*Geophys. J. Int.* (2006) **167,** 528–542 doi: 10.1111/j.1365-246X.2006.03155.x



# Trans-dimensional inverse problems, model comparison and the evidence

M. Sambridge,[1] K. Gallagher,[2] A. Jackson[3] and P. Rickwood[1]

[1]*Research School of Earth Sciences, Australian National University, Canberra,* ACT 0200, *Australia. E-mail: malcolm.sambridge@anu.edu.au*
[2]*Department of Earth Science and Engineering, Imperial College London, London,* SW7 2BP, *UK*
[3]*Institut für Geophysik, ETH Zürich,* CH-8093 *Zürich, Switzerland*

Accepted 2006 July 26. Received 2006 July 19; in original form 2005 November 9

## SUMMARY

In most geophysical inverse problems the properties of interest are parametrized using a fixed number of unknowns. In some cases arguments can be used to bound the maximum number of parameters that need to be considered. In others the number of unknowns is set at some arbitrary value and regularization is used to encourage simple, non-extravagant models. In recent times variable or self-adaptive parametrizations have gained in popularity. Rarely, however, is the number of unknowns itself directly treated as an unknown. This situation leads to a trans-dimensional inverse problem, that is, one where the dimension of the parameter space is a variable to be solved for.

This paper discusses trans-dimensional inverse problems from the Bayesian viewpoint. A particular type of Markov chain Monte Carlo (MCMC) sampling algorithm is highlighted which allows probabilistic sampling in variable dimension spaces. A quantity termed the evidence or marginal likelihood plays a key role in this type of problem. It is shown that once evidence calculations are performed, the results of complex variable dimension sampling algorithms can be replicated with simple and more familiar fixed dimensional MCMC sampling techniques. Numerical examples are used to illustrate the main points. The evidence can be difficult to calculate, especially in high-dimensional non-linear inverse problems. Nevertheless some general strategies are discussed and analytical expressions given for certain linear problems.

**Key words:** evidence, inverse problems, model comparison, parametrization.

## 1 INTRODUCTION

The study of inverse problems has a long history in the geosciences, dating back to the pioneering work of Backus & Gilbert (1967, 1968, 1970). Over the past 30 years there has been a strong focus on estimating parameters, that is, building models of the Earth which satisfy data and are in some sense 'close' to the real Earth, or have properties in common with it. The usual approach is to choose some suitable parametrization of the physical property of interest and then use the data to estimate its free parameters. If a suitable set of parameters exist and can be found, then an earth model has been built and there is a temptation to try and interpret its features. Geophysicists, however, have been aware since the work of Backus and Gilbert that uniqueness is not guaranteed (and indeed non-uniqueness almost always is). Furthermore Occam's razor suggests that simple models should be preferred over complex ones (Constable *et al.* 1987; Parker 1994), and so the focus is therefore on building models with as few degrees of freedom as necessary to fit the data. In many data-fitting problems it is common practice to minimize the number of unknowns required to fit the data. A range of statistical techniques have been developed for judging whether the introduction of extra unknowns is warranted by the data, for example, $F$-tests. One question that has received much less attention in the geosciences is asking whether the data itself can provide information on the number of unknowns, that is, treating the number of unknowns as one of the unknowns. This is what we mean by a trans-dimensional inverse problem.

In contrast to developments in the geosciences, Bayesian statisticians have considered the problem of a variable numbers of unknowns for some time, and more recently proposed innovative methods for its solution. The Bayesian, or probabilistic approach to inverse problems has its detractors. Central to the Bayesian approach is the idea that the state of knowledge about a set of unknowns can be described by a probability density function (PDF). Non-Bayesians would argue that this is inappropriate because the physical properties of interest are not random variables (there is only one Earth). Furthermore the dependence of Bayesian techniques on subjective prior information is a disadvantage and potentially unsafe, since extra information is injected into a problem that may not be intended or desired (see Backus 1988b, for an example). The Bayesian response might be that it is our state of knowledge about the unknowns which is being described by a PDF and the not the variables

themselves. Statisticians like Denison *et al.* (2002) readily acknowledge that there is no natural way to encode complete ignorance about model parameters within a Bayesian formulation, but many also argue that one nearly always has some form of prior information, however weak. It has even been claimed that the Bayesian approach is really just a mathematical formulation of logical scientific reasoning (Jaynes 2003). For a more expansive summary of the issues and arguments see (Backus 1988a; Scales & Snieder 1997; Scales & Tenorio 2001; Jaynes 2003).

While Bayesian and non-Bayesian approaches often result in the same answer for a data inference problem, what is not widely appreciated is that Bayesian methods offer novel solutions to the model comparison problem (Bernardo & Smith 1994; Denison *et al.* 2002; Mackay 2003). This is where two or more alternate ways of explaining data are compared, each being based on differing hypotheses, formulations or perhaps involving different assumptions in the model building process. Bayesian techniques allow quantitative assessment of the level of the support provided to each hypothesis by the data.

In this paper we concentrate on a particular example of the comparison problem. Specifically we discuss Bayesian approaches to trans-dimensional inverse problems, that is, where the number of unknowns is itself an unknown. We describe some probabilistic sampling approaches for dealing with this problem that have become popular in the statistical literature. We show how these trans-dimensional sampling techniques may be replicated using more familiar techniques for fixed dimensional sampling. In this discussion we highlight the role of the 'evidence' or 'marginal likelihood', a quantity which has largely been ignored in the geosciences and elsewhere. We argue, in agreement with others, that this is an oversight and that the evidence has a number of important uses, both in trans-dimensional inverse problems and more generally for model comparison. We highlight the role of the evidence and its interpretation through some simple but illustrative examples. We also discuss ways of calculating it for fully non-linear problems with few unknowns and for linear or linearizable problems with many unknowns.

## 2 BAYESIAN INFERENCE AND SAMPLING

Bayesian inference centres on the use of Bayes' theorem (Bayes 1763) which can be written as

$$p(\mathbf{x} \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid \mathbf{x}) p(\mathbf{x})}{p(\mathbf{d})}, \tag{1}$$

where $p(\mathbf{x} \mid \mathbf{d})$ is the *a posteriori* probability density of a vector of unknowns $\mathbf{x}$ given the data $\mathbf{d}$; $p(\mathbf{d} \mid \mathbf{x})$ is the likelihood of observing data $\mathbf{d}$ given a particular $\mathbf{x}$, and $p(\mathbf{x})$ is the *a priori* probability density of $\mathbf{x}$, that is, what we know about $\mathbf{x}$ before measuring the data $\mathbf{d}$. (For brevity the terms 'prior' and 'posterior' are often used to refer to these two probability density functions.) Note that the vertical bar in terms like $p(\mathbf{x} \mid \mathbf{d})$ indicate conditional dependence, which means that quantities to the left are variable and to the right are fixed. To make quantitative use of (1) knowledge of data error statistics is required, as well as prior information in the form of a probability density function. In geophysical inverse problems data errors may arise from a multitude of sources and their proper characterization in terms of a single PDF may be difficult, or impossible. Similarly representation of all prior knowledge in terms of PDFs can be problematic. In cases where some simple form of likelihood in (1) is appropriate but its defining parameters (e.g. variance) are unknown, then statisticians often make use of a hierarchical construction. In this case parameters controlling the type of prior and likelihood are themselves treated as variables (known as hyper-parameters) and assigned a prior PDF (see Malinverno & Briggs 2004, for a discussion and examples). In geophysical studies the importance of geological prior information has been widely recognized and some innovative ways of incorporating it into Bayesian methods have been proposed (Mosegaard & Tarantola 1995; Curtis & Lomax 2001; Curtis & Wood 2004).

In a similar vein, uncertainty in the forward problem can be incorporated into the inference problem by assigning PDFs associated with variables in the forward problem that are uncertain. As with the hyper-parameters, it is usually most convenient to marginalize these parameters out. An early geophysical example of such a procedure is given in Jackson (1995).

The denominator in (1) is the integral of the likelihood times the prior over the model space. The term $p(\mathbf{d})$ normalizes the posterior

$$p(\mathbf{d}) = \int p(\mathbf{d} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \tag{2}$$

and has been called the marginal likelihood or 'evidence'. In words (1) may be written

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{3}$$

Note from (2) that the evidence is not directly a function of the model parameters $\mathbf{x}$ because they are 'integrated out'. Possibly for this reason, and also that it can be difficult to calculate, it has largely been ignored in the treatment of inverse problems, especially within geophysics (see for example Tarantola 2005). Recently, however, there have been claims that this is an oversight and indeed the evidence is a useful quantity that should be calculated for all inverse problems (Malinverno 2000; Skilling 2004). In this paper we discuss the evidence further and show how it plays a crucial role in the model selection problem, which arises when deciding between competing theories or choices of parametrization in an inverse problem.

The primary objective of Bayesian inference is to learn about the model $\mathbf{x}$ from the data $\mathbf{d}$. Statisticians often do this through some form of sampling procedure, that is, draw random samples $\mathbf{x}_i^*$, $(i = 1, \ldots, n)$ whose density distribution follows that of the posterior. With the samples in hand, one can estimate any quantities of interest. Common choices are the model $\mathbf{x}_{\mathrm{MAP}}$ which maximizes the posterior in the model space. Others are the expected model,

$$E\{\mathbf{x}\} = \int \mathbf{x} p(\mathbf{x} \mid \mathbf{d})\, d\mathbf{x} \approx \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^*, \tag{4}$$

the model covariance matrix

$$C_{i,j} = \int x_i x_j p(\mathbf{x} \mid \mathbf{d})\, d\mathbf{x} - E\{x_i\} E\{x_j\}, \tag{5}$$

and marginals of $\mathbf{x}$

$$M(x_i) = \int \ldots \int p(\mathbf{x} \mid \mathbf{d}) \prod_{\substack{j=1 \\ j \neq i}}^{k} dx_j, \tag{6}$$

where $k$ is the number of unknowns represented by the vector $\mathbf{x}$, and $x_j$ is the $j$th component of $\mathbf{x}$. If samples from the posterior are available then all three types of quantity can be easily calculated through ensemble averages, as shown in (4) (see Sambridge 1999, for examples). The primary issue for statisticians is to design algorithms that efficiently sample the posterior and allow accurate estimation of the integrals (4)–(6) for a given number of samples, $n$.

In geophysics one is often concerned with model estimation, that is, finding some optimal model from data. In the special case where the likelihood and prior take Gaussian forms, and the data model relationship is linear, the posterior $p(\mathbf{x} \mid \mathbf{d})$ is a multidimensional Gaussian where the most likely model ($\mathbf{x}_{\mathrm{MAP}}$) equals the average or expected model ($E\{\mathbf{x}\}$). In this case some algebra shows that (4) and (5) correspond to the familiar least-squares solution and model covariance matrix, respectively (e.g. see Tarantola & Valette 1982; Tarantola 2005). While statisticians favour sampling from posteriors, geophysicists have tended to use optimization algorithms to locate single optimal solutions and then calculated model covariance matrices using local (derivative) information about the solution (see Menke 1989; Aster *et al.* 2005). The latter has the advantage of being practical when the number of unknowns is high (say $10^2$–$10^6$), but requires the inverse problem to be linear or linearizable. For inverse problems with fewer unknowns, say $10^2$–$10^3$, Bayesian sampling of posterior PDFs is a viable approach and has been used widely in geophysical studies (see Mosegaard & Sambridge 2002, for a review).

### 2.1 Model comparison and the evidence

The model comparison problem arises when two or more competing theories, or hypotheses, are to be tested against data. (Note that here the term 'model' is used to describe a mathematical formulation rather than a vector of unknowns). For example, in seismology a hypothesis might be that the global database of teleseismic traveltime phases can be adequately fit assuming seismic wave speed varies only with depth. In this case a competing hypothesis might be that lateral variations in wave speed are needed to fit the data. In geochronology, two competing hypotheses may involve differing numbers of geological components needed to fit a set of age measurements in a rock. The common theme in the hypotheses considered here is that they each have an associated number of parameters which differ between competing hypotheses.

