# AALBORG UNIVERSITY

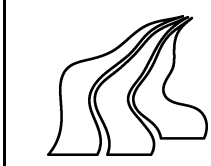
## Bayesian Model Discrimination for Inverse Problems

by

Kim E. Andersen, Stephen P. Brooks and Malene Højbjerre

March 2003 R-2003-06

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: www.math.auc.dk/research/reports/reports.htm



ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

# Bayesian Model Discrimination for Inverse Problems

Andersen, K.E.

*Department of Mathematical Sciences, Aalborg University*

Brooks, S.P.†

*Statistical Laboratory, University of Cambridge*

Højbjerre, M.

*Department of Mathematical Sciences, Aalborg University*

**Summary**. The Bayesian approach to the analysis of ill-posed inverse problems provides a number of distinct advantages over classical and deterministic alternatives. In this paper, we demonstrate how the Bayesian model determination problem can be tackled using trans-dimensional Markov chain Monte Carlo methods. In the context of several examples, we demonstrate how Markov chain mixing may be improved through the use of tempering-type methods and show how natural temperature schemes often arise in the context of many standard inverse problems. We use data arising from an intravenous glucose tolerance test to demonstrate the utility of these methods, comparing the standard so-called minimal model with a range of plausible alternatives.

*Keywords*: Markov chain Monte Carlo; Glucose-Insulin Homeostasis; Simulated Tempering; Parallel Tempering; Trans-dimensional MCMC

## 1. Introduction

Statistical inference was originally referred to as *inverse probability* relating, as it does, to the inversion of a joint probability distribution $f(\mathbf{\Phi} \mid \boldsymbol{\theta})$ for observed data $\mathbf{\Phi}$ given the value of model parameters $\boldsymbol{\theta}$, to obtain a function of those parameters given the data.

Though all of statistical inference is essentially concerned with the study of inverse problems, modern usage of the term *inverse problem* tends to refer to those that are defined to be *ill-posed* and which present special difficulties for the analyst. A general form for the deterministic inverse problem begins with a *forward problem* equating the data with some function of the parameters so that

$$\mathbf{\Phi} = F(\boldsymbol{\theta}). \tag{1}$$

The inverse problem thus involves the inversion of the function $F$ to obtain the value(s) of $\boldsymbol{\theta}$ as a function of the data. This problem is called *well-posed* if: (a) a solution for $\boldsymbol{\theta}$ exists for any $\mathbf{\Phi}$; (b) that solution is unique; and (c) the inverse mapping from $\mathbf{\Phi}$ to $\boldsymbol{\theta}$ is continuous. This latter condition is a necessary condition for the stability of the solution with respect to small errors in the data. In this paper we shall focus upon the analysis of the ill-posed stochastic inverse problem, noting that in most practical applications it is condition (b) that is violated.

There are at least two distinct approaches to the statistical inverse problem: the classical approach based upon the likelihood function $l(\boldsymbol{\theta} \mid \mathbf{\Phi})$ which is simply taken to be any function proportional to $f(\mathbf{\Phi} \mid \boldsymbol{\theta})$; and the Bayesian approach which is based upon a posterior distribution for the model parameters $\pi(\boldsymbol{\theta} \mid \mathbf{\Phi})$ which, from Bayes' theorem, is proportional to $f(\mathbf{\Phi} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ where $p(\boldsymbol{\theta})$ denotes a prior distribution representing the analyst's beliefs about the model parameters obtained independently from the data $\mathbf{\Phi}$. Ill-posed problems are often solved by imposing certain regularity conditions on the solution. This regularisation technique is equivalent to using a penalised likelihood, where the solution space is reduced by introducing a penalty function for implausible solutions. The Bayesian

†*Address for correspondence:* Steve Brooks, The Statistical Laboratory, Wilberforce Road, Cambridge, CB3 0WB, UK@. Email: steve@statslab.cam.ac.uk

approach provides a natural scheme for doing this, as implausible parameters are automatically penalised by the prior.

Bayesian approaches to statistical inverse problems have received a great deal of interest over the past few years (Kaipio *et al* 1999; Higdon *et al* 2001; Andersen *et al* 2001; Baussard *et al* 2001; Glad and Sebastiani 1995). There are several key advantages inherent in the Bayesian approach: (a) the availability of computational tools (such as Markov chain Monte Carlo – MCMC) which allow the construction and analysis of suitably complex models without the need for simplifying assumptions (e.g., generalising (1) to allow for stochastic functions $F$ and for complex forms of measurement error); (b) the ability to incorporate prior information into the analysis both to augment the observed data and as a very natural scheme for regularising the ill-posed problem; and (c) the potential for efficient model exploration via trans-dimensional (TD)MCMC. Whilst the first two of these potential advantages have already begun (though not yet fully) to be exploited within the inverse problems literature, the full potential of the last (perhaps the most important) remains largely unexplored.

In the context of inverse problems, where inference may be highly sensitive to model choice, model exploration and (potentially) model-averaging play pivotal roles in the Bayesian inferential process. Both are based upon the calculation of posterior model probabilities and require the use of TDMCMC algorithms to explore (and essentially weight) model space. One potential problem with the application of TDMCMC algorithms to inverse problems is that they tend to suffer from the (often severe) computational problem that likelihood calculations can be computationally very expensive and so the number of such calculations needs to be kept to a minimum. Thus, the construction of rapidly mixing chains (minimising run lengths) and efficient updating schemes are the keys to successful application of TDMCMC algorithms in this area – perhaps more so than in any other.

In this paper we discuss several techniques which use combinations of approximate forward solvers to improve within and between-model mixing as well as reducing computational expense per iteration. Drawing strongly on ideas such as simulated and parallel tempering, we develop generally applicable trans-dimensional sampling schemes that can be directly applied to the generic inverse problem. We demonstrate the utility of these methods, in terms of the efficiency of the resulting chains, as well as the advantages of the general Bayesian model averaging approach to estimation and prediction for inverse problems. Though several examples will be discussed, we will focus mainly upon the problem of modelling glucose disposal and insulin kinetics via the minimal model of Bergman *et al* (1979) and other related physiological models.

## 2. Bayesian Inference and Computation

If we assume that the data observed are described by some model $m$ with associated parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then the statistical problem is to estimate $\boldsymbol{\theta}$ given data $\boldsymbol{\Phi}$. In practice, Bayesian statistical inference is obtained through the calculation of posterior moments, for example the posterior mean of $\boldsymbol{\theta}$.

When the model itself is the subject of inference, the Bayesian posterior distribution can be extended to incorporate model as well as parameter uncertainty. By specifying a prior model probability, $p(m)$ for models $m \in \mathcal{M}$, the corresponding posterior distribution becomes

$$\pi(\boldsymbol{\theta}_m, m \,|\, \boldsymbol{\Phi}) \propto f_m(\boldsymbol{\Phi} \,|\, \boldsymbol{\theta}_m) p_m(\boldsymbol{\theta}_m) p(m),$$

where $f_m(\boldsymbol{\Phi} \,|\, \boldsymbol{\theta}_m)$ denotes the joint probability distribution of the data under model $m$ given parameter vector $\boldsymbol{\theta}_m \in \boldsymbol{\Theta}_m$, and $p_m(\boldsymbol{\theta}_m)$ denotes the corresponding prior for $\boldsymbol{\theta}_m$ under model $m$. Posterior inference is then often summarised in the form of the marginal posterior model probabilities

$$\pi(m \,|\, \boldsymbol{\Phi}) \propto \int \pi(\boldsymbol{\theta}_m, m \,|\, \boldsymbol{\Phi}) \, d\boldsymbol{\theta}_m.$$

These model probabilities may then be used either to discriminate between competing models using Bayes factors (Kass and Raftery 1995) or to provide model-averaged prediction for parameters that retain a coherent interpretation across models (Clyde 1999; Madigan *et al* 1996).

