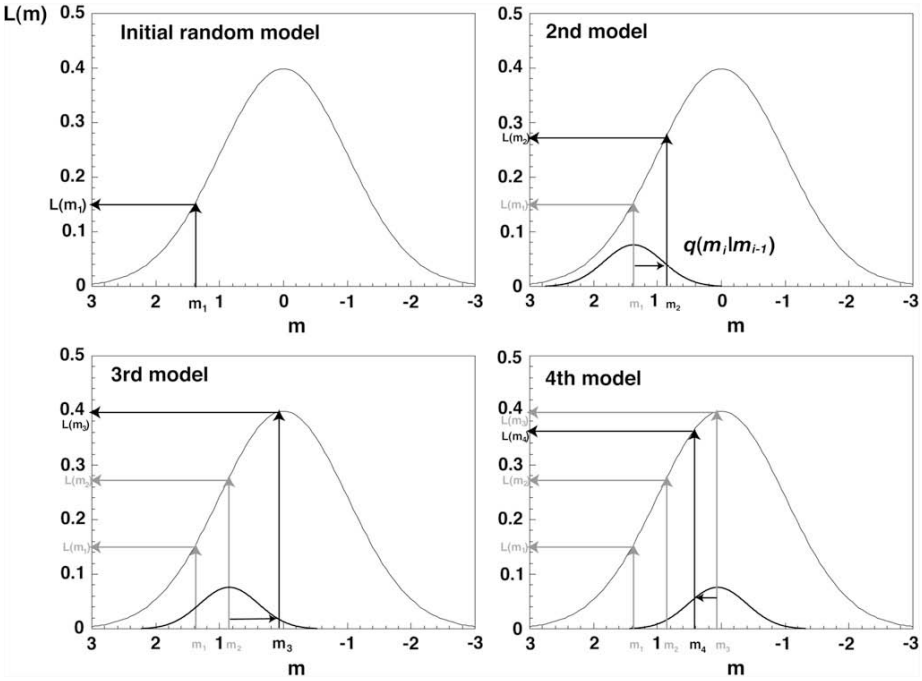

Fig. 2. An illustration of the implementation of the proposal function, $q(\boldsymbol{m}_i|\boldsymbol{m}_{i-1})$. In this case, we chose a normal distribution (thick line) to generate a sequence of models, in which the proposed model is conditional on the current model. The initial model is drawn from the prior (not shown here), and the likelihood is the broader scale distribution (thin line). Generally, a proposed model is always accepted if the likelihood is higher than the current model. If not, the acceptance criterion (see text) is used to determine if the proposed model should be accepted or not. Tuning of the proposal function is usually required for specific problems. If the proposal function is too narrow relative to the likelihood then the sampling involves small moves and we tend to accept the propose model nearly always. If the proposal function is too broad, then we tend to propose many models with a large difference in the likelihood, and these tend to be rejected if the likelihood is lower, leading to a low acceptance rate.

order, and then taking the lower limit as the $0.025N_a$ value, and the upper limit as the $0.975N_a$ value. Credible intervals reflect the Bayesian approach to uncertainty and although the calculated values may be similar to more conventional confidence intervals the interpretation differs (see Bernardo and Smith, 1994 for a discussion of credible intervals).

One final point is that the inference of optimal models and model uncertainty requires knowledge of the errors in the observations (and ideally any uncertainty in the theory itself). If there data errors are not well known, then these can be part of the inference process (e.g. Denison et al., 2002; Malinverno and Briggs, 2005; Malinverno and Parker, 2005). In general, this involves the use of hyper-parameters, and the details of this are beyond the scope of this paper, although one example we consider later from Jasra et al. (2006) does use such an approach.

4. Variable dimension MCMC

In most real world problems the unknown model is a continuous function, for example heat flow as a function of time. However, in practical applications, we typically need to deal with a simplified or often discretised version of the continuous function. For example, Gallagher (1998) used a discrete model for heat flow reconstruction, and chose a series of nodes to represent heat flow at

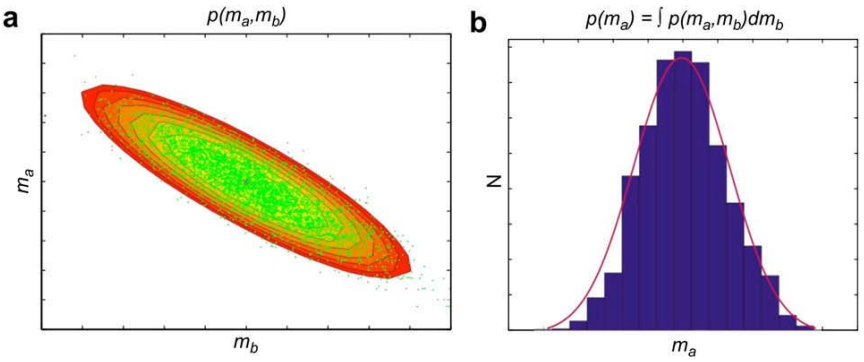

Fig. 3. (a) The joint posterior distribution, $p(m_a,m_b)$, for the linear regression problem, $y = ax + b$, where m_a and m_b are the model parameters. The contours are the analytical solution and the dots indicate the distribution of MCMC samples. Notice how the density of samples is greater where the analytical probability density is higher. (b) The marginal distribution, $p(m_a)$, for the slope parameter, m_a, represented as a histogram. The analytical solution is given as the continuous line.

specific times. Linear interpolation was used to obtain heat flow values between the nodes. In representing a continuous model like this, issues arise as to how we should best do it and more specifically, how complex we should make the discrete model, e.g. how many nodes? Here, the number of nodes is clearly not a physical parameter that can be measured. However, it can be used as an indicator of the complexity of a model, given a specific model structure, in that a more complex model will have more parameters. In other problems, the notion of a discrete model is more natural, such as estimating the number of components in a mixture or the number of coefficients in a polynomial. We return to this later.