If two hypotheses are labelled $\mathcal{H}_1$ and $\mathcal{H}_2$ then Bayes' theorem can be used to determine the relative plausibility of each given the data. We have

$$\frac{p(\mathcal{H}_1 \mid \mathbf{d})}{p(\mathcal{H}_2 \mid \mathbf{d})} = \frac{p(\mathbf{d} \mid \mathcal{H}_1)p(\mathcal{H}_1)}{p(\mathbf{d} \mid \mathcal{H}_2)p(\mathcal{H}_2)}. \tag{7}$$

In words (7) says that the posterior ratio equals the evidence ratio times the prior ratio. The prior ratio measures how much we favour one hypothesis over the other before we collect any data. This may or may not be unity depending on the problem in hand. The evidence ratio measures how well the two theories predict the data. As the left-hand side increases we prefer $\mathcal{H}_1$ over $\mathcal{H}_2$, given the data $\mathbf{d}$.

Hypotheses are often introduced in the initial formulation of an inverse problem, for example, in choosing an appropriate parametrization for $\mathbf{x}$, or by using simplifying approximations in the physics relating $\mathbf{d}$ and $\mathbf{x}$. Therefore, more strictly each term in (1) and (2) should have been written with a conditional dependence on the underlying hypotheses, that is,

$$p(\mathbf{x} \mid \mathbf{d}, \mathcal{H}) = \frac{p(\mathbf{d} \mid \mathbf{x}, \mathcal{H})p(\mathbf{x} \mid \mathcal{H})}{p(\mathbf{d}|\mathcal{H})}, \tag{8}$$

and

$$p(\mathbf{d} \mid \mathcal{H}) = \int p(\mathbf{d} \mid \mathbf{x}, \mathcal{H})p(\mathbf{x} \mid \mathcal{H})d\mathbf{x}. \tag{9}$$

In this case the vector $\mathbf{x}$ represents parameters resulting from hypothesis $\mathcal{H}$. We see that the likelihood term in (7) is the marginal likelihood or evidence given by (9), which is obtained by integrating the posterior over the model parameters. Hence $p(\mathbf{d} \mid \mathcal{H})$ measures how well the hypothesis $\mathcal{H}$ explains the data taking all possible combinations of its parameters into account. From (7) we see then that the ratio of the evidence values is the factor which converts the prior support between two competing theories to the posterior support, and in this way tells us whether the data have increased or decreased the support for $\mathcal{H}_1$ over $\mathcal{H}_2$. In the statistical literature this term is often called the *Bayes factor* (For discussions see Aitkin 1991; Bernardo & Smith 1994; Kass & Raftery 1995; Denison *et al.* 2002).

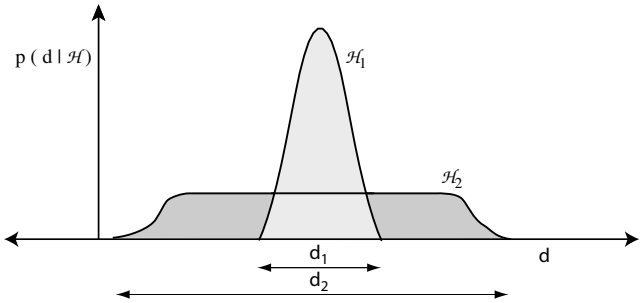


**Figure 1.** An illustration of natural parsimony explained in terms of predictive power of two hypotheses, $\mathcal{H}_1$ and $\mathcal{H}_2$. The horizontal axis is the range of the data space and the vertical axis is the value of the evidence $p(\mathbf{d} \mid \mathcal{H})$ Here the simpler theory, $\mathcal{H}_1$ makes precise predictions over a limited range of data ($d_1$), while the more complex theory $\mathcal{H}_2$ has a larger range in data space ($d_2$). Due to normalization the predictive capability of $\mathcal{H}_2$ has lower amplitude in the data range predicted by both theories. After Mackay (2003).

A property of Bayesian inference, perhaps not widely recognized within geophysics, is that of 'natural parsimony'. This means that it incorporates Occam's razor, that is, given a choice between a simple and complex model that provide similar fits to data, it will favour the simpler one. Note that this is without any preference for the simpler model being expressed in the prior on the number of unknowns. (Mackay 2003, Ch. 28) explains Bayesian parsimony by noting that simple models tend to make precise predictions while complicated ones can make a greater variety of predictions. If two competing theories are available, as in (7), and $\mathcal{H}_2$ is more complicated (has more free parameters) than $\mathcal{H}_1$, then $\mathcal{H}_2$ will be able to fit a wide range of data values and after normalization will have on average a lower predictive probability in the range covered by both theories (see Fig. 1). This suggests that the evidence will tend to be higher for the simpler model, $\mathcal{H}_1$, and lower for the more flexible model $\mathcal{H}_2$. We see then that the key factor which incorporates Occam's principle into Bayesian inference is the evidence.

Skilling (2004, 2005) has argued that the evidence is a useful transportable quantity, in that two different analyses of data (i.e. involving different assumptions) may be performed years apart, but so long as the evidence values are available the results may be quantitatively compared. The study with the higher evidence corresponds to the more successful fitting of the data. Examples of the parsimonious nature of Bayesian inference can be seen in the numerical results presented later in this paper. For more detailed discussions of parsimony and examples the reader is referred to Malinverno (2000, 2002), Denison *et al.* (2002) and Mackay (2003).

In contrast to model estimation, techniques for model comparison are much less used by geophysicists. However, since inversion studies are always based on assumptions, (e.g. to simplify the physics of the forward problem, or to parametrize the inverse problem with a fixed number of unknowns) then in principle one

might expect Bayesian model comparison to find many applications. Malinverno (2000), Malinverno & Briggs (2004) have argued along similar lines and showed how the evidence term in (7) may be used to perform model comparison over differing parametrizations in an inverse problem. In this paper we concentrate on a particular example of model comparison, the case of choosing the number of unknowns in the parametrization of an inverse problem. The philosophy is on letting the data decide between models with differing numbers of unknowns, rather than having the choice fixed in advance. This has been an area of growing interest within the Earth Sciences, for example through adaptive parametrizations in seismic tomography (Sambridge & Rawlinson 2005), and thermochronology (Stephenson *et al.* 2006), as well as in mixture modelling in geochronology (Jasra *et al.* 2006). In addition the Bayesian sampling algorithm described below has already begun to make its mark in both seismic and electrical resistivity sounding problems where the numbers of layers in a 1-D model is treated as an unknown (see Malinverno 2000; Malinverno & Leaney 2005, for details).

## 3 TRANS-DIMENSIONAL INVERSE PROBLEMS

We define a trans-dimensional inverse problem as one where the dimension of the model space, $k$, is an unknown. Here we do not distinguish between the cases where $k$ is a measurable physical parameter, for example, the number of mineralogical components in a rock, and where it merely represents the number of basis functions chosen for a model. Some would argue that it is inappropriate to apply Bayesian methodology to the latter case, as a prior on $k$ might have little meaning and be difficult to quantify. These points are certainly debatable; however, in both cases common practice in geophysics is to fix the number of unknowns either arbitrarily, or using physical considerations. We view the trans-dimensional approach described here as a natural extension of common practice, that is, where the hard constraint of imposing a fixed $k$ is relaxed and one asks the data to decide on the number of unknowns.

To formulate the problem from a Bayesian viewpoint we use the conditional bar notation, '|' discussed above. For example, if we let $k$ be the number of unknowns in the model vector $\mathbf{x}$ then we can rewrite (1) and (2) as

$$p(\mathbf{x} \mid k, \mathbf{d}) = \frac{p(\mathbf{d} \mid \mathbf{x}, k) p(\mathbf{x} \mid k)}{p(\mathbf{d} \mid k)}, \tag{10}$$

and

$$p(\mathbf{d} \mid k) = \int p(\mathbf{d} \mid \mathbf{x}, k) p(\mathbf{x} \mid k) d\mathbf{x}, \tag{11}$$

to indicate that this is for the case of a fixed $k$-dimensional model space. A property of PDFs is

$$p(x, y) = p(x \mid y) p(y), \tag{12}$$

which says that the joint probability density for variables $x$ and $y$ is equal to the probability of $x$ given a particular value of $y$, times the probability of that value of $y$ occurring. Using (12) we have

$$p(\mathbf{x}, k \mid \mathbf{d}) = p(\mathbf{x} \mid k, \mathbf{d}) p(k \mid \mathbf{d}), \tag{13}$$

and from Bayes' theorem applied to the dimension we also have

$$p(k \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid k) p(k)}{p(\mathbf{d})}. \tag{14}$$

Combining (10), (13) and (14) we get

$$p(\mathbf{x}, k \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid k, \mathbf{x}) p(\mathbf{x} \mid k) p(k)}{p(\mathbf{d})}, \tag{15}$$

which is Bayes' theorem for the variable dimension problem. Here the left-hand side is the joint posterior for the vector of unknowns $\mathbf{x}$ and its dimension $k$, while $p(k)$ is the prior on the number of unknowns; and $p(\mathbf{d})$ is the normalization factor (or total evidence ). The variable dimension posterior in (15) can be used for Bayesian inference in inverse problems where 'one of the unknowns is the number of unknowns'. Finding ways of sampling from trans-dimensional posteriors (15) has been an active area of research in statistics culminating with the breakthrough papers of Geyer & Møller (1994) and Green (1995). The latter introduced what is became known as the reversible jump Markov chain Monte Carlo (MCMC) algorithm. This extended the familiar MCMC method for sampling a fixed dimensional space into one for a general trans-dimensional problem. To understand the reversible jump algorithm first requires a discussion of the fixed dimension case.

### 3.1 Fixed dimensional MCMC

The modern workhorse technique for sampling of arbitrary (fixed dimension) posterior PDFs as in (10) has been the MCMC algorithm. (For summaries see Smith 1991; Gelfand & Smith 1990; Smith & Roberts 1993). This has been used extensively for Bayesian approaches to geophysical inverse problems (Mosegaard & Sambridge 2002). To generate independent samples from an arbitrary posterior, a random walk is performed. At each step in the chain a move is proposed from a current model $\mathbf{x}^p$ to a new model $\mathbf{x}^q$ which is either accepted or rejected. The new model $\mathbf{x}^q$ is generated from $\mathbf{x}^p$ in a probabilistic manner using a (chosen) proposal distribution, which we write as $q(\mathbf{x}^q \mid \mathbf{x}^p)$. (As before terms to the right of the bar are fixed and to the left are variable.) The model, $\mathbf{x}^q$ is then accepted with probability, $\alpha$, given by

$$\alpha = \mathrm{Min}\left[1, \frac{p(\mathbf{x}^q \mid k, \mathbf{d})\, q(\mathbf{x}^p \mid \mathbf{x}^q)}{p(\mathbf{x}^p \mid k, \mathbf{d})\, q(\mathbf{x}^q \mid \mathbf{x}^p)}\right]. \tag{16}$$

In practice this means we generate a uniform random number between zero and one ($U[0, 1]$) and take the step from $\mathbf{x}^p$ to $\mathbf{x}^q$ only if $u < \alpha$. If $u > \alpha$ then the model position is rejected, and the algorithm stays at $\mathbf{x}^p$. Clearly the acceptance value, $\alpha$, is at the centre of the whole algorithm. Expanding (16) we have

$$\alpha = \mathrm{Min}\left[1, \frac{p(\mathbf{d} \mid \mathbf{x}^q, k)\; p(\mathbf{x}^q \mid k)\; q(\mathbf{x}^p \mid \mathbf{x}^q)}{p(\mathbf{d} \mid \mathbf{x}^p, k)\; p(\mathbf{x}^p \mid k)\; q(\mathbf{x}^q \mid \mathbf{x}^p)}\right]. \tag{17}$$

This is known as the Metropolis–Hasting rule (Metropolis & Ulam 1949; Hastings 1970). The first term in the quotient (16) is the ratio of the posterior probability of the proposed point, $\mathbf{x}^q$, to the current point, $\mathbf{x}^p$; while the second term is the ratio of the proposal probabilities.