The computational problem is thus the integration of the posterior density function over what is typically a large and complex parameter space. MCMC methods overcome this problem by simulating realisations from the posterior distribution so that empirical estimates can be calculated for any statistic of interest. These realisations are obtained by simulating a Markov chain with the required stationary distribution. See Gilks *et al* (1996) and Brooks (1998), for example.

For the purposes of this paper, we shall distinguish between three separate simulation algorithms that will be combined to provide a technique suitable for parameter estimation and model exploration in the context of ill-posed inverse problems. We begin by briefly describing the Metropolis Hastings updating scheme for within-model exploration.

### 2.1. Metropolis Hastings Updates

Metropolis Hastings updates are used to move around parameter space by proposing moves which are subsequently either accepted or rejected. Suppose that we are currently in state $\boldsymbol{\theta}$, then we draw a new state $\boldsymbol{\theta}'$ from some proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$. This proposal is then subsequently accepted with probability

$$\alpha(\boldsymbol{\theta},\ \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}' \mid \boldsymbol{\Phi})q(\boldsymbol{\theta}', \boldsymbol{\theta})}{\pi(\boldsymbol{\theta} \mid \boldsymbol{\Phi})q(\boldsymbol{\theta},\ \boldsymbol{\theta}')}\right\}.$$

However, if the proposal is rejected, the chain remains in the current state. The choice of proposal distribution $q$ is essentially arbitrary, though several forms possess useful analytic properties. One particularly useful proposal scheme updates a subset of the parameters by sampling new values from their corresponding posterior conditional. This special case is known as a Gibbs sampler update (Casella and George 1992; Brooks 1998) and is often used when the corresponding posterior conditional distributions are of standard form.

### 2.2. Trans-Dimensional MCMC

Suppose that we wish to construct a Markov chain with stationary distribution $\pi(\boldsymbol{\theta}_m, m \mid \boldsymbol{\Phi})$, then the Metropolis Hastings updates described above can be used to perform within-model moves that keep the model fixed whilst updating the state vector. In order to move between models, we typically require what are known as trans-dimensional MCMC updates, since such updates often involve moving between states of different dimensions i.e., the dimension of the parameter vector under one model may be different from that under another.

Though other schemes exist (see e.g., Stephens 2000) reversible jump (RJ)MCMC updates (Green 1995) provide a natural extension to the basic Metropolis Hastings algorithm to allow for moves that change the dimension of the state vector. Suppose that we are currently in state $(\boldsymbol{\theta}_m, m)$ where $\boldsymbol{\theta}_m$ has dimension $n_m$ and that we wish to propose a move to model $m'$. Suppose also that we have defined one or more different move types allowing transitions between these two models spaces. Then the RJMCMC update proceeds via four distinct steps.

(a) With probability $p_m(r)$ we choose to perform move type $r$.
(b) Generate $\boldsymbol{u}$ from a specified proposal density $q_{r,m,m'}(\boldsymbol{u})$.
(c) Set $(\boldsymbol{\theta}_{m'}, \boldsymbol{u}') = g_{r,m,m'}(\boldsymbol{\theta}_m, \boldsymbol{u})$ where $g$ is a pre-specified invertible map between the parameter spaces under the two models, $n_m + n_{\boldsymbol{u}} = n_{m'} + n_{\boldsymbol{u}'}$, and $n_{\boldsymbol{u}}$ denotes the dimension of the vector $\boldsymbol{u}$.
(d) Accept $(m', \boldsymbol{\theta}_{m'})$ as the new state of the chain with probability $\alpha(m, m') = \min(1, A(m, m'))$ where

$$A(m, m') = \frac{\pi(\boldsymbol{\theta}_{m'}, m' | \boldsymbol{\Phi}) p_{m'}(r') q_{r',m',m}(\boldsymbol{u}')}{\pi(\boldsymbol{\theta}_m, m | \boldsymbol{\Phi}) p_m(r) q_{r,m,m'}(\boldsymbol{u})} \left| \frac{\partial g(\boldsymbol{\theta}_m, \boldsymbol{u})}{\partial(\boldsymbol{\theta}_m, \boldsymbol{u})} \right| \tag{2}$$

is called the acceptance ratio and $r'$ denotes the reverse move to $r$.

In practice, the implementation of the RJMCMC scheme can be rather problematic because of the need to specify the map $g$ between the parameter spaces of different models and the proposal, $q$.

For problems with only a small number of competing models the between-model transitions can be pilot-tuned beforehand (King and Brooks 2001; Richardson and Green 1997). However, for more general modelling problems more complex automated schemes are required (see Brooks *et al* 2003; Green 2000).

Alternatively, these transition problems can be overcome by periodically "relaxing" the algorithm so that the accept-reject step becomes more lenient, allowing a wider variety of proposals to be accepted. This can be achieved through the use of tempering-type approaches.

### 2.3. Tempering Algorithms

Marinari and Parisi (1992) suggest the use of simulated tempering as a means of improving the mixing rate of fixed-dimensional Markov chains. Here, we introduce a series of stationary distributions $\pi_1, \ldots, \pi_T$ each with support $\boldsymbol{\Theta}$, and augment the state vector to include an indicator variable signalling which distribution is being used at any time. If we let $\pi = \pi_1$ and choose $\pi_\tau,\ \tau = 2, \ldots, T$ so that movement within the distribution becomes easier as the temperature $\tau$ increases, then we can run a chain on the augmented state space $\boldsymbol{\Theta} \times \{1, \ldots, T\}$ with distribution $\pi(\boldsymbol{\theta}, \tau) = c_\tau \pi_\tau(\boldsymbol{\theta})$, where the $c_\tau$ denote weights incorporating the unknown normalisation constants associated with the $\pi_\tau$. Inference is then based only upon observations in the chain attributed to distribution $\pi_1$. The $\tau$'s are often referred to as temperatures with $\tau = T$ being the so-called hot temperature where movement is easiest and $\tau = 1$ the cold temperature, usually corresponding to the distribution of primary interest. Since movement is easier in the hotter temperatures, we obtain a more rapidly mixing chain and movement within the target distribution $\pi_1$ is facilitated by brief tours in other temperatures.

Given an appropriate collection of stationary distributions, the chain accepts swaps between temperatures $\tau$ and $\tau'$ in state $\boldsymbol{\theta}$ with probability

$$\alpha(\tau, \tau' \,|\, \boldsymbol{\theta}) = \min\left\{1, \frac{\pi_{\tau'}(\boldsymbol{\theta} \,|\, \boldsymbol{\Phi}) c_{\tau'} p_{\tau' \to \tau}}{\pi_\tau(\boldsymbol{\theta} \,|\, \boldsymbol{\Phi}) c_\tau p_{\tau \to \tau'}}\right\},$$

where $p_{\tau \to \tau'}$ denotes the probability of proposing to move to $\tau'$ from $\tau$. Once again, any general transition scheme is possible but experience suggests that schemes in which transitions are made only between neighbouring temperatures appears to work well. Thus, we might set $p_{\tau \to \tau+1} = p_{\tau \to \tau-1} = \frac{1}{2}$ for $\tau = 2, \ldots, T-1$, and $p_{1 \to 2} = p_{T \to T-1} = 1$. The normalisation constants $c_i$ are chosen so as to ensure that the chain divides its time roughly equally among the $T$ different samplers (Geyer and Thompson 1995) and often require some degree of pilot-tuning. We shall discuss the implementation of an automatic pilot-tuning procedure for selecting suitable normalisation constants in Section 3.