In these cases, we can think of the model as being parameterised in terms of a number of components, but that number is not known. Put another way, we can say that one of the things we don't know is the number of things we don't know. Green (1995, 2001, 2003) presented a generalisation of MCMC, commonly known as reversible jump (RJ) MCMC, to this variable dimension problem. Sambridge et al. (2006) compared the performance of variable dimension RJMCMC to a series of fixed dimension MCMC simulations and showed that provided the fixed dimension sampling is done appropriately, then the two approaches produced the same results.

If we are dealing with two models with dimensions k, and k', then acceptance criterion can be written as

$$\alpha = \min\left\{1, \frac{p(k')p(\boldsymbol{m}'|k')p(\boldsymbol{d}|\boldsymbol{m}',k')q(\boldsymbol{m}|\boldsymbol{m}')}{p(k)p(\boldsymbol{m}|k)p(\boldsymbol{d}|\boldsymbol{m},k)q(\boldsymbol{m}'|\boldsymbol{m})}\right\} \tag{10}$$

Here we separate the prior on the number of dimension, $p(k)$, from the model parameter prior, $p(\boldsymbol{m}|k)$. However, the form of the proposal function $q()$ is more complex as we need to propose models with different dimensions. Furthermore, we need to allow for the transformation from one model to another to ensure that the theoretical probability requirements are maintained. If we dealing with a situation where we are simply increasing or decreasing the number of parameters, then we can write

$$\alpha = \min\left\{1, \frac{p(k')p(\boldsymbol{m}'|k')p(\boldsymbol{d}|\boldsymbol{m}',k')g\left(\boldsymbol{u}^{k'}\right)}{p(k)p(\boldsymbol{m}|k)p(\boldsymbol{d}|\boldsymbol{m},k)g(\boldsymbol{u}^{k})}|J|\right\} \tag{11}$$

Here $\boldsymbol{u}^{k'}$ and $\boldsymbol{u}^{k}$ are vectors of random numbers of length r' and r, respectively, and used to transform from one model to another, such that $r+k=r'+k'$, and $g(\cdot)$ is the probability distribution used to generate these random numbers. The term $|J|$ is the Jacobian, and allows for the transformation between the 2 models, i.e.

$$|J| = \frac{\partial\left(\boldsymbol{m}', \boldsymbol{u}^{k'}\right)}{\partial\left(\boldsymbol{m}, \boldsymbol{u}^{k}\right)} \tag{12}$$

Equation (11) is actually a general form for the acceptance criterion, although for fixed dimensional problems the Jacobian is generally 1, and the proposal functions are of the form as described earlier. The details of the reversible jump acceptance criterion are discussed in mode detail by Green (1995, 2001, 2003), Malinverno (2002), and Sambridge et al. (2006) and examples of the implementation algorithms for variable dimension problems are given in Charvin et al. (in press-a) and Hopcroft et al. (2007).

4.1. Model choice

In many real world problems, we often have different models or theories to consider. These may involve different numbers of parameters, or different model formulations or even physical hypotheses. Here we may face a problem of choosing a model. Just as in optimisation and inversion, we can use some statistical measures of fit to data. However, various models will often fit the data equal well. A common issue arises with hierarchical models, developed by increasing the number of parameters while maintaining a similar mathematical structure (e.g. a polynomial of increasing order). Here the problem is to decide how much complexity is justified in the model, given that we must reach a point of diminishing returns in terms of fitting the data with more and more complex models.

One approach to this leads us to consider Bayes' theorem again, in a form slight different to that given in equation (1), i.e.

$$p(\boldsymbol{m}|\boldsymbol{d},H)p(\boldsymbol{d}|H) = p(\boldsymbol{d}|\boldsymbol{m},H)p(\boldsymbol{m}|H) \tag{13}$$

where H is the underlying hypothesis or theory for a model. This incorporates various aspects such as the assumptions we have made in formulating the model structure, how we have parameterised the model and how many parameters. Clearly, the inference is conditional on the hypothesis when defined like this. The 2nd term on the right, $p(\boldsymbol{d}|H)$ is sometimes referred to as the evidence (e.g. Mackay, 2003; Sambridge et al., 2006), and the evidence measures how well a given hypothesis can explain the data.

The evidence is the normalising constant implied by the proportionality in equation (2), and is given by

$$p(\boldsymbol{d}|H) = \int p(\boldsymbol{d}|\boldsymbol{m},H)p(\boldsymbol{m}|H)\mathrm{d}\boldsymbol{m} \tag{14}$$

Here we can see that the evidence is calculated by integrating over all values of the model parameter vector, given the prior and likelihoods, for a given hypothesis.

If we consider two competing hypotheses, H_1 and H_2, then we can write the posterior probability for a given hypothesis as

$$p(H_1|\boldsymbol{d}) = \frac{p(H_1)p(\boldsymbol{d}|H_1)}{p(\boldsymbol{d})} \tag{15}$$

Given $p(\mathbf{d})$ is constant, then we can write the ratio of posterior probability for 2 competing hypotheses as

$$\frac{p(H_1|\boldsymbol{d})}{p(H_2|\boldsymbol{d})} = \frac{p(H_1)p(\boldsymbol{d}|H_1)}{p(H_2)p(\boldsymbol{d}|H_2)} \tag{16}$$

where we see the evidence for each hypothesis as the 2nd term on the right handside. This is also known as the posterior odds ratio (e.g. Jaynes, 2003, chapter 20). We can also calculate the ratio of the posterior to prior odds ratio, which is also known as the Bayes Factor (e.g. Denison et al., 2002, chapter 2), and is given as

$$\mathrm{BF}(H_1,H_2) = \frac{p(H_1|\boldsymbol{d})}{p(H_2|\boldsymbol{d})}\Big/\frac{p(H_1)}{p(H_2)} \tag{17}$$

This provides a measure of whether the data have changed the prior odds of one hypothesis relative to the other. If the ratio is greater than 1, then the odds of the numerator hypothesis (i.e. H_1 above) have been increased relative to the denominator (i.e. H_2 above). If the priors are the same for both hypotheses, then the Bayes Factor is the same as the posterior.