Note that if the proposal distribution is chosen to be symmetric (the usual case), then the third term in (17) cancels. Also if the proposed model $\mathbf{x}^q$ has a higher value of the posterior than the starting model $\mathbf{x}^p$, then $\alpha$ will be equal to one. Hence a step to a model that increases the posterior will always be accepted, and a step to a lower value will sometimes be accepted, depending on the random number $u$ and ratio of the posterior PDF values. The 'output distribution' of the algorithm is made up of the set of models at the end of each step, or a fixed number of steps if 'chain thinning' is used. (The latter being the process by which samples are collected only intermittently from the chain, to reduce unwanted correlation between samples, see Smith (1991), Smith & Roberts (1993).) Normally a 'burn in' period is also employed. Which means that the early models entering into the output distribution are thrown away, because they are biased by the starting position chosen for the random walk.

It can be proven that using the Metropolis–Hastings rule the random walk relaxes to the target distribution $p(\mathbf{x} \mid k, \mathbf{d})$, that is, the density of the output population of samples will asymptotically follow the fixed dimensional posterior. In practice multiple random walks can be used from differing starting points and their results combined.

Since $\alpha$ in (16) only depends on the ratio of the posteriors between any two points in the model space, $p(\mathbf{x} \mid k, \mathbf{d})$ need only be known up to a multiplicative constant. This is of practical importance and means that the evidence term in (1) need not be evaluated in order to make use of MCMC sampling, which in part explains why it has been ignored so often. The choice of proposal distribution, $\mathbf{x}^p \rightarrow \mathbf{x}^q$, affects the efficiency of the algorithm, that is, how many steps are required before it relaxes to the target posterior, $p(\mathbf{x} \mid k, \mathbf{d})$. Efficient and practical choices of proposal distribution are the subject of much research (Godsill 2003). A simple choice is to perturb each component of $\mathbf{x}$ in succession in a fixed range,

$$x_j^q = x_j^p + u_j \Delta x_j, \quad (j = 1, \ldots, k), \tag{18}$$

where $x_j^p$ and $x_j^q$ are the components of $\mathbf{x}^p$ and $\mathbf{x}^q$, respectively, $u_j$ is a $U[0, 1]$ random number and $\Delta x_j$ is a constant scale factor (e.g. the prior range) for the $j$th component of $\mathbf{x}$. A more complicated proposal is to use a multivariate Gaussian

$$\mathbf{x}^q = N[\mu, M], \tag{19}$$

where $(\boldsymbol{\mu}, M)$ are the mean and covariance matrix, respectively, of the multivariate Gaussian. (Note that the centre of the Gaussian need not be at the current point $\mathbf{x}^p$, nor aligned with the axes.) In cases (18) and (19) the new point $\mathbf{x}^q$ would be accepted probabilistically using (16). Yet another class of move is to again cycle through each component of $\mathbf{x}$ but set $q(\mathbf{x}^q \mid \mathbf{x}^p)$ equal to the conditional of the posterior along the axis through that point, resulting in an algorithm known as the Gibbs sampler (see Smith 1991). In this case $\alpha$ becomes unity and hence the move is always accepted, but the cost is that the posterior needs to be evaluated, along the entire axis through $\mathbf{x}^p$. (In geophysics the Gibbs sampler was used by Rothman (1985, 1986) for global optimization in a residual statics problem.) Note that if the proposal distribution were equal to the posterior

$$q(\mathbf{x}^q \mid \mathbf{x}^p) = p(\mathbf{x}^q \mid k, \mathbf{d}), \tag{20}$$

then $\alpha$ would always equal 1, which would be ideal since all steps would be accepted. However, this defeats the object of the procedure in the first place, since a pre-requisite for using Metropolis–Hastings is that we must have an algorithm to generate new proposal points $\mathbf{x}^q$ from $q(\mathbf{x}^q \mid \mathbf{x}^p)$. Mosegaard & Tarantola (1995) discuss these issues further, and provide geophysical examples of the use of (16).

### 3.2 Reversible jump MCMC

The reversible jump algorithm of Green (1995) extends the use of the Metropolis–Hastings rule to cases where the proposal distribution not only moves a point within the current model (or state) space, but also between state spaces, of different dimension or type, that is, the step can include movement from a vector $\mathbf{x}^p$ of length $k$ to a vector $\mathbf{x}^q$ of length $k'$. Fig. 2 illustrates schematically the general situation of moving between each pair of state spaces.

In the reversible jump algorithm the proposal distribution in (16) is replaced with a two step procedure. We first generate $r$, random numbers (represented by the vector $\mathbf{u}$), using a chosen distribution $g(\mathbf{u})$, and then calculate the proposed model $\mathbf{x}^q$, using $\mathbf{u}$ and the current model $\mathbf{x}^p$ and some chosen (one to one invertible) function, $h$,

$$\mathbf{x}^q = h(\mathbf{x}^p, \mathbf{u}). \tag{21}$$

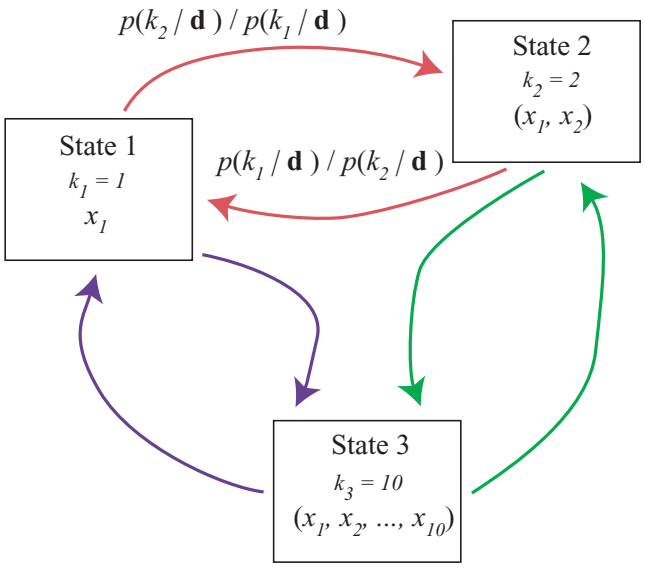


**Figure 2.** A schematic illustration of the reversible jump MCMC algorithm sampling across three independent state spaces. The number of variables in each ($k_j$, $j = 1, \ldots, 3$) differs between states. In general the meaning of each variable, $x_j$, could differ as well between states, but here only a simple addition and subtraction is shown. The arrows represent the jumps between states. The average number of successful jumps is controlled by the ratio of the total posterior probability in each state.

The only restriction on the transformation $h$ is that it is a diffeomorphism (i.e. both $h$ and its inverse are differentiable). Note that this formulation is quite general and each of the simple proposal distributions in (18) and (19), for the fixed dimension sampling, correspond to particular choices for $h$, $r$ and $g(\mathbf{u})$. For example, in (18) the right-hand side corresponds to the transformation function $h(\mathbf{x}^p, \mathbf{u})$, the distribution for the random number generation, $g(\mathbf{u})$ is uniform, $U[0, 1]$; and the number of random numbers generated, $r = k$. The introduction of the transformation (21) may seem unnecessarily awkward; however, the real power is seen when one realizes that now $\mathbf{x}^p$ need not be the same dimension as $\mathbf{x}^q$. Furthermore the underlying parametrization associated with $\mathbf{x}^p$ and $\mathbf{x}^q$ need not be the same either. With (21) we may step between any two parameter spaces, that is, $\mathbf{x}^q$ may have more, the same, or less variables than $\mathbf{x}^p$ and each component of $\mathbf{x}^q$ may be a function of all components of $\mathbf{x}^p$. Note that if between any two states the dimension changes by the simple addition of new components then a random number is generated for each new variable and hence $r = k' - k$, that is, one random number is required for each new component. Green (1995) showed that within this general framework the Metropolis–Hasting rule becomes

$$\alpha = \text{Min}\left[1, \frac{p(\mathbf{x}^q, k' \mid \mathbf{d})\, g'(\mathbf{u}^q)}{p(\mathbf{x}^p, k \mid \mathbf{d})\, g(\mathbf{u}^p)} |J|\right], \tag{22}$$

where $g(\mathbf{u}^p)$ and $g'$ $(\mathbf{u}^q)$ are the PDFs of the random numbers used in the forward step $\mathbf{x}^p \rightarrow \mathbf{x}^q$ and the corresponding reverse step $\mathbf{x}^q \rightarrow \mathbf{x}^p$, respectively. More specifically we can write the forward and reverse transformations as,

$$\mathbf{x}^q = h(\mathbf{x}^p, \mathbf{u}^p) \tag{23}$$

and

$$\mathbf{x}^p = h'(\mathbf{x}^q, \mathbf{u}^q) \tag{24}$$

where random vectors $\mathbf{u}^p$ and $\mathbf{u}^q$ are of size $r$ and $r'$, respectively, and we have $k + r = k' + r'$. (Note: in the common case of state $\mathbf{x}^q$ being an addition of one extra variable to $\mathbf{x}^p$ then $r = 1, r' = 0$.) The Jacobian of this transformation is needed in the Metropolis–Hastings rule (22). Since the transform and its inverse in (23) and (24) exist by definition, then the Jacobian also exists and is given by

$$|J| = \left| \frac{\partial(\mathbf{x}^q, \mathbf{u}^q)}{\partial(\mathbf{x}^p, \mathbf{u}^p)} \right| \tag{25}$$

The matrix $J$ is of size $(k + r) \times (k' + r')$, and its role is to account for the scale changes in the transformation from $\mathbf{x}^p \rightarrow \mathbf{x}^q$. Note that the posterior in (22) is now the variable dimension posterior given by (15) and not the fixed dimension one used previously. Since the more general formalism describes both the fixed and variable dimension cases, (22) reduces to (16) with appropriate choices of $(h, r, g(\mathbf{u}))$. For more discussion on the details of the algorithm and form of the Jacobian in special cases the reader is referred to Denison *et al.* (2002) and Green (2003).

As an example, imagine that we have three possible states with 1, 2 and 10 unknowns, respectively. To simplify notation we use $k = 1$, 2, 10 to label the states as well as represent the number of unknowns in each state. (In general, however, more than one state may have the same number of unknowns.) If the relationship between states $k$ and $k'$ is a simple addition of variables (i.e. $k' > k$), then the forward transformation (23) from model $\mathbf{x}^p$ (of length $k$) to model $\mathbf{x}^q$ (of length $k'$) is given by

$$x_i^q = x_i^p, \quad (i = 1, \ldots, k) \tag{26}$$

$$x_i^q = u_{i-k}^p \, \Delta x \quad (i = k + 1, \ldots, k'), \tag{27}$$

where $x_i^p$ is the $i$th component of $\mathbf{x}^p$, $\Delta x$ is a constant scale length associated with each state (e.g. the range of allowed values of $x_i$) and $u_j^p$ is a uniform random number $U[0, 1]$. The reverse transformation (24) is then given by

$$x_i^p = x_i^q, \quad (i = 1, \ldots, k). \tag{28}$$

In this case $r$ is the number of random numbers required in the forward transformation

$$r = k' - k, \tag{29}$$

and $r'$ is the number of random numbers needed in the reverse transformation, which is zero. Hence $J$ is a matrix of size $k' \times k'$. Some simple algebra shows that the Jacobian of the forward transformation (27) is given by

$$|J| = (\Delta x)^r = \frac{(\Delta x)^{k'}}{(\Delta x)^k}. \tag{30}$$

If we let the prior for $\mathbf{x}$ be a constant over the interval $\Delta x$ in all variables, then we have

$$p(\mathbf{x} \,|\, k) = \frac{1}{(\Delta x)^k}, \quad (k = 1, 2, 10), \tag{31}$$

and if we also let the prior for the dimension, $k$ be

$$p(k) = \frac{P_k}{(P_1 + P_2 + P_{10})}, \tag{32}$$

where $P_k$ is the prior probability for dimension $k$. Using (15) we see the posterior ratio for the jump from $k$ to $k'$ is

$$\frac{p(\mathbf{x}, k' \,|\, \mathbf{d})}{p(\mathbf{x}, k \,|\, \mathbf{d})} = \frac{p(\mathbf{d} \,|\, \mathbf{x}^q, k')(\Delta x)^k P_{k'}}{p(\mathbf{d} \,|\, \mathbf{x}^p, k)(\Delta x)^{k'} P_k}. \tag{33}$$

If the forward and reverse proposal distributions $g(\mathbf{u})$, $g'(\mathbf{u})$ are the same, then inserting (13) and (20) into (22) we get the acceptance probability for the jump

$$\alpha = \text{Min}\left[1, \frac{p(\mathbf{d} \,|\, \mathbf{x}^q, k')\,(\Delta x)^k \; P_{k'}\,(\Delta x)^{k'}}{p(\mathbf{d} \,|\, \mathbf{x}^p, k)\,(\Delta x)^{k'} \; P_k\,(\Delta x)^k}\right], \tag{34}$$

which simplifies to

$$\alpha = \left[1, \frac{p(\mathbf{d} \,|\, \mathbf{x}^q, k')\; P_{k'}}{p(\mathbf{d} \,|\, \mathbf{x}^p, k)\; P_k}\right]. \tag{35}$$

Therefore the probability of a jump between different dimensional states being accepted does not depend on the size of the space itself, since the $\Delta x$ terms cancel. Hence all states are treated equally by the algorithm and there is no 'inbuilt' bias of accepting moves to one state or the other.