More general simulated tempering schemes are possible. In the case of model discrimination, the interpretation of the different parameters may change from model to model and this might change their values. Therefore, keeping the parameters fixed whilst the temperature is updated is not always appropriate. Suppose that at temperature $\tau$, $\pi_\tau(\boldsymbol{\theta})$ is defined upon some $\boldsymbol{\Theta}_\tau \subseteq \mathbb{R}^n$, so that the range of $\boldsymbol{\theta}$ may differ between temperatures. In that case, we may wish to alter the temperature update so that the parameter vector does not remain constant, but rather is updated together with the temperature. In this case, we might propose to move to $\tau'$ with probability $p_{\tau \to \tau'}$ and take $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}, \boldsymbol{\theta}'; \tau')$. This move would be accepted with probability $\alpha[(\boldsymbol{\theta}, \tau), (\boldsymbol{\theta}', \tau')] = \min(1, A[(\boldsymbol{\theta}, \tau), (\boldsymbol{\theta}', \tau')])$, where

$$A[(\boldsymbol{\theta}, \tau), (\boldsymbol{\theta}', \tau')] = \frac{\pi_{\tau'}(\boldsymbol{\theta}' | \boldsymbol{\Phi}) c_{\tau'} p_{\tau' \to \tau} q(\boldsymbol{\theta}', \boldsymbol{\theta}; \tau)}{\pi_\tau(\boldsymbol{\theta} | \boldsymbol{\Phi}) c_\tau p_{\tau \to \tau'} q(\boldsymbol{\theta}, \boldsymbol{\theta}'; \tau')}. \tag{3}$$

Alternatively if $\boldsymbol{\Theta}_\tau \subseteq \mathbb{R}^{n_\tau}$ i.e., at each temperature $\tau$ the corresponding density $\pi_\tau(\boldsymbol{\theta})$ is defined upon a different dimensional space, then the simulated tempering scheme essentially reduces to the RJMCMC scheme described above.

Parallel tempering provides an alternative to the simulated tempering scheme described above. Known also as *exchange Monte Carlo* (Hukushima and Nemoto 1996) and *Metropolis coupled* MCMC (Geyer 1991), the parallel tempering scheme constructs a Markov chain on the product space $\boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_T$ with stationary distribution

$$\pi^*(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T \,|\, \boldsymbol{\Phi}) = \prod_{\tau=1}^{T} \pi_\tau(\boldsymbol{\theta}_\tau \,|\, \boldsymbol{\Phi}),$$

so that parallel chains are run, each with a different stationary distribution. Instead of introducing a temperature update within each iteration, we propose instead to "swap" the states of any two chains.

More formally, at each iteration we randomly pick a pair of chains $\tau_1$ and $\tau_2 \neq \tau_1$, say, and propose swapping the states of these two chains. Suppose that chain $i$ is currently in state $\boldsymbol{\theta}_{\tau_i} \in (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$, then this proposal is accepted with probability

$$\min\left\{1, \frac{\pi_{\tau_1}(\boldsymbol{\theta}_{\tau_2} \mid \boldsymbol{\Phi})\pi_{\tau_2}(\boldsymbol{\theta}_{\tau_1} \mid \boldsymbol{\Phi})}{\pi_{\tau_1}(\boldsymbol{\theta}_{\tau_1} \mid \boldsymbol{\Phi})\pi_{\tau_2}(\boldsymbol{\theta}_{\tau_2} \mid \boldsymbol{\Phi})}\right\}.$$

Otherwise, the swap is rejected. This swap transition is then followed by a series of Metropolis Hastings transitions to update the state of each chain, preserving the corresponding stationary distribution. Realisations from the chain corresponding to the stationary distribution of interest may then be used for inference.

The advantage of parallel tempering over the simulated tempering scheme is that no pilot tuning of normalisation constants is required and the parallel tempering scheme works well whenever the stationary distributions for each of the chains are fairly similar in terms of the range of values with reasonable posterior mass. In fact the parallel tempering scheme even works well when the parallel chains all have the same stationary distribution since it is the ability to make "large" jumps through the swap transition that improves the mixing properties of the chain of primary interest. On the other hand, if the stationary distributions of the different chains are quite different with, for example, little or no overlap between them, then the swap transitions within the parallel tempering chain will rarely (if ever) be accepted and so no improvement in mixing is observed.

In contrast, the simulated tempering scheme works poorly when the stationary distributions at each temperature are too similar, since it is the ability of the chain to be able to mix better in at least one temperature that leads to an overall improvement in mixing. If the stationary distributions are too similar, the mixing rates will also be similar and no improvement will be observed. On the other hand, if the stationary distributions under the different chains differ greatly with at least one temperature providing a Markov chain with good mixing properties, then the whole chain will mix well. Of course, the transition between temperatures will be difficult to accept, but the simulated tempering algorithm can compensate for this via the normalisation constants. Thus, though many people have suggested that the simulated tempering scheme is hampered by the need to specify normalisation constants they do in fact provide an additional level of flexibility which facilitates jumps between quite different stationary distributions.

In practice, when we are interested in Bayesian model discrimination, even an apparently fine grid of temperatures may lead to stationary distributions with quite different posterior model probabilities. Thus, though the parallel tempering may work well in fixed-dimensional contexts, it will often fail when applied to trans-dimensional chains. The simulated tempering algorithm overcomes these problems both through the additional flexibility that the normalisation constants provide and due to the fact that only moves that update the model *or* the temperature are usually employed as opposed to the update of both simultaneously that occurs with the parallel tempering scheme.

In the next section, we will argue that for inverse problems in particular, natural temperature schemes often arise and we will discuss which of the two tempering schemes may be most appropriately applied in different general contexts.

## 3. The General Inverse Problem

A general form of the stochastic forward model is of the form

$$\boldsymbol{\Phi} = F(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ denotes some stochastic process accounting for errors made in measuring the data $\boldsymbol{\Phi}$ and $F$ may denote either a deterministic or stochastic function of the model parameters. For many problems, the function $F$ cannot be expressed in closed form and so approximations are often required in order to construct the likelihood necessary for statistical inference. Whether the function $F$ is available in

closed form or not, a characteristic feature of ill-posed inverse problems is that the evaluation of $F$ is typically computationally very expensive and so the number of evaluations needs to be kept to a minimum.