As described by Malinverno (2000), Denison et al. (2002) and Mackay (2003), the Bayesian formulation for inference is naturally parsimonious. This means that if we have 2 competing models with different numbers of parameters that both fit the data equally well, then the evidence will favour the simpler model. As shown in Fig. 4, this is because the probability density is more concentrated for the simpler model (fewer parameters will predict a smaller range in terms of possible realisations of data), and so will have a higher

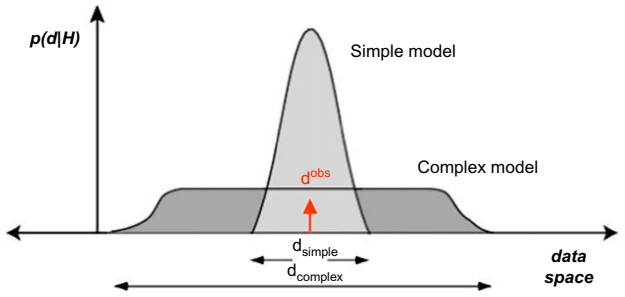

Fig. 4. Illustrating how the Bayesian approach naturally favours a simpler model due to the lower predictive range (in the model space) compared to a complex model, even though both fit or predict the observed data equally well. As we deal with probability distributions, the area under the curves for the evidence, $p(\boldsymbol{d}|H)$, is the same (and is equal to 1 by definition), so the amplitude of the simple model curve must be higher in the region of the observed data ($\boldsymbol{d}^{obs}$) (after Mackay, 2003).

probability in the region of the data space where our actual observations lie. Alternatively, we can think of the prior penalising the more complex model relative to the simpler model. Each parameter has an associated prior probability which is a value between 0 and 1. If the parameters are independent (as is generally assumed), then the individual prior probabilities multiply together to form the joint prior probability. Assuming a complex and a simple model both data fit the data equally well (i.e. the same likelihood) and have similar priors, then the model with more parameters will tend to have a lower joint prior probability than one with few parameters, as we multiply together more numbers less than 1. This then leads to a lower overall posterior probability.

A simpler, but approximate, approach to model choice, also based on the Bayesian philosophy, was given by Schwarz (1978). This is known as the Bayesian Information Criterion (BIC) and is given as

$$\mathrm{BIC}(\boldsymbol{m}) = -2\mathrm{LL}(\boldsymbol{m}) + k\,\mathrm{Ln}(N)$$

where LL is the log likelihood (log of $p(\boldsymbol{d}|\boldsymbol{m},k)$) and k and N are the number of model parameters and data, respectively. When comparing 2 models of different dimensions, then we choose the model with the lower value of the BIC (e.g. Gallagher et al., 2005; Sambridge et al., 2006). However, these point estimates like the BIC and the full integration criteria represented by (14) are different measures for model selection. The latter is a more complete answer to the model selection problem but requires integration over the model space, whereas the former is just a point estimate in model space evaluated at the 'best fit', or optimal, model. In addition to the references already cited in this section, we refer the reader to Burnham and Anderson (2002) and Spiegalhalter et al. (2002) for further discussion regarding issues of model choice (the latter is particularly relevant to MCMC sampling methods).

5. Examples of MCMC applied to Earth Science problems

Here we briefly present 3 examples from some of our earlier work in applications of variable dimension MCMC. Other Earth

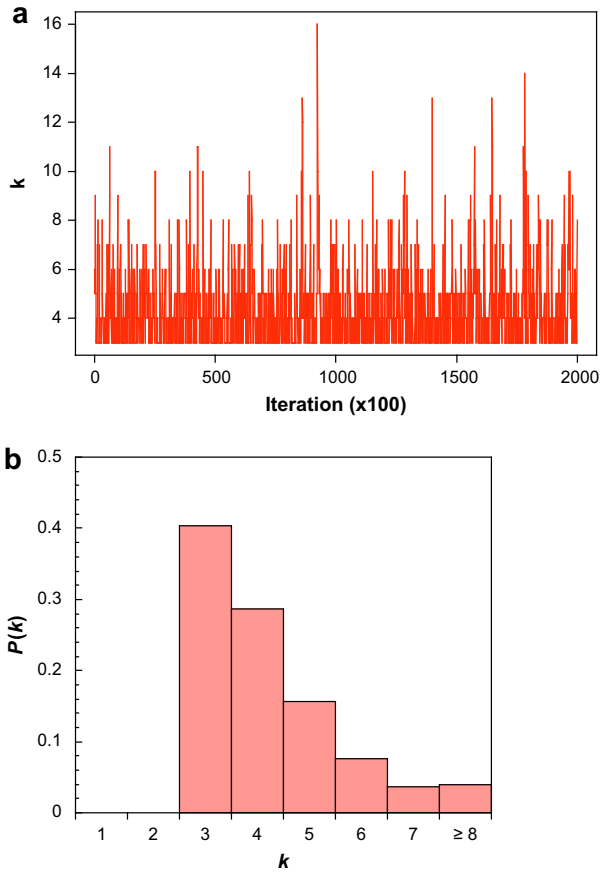

Fig. 5. (a) Sampling history for k, the number of components in the distribution of U-Pb ages from Carrickalinga Head, South Australia, as a function of iteration. (b) Posterior distribution on k.