By setting the likelihood to one in (35) we get the criterion for sampling from the prior

$$\alpha = \left[1, \frac{P_{k'}}{P_k}\right]. \tag{36}$$

This tells us that when applying the reversible jump algorithm to the prior the number of transitions between any two states will be equal to the ratio of the prior probabilities for those states, which by definition should be the case. For the posterior the same step is modulated by the likelihood ratio. For the case $k = k'$ sampling is within a single state space and we get

$$\alpha = \left[1, \frac{p(\mathbf{d} \,|\, \mathbf{x}^q, k)}{p(\mathbf{d} \,|\, \mathbf{x}^p, k)}\right], \tag{37}$$

which is exactly the same as the fixed dimension Metropolis algorithm encountered earlier with symmetric proposal distributions.

A point to note about the reversible jump algorithm is that since the acceptance probability in (22) uses the variable $k$ posterior in both the numerator and denominator then the normalizing constant, $p(\mathbf{d})$ (i.e. the total evidence), cancels just as the conditional evidence, $p(\mathbf{d} \,|\, k)$, cancelled in the fixed dimension case. Hence, again the posterior only needs to be known up to a single constant of proportionality across all dimensions. Another point to note is that there is some freedom in choosing the combination $(h, r, g(\mathbf{u}))$ and different choices can lead to the same algorithm. For example, in the above problem one could redefine the random numbers, $u_j$ in (18) as the complete perturbation to $\mathbf{x}_j^p$ rather than a $U[0, 1]$ random variable multiplied by the scale factor $\Delta x$. By repeating the algebra above we find that this trivial change results in a Jacobian of unity, but as a consequence the ratio of the PDFs of the random numbers is changed. We get

$$g(\mathbf{u}^p) = \frac{1}{(\Delta x)^r}, \qquad g'(\mathbf{u}^q) = 1, \tag{38}$$

which gives,

$$\frac{g'(\mathbf{u}^q)}{g(\mathbf{u}^p)} = \frac{(\Delta x)^{k'}}{(\Delta x)^k}, \tag{39}$$

and so overall acceptance probability in (35) is unchanged. By arranging for the Jacobian to be unity in this way, the whole process looks much more like the familiar fixed dimension MCMC sampler. Exploiting this flexibility can be of help in designing particular jump proposals.

A useful special case of the reversible jump algorithm is where the only transitions allowed are between states with just one unknown extra or one less. This is often referred to as *birth–death* MCMC.

For this case the Jacobian is again unity and at each time step a birth is proposed and accepted with probability

$$\alpha = \text{Min}\left[1, \frac{p(\mathbf{d}|\mathbf{x}^q, k+1)}{p(\mathbf{d}\,|\,\mathbf{x}^p, k)}\frac{d_{k+1}}{b_k}\right]. \tag{40}$$

and then a death is proposed and accepted with probability

$$\alpha = \text{Min}\left[1, \frac{p(\mathbf{d}|\mathbf{x}^q, k-1)}{p(\mathbf{d}\,|\,\mathbf{x}^p, k)}\frac{b_{k-1}}{d_k}\right]. \tag{41}$$

where $d_k$ is the assigned probability of a death when the model has $k$ unknowns (jump to the state with $k-1$ unknowns), and $b_k$ is the probability of a birth when the model has $k$ unknowns (jump to a state with $k+1$ unknowns). Typically one has $b_k = d_k = 1/2$ for $k = 2, \ldots, k_{\max} - 1$ with $b_1 = d_{\text{kmax}} = 1$, $b_{\text{kmax}} = d_1 = 0$. Here it has been assumed that the value of the new parameter during the birth step has been generated according to the conditional prior (see Denison *et al.* 2002, p36 for details). This case is the most easiest to implement since it is very similar to the the fixed dimension MCMC algorithm in (16). See Denison *et al.* (2002) and Carlin & Chib (1995) for examples.

While trans-dimensional MCMC is well established in the statistical community, awareness in the Earth Sciences is currently not widespread, especially in the context of inverse problems. Reversible jump MCMC was first applied in the geophysical literature by Malinverno & Leaney (2000) to the inversion of zero-offset vertical seismic profiles. (For a more complete treatment see Malinverno & Leaney 2005). Subsequent work was by Malinverno (2002) who applied it to electrical resistivity sounding. More recently, it has been used as the basis of new approaches to modelling in thermochronology (Gallagher *et al.* 2005; Stephenson *et al.* 2006) and spatial statistics (Stephenson *et al.* 2005). Although the approach is gaining recognition there remain important practical issues to be addressed, not least of which is the design of efficient proposal distributions for the dimension jump (see Brooks *et al.* 2003). Furthermore, translating a particular algorithm into computer software is often a non trivial task. As a first step in addressing these issues Green (2003) proposed an 'Automatic' reversible jump MCMC algorithm for general use, and also made a computer code available (see below). The power of that implementation is that the transformation between state-space variables in (21) is completely at the discretion of the user, and the Jacobian and proposal distributions are determined automatically.

Green (2003) claims that the automatic sampler implementation of reversible jump is efficient for relatively small dimensional problems, say up to 30 unknowns. However, Malinverno & Leaney (2005) were able to apply their reversible jump scheme to an earth model with up to 100 layers. Below we compare the reversible jump sampler embodied in (22) to the more familiar fixed dimension MCMC sampler (16).

### 3.3 Trans-dimensional inference using a fixed dimension sampler

As has been noted above, fixed dimensional inverse problems are common in the geosciences and fixed dimensional MCMC sampling algorithms are well established in the geophysics literature. The reversible jump algorithm, on the other hand, is both new to geophysics and arguably more complicated to implement. Certainly there are as yet no general purpose computer codes with a wide range of applicability. In this situation we might ask then whether it is possible to replicate the results of the reversible jump algorithm with more familiar fixed dimensional methodologies. Here we show how this can be done without specific introduction of the reversible jump machinery.

Let's assume we have an inverse problem which can be parametrized with up to $k_{\max}$ unknowns, and further assume that we have performed $k_{\max}$ independent runs of a standard MCMC sampler each from the fixed dimension posterior (10) for $k = 1, \ldots, k_{\max}$. Let the models generated be $\mathbf{x}_i^{(k)}$, $(i = 1, \ldots, n_k)$, where $\mathbf{x}^{(k)}$ represents a vector with $k$ components and $n_k$ is the number of samples generated in each case. The samples we have are drawn from the fixed dimension posterior, $p(\mathbf{x}\,|\,k, \mathbf{d})$, and what we want are samples drawn from the variable dimension posterior, $p(\mathbf{x}, k\,|\,\mathbf{d})$. The expression that connects these two is (13). It shows that they differ only by the factor $p(k\,|\,\mathbf{d})$ which is the posterior on the dimension $k$. Hence the value of the posterior on $k$ for each dimension is the required weight we must give to the fixed dimension samples to achieve variable $k$ sampling. If we assume, for the moment, that we have the posterior for the dimension $p(k\,|\,\mathbf{d})$ for each $k$, then a simple hierarchical approach is suggested. First we choose $k$ probabilistically with weights $P_k$

$$P_k = \frac{p(k\,|\,\mathbf{d})}{\sum_{k'=1}^{k_{\max}} p(k'\,|\,\mathbf{d})}, \tag{42}$$

and then for the selected $k$, choose a model at random from the corresponding population with uniform weight from $\mathbf{x}_i^{(k)}$, $(i = 1, \ldots, n_k)$. In this way any number of samples from the variable dimension posterior can be obtained using samples drawn from the fixed dimension sampler.

Now we return to the question of calculating the posterior on the number of unknowns, $k$. From (14) we see that this is given by the prior $p(k)$ multiplied by the term $p(\mathbf{d}\,|\,k)/p(\mathbf{d})$, which is the ratio of the conditional evidence (for fixed $k$) divided by the total evidence. Combining (14) with (42) we get values for the probabilities, $P_k$,

$$P_k = \frac{p(\mathbf{d}\,|\,k)p(k)}{\sum_{k'=1}^{k_{\max}} p(\mathbf{d}\,|\,k')p(k')}. \tag{43}$$

Hence the total evidence cancels, and again need not be calculated. Expression (43) shows that the key factor in connecting trans-dimensional and fixed dimension MCMC sampling is the conditional evidence for each dimension, $p(\mathbf{d}\,|\,k)$. As long as this quantity is known, it is possible to replicate the output population of a reversible jump algorithm, by probabilistically resampling, using weights $p_k$, from the output population of a fixed dimensional MCMC sampler.

#### *3.3.1 A linear example*

To illustrate the above ideas we consider a simple example of regression. Fig. 3 shows a set of 20 synthetically generated $(x, y)$ pairs and associated error bars. The $y$ values are the data and were generated from a straight line with intercept 0.3 and gradient 0.6, using $N = 20$ random $x$ values between 0 and 1, after which Gaussian noise ($\sigma = 0.2$) was added. Fig. 3 shows best-fit curves for a first, second, third and fourth order polynomial. For reference Fig. 4 shows three measures often used to estimate the number of unknowns needed to fit data. These are the Chi-square values, the probability of Chi-square and the Bayesian information criterion (BIC) (Schwartz 1978). The Chi-square is defined as

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(d_i - \sum_j^k \lambda_j x^{j-1}\right)^2}{\sigma^2}, \tag{44}$$

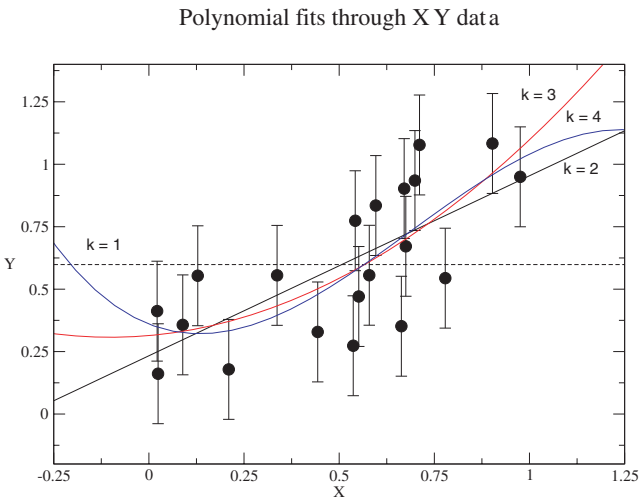


**Figure 3.** Input data used in the regression example. 20 pairs of $(x, y)$ variables with error bars. The lines show best-fit constant, linear, quadratic, and cubic polynomials.

where the unknowns are $\lambda_i$, $(i = 1, \ldots, k)$. The BIC is given by

$$\mathrm{BIC} = -2 * \log(p(\mathbf{d} \mid k, \mathbf{x}_{\mathrm{MAP}}) + k \log N. \tag{45}$$

The probability of Chi-square is determined from standard statistical tables (Press *et al.* 1992). Each of these are 'point' estimates, in that they are evaluated at a single best-fit model. For Gaussian data errors the first term in the BIC is $-2N$ times the Chi-square measure of data fit in (44) while the second term ($k \log N$) is a penalty against increasing the numbers of parameters used to achieve the fit. In Fig. 4 each quantity is plotted as a function of the number of unknowns in the polynomial, $k$. The situation is ideal for use of a Chi-square measure in that the noise is Gaussian and the variance known. The correct number of unknowns ($k = 2$) gives a Chi-square of one, however the probability value, which measures the significance of the fit has its peak at the incorrect value of $k = 3$. Using an $F$-test one can calculate how much the Chi-square can change for a given level of significance. In this case a variation of up to 0.45 is allowed before reaching the 95 per cent significance level. Since the minimum Chi-square for $k = 3$ and $k = 4$ vary less than this, one could conclude that they do not provide significant improvements in fit, and preference should be given to the $k = 2$. The BIC finds the solution directly with a peak at the correct value. In this simple problem then standard point estimates are able to pick the best solution, but they do not provide any information on the relative levels of support for each case.