*Example I: Image Analysis*

Suppose we wish to model the number of photons hitting an $n \times n$ grid of detectors over a fixed period of time. We record the number of hits at each detector by $\mathbf{\Phi} = \{\phi_{1,1}, \ldots, \phi_{n,n}\}$ and assume that these observations follow a Poisson distribution with intensity $\lambda_{ij}$ at cell $(i, j)$, so that

$$f(\mathbf{\Phi} \,|\, \boldsymbol{\lambda}) \propto \prod_{i=1}^{n} \prod_{j=1}^{n} \lambda_{ij}^{\phi_{ij}} \exp(-\lambda_{ij}).$$

Finally, we take a simple Gaussian Markov random field prior for the unknown image intensities,

$$p(\boldsymbol{\lambda} \,|\, \psi) \propto \psi^{n^2/2} \exp(-\frac{1}{2}\psi \boldsymbol{\lambda}^T C \boldsymbol{\lambda}),$$

where $C$ denotes an $n^2 \times n^2$ pre-specified precision matrix and $\psi$ a smoothness parameter to be estimated. In practice $C$ usually imposes some form of neighbourhood structure so that knowledge about the intensity at one pixel will be related to that at its neighbours. For example, we might take $C_{i,i}$ equal to the number of pixels adjacent to pixel $(i, i)$ and $C_{i,j}$ equal to 1, whenever pixels $i$ and $j$ are adjacent and zero otherwise. The Bayesian analysis is therefore based upon the posterior distribution

$$\pi(\boldsymbol{\lambda}, \psi \,|\, \mathbf{\Phi}) \propto f(\mathbf{\Phi} \,|\, \boldsymbol{\lambda}) p(\boldsymbol{\lambda} \,|\, \psi) p(\psi),$$

with $p(\psi)$ denoting the prior placed on $\psi$, often taken to be a gamma distribution for conjugacy. Here, we have $\mathbf{\Phi} = F(\boldsymbol{\lambda}, \psi)$ where $F$ is a stochastic function and we have no observation error. Different models can be expressed in terms of restrictions placed upon the $\lambda_{ij}$. For example we might take $\lambda_{ij} \equiv \lambda_i \, \forall j$, suggesting that the intensity is constant across rows. Under any model, it is clear that for large $n$, evaluation of the posterior distribution can be extremely computationally intensive.

*Example II: Electrical Impedance Tomography*

Andersen *et al* (2001) discuss the problem of locating perfectly insulating line segments (cracks) within a solid electrically conducting media $\Omega$. They show that given an induced current and observed voltage readings $\Phi(t_i)$ at points $t_i$, $i = 1, \ldots, m$ at the surface, the joint probability distribution of the data, given a particular configuration of linear perfectly insulating cracks within the media, $\boldsymbol{\theta}$ is given by

$$f(\mathbf{\Phi} \,|\, \boldsymbol{\theta}, \gamma) \propto \exp\left( -\sum_{i=1}^{m} [\mathbf{\Phi}(t_i) - D_{\boldsymbol{\theta},\gamma}(t_i)]^2 / 2\sigma^2 \right),$$

where $D_{\boldsymbol{\theta},\gamma}(t_i) = S_{\boldsymbol{\theta},\gamma}(t_i) - S_{\boldsymbol{\theta},\gamma}(t_{i-1})$, and $S_{\boldsymbol{\theta},\gamma}$ is the solution operator to the differential equation

$$\nabla \cdot (\gamma^{-1} \nabla \boldsymbol{u}) = 0 \quad \text{in } \Omega \setminus \boldsymbol{\theta},$$

with respect to $\boldsymbol{u}$ and subject to a series of boundary conditions. Both the background conductivity $\gamma$ and the measurement error variance $\sigma^2$ are unknown quantities subject to estimation. Here, we have model $\mathbf{\Phi}(t_i) = D_{\boldsymbol{\theta},\gamma}(t_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ and $D_{\boldsymbol{\theta},\gamma}$ is the deterministic function described above. The solution to the steady state partial differential equations cannot be written in closed form and so a discretized approximation is generally used (Bryan and Vogelius 1994). As with the image analysis example above, the solution of the partial differential equation at each of the data points is highly computationally intensive and so considerable computational expense is involved in evaluating the likelihood. Model choice questions arise when the number of cracks in the media is unknown in which case the dimensionality of $\boldsymbol{\theta}$ will change as we add and delete cracks to the model.

The Bayesian analysis of these systems involves the repeated evaluation of the likelihood function with usually at least two evaluations per update. Computational expense can therefore be minimised in any one of three ways: (a) by minimising the expense per evaluation; (b) by minimising the number of evaluations per iteration; and (c) by minimising the number of iterations. All of these are important considerations for any MCMC simulation, but the characteristically high expense of evaluating the likelihood for inverse problems such as these, makes them a necessity rather than a convenience.

### *3.1. Minimising the Expense per Evaluation*

For many inverse problems, and certainly in the case of the two described above, the expense of evaluating the likelihood can be reduced by using an appropriate approximation. Such approximations can take a variety of forms. For example we might reduce the number of data points and use the likelihood for these reduced data as an approximation to the true likelihood. Alternatively, since for many inverse problems the likelihood is already an approximation, the resolution of that approximation can be reduced to improve efficiency.

*Example I continued*

In the context of the image analysis example, we can approximate the full likelihood by combining pixels to form a lower resolution image. For example an $n/2 \times n/2$ image can be formed by arranging pixels in groups of four. This reduces the size of the $C$ matrix by a factor of four, thereby reducing computation time by a similar factor. Coarser approximations can be achieved by grouping larger collections of pixels. See Higdon *et al* (2001), for example.

*Example II continued*

In the context of the electrical impedance tomography (EIT) example, Andersen *et al* (2001) suggest solving the differential equations by integrating both sides to form a series of integral equations. Nystrøm's method (see e.g. Atkinson 1976) then discretises these integral equations by quadrature rules to obtain a set of linear equations, which can be solved. Bryan and Vogelius (1994) use $k$ nodes within the quadrature rule so that the accuracy of the approximation improves with increasing $k$. In practice, we would identify a particular resolution level, $k^*$ providing satisfactory inference, but lower resolution levels (i.e., smaller values of $k$) would provide coarser approximations that would be faster to compute.

Of course, by approximating the likelihood in this way, we obtain a different posterior distribution which, typically, will be less strongly influenced by the data than that based upon the full likelihood. In practice, these coarser likelihood approximations will generally be of little or no inferential value in themselves. However, as we shall see below, they may be used in combination with the full likelihood in order to provide useful inference with less computational expense.

### *3.2. Minimising the Expense per Iteration*

Minimising the expense per iteration typically means fewer likelihood evaluations per iteration. A simple way to halve the number of likelihood evaluations is to temporarily store the likelihood corresponding to the current state at each iteration so that it can be used in the denominator for the next accept-reject step. More generally, fewer likelihood evaluations per iteration, in turns, means fewer Metropolis Hastings steps. The number of Metropolis Hastings steps can be minimised in two ways:

first, by making as much use of Gibbs sampler updates as possible (which do not require the accept-reject step and therefore require no likelihood calculations); and second, by updating parameters in groups rather than individually.

It is well known that grouping highly correlated parameters and updating them together using a single Metropolis Hastings updates can improve the overall mixing rate of the resultant Markov chain (Roberts and Sahu 1997). However, it is also possible to update uncorrelated parameters as groups. Using independent proposals, updating two parameters together will lead to an acceptance ratio equal to the product of the two corresponding acceptance ratios, but an acceptance probability that is greater than or equal to the product of the two corresponding acceptance probabilities, since $\min(1, p_1 p_2) \leq \min(1, p_1)\min(1, p_2)$. Thus, the chance of accepting both moves is greater if they are combined. Of course, the overall probability of accepting any jump at all decreases, since by combining the updates it is no longer possible to accept just one of the two proposals. Thus, Markov chain mixing may not be improved by combining uncorrelated parameters. However, the saving in computational expense per iteration may more than offset the increase in run lengths required to obtain similar levels of Monte Carlo errors.

### *3.3. Minimising the Number of Iterations*

Minimising the number of iterations essentially means improving the mixing rate of the Markov chain. As we discuss above, grouping correlated parameters may well improve mixing but, in practice, finding suitable proposals for multivariate updates may be difficult. In addition between-model mixing is often very slow and with likelihoods (and, correspondingly, posteriors) that cannot be written in closed form, it is often very difficult to find proposals that lead to satisfactorily high acceptance rates.