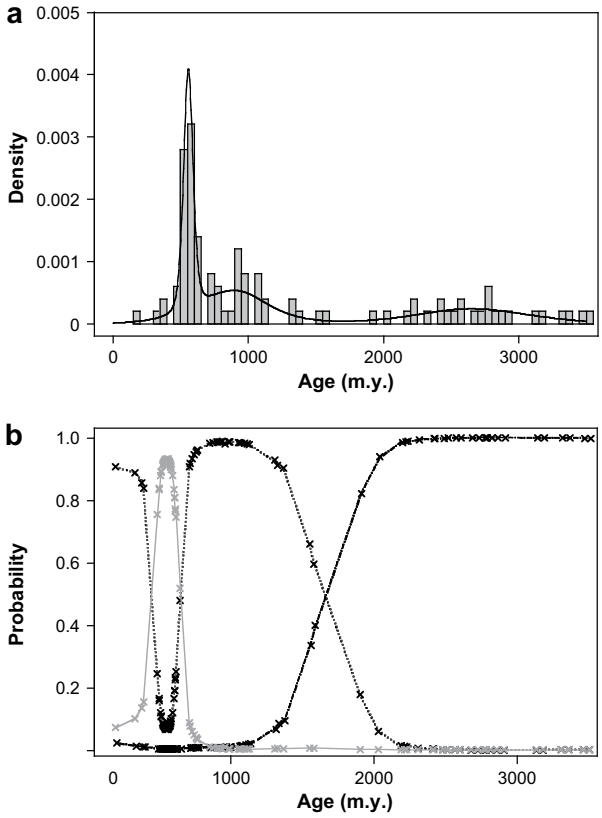

Fig. 6. (a) Inferred distribution, conditioned on 3 components (solid line), for the observed Carrickalinga Head data (shown as histogram). (b) Classification probabilities of individual age data for each of the 3 components shown in (a).

Science related applications of MCMC include Mosegaard and Tarantola (1995), Ferrero and Gallagher (2002), Malinverno (2000, 2002), Stephenson et al. (2004, 2006a,b), Sambridge et al. (2006), and Hopcroft et al. (2007).

The first example is taken from Jasra et al. (2006) who developed an approach for modelling geochronological ages in terms of a finite number of component distributions mixed together. The problem is to unmix the mixed distribution into individual component distributions, and identify the proportions and statistical properties of these distributions. In this case, the variable dimension comes through the unknown number of discrete components, k, and as this changes, the number of distribution related parameters change. These additional parameters include the proportion of each component, and the statistical properties (mean, variance, and skew) of the component distributions. Additionally, the algorithm was formulated to allow for relatively vague prior information on the component properties, and this lead to additional unknown parameters arising in the prior distributions reflecting uncertainty in the model. These extra parameters are of little intrinsic interest but are varied as part of the MCMC sampling process. As mentioned earlier, these are known as hyper-parameters and can also be used when the uncertainty on the observations is not well known also (e.g. Denison et al., 2002; Malinverno and Briggs, 2004; Malinverno and Parker, 2005). In this example, the reported errors on individual age measurements were used. Here we present representative results to illustrate the output from the approach while the details are explained more fully in Jasra et al. (2006).

The particular example we choose involves 100 individual U-Pb zircon ages in a sediment from South Australia. In Fig. 5, we show the sampling history for the number of components (k), together with the posterior density distribution on k. The prior on k was uniform with a range between 1 and 20. As can be seen from Fig. 5, the sampler examined up to 16 components, but quickly settles to much fewer than that. The posterior distribution clearly favours 3 components ($p(3) \sim 0.4$), although a 4 component model has relatively similar support probabilistically, while more than 7 components is inferred to be unlikely. In Fig. 6a we show the distribution of observed ages together with the inferred distribution, averaged over all models with 3 components. There is a well defined peak around 550 Ma, while the other 2 inferred distributions are relatively broad and show some skewness towards younger ages. The reported error on each age measurement is incorporated into the inference process. These are variable but have an average value around 30 m.y., and tend to increase with increasing age. As the approach is formulated in probabilistic framework, we can naturally examine the probability of a given age belonging to a particular component distribution. This classification is shown in Fig. 6b, and most individual ages are well identified with a particular component. Interestingly, in this case, we can see that the youngest individual ages are actually classified with the middle age component distribution. This is because the youngest component distribution is inferred to have a narrow spread, and the youngest ages have a low probability of being sampled from that distribution. In this particular simulation, the prior information on the nature of the component distributions was specified to be fairly vague and allows for very general forms for the distribution, hence the broad skewed shapes, with relatively high amplitude probability in the tails. As reported in Jasra et al. (2006), when running the same dataset allowing only normal distributions (i.e. no skewness, and low amplitude tails), up to 22 individual components can be inferred, but leads to a considerably more complex model with 65 distribution related parameters, compared to the 3 component skew model which has 15.