Applying trans-dimensional Bayesian inference to this problem we use the likelihood function

$$p(\mathbf{d} \mid \lambda, k) = \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\left\{-\frac{N}{2}\chi^2\right\}, \tag{46}$$

and the flat priors

$$p(\lambda_j \mid k) = \begin{cases} \frac{1}{(\lambda_j^U - \lambda_j^L)} & : \quad \lambda_j^L \le \lambda_j \le \lambda_j^U \\ 0 & : \quad \text{otherwise,} \end{cases} \qquad (j = 1, \ldots, k), \tag{47}$$

$$p(k) = \frac{1}{k_{\max}}, \tag{48}$$

where $(\lambda_j^L, \lambda_j^U)$ are the upper and lower values imposed for $\lambda_j$, and $k_{\max} = 4$. We set the coefficient ranges to be quite broad using lower limits of $(0, -2, -10, -30)$ and upper limits of $(1.2, 2, 10, 30)$. The posterior for the variable $k$ case is then the product of (46), (47) and (48) divided by the (unknown) total evidence, as shown in (15). The total evidence, $p(\mathbf{d})$, is a constant factor for all $k$ and hence cancels

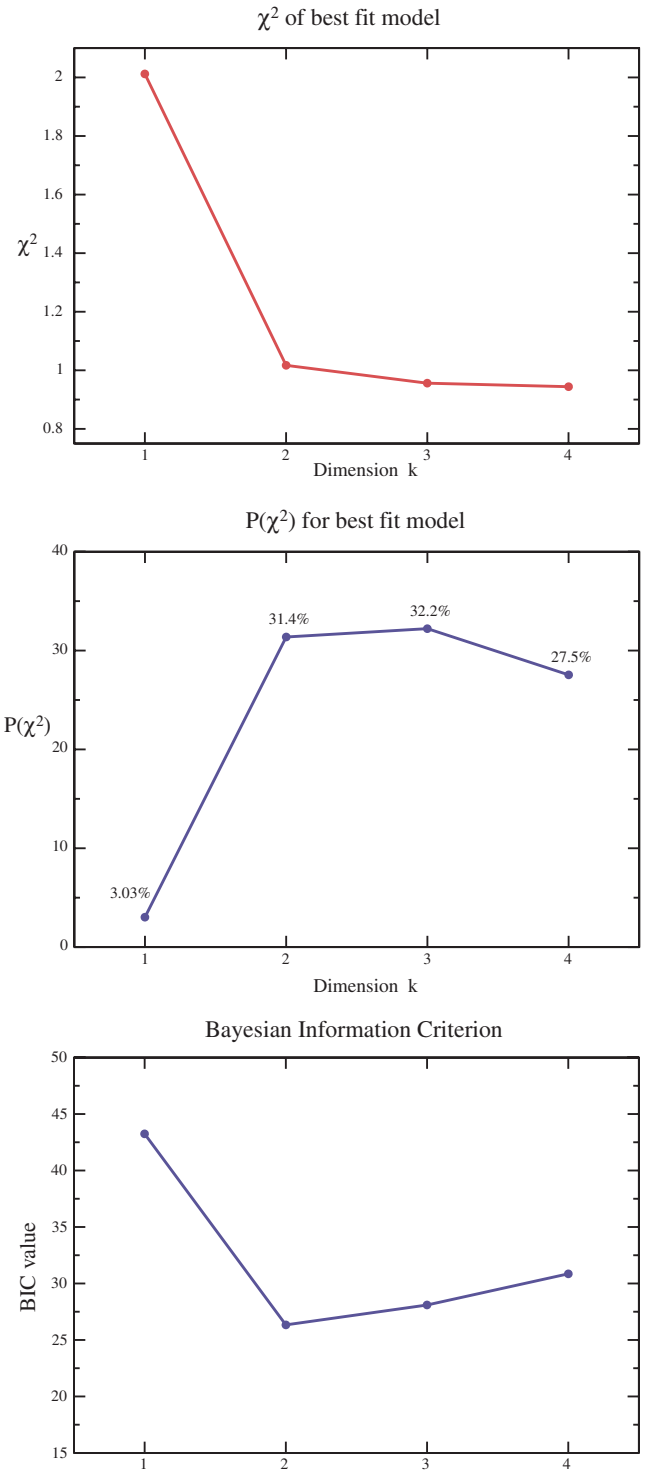


**Figure 4.** (a) Chi-squared values for maximum likelihood solutions to fitting the data in Fig. 3 using a linear polynomial with one, two, three and four parameters; (b) The probability of the Chi-square value for each case (note that this measure peaks at the incorrect value of $k = 3$); (c) The Bayesian information Criterion (see text) for each case in a).

out in the calculation of the acceptance probability (22). [Note that we use a flat prior here simply for illustrative purposes. Jeffreys (1939) argued that the only way to encode complete ignorance for a positive quantity was to set $p(\log k) = \text{const}$. This has the advantage of invariance under certain parameter transformations, for example, setting $k$ to any power results in a new variable with the same probability density, which is not true of the flat prior (48). However, the Jeffreys prior cannot be normalized over $0 < k < \infty$. For a detailed justification for the Jeffrey's prior the reader is referred to Jaynes (2003).]

We use the automatic reversible jump implementation of Green (2003). In this the proposal distributions for walks within a model space and between model spaces are based on multivariate Gaussians centred at the point $\mu_k$ with covariance matrix $B_k B_k^T$,

$$q\left(\mathbf{x}^q \mid \mathbf{x}^p\right) = N\left(\mu_k, B_k B_k^T\right). \tag{49}$$

The covariance matrices, $B_k B_k^T$ and means $\mu_k$ for each model space are estimated by first sweeping through each dimension separately and applying a standard MCMC sampler [see (4) and (5)]. Implementation details of the scheme are omitted here, for full details readers are referred to Green (2003).

The reversible jump algorithm results in a chain of samples, each of which is a vector of $k$-coefficients, with $k$ itself varying between samples. Fig. 5 shows the prior, $p(k)$, and posterior, $p(k \mid \mathbf{d})$, for the number of components, $k$. The posterior was obtained by running the reversible jump sampler for $10^6$ steps, collecting every 200th sample and tabulating the frequency of $k$ values. The prior was set in advance to be uniform across dimension, and Fig. 5(b) shows the result of sampling with the likelihood in (46) set to unity. The fact that this recovers the imposed prior is a useful check on the algorithm, and shows that in the absence of any data there is no preference for any dimension. Results of the reversible jump algorithm suggest support in the ratio of 0.04, 87.02, 11.89 and 1.05 per cent for $k = 1, \ldots, 4$, respectively, clearly favouring the linear fit model. We see then that when the data are introduced there is a preference for simple rather than complex models. This is an example of the natural parsimony property of Bayesian inference discussed in Section 2.1.

We now apply the fixed dimension MCMC sampler to the problem and use the algorithm described in Section (3.3) to resample the results and simulate the variable dimension posterior. In this case each value of $k$ is considered independently, and the posterior is given by the likelihood (46) times the prior (47) divided by the conditional evidence, $p(\mathbf{d} \mid k)$, which is again unknown. As in the reversible jump case the evidence cancels out in the calculation of $\alpha$, see (16). A simple axi-symmetric Gaussian proposal distribution is used with variances calculated from the prior (i.e. $\sigma_j = (\lambda_j^U - \lambda_j^L)/\sqrt{12}$). Note that this is not the same proposal distribution used by the reversible jump algorithm in (49). To estimate the conditional evidence $p(\mathbf{d} \mid k)$ for each $k$, we generate $N_s$ samples from the prior (47) and numerically estimate the integral in (11) using

$$\tilde{p}(\mathbf{d} \mid k) = \frac{1}{N_s} \sum_{i=1}^{N_s} p(\mathbf{d} \mid \boldsymbol{\lambda}_i, k), \tag{50}$$

where $\boldsymbol{\lambda}_i$ represents the coefficients of the $i$th sample. With $N_s = 10^6$ samples the evidence values obtained for $k = 1, \ldots, 4$ were (0.000170, 0.36976, 0.05069, 0.00271), respectively, which are in the ratio 0.04, 87.34, 11.98 and 0.64 per cent. (Errors on the relative evidence are less than 0.003 per cent in all cases.) These values for the relative evidence compare very well with the posteriors found from the reversible jump algorithm, and (14) says they should be the same for this problem. [Note that the evidence has a maximum at $k = 2$ and then decreases which is in contrast to the best data fit (shown in Fig. 4) which continues to increase with $k$.]

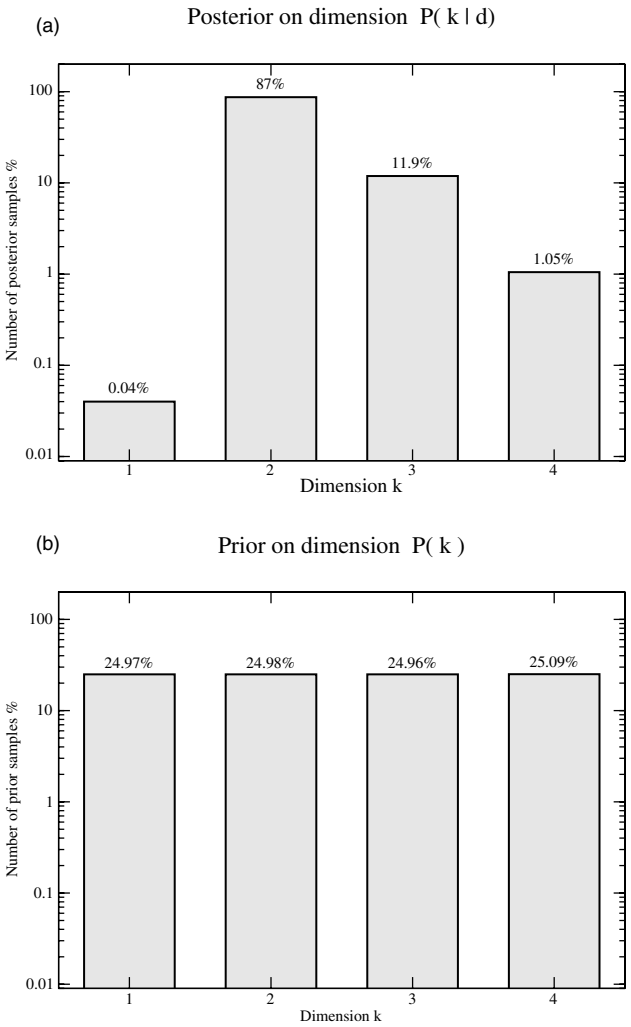


**Figure 5.** (a) The posterior probability density for the number of unknowns $k$ in the regression model calculated using the reversible jump algorithm. Note the clear preference for two unknowns with other values of $k$ receiving much less support. (b) same as a) but when the prior is sampled. Note the equal frequencies for $k$ in the simulated samples reflect the uniform prior imposed. Comparing (a) to (b) shows how the data increase support for the $k = 2$ case at the expense of the others. The frequency scale is logarithmic.