One way around this problem is to use one of the two tempering schemes introduced in Section 2. The simulated tempering scheme improves mixing by introducing temperatures with flatter posterior distributions within which it is easier to move. Similarly, between model transitions are often easier within higher temperatures as the corresponding posterior distributions are less concentrated and more forgiving of poorly chosen proposal distributions. Parallel tempering can also improve mixing since it provides an essentially automatic proposal generating mechanism which, under the right conditions, is capable of proposing large jumps with reasonably high acceptance rates, even for trans-dimensional jumps.

Within the general inverse problem framework these tempering algorithms are made all the more attractive by the existence, for many problems, of a very natural series of stationary distributions. As we discussed above, the computational expense of evaluating the likelihood can often be decreased by either decreasing the sample size or by using increasingly coarse approximations to the true likelihood.

*Example I continued*

In the context of the image analysis example, a suitable set of stationary distributions might be obtained by taking

$$\pi_\tau(\boldsymbol{\lambda}_\tau, \psi \mid \boldsymbol{\Phi}_\tau) \propto f(\boldsymbol{\Phi}_\tau \mid \boldsymbol{\lambda}_\tau) p(\boldsymbol{\lambda}_\tau \mid \psi) p(\psi),$$

where $\boldsymbol{\Phi}_\tau$ corresponds to an $n/2^\tau \times n/2^\tau$ matrix of observations formed by grouping the original data into squares comprising $2^\tau$ neighbouring pixels, with $\boldsymbol{\lambda}_\tau$ denoting the corresponding vector of intensities. Obviously, moves from one temperature to the next may involve a change in dimension for the state vector $\boldsymbol{\lambda}_\tau$, depending on the model. This causes no problems for the parallel tempering algorithm, because the presence of the posterior distribution for both temperatures in both the numerator and denominator of the acceptance ratio ensures that the dimension matching condition (Green 1995) is automatically satisfied. In order to implement the simulated tempering algorithm, we require the trans-dimensional equivalent of the simulated tempering scheme, proposing values for the additional intensity parameters necessary to jump to the stationary distribution with the finer resolution data. Thus, the parallel tempering scheme might be the simplest to implement for this example, particularly, if we were to consider only one model.

*Example II continued*

In the context of the EIT example, a suitable set of stationary distributions might be obtained by taking

$$\pi_\tau(\boldsymbol{\theta} \,|\, \boldsymbol{\Phi}) \propto f_\tau(\boldsymbol{\Phi} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where $f_\tau(\boldsymbol{\Phi} \,|\, \boldsymbol{\theta})$ denotes an approximation to the joint probability distribution of the data given the model parameters using $\tau \leq k^*$ nodes within the quadrature procedure. Here, the dimensionality of the state vector remains the same in all temperatures and therefore both tempering schemes could be used here for fixed-model simulations.

In some cases, it may not be possible to either reduce or combine the data, or to use different levels of approximation to the likelihood. In such cases, a suitable set of stationary distributions may be obtained by taking

$$\pi_\tau(\boldsymbol{\theta}|\boldsymbol{\Phi}) \propto f(\boldsymbol{\Phi} \,|\, \boldsymbol{\theta})^{1/\tau} p(\boldsymbol{\theta}).$$

Clearly, as $\tau$ increases, the influence of the data on the posterior is diminished and, as $\tau \to \infty$, the posterior converges to the prior. This provides a generic set of tempering distributions that can be used for almost any fixed or variable dimensional problem, so long as the prior distribution is proper.

In terms of which of the two tempering algorithms are best applied in which context, our experience suggests that using the final set of tempering distributions, in which the likelihood contribution is weakened as $\tau$ increases, the posterior model probabilities in particular can change quite dramatically with temperature and so the parallel tempering scheme tends to perform poorly. On the other hand, with tempering distributions based upon the coarseness of the likelihood, the change in the posterior tends to be less dramatic and the performance of the parallel tempering scheme improves considerably.

### 3.4. Tuning the Proposals

With any (TD)MCMC algorithm, an element of pilot-tuning of the proposals is nearly always necessary. Our proposal for an algorithm combining Metropolis Hastings, tempering and reversible jump MCMC updates therefore requires a degree of pilot-tuning for each of these three transition types.

#### 3.4.1. Tuning the Tempering Schemes

The tempering schemes require the specification of an appropriate set of temperatures and, for the simulated tempering scheme, the corresponding normalisation constants.

In specifying the number of temperatures, a trade-off is sought between the number of temperatures used (which should be kept as small as possible) and the ability of the chain to move adequately around the state space. The latter requires that a reasonably high temperature is included which allows rapid movement within that temperature and then as few temperatures must be added in between this temperature and the cold distribution to allow the chain to jump from one temperature to the next. In practice, tuning the temperature schedule is somewhat more critical for the parallel tempering scheme, since for the simulated tempering scheme the normalisation constants can be chosen so as to compensate for a poor choice of temperatures. A suitable hot distribution can usually be found by using pilot runs and then, by sequentially adding intermediate temperatures until satisfactory performance is achieved, the full temperature schedule can be obtained.

As discussed in Section 2, the parallel tempering scheme does not require pilot tuning as the swap transitions are completely defined by the tempering distributions chosen. However, in the simulated tempering scheme, normalisation constants which allow rapid transitions between temperatures need to be found. This can often be a time-consuming process and so we propose the following automated scheme. Given a simulated tempering scheme with temperatures $\tau = 1, \ldots, T$, we begin by setting $c_\tau = 1$ for all $\tau = 1, \ldots, T$. We fix $c_1 = 1$ to ensure a unique solution and, without loss of generality, restrict attention to tempering schemes under which transitions may only be made

between neighbouring temperatures. We begin our simulated tempering scheme in some arbitrary starting position and after each temperature transition, we update the normalisation constants as follows. Suppose we currently have normalisation constants $c_1, \ldots, c_T$ and that we have just proposed a move from temperature $\tau$ to $\tau + 1$, say. Whether or not we accept this move we now divide all constants $c_{\tau+1}, \ldots, c_T$ by the acceptance ratio for the proposed move. Thus, if the move attracts a low acceptance probability, the normalising constant for the proposed model is increased. Similarly, since we have learnt only about the $\tau \to \tau + 1$ transition from this proposed move, all of the normalisation constants to the right of $\tau$ are increased to preserve their relative size. Similarly, if we propose to move from $\tau$ to $\tau - 1$, we would *multiply* the constants $c_\tau, \ldots, c_T$ by the corresponding acceptance ratio. This procedure is repeated until the weights settle to roughly constant values and, once the weights have been determined, they can then be used as inputs for the main simulation.

Of course, when we wish to simulate trans-dimensional chains, we may need to specify different weights under different models. Thus, we obtain not just $c_\tau$, but $c_{\tau,m}$. In order to ensure that the prior model probabilities are not affected, we set $c_{1,m} = 1$ for all models $m$ and then estimate the remaining weights either using a separate simulation for each model or using a single pilot-tuning run of the trans-dimensional chain.

### *3.4.2. Within-Model Updates*

Within-model updates are also tuned via an initial pilot study. Here, we adapt the proposal scales as follows. We run the simulation for an initial $N$ iterations and then calculate the mean acceptance ratio for the updates for each parameter in the model. For any parameter with a mean acceptance ratio less than 0.2, we divide the current proposal variance by $\delta$. For any parameter with a mean acceptance ratio greater than 0.5, we multiply the current proposal variance by $\delta$. Taking $N = 100$ and $\delta = 1.1$ appears to perform well in practice. This process is continued until all mean acceptance ratios lie within $(0.2, 0.5)$, see Gelman *et al* (1996).