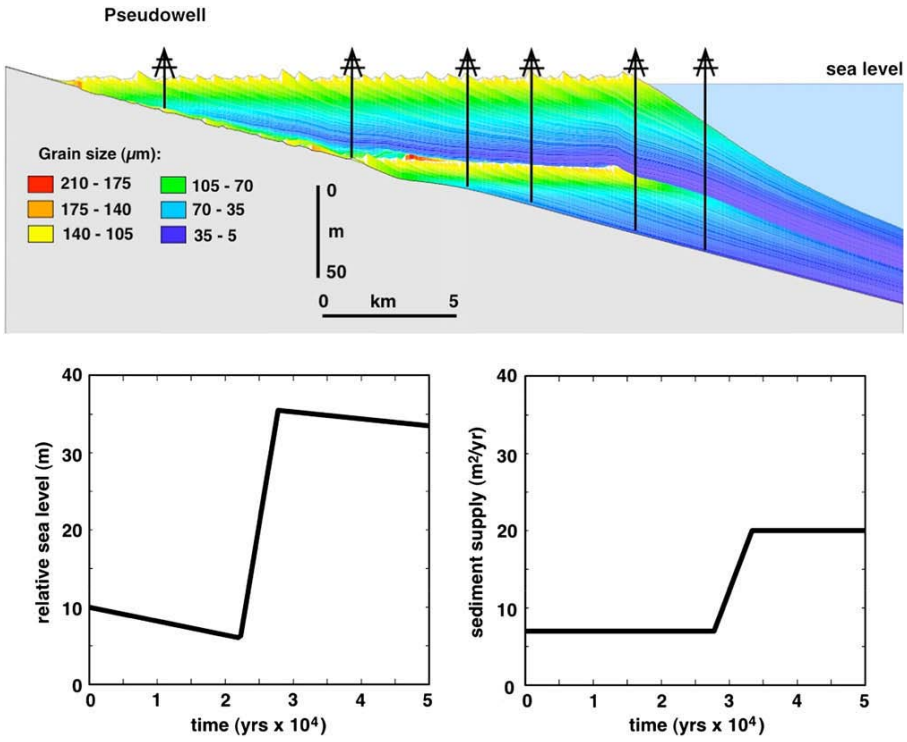

Fig. 7. Input model (sea-level and sediment supply histories in the lower panels) and the 2D stratigraphic cross-section generated from this model. We selected the vertical distribution of grain size at 6 pseudo-well locations, and the total sediment thickness between the 2 pseudo-wells at each end as input data for the MCMC sampler.

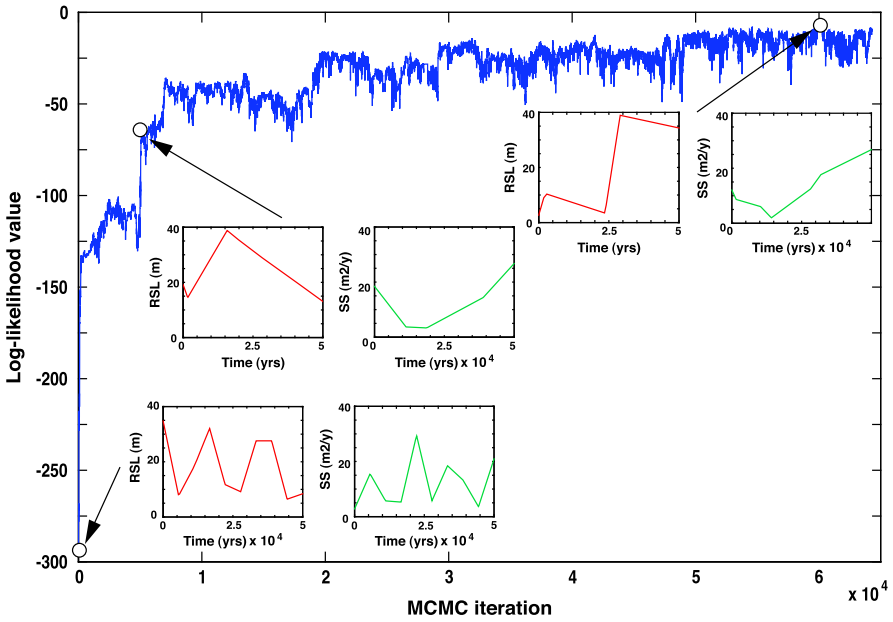

Fig. 8. Evolution of the log of the likelihood (values closer to zero are better data-fitting models) as a function of MCMC iteration, together with 3 models. The first is the initial random starting model which has 10 nodes for both the relative sea-level (RSL) and sediment supply (SS) histories. The middle model is selected at random, and shows how the sampler prefers fewer parameters and can fit the data better (higher likelihood) than the initial model. The model on the right is the highest likelihood model which captures the variation in the RSL curve, but the data are not so sensitive to the SS curve. In this run, the burn-in period was taken as 3×10^4 models, after which the sampler is relatively stable in terms of the variation in likelihood.

The second example we consider is taken from Charvin et al. (in press-a,b). Here the problem was to infer time dependent environmental variables (sea-level, sediment supply) directly from the stratigraphic record. Given these environmental variables, the forward model allows us to predict a 2D stratigraphic section in which grain size varies (see Charvin et al., in press-a,b for details). The study considers synthetic cases in 2-D, and uses realistic types of data, such as grain size distribution and stratigraphic unit thicknesses from pseudo-wells and total sediment thickness in a 2-D cross-section. Here we are dealing with various different types of data, sensitive to different parts of the process-based model, and so the likelihood function needed to be appropriately scaled to avoid one type of data dominating the inference of model parameters. The unknown sea-level and sediment supply histories were parameterised as a series of nodes representing each variable as a function of time, with linear interpolation between each node, and the number of nodes was allowed to vary.

In Fig. 7 we show the typical synthetic 2-D stratigraphic cross-section, together with the input sea-level and sediment supply histories. The grain size data were taken from the 6 pseudo-wells across the cross-section, and we also use the total sediment thickness between the pseudo-wells. In Fig. 8 we show the evolution of the sampler, showing selected models for the sea-level and sediment supply histories, and predicted stratigraphic cross-sections. As can be seen, the sampler considers relatively complex models early on (during the burn-in period), although these do not fit the data particularly well. It soon learns that it does not need particularly complex models and focuses more on simpler, but better data-fitting models.

As mentioned earlier, we can always choose the best data-fitting model (the maximum likelihood model), although with MCMC there is no guarantee that we will have sampled the true optimal model. Rather, MCMC samples the model space, and so one appropriate model to examine is the expected model, determined by averaging over all sampled models as described by equation (7).