Using the relative evidence values as weights we then randomly selected 5000 samples from the $4 \times 10^6$ fixed-$k$ samples, to simulate a variable-$k$ sample. A useful way to compare the combined sample with those from reversible jump algorithm, is to calculate a density plot of the corresponding data fit curves, that is, for each sample and associated $k$ value we calculate the curve using

$$y(x) = \sum_{j=1}^{k} \lambda_j x^{j-1}. \tag{51}$$

Figs 6(a)–(d) show the density plots for each of the four fixed-$k$ samples, while Figs 6(e) and (f), show the variable-$k$ posterior from the combined and reversible jump samples, respectively. The range in each panel is the same as Fig. 3. It is clear that the two posterior simulations are almost identical, indicating that the density of the reversible jump samples have been successfully recreated by

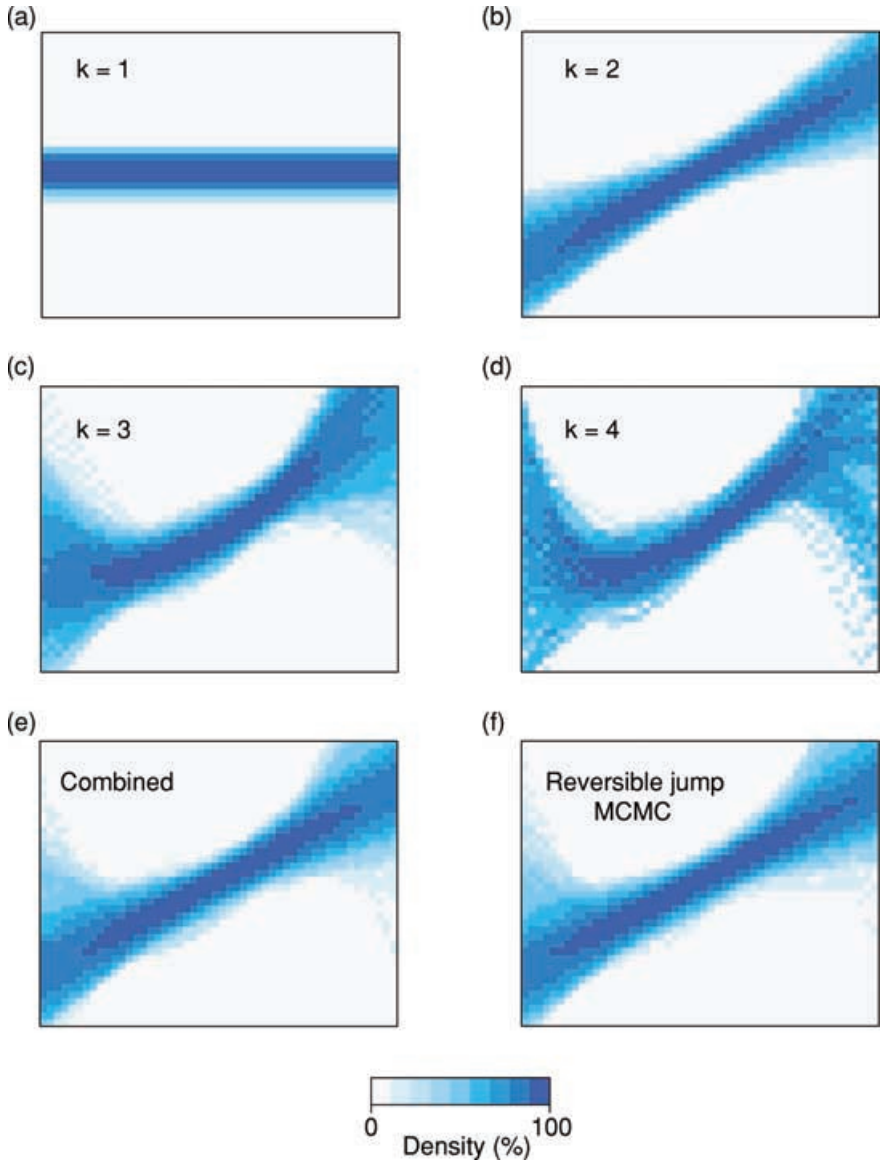


**Figure 6.** Density of 5000 predicted regression curves for fixed dimension samples, (a) $k = 1$, (b) $k = 2$, (c) $k = 3$, (d) $k = 4$. (e) shows the density of 5000 curves calculated from randomly selected fixed $k$ curves using the evidence values as weights (see text). (f) is for 5000 reversible jump samples with variable $k$. Axes limits are the same as in Fig. 3. The similarity of (e) and (f) shows that the variable dimension sample density can be recreated from the fixed dimension samples.

combining the fixed dimension samples. Furthermore, one sees how the density pattern of the combined samples reflects the individual sample densities. Clearly the pattern is dominated by $k = 2$ with contributions from $k = 3$, and, to a much lesser extent $k = 4$, apparent at the two extremes. This example clearly demonstrates that in a simple linear problem the trans-dimensional sampling with the reversible jump algorithm can be recreated with the more familiar fixed dimensional MCMC algorithm. For further details of how reversible jump algorithms can be applied to a range of more complicated regression problems the reader is referred to the comprehensive study by Denison *et al.* (2002).

### *3.3.2 A non-linear example*

We now illustrate the role of the evidence in the non-linear problem of mixture modelling (McLachlan & Basford 1987). This trans-dimensional inverse problem arises in several areas of the Earth Sciences, including fission track studies (Galbraith & Green 1990) and other areas of geochronology (Sambridge & Compston 1994; Jasra *et al.* 2006). In these problems observations may reflect an overlapping combination of distinct geological ages, and it is important to try and recover individual components from the combined distribution. Mixture modelling is also a problem of recurring interest in statistics and the reversible jump algorithm has been successfully applied across a range of applications (see Richardson & Green 1997; Jasra *et al.* 2006, for details). Fig. 7 shows ($N_d = 100$) data generated from two equally weighted Gaussian distributions with means (540, 570) and standard deviation $\sigma = 30$. Here we consider a simplified version of the mixture modelling problem, where only the number of components and the means of those components are unknown. (The relative weights and standard deviations being kept fixed.) For this case the likelihood function for $k$ Gaussian

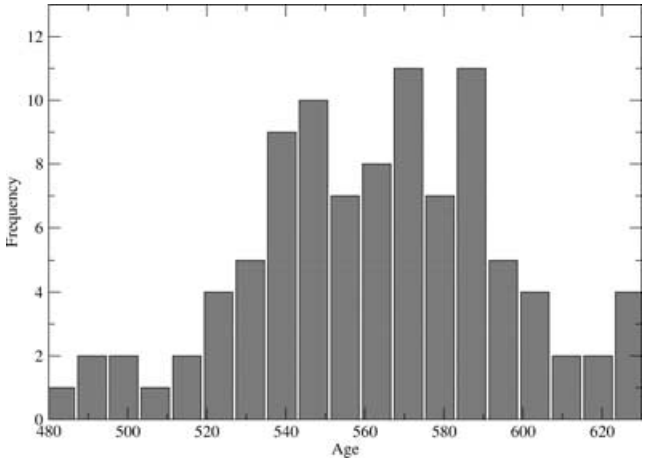


**Figure 7.** Histogram of 100 synthetically created age data used in the mixture modelling example. The age values were generated from two equal Gaussian centred at 540 and 570 Ma and with standard deviations of 30 Ma. The number of components and the mean ages are treated as unknowns in the example. See text for details.

components is,

$$p(\mathbf{d} \mid \mathbf{x}, k) = \frac{1}{(\sigma\sqrt{2\pi})^{N_d}} \prod_{i=1}^{N_d} \sum_{j=1}^{k} \frac{1}{k} \exp\left\{ \frac{-(d_i - x_j)^2}{2\sigma^2} \right\}. \qquad (52)$$

The prior on $x_j$, $(j = 1, \ldots, k)$ is again chosen to be uniform as in (47) with bounds equal to that of the data. The prior on the number of components, $p(k)$, is also uniform, as in (48), with a maximum of 7. For variable $k$ the posterior is the product of likelihood (52) and the two priors divided by the evidence. The resulting posterior for this problem can be multimodal due to the more complex likelihood function (52). For the fixed $k$ case the posterior is proportional to the product of the likelihood (52) and the uniform prior on the component means only.

Both sampling algorithms are exactly as in the regression example. Fig. 8 shows the results of the posterior on the number of components, $p(k \mid \mathbf{d})$, calculated with the reversible jump algorithm compared to the evidence values, $p(\mathbf{d} \mid k)$ found with the fixed dimension sampling and numerical integration (50). The maximum dimension is now 7 and the problem more complex, but the evidence values for each $k$ clearly give the same trend as the posterior on $k$. In this case we see a greater spread of support across the number of components, which is due to the relatively large noise in the data (with noise standard deviation being equal to the separation of the two real components). Nevertheless the trans-dimensional samples have clearly picked out $k = 2$ as the most preferred solution.

To examine the samples produced we plot the density of all ages on a single axis in Fig. 9. Fig. 9(a) shows density curves for all seven fixed $k$ runs, while Fig. 9(b) shows the variable-$k$ posterior from both the combined and reversible jump samples. The two variable $k$ samples are identical within sampling error, and again one sees how the combined sample is made up of fixed dimension components. Here the central peak is due to the $k = 1$ contribution, the high- and low-age peaks are due to $k = 2$, and the skewness toward the higher ages is due to the combination of cases $k = 3, \ldots, 7$. Again we see the preference toward simpler models in explaining the data.

Both the linear and non-linear examples verify that it is possible to simulate trans-dimensional MCMC sampling using a fixed dimension sampler and the relative evidence values for each $k$. While the latter approach is conceptually simple, we are not advocating

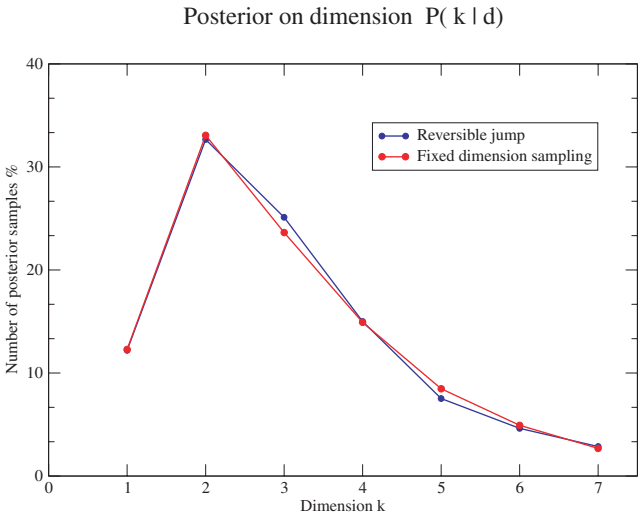


**Figure 8.** Information on the number of components in the mixture modelling problem. Blue curve shows the posterior probability density for $k$ found with the reversible jump algorithm. The red curve is plotted using the evidence values found from each of the fixed dimension MCMC runs. This shows how the fixed dimension sampling can recreate the reversible jump posterior on the number of components.

that it is superior to the former. Computational costs are comparable in these examples, but we have not experimented with efficiency issues here. If the maximum number of unknowns is large, it is seems likely that simply sampling each parameter space in turn using fixed dimensional MCMC will be prohibitive, since the cost of each integration would increase significantly with dimension. A trans-dimensional MCMC algorithm is likely to be much more efficient since, by definition, it aims to spend only the amount of time necessary in each dimension as needed to calculate the posterior on $k$, $p(k \mid \mathbf{d})$. No doubt more efficient use of fixed dimensional sampling could be devised whereby the number of sweeps at each fixed $k$, $N_k$, is adjusted in an iterative manner, using the relative evidence estimates obtained. In addition, for real problems there is no need for the prior on the dimension to be uniform. Indeed as we have argued above a more sensible choice is, $p(k) \propto 1/k$ (equivalent to the Jeffreys prior on $k$), which would help reduce the number of samples required as the dimension increased. We have not explored any of these options. Our intention is simply to show the connection between the trans-dimensional sampler and a standard technique which may be more familiar to Earth scientists. This provides a reference method with which to test implementations of reversible jump algorithms. At the same time we have demonstrated that the evidence is the key quantity that connects the two.