### *3.4.3. Between-Model Updates*

The between model updates are performed by using uniform proposals for any parameters being added to the current model. The end points of the uniform distribution are taken to correspond to a roughly 95% interval under the prior. In fact, because most of the between-model transitions are made at higher temperatures, where the target distribution is fairly flat, the performance of the algorithm is not greatly affected by the specification of the between-model proposals, which makes the use of tempering schemes for TDMCMC a particularly useful device for facilitating between-model mixing.

## 4. Analysing Glucose-Insulin Homeostasis

Here we consider the analysis of data arising from an intravenous glucose tolerance test (IVGTT) study in which a dose of glucose is administered intravenously to a patient over a 60-second period. The glucose and insulin concentrations within the bloodstream are then recorded at pre-specified and irregular times, see Pacini and Bergman (1986), for example. A sample of observations provided by Novo Nordisk A/S is plotted in Figure 1.

The intravenous glucose dose rapidly elevates the glucose concentration in the patient's bloodstream initiating the pancreatic $\beta$-cells in a healthy patient to secrete insulin. The level of insulin within the bloodstream therefore increases, triggering the absorption of glucose from the bloodstream into the adipose tissue, liver and muscles. This absorption process lowers the glucose concentration, affecting a reduction in pancreatic insulin production. This, in turn reduces the rate of glucose absorption and the cycle continues until the levels of both glucose and insulin within the bloodstream return to normal. In a healthy individual the entire process should normally take less than three hours.

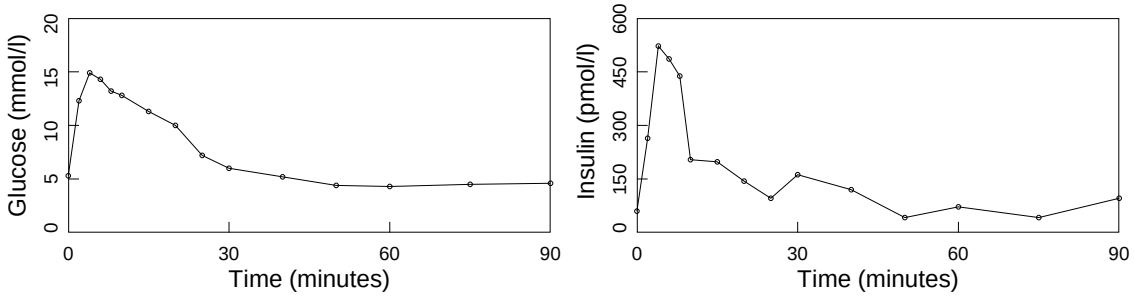



**Fig. 1.** Glucose and insulin concentrations in a healthy patient recorded over a 90-minute period after an intravenous glucose injection.

### *4.1. The Minimal Model*

The medical literature provides a range of alternative models to describe the glucose-insulin system. However, the minimal model proposed by Bergman *et al* (1979) is by far the most popular of the various alternatives. The minimal model is wholly deterministic and assumes that the system can be described by the following set of differential equations.

$$\begin{aligned}
\dot{G}(t) &= -S_G(G(t) - G_b) - X(t)G(t), & G(0) &= G_0,\\
\dot{X}(t) &= -p_2(X(t) + S_I(I(t) - I_b)), & X(0) &= 0,\\
\dot{I}(t) &= -n(I(t) - I_b) + \gamma J_+(G(t) - h)t, & I(0) &= I_0,
\end{aligned}$$

where $t = 0$ is the glucose injection time; $J_+(x) = x$ if $x > 0$ and zero otherwise; $G(t)$ denotes the glucose concentration at time $t$; $X(t)$ denotes the remote insulin action at time $t$ (i.e., the action of the insulin upon the absorption rate of glucose); $I(t)$ denotes the insulin concentration at time $t$; and the remaining parameters are unknowns to be estimated (see Appendix A for an interpretation). We label the insulin model above as $I_1$ and the combined glucose and insulin action model as $G_1$.

This deterministic model is usually estimated by non-linear weighted least squares estimation in a two-step procedure, described by Pacini and Bergman (1986) for example. In the first stage the parameters in the first two equations are estimated treating insulin as a known forcing function. In the second stage the third equation is solved treating glucose as a known forcing function. There are many problems with this approach, not least the fact that negative confidence intervals are often obtained for strictly positive parameters, such as $S_I$. Pillonetto *et al* (2002) adopt a Bayesian approach, but consider fitting only the system of $G$ and $X$ equations, treating $I$ as known. The basis for the minimal model assumes that the parameters $G$, $X$ and $I$ constitute a single dynamical system and important information is lost in trying to treat the system in two separate stages or in only analysing parts of it. Further, as pointed out by de Gaetano and Arino (2000), the coupled model may, for even commonly observed combinations of parameter values, fail to admit an equilibrium, i.e. the coupled system becomes highly ill-posed.

Andersen and Højbjerre (2003) demonstrate how a Bayesian approach can be used to solve the ill-posed inverse problem of the entire set of system equations by log-transforming the system to be on the same scale, discretising and then embedding it within a stochastic state space modelling framework allowing for errors both within the system itself and within the measurement process. In particular, they show that if we let $g_t^s = \log G(t)$, $x_t^s = \log X(t)$ and $i_t^s = \log I(t)$ denote the true system values of the log glucose, log remote insulin action and log insulin levels at time $t$; adopt a Brownian motion error process within the system; let $g_t^o$ and $i_t^o$ denote the observed log-levels of two corresponding system values; and adopt a normal measurement error process; then the corresponding likelihood for the system can be written as

$$L(\boldsymbol{\theta}, \boldsymbol{\Phi}^s \mid \boldsymbol{\Phi}^o) = f(\boldsymbol{\Phi}^s \mid \boldsymbol{\theta}) f(\boldsymbol{\Phi}^o \mid \boldsymbol{\Phi}^s, \boldsymbol{\theta}), \tag{4}$$

where $\boldsymbol{\Phi}^s = \{g_t^s, x_t^s, i_t^s\}_{t\in\Lambda}$; $\Lambda$ denotes the set of time points chosen for discretising the system; $\boldsymbol{\Phi}^o = \{g_t^o, i_t^o\}_{t\in\mathcal{T}}$; $\mathcal{T} \subseteq \Lambda$ denotes the set of observation times; $\boldsymbol{\theta} = (S_G, p_2, S_I, n, \gamma, h, G_b, I_b, g_0, i_0, \nu_{g^s}, \nu_{x^s},$

$\nu_{i^s}, \nu_{g^o}, \nu_{i^o}$); and the $\nu$'s denote the precision of the corresponding error processes on the system and measurements. The larger the value of $|\Lambda|$, the smaller the time intervals in the discretisation and the more accurate the approximation. Thus, the time intervals indicates a level of coarseness of the solution. In practice, at resolution level $\tau$, we divide the interval $[t_j, t_{j+1}]$ into $k_{j\tau}$ sections so that $k_{j\tau}$ quantifies the coarseness and $\tau$ indexes the coarseness level. As an example, we might set $k_{j\tau} = \tau$, so that each time period is divided into an equal number parts. We return to the definition of $k$ in the next section. Of course, for large values of $|\Lambda|$ the likelihood evaluations are extremely computational expensive.