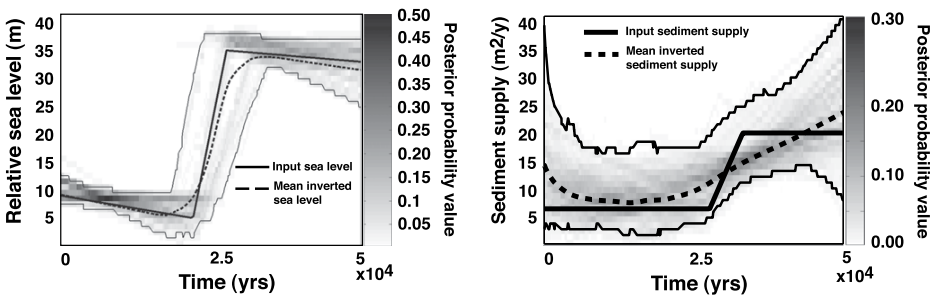

Fig. 9. Results of the MCMC sampler, post-burn-in. The input (or true) model is shown, together with the posterior mean (expected) models for relative sea-level and sediment supply. The mean is calculated from 30,000 samples. We have represented the posterior density function as a 2D histogram, and the thin black lines are the 95% credible intervals.

This averaged model is shown in Fig. 9, together with the 95% credible intervals about this model, and a representation of the probability density function for the posterior distribution on the model parameters. In this particular example, the sea-level history is better resolved than the sediment supply history. One feature of the averaged model is that it can be both smoother, yet potentially more structurally complex, than any individual model. The smoothness is expected as this is a consequence of averaging. However, the greater complexity arises as any single model must be comprised of nodes, joined by linear segments. Thus any changes in time must be linear. However, the averaged model can produce more local variation over time as it incorporates small changes in rates of change between different individual models. While the averaged model from a complex problem like this will generally not fit the observed data as well as the best model, it does provide a representative 'average' prediction, given the uncertainty in the inferred model parameters. Furthermore, in this particular application, the ability to produce such a model, and also sample a range of different individual models, provide a natural complement to the commonly adopted geostatistical methods when assessing uncertainty in reservoir simulation.

The final example we show is from Stephenson et al. (2006b). Here, the problem addressed involved inferring a 1-D function, which was the thermal history from apatite fission track data from a series of samples distributed over the surface, but at different elevations. In this we wanted to allow for discontinuous spatial variations in the thermal history function, as could arise if major faults are present and lead to differential cooling either side of a fault. However, the difficulty is, in general, the lack of knowledge about where these discontinuities may be, how many there are and when they were active. In this case, we parameterised the problem in terms of an unknown number of 2D regions, or partitions, within which the unknown thermal history was the same everywhere, but independent of adjacent regions. We could have chosen to vary the thermal history within a partition, and estimate the parameters associated with the spatial variation, while still maintaining the independence between regions. The approach we used did allow for variations in the thermal history with elevation, following Gallagher et al. (2005). This allows us to make an inference of the palaeogeotherm in each region. This problem then involves a discrete model (the number of partitions) coupled with a discretized continuous model (the thermal history in each partition). The thermal histories were of fixed dimensions, and parameterised as a series of nodes (representing a time–temperature point), and linear interpolation was used to predict the apatite fission track parameters.

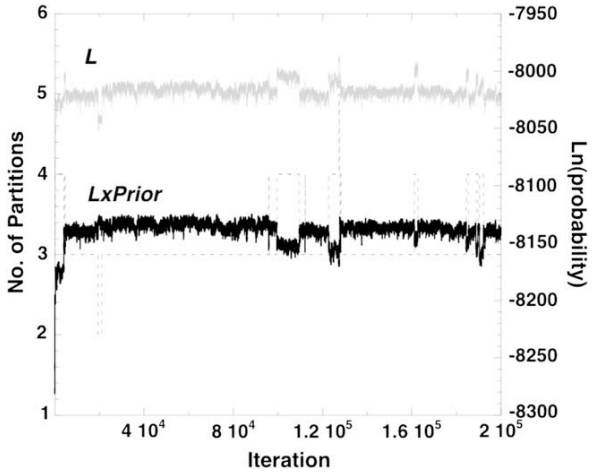

Fig. 10. Likelihood (L) and the product of the likelihood and prior (which is proportional to the posterior used in the MCMC acceptance criterion), and the number of partitions (dashed line) as a function of iteration during the MCMC sampling. This illustrates that although increasing the number of partitions (more complex model) improves the fit to the data (increases the likelihood), the increasing complexity is effectively penalised through the prior.

In Fig. 10, we show the likelihood, and the posterior (α likelihood $\times$ prior) as a function of iteration, together with the number of partitions for a problem with 20 samples with 4 discrete regions with different thermal histories. This illustrates the point made earlier about the prior effectively penalising the more complex models. Thus, a model with more partitions has more independent thermal histories, and is more likely to provide a better fit to the data (higher likelihood). However, the fact that we need more parameters leads to the higher dimensional prior acting to reduce the posterior, which is used to in the acceptance criterion during the sampling. In Fig. 11 we show the inferred distribution of the

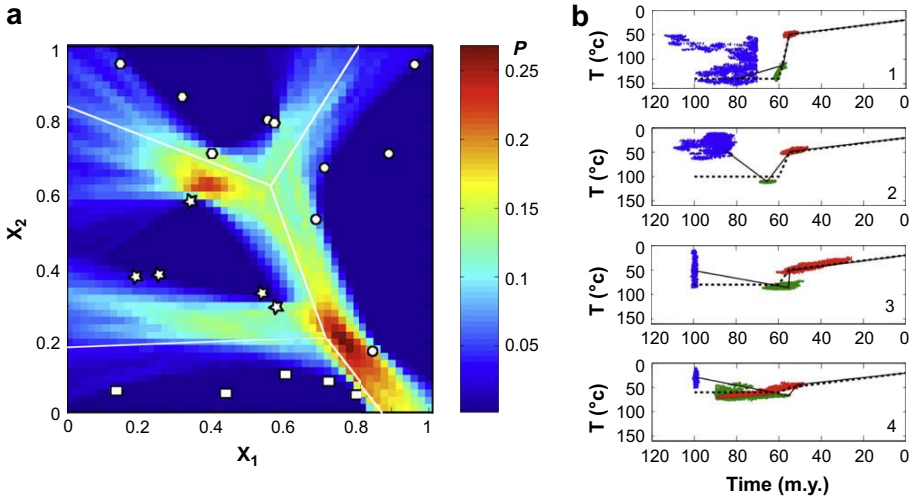

Fig. 11. (a) Probability density distribution of inferred partition boundaries and the input locations (white lines). The sample locations are the open circles. P is probability. (b) The inferred thermal history (solid line) and the input history (dashed line) in each partition, identified by the number in the bottom right corner. The clusters of points show the distribution of MCMC samples for each of the 3 nodes in the thermal history.