## 4 CALCULATING THE EVIDENCE

We conclude this paper with a brief discussion of how the evidence might be calculated in different situations. In the two examples above a simple integration method is sufficient. Likelihood values are averaged at samples drawn from the uniform prior. This is not likely to be practical for higher-dimensional problems or where the posterior is more complex. An alternative approach can be found by rearranging (10) to give

$$\frac{p(\mathbf{x} \mid k)}{p(\mathbf{d} \mid k)} = \frac{p(\mathbf{x} \mid k, \mathbf{d})}{p(\mathbf{d} \mid \mathbf{x}, k)}. \qquad (53)$$

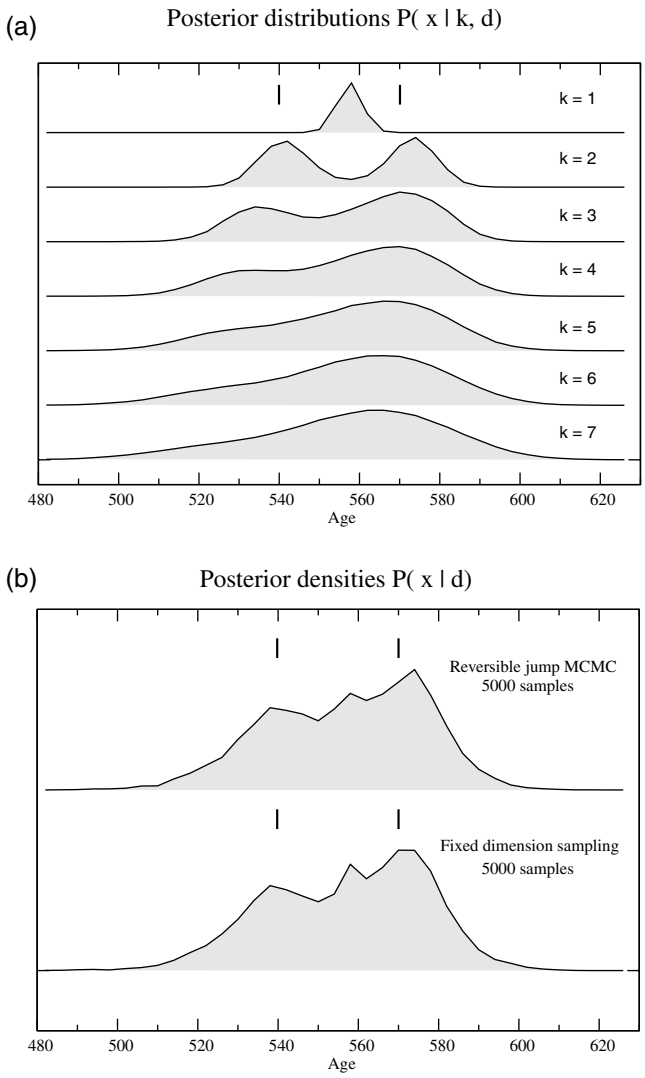


**Figure 9.** A comparison of posterior densities in the mixture modelling example. (a) shows the density of the fixed $k$ MCMC runs with up to seven components and (b) the weighted fixed dimension samples and the reversible jump samples. For each curve the density plots are for all age components plotted on a single axis. The reversible jump posterior is identical to the weighted fixed dimension samples within sampling error. Comparing (a) and (b) one can also see how structural features of the fixed dimension posteriors contribute to the peaks in the variable dimension posterior. In both cases the vertical bars represent the true values of each component age.

Since the prior $p(\mathbf{x} \mid k)$ is normalized (i.e. integrates to one), we can integrate both sides over $\mathbf{x}$ to give

$$[p(\mathbf{d} \mid k)]^{-1} = \int \frac{p(\mathbf{x} \mid k, \mathbf{d})}{p(\mathbf{d} \mid \mathbf{x}, k)} d\mathbf{x}. \tag{54}$$

A Monte Carlo estimate of (54) can be found by drawing samples, $\tilde{\mathbf{x}}_i, (i = 1, \ldots, n)$ from the posterior, $p(\mathbf{x} \mid k, \mathbf{d})$. We get

$$p(\mathbf{d} \mid k) \approx \frac{1}{\sum_{i=1}^{n} [p(\mathbf{d} \mid \tilde{\mathbf{x}}_i, k)]^{-1}}. \tag{55}$$

Green (2003) expects the estimator (55) to have high variance, and hence for many situations its numerical error will only slowly decrease with $n$. We evaluated (55) for the both of the numerical examples above and found convergence to be much slower than for the rather simple procedure used in Section 3. A number of importance sampling approaches to calculating the evidence have been proposed by Newton & Raftery (1994); Gelfand & Dey (1994); Chib (1995); Chib & Jeliazkov (2001). More recently Skilling (2004, 2005) has proposed a new approach for general Bayesian sampling and in particular evidence calculations, known as 'Nested sampling'. (General use software has been made available implementing the latter, see below). None of these approaches can yet be described as definitive, and it seems sampling based methods for evidence calculations will be an active area of research in the future.

As the number of unknowns becomes large all Bayesian sampling techniques tend to suffer from the curse of dimensionality (see Curtis & Lomax 2001, for a discussion), and evidence calculations are likely to become impractical with a sampling technique. It is worth noting, however, that in some special cases analytical expressions are available for the evidence. For example, this is the case for a linear inverse problem with Gaussian noise and a quadratic prior. We have

$$\mathbf{d} = \mathrm{A}\mathbf{x}, \tag{56}$$

with likelihood

$$p(\mathbf{d} \mid \mathbf{x}, k) = \frac{1}{(2\pi)^{N_d/2} |C_D|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{d} - \mathrm{A}\mathbf{x})^T C_D^{-1} (\mathbf{d} - \mathrm{A}\mathbf{x})\right\}, \tag{57}$$

where $C_D$ is the data covariance matrix of the data noise, and prior

$$p(\mathbf{x} \mid k) = \frac{1}{(2\pi)^{k/2} |C_M|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}_o)^T C_M^{-1} (\mathbf{x} - \mathbf{x}_o)\right\}, \tag{58}$$

where $C_M$ is the prior model covariance matrix. To make the connection to large-scale inverse problems clearer it is worth noting that the terms within the curly brackets of (57) and (58) depend on the chi-square of data fit and a quadratic model penalty term. Specifically, we can define $\chi^2(\mathbf{d}, \mathbf{x})$ and $\phi(\mathbf{x}, \mathbf{x}_o)$ as,

$$\chi^2(\mathbf{d}, \mathbf{x}) = \frac{1}{N}(\mathbf{d} - \mathrm{A}\mathbf{x})^T C_D^{-1} (\mathbf{d} - \mathrm{A}\mathbf{x}), \tag{59}$$

and

$$\phi(\mathbf{x}, \mathbf{x}_o) = (\mathbf{x} - \mathbf{x}_o)^T C_M^{-1} (\mathbf{x} - \mathbf{x}_o). \tag{60}$$

If the (fixed) number of unknowns in the inverse problem is $k$ then the evidence can be found by integrating the product of the likelihood (57) and the prior (58) over the parameter space, as in (11). We get

$$p(\mathbf{d} \mid k) = \frac{\rho(\hat{\mathbf{x}}, \mathbf{d}, \mathbf{x}_o)}{((2\pi)^{(N_d+k)} |C_D||C_M|)^{1/2}} \times \int \exp\left\{-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^T C_M'^{-1} (\mathbf{x} - \hat{\mathbf{x}})\right\} d\mathbf{x}, \tag{61}$$

where $\hat{\mathbf{x}}$ is the usual damped least-squares solution

$$\hat{\mathbf{x}} = \left(\mathrm{A}^T C_D^{-1} \mathrm{A} + C_M^{-1}\right)^{-1} \left(\mathrm{A}^T C_D^{-1} \mathbf{d} + C_M^{-1} \mathbf{x}_o\right), \tag{62}$$

$C_M'$ is the (posterior) model covariance matrix

$$C_M' = \left(\mathrm{A}^T C_D^{-1} \mathrm{A} + C_M^{-1}\right)^{-1}, \tag{63}$$

and

$$\rho(\hat{\mathbf{x}}, \mathbf{d}, \mathbf{x}_o) = \exp\left\{\frac{-N}{2}\chi^2(\mathbf{d}, \hat{\mathbf{x}}) - \frac{1}{2}\phi(\hat{\mathbf{x}}, \mathbf{x}_o)\right\}. \tag{64}$$

Note that the term $\rho(\hat{\mathbf{x}}, \mathbf{d}, \mathbf{x}_o)$ depends only on the data fit and model penalty at the least-squares solution. The evidence can be found by noting that the integral in (61) is a $k$-dimensional Gaussian. Assuming the parameter space is unbounded we have

$$\int \exp\left\{-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^T C_M'^{-1} (\mathbf{x} - \hat{\mathbf{x}})\right\} d\mathbf{x} = \left((2\pi)^k |C_M'|\right)^{1/2}. \tag{65}$$

Substituting into (61) gives

$$p(\mathbf{d} \mid k) = \frac{\rho(\hat{\mathbf{x}}, \mathbf{d}, \mathbf{x}_o)}{(2\pi)^{(N_d)/2}} \left( \frac{|C'_M|}{|C_D||C_M|} \right)^{1/2}. \quad (66)$$

Therefore the evidence for a particular number of unknowns, $k$, depends on the data fit at the solution and a ratio of determinants of the posterior to prior covariance matrices and as such measures the change in shape of the PDF as the data are added. The result (66) has been stated recently by Malinverno & Briggs (2004) in their Appendix B; however, the result is much older than this. Its first statement is difficult to determine, but certainly it has been derived and used by Gull (1989) in calculating the evidence in his work in maximum entropy, and a few words are in order to describe this.

Gull (1989) is concerned with a prior PDF which is not Gaussian, but which is associated with the entropy of the image with regard to which he is trying to make inferences. He applies the so-called Laplace approximation for this prior PDF, which means that he approximates it as being Gaussian; the relevant matrix $C_M^{-1}$ is associated with the second term in the Taylor expansion of the logarithm of the prior PDF (see e.g. Mackay 2003, p. 341). He then performs the multidimensional integral to find the evidence, giving a formula equivalent to (66). His developments go further, however, because he discusses the situation when one does not know $C_M^{-1}$, but instead only a scalar multiple of it. In other words there is an unknown parameter that scales the width of the prior PDF. Gull goes on to show how this parameter can be chosen so as to maximize the evidence.

To help interpret (66) consider a problem with one unknown, and let the prior be uniform over some finite interval, $\sigma_{\text{prior}}$ and zero elsewhere. In this case $\phi(\hat{\mathbf{x}}, \mathbf{x}_o) = 0$, $C_M = \sigma^2_{\text{prior}}$, the prior model variance, and $C'_M = \sigma^2_{\text{post}}$, the posterior variance, and the maximum likelihood value of the best-fit model $\hat{\mathbf{x}}$ is

$$p(\mathbf{d} \mid \hat{\mathbf{x}}, k) = \rho(\hat{\mathbf{x}}, \mathbf{d}, \mathbf{x}_o) \left( (2\pi)^{N_d} |C_D| \right)^{-1/2}. \quad (67)$$

Substituting into (66) gives

$$p(\mathbf{d} \mid k) \approx p(\mathbf{d} \mid \hat{\mathbf{x}}, k) \times \left( \frac{\sigma_{\text{post}}}{\sigma_{\text{prior}}} \right). \quad (68)$$

Note the expression is approximate because the prior is only non-zero over a restricted region and hence the integration in (65) is no longer over an unbounded parameter space. Expression (68) shows that the evidence is approximately the best-fit likelihood value multiplied by the ratio of the posterior PDF width to the prior PDF width. The latter term is often called the 'Occam factor'. It is always less than one and measures the relative amount the PDF has narrowed due to the data (see Malinverno 2002; Mackay 2003, for useful discussions of the Occam factor).

Comparing inversion solutions with different numbers of unknowns $k_1$ and $k_2 (k_2 > k_1)$, the best-fit (maximum) likelihood value will tend to increase (favouring the larger $k$); however; the Occam factor will tend to decrease (favouring the smaller $k$). The overall value of the evidence will be a trade-off of these two quantities. In principle then one could resolve a linear inverse problem with differing numbers of unknowns and calculate the evidence for each. An alternative expression to (66), and perhaps simpler for this purpose is found by making use of (57), (58) and (64). In this case (66) becomes

$$p(\mathbf{d} \mid k) = p(\mathbf{d}|\hat{\mathbf{x}}, k) p(\hat{\mathbf{x}} \mid k) (2\pi)^{k/2} |C'_M|^{1/2}. \quad (69)$$

This says that the evidence is the product of the likelihood and the prior evaluated at the posterior maximum (or mean), multiplied by the root of the determinant of the posterior covariance matrix. Strictly speaking (69) is only valid in the linear case with Gaussian prior and likelihood in an unbounded parameter space. Provided all three quantities have been calculated (which is relatively standard in linear inverse problems) the evidence can be calculated using only information at the solution. The linear example problem in Section 3.3.1 has simple prior bounds on all unknowns and so (69) will only be an approximation to the evidence. Nevertheless it gives values in the ratio 0.04, 87.07, 11.67 and 1.21 per cent, for $k = 1, \ldots, 4$, which are in good agreement with the values found through sampling.