The two contributions to the likelihood in (4) are

$$f(\boldsymbol{\Phi}^s \mid \boldsymbol{\theta}) \propto (\nu_{g^s}\nu_{x^s}\nu_{i^s})^{|\Lambda|/2} \exp(-V(\boldsymbol{\Phi}^s, \boldsymbol{\theta})),$$
$$f(\boldsymbol{\Phi}^o \mid \boldsymbol{\Phi}^s, \boldsymbol{\theta}) \propto (\nu_{g^o}\nu_{i^o})^{|\mathcal{T}|/2} \exp(-W(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta})),$$

where

$$V(\boldsymbol{\Phi}^s, \boldsymbol{\theta}) = \frac{1}{2}\sum_{t_k \in \Lambda} \nu_{g^s}(g^s_{t_k} - f^g_{t_k})^2 + \nu_{x^s}(x^s_{t_k} - f^x_{t_k})^2 + \nu_{i^s}(i^s_{t_k} - f^i_{t_k})^2,$$
$$W(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}) = \frac{1}{2}\sum_{t_k \in \mathcal{T}} \nu_{g^o}(g^o_{t_k} - g^s_{t_k})^2 + \nu_{i^o}(i^o_{t_k} - i^s_{t_k})^2,$$

and

$$f^g_{t_k} = g_{t_{k-1}} - (t_k - t_{k-1})\big(S_G(1 - G_b\, e^{-g_{t_{k-1}}}) + e^{x_{t_{k-1}}}\big),$$
$$f^x_{t_k} = x_{t_{k-1}} - (t_k - t_{k-1})p_2\big(1 - S_I(e^{i_{t_{k-1}}} - I_b)e^{-x_{t_{k-1}}}\big),$$
$$f^i_{t_k} = i_{t_{k-1}} + (t_k - t_{k-1})\big(-n(1 - e^{-i_{t_{k-1}}} I_b) + e^{-i_{t_{k-1}}}\gamma J_+(e^{g_{t_{k-1}}} - h)t_{k-1}\big).$$

Space constraints prevent us from describing the likelihood derivation in greater detail, but the interested reader is referred to Andersen and Højbjerre (2003) for further details and for discussion on the implementation of an MCMC algorithm to perform reliable inference for this model. Here, we extend the analysis of Andersen and Højbjerre (2003) to consider the problem of model selection, and investigate a variety of plausible alternatives to the minimal model.

### *4.2. Alternative Models*

The glucose-insulin model has attracted many mathematical descriptions. Using a combination of Akaike information criteria and arguments based upon the minimisation of residual sums of squares, Bergman *et al* (1979) showed that the minimal model dominated a large range of alternative models. Nevertheless, several criticisms of the model remain, namely: (a) that the positive truncation $J$ is physiologically questionable; and (b) that the multiplicative effect of time $t$ in system $I_1$ suggests that the effect of circulating hyperglycemia on the rate of pancreatic secretion of insulin is proportional to the time elapsed from the glucose stimulus (Toffolo *et al* 1980), which is difficult to justify biologically. We therefore consider three additional variants of the insulin component of the model:

$$\dot{I}(t) = -n(I(t) - I_b) + \gamma J_+(G(t) - h), \quad I(0) = I_0,$$
$$\dot{I}(t) = -n(I(t) - I_b) + \gamma(G(t) - h)t, \quad I(0) = I_0,$$
$$\dot{I}(t) = -n(I(t) - I_b) + \gamma(G(t) - h), \quad I(0) = I_0.$$

We label these alternative insulin models $I_2$, $I_3$ and $I_4$ respectively.

In addition, we also consider the delay-differential model recently proposed by de Gaetano and Arino (2000) in which the coupling of the insulin and glucose processes is made without the use of non-observable state variables $X(t)$. The model is given by the following differential equations

$$\dot{G}(t) = -S_G(G(t) - G_b) - S_I(I(t)G(t) - I_bG_b), \quad G(0) = G_0,$$
$$\dot{I}(t) = -n\left[I(t) - \frac{I_b}{b_5 G_b}\int_{t-b_5}^{t} G(s)\, ds\right], \quad I(0) = I_0,$$

where $G(t) \equiv G_b$ for $t \in [-b_5, 0)$. We label the glucose model above as $G_2$ and the insulin model as $I_5$.

### 4.3. Implementation

Combining the two glucose models with the five different models for the insulin process, we obtain ten distinct models between which we wish to discriminate. A uniform prior model probability is specified for the ten models and we adopt a vague prior distribution on the parameter vector $\boldsymbol{\theta}_m$ in order to ensure that the posterior distribution is dominated by the likelihood. For mathematical convenience we assume that the precisions of the error processes in the system are identical, i.e. $\nu = \nu_{g^s} = \nu_{x^s} = \nu_{i^s}$ and that the precisions $\nu$, $\nu_{g^o}$ and $\nu_{i^o}$ are gamma distributed with large variances. For the remaining parameters, we adopt log-normal priors with large variances.

As a tempering scheme, we take

$$\pi_\tau(\boldsymbol{\theta}_m, m \mid \boldsymbol{\Phi}) \propto \left[ f_{m,\tau}(\boldsymbol{\Phi}^s \mid \boldsymbol{\theta}_m) f_{m,\tau}(\boldsymbol{\Phi}^o \mid \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m) \right]^{g(\tau)} p_m(\boldsymbol{\theta}_m) p(m), \tag{5}$$

where $g(\tau) = 1/2^{(\tau-1)n}$ for $n > 0$ and $f_{m,\tau}$ denotes the corresponding likelihood under model $m$, at coarseness level $\tau$. Here, we take $k_{j\tau} = 2^{a_j}/\tau - 1$, where $a_j$ is chosen sufficiently large so that all observations in $\boldsymbol{\Phi}$ are used for approximating the joint probability distribution. In order to account for the perturbation of the glucose-insulin system at time $t = 0$, we let $a_j$ decrease gradually over time as to allow for better approximation at the beginning of the experiment than at the end, i.e. we choose $a_j = 5$ for $t_{j+1} \leq 2$, $a_j = 4$ for $2 < t_{j+1} \leq 10$ and $a_j = 3$ otherwise. Here, we take $T = 4$ temperatures. Since we are powering down the likelihood in (5) and proposing trans-dimensional jumps within the simulation, we use the simulated tempering scheme (as opposed to the parallel tempering algorithm) and propose updating the temperature every 50 iterations. Trans-dimensional moves are proposed every 500 iterations and, in order to allow the perturbed glucose-insulin system to stabilise, we zero-weight observations taken at time $t < 8$ minutes.

An initial pilot-run of 1 million iterations was used to determine the within-model proposals and the normalisation constants for the tempering scheme. A second MCMC simulation comprising 2 million iterations was then used for inference. Specific details of the various transition schemes used in the simulation are given in Appendix B.

### 4.4. Results

Analysing the data summarised in Figure 1, we obtain positive posterior mass on only three models: $G_1/I_1$, $G_1/I_2$ and $G_1/I_4$. Here model $G_i/I_j$ denotes the model with glucose model $i$ and insulin model $j$. The corresponding posterior model probabilities are 0.88, 0.11 and 0.01 respectively. Clearly, Bergman's minimal model dominates all others for this particular data set and there is no posterior support for the glucose model of de Gaetano and Arino (2000). There is some support for the idea that the time factor in the glucose component of the minimal model might be removed but, with a Bayes factor of 7.6, the evidence is very weak indeed. Overall, we would conclude that the minimal model is most appropriate for this particular data set though, with over 12% of the posterior mass on alternative models, model averaged parameter estimates may be useful as they would properly reflect this element of uncertainty.