partitions, shown as a density distribution on the location of the boundaries, together with the maximum posterior and input thermal histories for each partition. The clouds of points around each node represent the sampling of the thermal histories, given the partition configuration, and then provide an estimate of the thermal history resolution. The 95% credible region is approximated by the distribution of the sampled points in 2D.

We recover the position of the partition boundaries and also the thermal history within each partition well. The former is clearly a function of the data locations, as a boundary can be placed anywhere between two clusters of locations and we will achieve the same result. This is reflected in density distribution of the partition boundaries. For example, the bottom right boundary is well constrained by a few relatively nearby locations, while the top right and bottom left are less well constrained as the locations are more spread out.

6. Concluding statements

We have presented a brief overview of Markov chain Monte Carlo methods, including the variable dimension approach, with a series of examples taken from our previous work. This is not a comprehensive review, and we have provided references with more detailed expositions on the theory and implementation. The attractive features of these sampling-based approaches as presented is that all we really need is to be able to solve the forward problem, and make a quantitative measure of how well we fit the observations (i.e. define an appropriate likelihood). There are different ways of implementing the sampling strategies, and from our experience, each problem requires some specific attention, particularly for the proposal functions. If the proposal function is too conservative (i.e. very small changes to a model) then the sampler moves very slowly around the model space and has an acceptance which is too high (a rule of thumb is about 25–30% of models should be accepted, Gilks et al., 1996; Brooks et al., 2003). If the proposal function is too disperse, i.e. proposes large relatively jumps in the model space, then many samples are rejected and the sampler effectively stalls.

Provided we tune the sampler appropriately, then the final collection of models is a good approximation of the joint probability distribution. As mentioned earlier, this allows us to derive useful summary statistics such as the expected or average model. Depending on the formulation on the individual models, the average model is generally smoother, but can also incorporate more complex features than any single model might be able to capture. Furthermore, we can quantify the uncertainty on the model parameters which allows us to determine predictive uncertainty, i.e. use the full range of models in forecasting in reservoir simulation or maturity modelling/prospect evaluation. Clearly some problems will be too computationally intensive to allow sufficient sampling, but often faster, simplified formulations are acceptable, given the uncertainty in many real world physical parameters. In the context of model choice, the Bayesian approach discussed here will naturally tend towards simpler models (at least in the context of hierarchical models with a finite number of parameters) as a consequence of the probability formulation.

In terms of comparing models with different formulations or different hypotheses, the normalising constant or evidence, $p(\boldsymbol{d})$, is one way of addressing this question. In principle, the evidence provides a means of choosing quantitatively between competing models or hypotheses. In practice, the evidence is difficult to calculate reliably for high dimensional problems. However, Skilling (2005) (and see also Sivia and Skilling, 2006) has suggested a novel method, nested sampling, to potentially deal with this in a practical way. Chopin and Robert (in press) have recently examined the underlying assumptions behind nested sampling and imply that the method has an approximation error, which increases as the dimension of the problem increases (the same as standard Monte Carlo methods). They propose some modifications and future work will no doubt establish the validity and practical applicability of such approaches.

Irrespective of the complications in calculating the evidence, MCMC allows us to generate samples from the posterior without knowing this normalising constant as we deal with ratios (and so this constant cancels out). As such, MCMC is a simple, but powerful sampling strategy to deal with complex modelling and inference problems, including those in which the dimensions of the model are not known a priori. One limitation is that we do need to solve the forward problem many times (thousands to millions), and therefore these methods will sometimes not be practical for problems with high computational overheads.

Acknowledgements

We would like to thank Total and NERC for financial support for some of this work. Chris Holmes, David Stephens and Ajay Jasra have provided many useful insights to many aspects of MCMC over the last few years. Also, we would like to acknowledge an anonymous reviewer and Manuel Nepvue for comments on the original manuscript.

References

Al-Awadhi, F., Hum, M., Jennison, C., 2002. Improving the acceptance rate of reversible jump MCMC proposals. Statistics and Probability Letters. 69, 189–198.

Ballester, P.J., Carter, J.N., 2007. A parallel real-coded genetic algorithm for history matching and its application to a real petroleum reservoir. Journal of Petroleum Science and Engineering 59, 157–158.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53, 370–418 (reprinted with biographical note by G.A. Barnard, in Biometrika, 45, 293–315).

Bernardo, J., Smith, A.F.M., 1994. Bayesian Theory. John Wiley and Sons, Ltd., Chichester.

Brooks, S.P., Giudici, P., Roberts, G.O., 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distribution. Journal of the Royal Statistical Society, Series B 65, 3–39. 90.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference – a practical information theoretic approach, second ed. Springer-Verlag, New York. 488.

Charvin, K., Gallagher, K., Hampson, G.J. A Bayesian approach to infer environmental parameters from stratigraphic data: method, Part 1. Basin Research, in press-a.

Charvin, K., Hampson, G.J., Gallagher, K. A Bayesian approach to infer environmental parameters from stratigraphic data, Part Ii: validation and sensitivity tests. Basin Research, in press-b.