Using (14) we can then calculate the posterior ratio for any two values of $k$

$$\frac{p(k_1 \mid \mathbf{d})}{p(k_2 \mid \mathbf{d})} = \frac{p(\mathbf{d}|\hat{\mathbf{x}}, k_1)}{p(\mathbf{d}|\hat{\mathbf{x}}, k_2)} \times \frac{p(\hat{\mathbf{x}}|k_1)}{p(\hat{\mathbf{x}}|k_2)} \times \left[ \frac{(2\pi)^{k_1} |C'_M(k_1)|}{(2\pi)^{k_2} |C'_M(k_2)|} \right]^{1/2} \times \frac{p(k_1)}{p(k_2)}. \quad (70)$$

By repeating an inversion for selected values of $k$ and using (70) we may hope to map out the posterior for $p(k|\mathbf{d})$ (as in Figs (5) and (8)) for larger problems where direct numerical integration is impractical. This approach might be useful for a range of linear problems and even non-linear ones that can be linearized about an optimal solution.

## 5 CONCLUSIONS

In this paper we have focused on trans-dimensional inverse problems, which we view as a natural extension to the common practice of fixing the number of unknowns in advance. We have shown how Bayesian tools may be used to quantitatively estimate the number of unknowns needed in the inverse problem. The theory leads directly to a posterior on the number of free parameters in the unknown model. Our main aim here is to highlight the role played by the 'evidence', a statistical quantity, awareness of which is rather limited in the geosciences. We have shown how the evidence links inverse problems in which the number of unknowns is fixed, with those where the number of unknowns is a variable. We have also shown how trans-dimensional Bayesian sampling techniques such as the reversible jump algorithm, can be replicated using traditional fixed dimensional sampling techniques and calculation of the evidence. Simple examples have been used to illustrate the main points. Some general approaches for calculating the evidence in non-linear inverse problems are given. These are far from definitive at present and research in this area is set to continue. For linear inverse problems analytical expressions for the evidence and the posterior on the dimension are available. We argue that these constitute a useful diagnostic tool for investigating the number of degrees of freedom in the model supported by the data.

## ACKNOWLEDGMENTS

P. J. Green is acknowledged for use of his Automatic reversible jump MCMC code (available from www.stats.bris.ac.uk/~peter/AutoRJ). Comments on earlier drafts of this manuscript were received by Jeannot Trampert and Brian Kennett. We also thank Alberto Malinverno and an anonymous reviewer for constructive reviews. J. Skilling's general Bayesian sampling package with evidence estimator 'Bayesys' is available from http://www.inference.phy.cam.ac.uk/bayesys.

## REFERENCES


Aitkin, M., 1991. Posterior Bayes factors, *J. of the Royal Stat. Soc., B*, **53**, 111–142.

Aster, R., Borchers, B. & Thurber, C.H., 2005. *Parameter estimation and inverse problems*, Vol. 90 of International Geophysics Series, Elsevier, Amsterdam.

Backus, G.E., 1988a. Bayesian inference in geomagnetism, *Geophys. J.,* **92,** 125–242.

Backus, G.E., 1988b. Comparing hard and soft prior bounds in geophysical inverse problems, *Geophys. J.,* **94,** 249–261.

Backus, G.E. & Gilbert, J.F., 1967. Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. astr. Soc.,* **13,** 247–276.

Backus, G.E. & Gilbert, J.F., 1968. The resolving power of gross earth data, *Geophys. J. R. astr. Soc.,* **16,** 169–205.

Backus, G.E. & Gilbert, J.F., 1970. Uniqueness in the inversion of inaccurate gross earth data, *Phil. Trans. R. Soc. Lond.,* A, **266,** 123–192.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances, *Philos. Trans. R. Soc. London,* **53,** 370–418. Reprinted, with biographical note by G. A. Barnard, 1958, in *Biometrika,* **45,** 293–315.

Bernardo, J.M. & Smith, A.F.M., 1994. *Bayesian Theory*, John Wiley & Sons, Chichester.

Brooks, S.P., Giudici, P. & Roberts, G.O., 2003. Efficient construction of reversible jump MCMC proposal distributions (with discussion), *J. of the Royal Stat. Soc.,* B, **65,** 3–55.

Carlin, B.P. & Chib, S., 1995. Bayesian model choice via Markov chain Monte Carlo, *J. of the Royal Stat. Soc., B,* **57,** 473–484.

Chib, S., 1995. Marginal likelihood from the Gibbs output, *J. Am. statist. Ass.,* **90,** 1313–1321.

Chib, S. & Jeliazkov, I., 2001. Marginal likelihood from the metropolis Hastings output, *J. Am. statist. Ass.,* **96,** 270–281.

Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52,** 289–300.

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions and the curse of dimensionality, *Geophysics,* **66,** 372–378.

Curtis, A. & Wood, R., 2004. Optimal elicitation of probabilistic information from experts, in *Geological Prior Information*, Publication 239, pp. 1–14, eds Curtis, A. & Wood, R., Geol. Soc. Lond.

Denison, D.G.T., Holmes, C., Mallick, B. & Smith, A.F.M., 2002. *Bayesian Methods for Nonlinear Classification and Regression*, John Wiley & Sons, Hoboken.

Galbraith, R.F. & Green, P.F., 1990. Estimating the component ages in a finite mixture, *Nucl. Tracks Radiat. Meas.,* **17,** 197–206.

Gallagher, K., Stephenson, J.A., Brown, R.W., Holmes, C.C. & Ballester, P., 2005. Exploiting 3-D spatial sampling in inverse modeling of thermochronological data, *Rev. in Mineral. and Geochem.,* **58,** 375–387.

Gelfand, A.E. & Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations, *J. of the Royal Stat. Soc.,* B, **56,** 501–514.

Gelfand, A.E. & Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities, *J. Am. statist. Ass.,* **85,** 398–409.

Geyer, C.J. & Møller, J., 1994. Simulation procedures and likelihood inference for spatial point processes, *Scand. J. Stats,* **21,** 359–373.

Godsill, S.J., 2003. Proposal densities and product-space methods, in *Highly Structures Stochastic Systems*, Oxford Statistical Science Series, chap. 6A, pp. 199–203, eds Green P.J., 0Hjort, N.L. & Richardson, S., O.U.P.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika,* **82,** 711–32.

Green, P.J., 2003. Trans-dimensional Markov chain Monte Carlo, in *Highly Structured Stochastic Systems*, Oxford Statistical Science Series, chap. 6, pp. 179–198, eds Green P.J., Hjort, N.L. & Richardson, S., O.U.P.

Gull, S.F., 1989. Developments in Maximum Entropy Data Analysis, in *Maximum Entropy and Bayesian Methods*, ed. Skilling, J., Kluwer.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chain and their applications, *Biometrika,* **57,** 97–109.

Jackson, A., 1995. An approach to estimation problems containing uncertain parameters, *Earth planet. Sci. Lett.,* **90,** 145–156.

Jasra, A., Stephens, D., Gallagher, K. & Holmes, C., 2006. Bayesian Mixture Modelling in Geochronology via Markov Chain Monte Carlo, *Math. Geol.,* **38,** 269–300, doi: 10.1007/s11004-005-9109-3.

Jaynes, E.T., 2003. *Probability Theory The Logic of Science*, Cambridge Univ. Press, Cambridge.

Jeffreys, H., 1939. *Theory of Probability*, Clarendon Press, Oxford.

Kass, R.E. & Raftery, A.E., 1995. Bayes factors, *J. Am. statist. Ass.,* **90,** 773–795.

Mackay, D.J.C., 2003. *Information Theory, Inference, and Learning algorithms*, C.U.P., Cambridge.

Malinverno, A., 2000. A Bayesian criterion for simplicity in inverse problem parametrization, *Geophys. J. Int.,* **140,** 267–285.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.,* **151,** 675–688.

Malinverno, A. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes, *Geophysics,* **69,** 1005–1016.

Malinverno, A. & Leaney, W.S., 2000. A Monte Carlo method to qantify uncertainty in the inversion of zero-offset vsp data, *70th SEG Annual Meeting Expanded Abstracts, Tulsa Oklahoma*, pp. 2392–2396.

Malinverno, A. & Leaney, W.S., 2005. Monte carlo bayesian look-ahead inversion of walkaway vertical seismic profiles, *Geophys. Prospect.,* **53,** 689–703.

McLachlan, G.J. & Basford, K.E., 1987. *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press, San Diego, revised edition.

Metropolis, N. & Ulam, S.M., 1949. The Monte Carlo method, *J. Am. Stat. Assoc.,* **44,** 335–341.

Mosegaard, K. & Sambridge, M., 2002. Monte Carlo analysis of inverse problems, *Inverse Problems,* **18,** R29–R54.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100,** 12 431–12 447.

Newton, M.A. & Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussions), *J. of the Royal Stat. Soc.,* B, **56,** 3–48.

Parker, R.L., 1994. *Geophysical Inverse Theory*, Princeton University Press, Princeton.

Press, W.H., Tuekolsky, S.A., Vetterling, W.T. & Flannery, B.P., 1992. *Numerical Recipes in FORTRAN*, Cambridge University Press, Cambridge.

Richardson, S. & Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion), *J. of the Royal Stat. Soc.,* B, **59,** 731–792.

Rothman, D.H., 1985. Nonlinear inversion statistical mechanics and residual statics corrections, *Geophysics,* **50,** 2784–2796.

Rothman, D.H., 1986. Automatic estimation of large residual statics corrections, *Geophysics,* **51,** 332–346.

Sambridge, M., 1999. Geophysical inversion with a Neighbourhood algorithm -II. Appraising the ensemble, *Geophys. J. Int.,* **138,** 727–746.

Sambridge, M. & Rawlinson, N., 2005. Tomography with irregular meshes, in *Seismic Earth: Array Analysis of Broadband Seismograms*, pp. 1–13, eds Lavender, A. & Nolet, G., AGU.

Sambridge, M.S. & Compston, W., 1994. Mixture modeling of multi-component data sets with application to ion-probe zircon ages, *Earth planet. Sci. Lett.,* **128,** 373–390.

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,** 389–397.

Schwartz, G., 1978. Estimating the dimension of a model, *Ann. Statist.,* **6,** 461–464.

Skilling, J., 2004. Nested sampling, in *Bayesian inference and maximum entropy methods in science and engineering*, AIP Conference Proceedings 735, New York.

Skilling, J., 2005. Nested sampling for general Bayesian computation, in *Bayesian inference and maximum entropy methods in science and engineering*, AIP Conference Proceedings 735, New York.

Smith, A.F.M., 1991. Bayesian computational methods, *Phil. Trans. R. Soc. Lond.,* A, **337,** 369–386.

Smith, A.F.M. & Roberts, G.O., 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *J. R. Statist. Soc.,* B., **55,** 3–23.

Stephenson, J., Gallagher, K. & Holmes, C.C., 2005. Beyond kriging ? dealing with discontinuous spatial data fields using adaptive prior information and Bayesian partition modeling, in *Geological Prior Information: informing Science and Engineering*, pp. 195–209, eds Curtis, A. & Wood, R., *Geol. Soc. Lond. Spec. Publication* 239.

Stephenson, J., Gallagher, K. & Holmes, C.C., 2006. Low temperature thermochronology and strategies for multiple samples 2: Partition Modelling for 2D/3D distributions with discontinuites, *Earth planet. Sci. Lett.* **241,** 557–570.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*, Siam, Philadelphia.

Tarantola, A. & Valette, B., 1982. Inverse problems = quest for information, *J. Geophys.,* **50,** 159–170.