Chopin, N., Robert, C.P. Contemplating evidence: properties, extensions of, and alternatives to nested sampling. Biometrika, in press.

Curtis, A., Wood, R. (Eds.), 2004. Geological Prior Information: Informing Science and Engineering. Geological Society, London, Special Publications, vol. 239.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M., 2002. Bayesian Methods for Nonlinear Classification and Regression. Wiley, Chichester.

Ferrero, C., Gallagher, K., 2002. Stochastic thermal history modeling, 1. Constraining heat flow histories and their uncertainty. Marine and Petroleum Geology 19, 633–648.

Gallagher, K., 1998. Inverse thermal history modelling as a hydrocarbon exploration tool. Inverse Problems 14, 479–497.

Gallagher, K., Morrow, D.W., 1998. A novel method for constraining heat flow histories in sedimentary basins. In: Düppenbecker, S.J., Illiffe, J.E. (Eds.), Basin Modelling: Practice and Progress. Geological Society London Special Publication, vol. 141, pp. 223–240.

Gallagher, K., Sambridge, M., 1992. The resolution of past heat flow in sedimentary basins from non-linear inversion of geochemical data: the smoothest model approach, with synthetic examples. Geophysical Journal International 109, 78–95.

Gallagher, K., Stephenson, J., Brown, R., Holmes, C., Fitzgerald, P., 2005. Low temperature thermochronology and modelling strategies for multiple samples 1: vertical profiles. Earth Planetary Science Letters 237, 193–208.

Gilks, W.R., Richardson, S., Spiegalhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (4), 711–732.

Green, P.J., 2001. A primer on Markov Chain Monte Carlo. In: Barndorff-Nielsen, O.L., Cox, D.R., Kluppelberg, C. (Eds.), Complex Stochastic Systems. Chapman and Hall.
Green, P.J., 2003. Trans-dimensional MCMC. In: Green, P.J., Hjort, N., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford Statistical Sciences Series, Ch.6, pp. 179–196.
Green, P.J., Mira, A., 2001. Delayed rejection in reversible jump metropolis-hastings. Biometrika 88, 1035–1053.
Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.
Hopcroft, P., Gallagher, K., Pain, C.C., 2007. Inference of past climate from boreholes using Bayesian reversible jump Markov chain Monte Carlo. Geophysical Journal International 171, 1430–1439.
Jasra, A., Stephens, D.A., Gallagher, K., Holmes, C.C., 2006. Analysis of geochronological data with measurement error using Bayesian mixtures. Mathematical Geology 38, 269–300.
Jaynes, E.T., 2003. Probability Theory – The Logic of Science. Cambridge University Press, p. 727.
Lee, P.M., 2004. Bayesian Statistics: An Introduction, third ed. Hodder Arnold. 368.
Lerche, I., 1990. Basin Analysis: Quantitative Methods, vol. 1. Academic Press. 562 pp.
Lerche, I., 1991. Basin Analysis: Quantitative Methods, vol. 2. Academic Press. 570 pp.
Lerche, I., Yarzab, R.F., Kendall, C.G., 1984. Determination of paleoheatflux from vitrinite reflectance data. Bulletin of the American Association Petroleum Geologists 68, 1704–1717.
Mackay, D.J.C., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press, p. 640.
Malinverno, A., 2000. A Bayesian criterion for simplicity in inverse problem parametrization. Geophysical Journal International 140, 267–285.
Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a non linear geophysical problem. Geophysical Journal International 151, 675–688.
Malinverno, A., Briggs, V.A., 2005. Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes. Geophysics 69, 1005–1016.
Malinverno, A., Parker, R.L., 2005. Two ways to quantify uncertainty in geophysical inverse problems. Geophysics 71, 15–27.
Mosegaard, K., Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. Journal of Geophysical Research 100 (B7), 12431–12447.
Nielsen, S.B., 1995. An upper limit to palaeoheat flow: theory and examples from the Danish Central Trough. Tectonophysics 244, 137–152.
Nielsen, S.B., 1996. Sensitivity analysis in thermal and maturity modelling. Marine and Petroleum Geology 13, 415–425.
Sambridge, M., Gallagher, K., Jackson, A., Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence. Geophysical Journal International 167, 528–542.
Scales, J.A., Snieder, R., 1997. To Bayes or not to Bayes. Geophysics 62, 1045–1046.
Scales, J.A., Tenorio, L., 2001. Prior information and uncertainty in inverse problems. Geophysics 66, 373–390.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.
Sivia, D., Skilling, J., 2006. Data Analysis: A Bayesian Tutorial. Oxford University Press, p. 246.
Skilling, J., 2005. Nested sampling. In: Bayesian Inference and Maximum Entropy Methods, AIP Conference Proceedings 735, New York.
Spiegalhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society 64, 583–639.
Stephenson, J., Gallagher, K., Holmes, C.C., 2004. Beyond kriging: dealing with discontinuous spatial data fields using adaptive prior information and Bayesian partition modelling. In: Curtis, A., Wood, R. (Eds.), Geological Prior Information: Informing Science and Engineering. Geological Society, London, Special Publications, vol. 239, pp. 195–209.
Stephenson, J., Gallagher, K., Holmes, C., 2006a. A Bayesian approach to calibrating apatite fission track annealing models for laboratory and geological timescales. Geochimica Cosmochimica Acta 70, 5183–5200.
Stephenson, J., Gallagher, K., Holmes, C., 2006b. Low temperature thermochronology and modelling strategies for multiple samples 2: partition modelling for 2D and 3D distributions with discontinuities. Earth Planetary Science Letters 241, 557–570.
Tarantola, A., 2005. Inverse Problem Theory and Methods for Models Parameter Estimation. SIAM. 358p.
Tarantola, A., Valette, B., 1982. Inverse problems = quest for information. Journal of Geophysics 50, 159–170.