Table 1 provides the posterior means and credible intervals of the parameters under the three models identified, together with model-averaged values. Each of these parameters retain a consistent interpretation across the three models identified and we can see that most parameters vary very little between models. An exception is the parameter $p_2$ which, under the *a posteriori* most probable model has a credible interval of $(139.7, 16\,313.0)$, but a model-averaged interval of $(157.1, 17\,075.5)$. The moderate increase in width here reflects the additional uncertainty due to the model, though clearly the data provides rather little information as to the true value of this parameter.

Perhaps one of the most important outcomes of the analysis, is the estimation of the glucose process, together with a measure of our associated uncertainty. A graphical representation of the true

**Table 1.** Posterior mean and 95% credible intervals for the parameters in model $G_1/I_1$, $G_1/I_2$ and $G_1/I_4$ together with a model-averaged posterior mean and a model-averaged 95% credible intervals for the same parameters.

| | Model $G_1/I_1$ | | | Model $G_1/I_2$ | | |
|---|---|---|---|---|---|---|
| Parameters | Posterior mean | 95% credible interval | | Posterior mean | 95% credible interval | |
| $S_G$ | 0.072 | 0.064 | 0.082 | 0.072 | 0.065 | 0.079 |
| $S_I \cdot 10^5$ | 50.667 | 3.616 | 115.041 | 33.894 | 1.710 | 104.632 |
| $n$ | 0.083 | 0.052 | 0.127 | 0.087 | 0.055 | 0.141 |
| $\gamma \cdot 10^4$ | 331.194 | 3.896 | 1 016.415 | 378.168 | 11.131 | 939.292 |
| $G_b$ | 4.340 | 4.149 | 4.552 | 4.325 | 4.133 | 4.519 |
| $I_b$ | 61.725 | 52.281 | 71.891 | 61.955 | 52.046 | 72.445 |
| $p_2$ | 5 252.853 | 139.668 | 16 312.985 | 7 897.193 | 609.328 | 17 182.223 |
| $h$ | 531.065 | 22.407 | 1 858.416 | 207.807 | 7.905 | 1 062.872 |
| $G_0$ | 22.496 | 20.195 | 25.475 | 22.359 | 20.904 | 24.550 |
| $I_0$ | 583.127 | 345.881 | 976.437 | 635.645 | 353.219 | 1 257.423 |
| | Model $G_1/I_4$ | | | Model averaged | | |
| Parameters | Posterior mean | 95% credible interval | | Posterior mean | 95% credible interval | |
| $S_G$ | 0.071 | 0.067 | 0.074 | 0.072 | 0.064 | 0.081 |
| $S_I \cdot 10^5$ | 29.881 | 24.752 | 33.811 | 48.221 | 2.935 | 114.305 |
| $n$ | 0.080 | 0.058 | 0.111 | 0.084 | 0.053 | 0.129 |
| $\gamma \cdot 10^4$ | 3.725 | 0.184 | 9.177 | 336.261 | 3.955 | 1 010.129 |
| $G_b$ | 4.371 | 4.193 | 4.450 | 4.338 | 4.145 | 4.549 |
| $I_b$ | 62.387 | 54.342 | 72.480 | 61.761 | 52.255 | 72.000 |
| $p_2$ | 2 559.485 | 2 536.075 | 2 587.847 | 5 610.851 | 157.155 | 17 075.480 |
| $h$ | 9.653 | 6.646 | 13.669 | 483.368 | 13.946 | 1 857.206 |
| $G_0$ | 22.052 | 21.756 | 22.248 | 22.474 | 20.247 | 25.372 |
| $I_0$ | 525.016 | 384.014 | 733.127 | 590.216 | 346.921 | 1 019.151 |

underlying glucose process is provided in Figure 2 under the three top models, together with a model-averaged estimate. It is apparent that the three models with positive posterior mass all provide similar descriptions of the glucose disposal process. (Recall that the glucose-insulin system takes a few minutes to settle down after the initial injection and that we zero-weight observations during the first 8 minutes. Thus, the model does not attempt to fit observations during this period.)

Curves such as these can be used to monitor a particular patients' reaction to the IVGTT test to determine whether or not the patient is in the risk group for developing type 2 diabetes. Gaining a reliable estimate of the true underlying glucose disposal process facilitates a more robust classification procedure which is further improved by the provision of the corresponding error bounds. It is on the basis of such plots that the reaction to treatment, for example, is judged and so obtaining reliable estimates is of paramount importance.

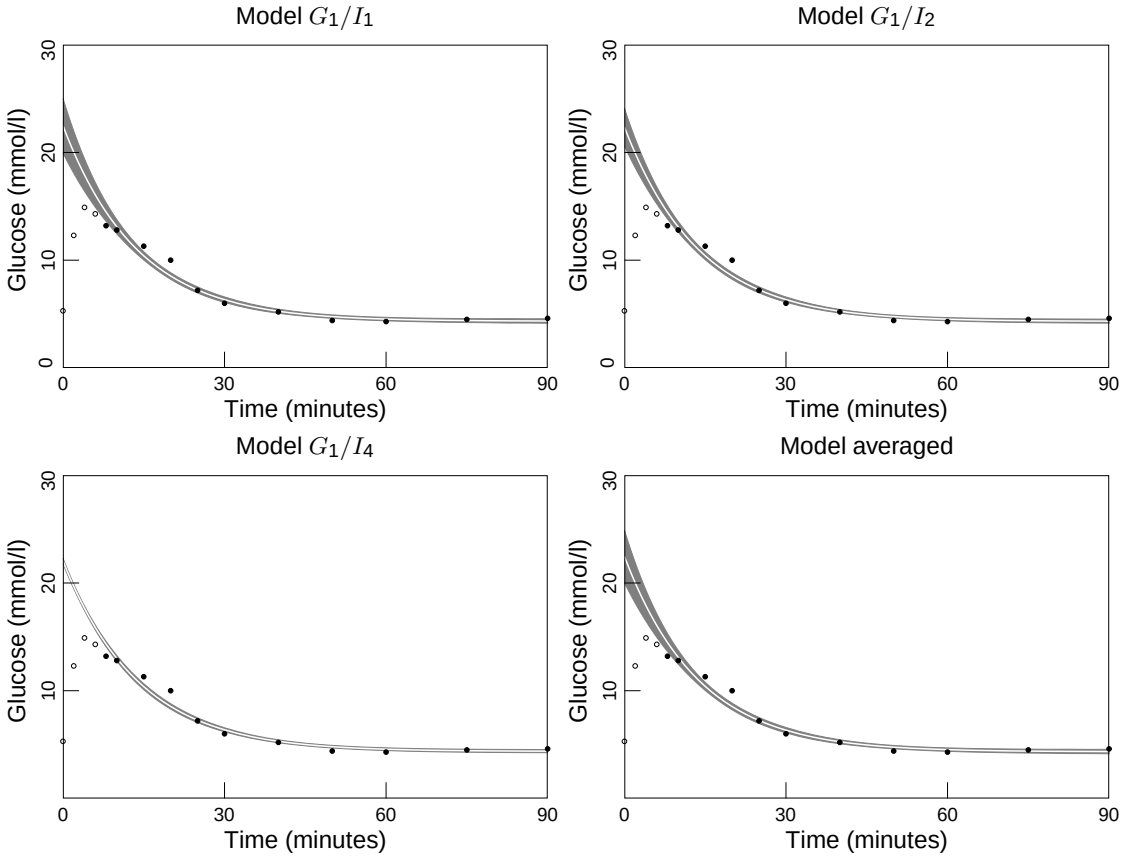



**Fig. 2.** *Posterior mean (white line) and 95% credible intervals superimposed in grey for the glucose process: (a) under model $G_1/I_1$; (b) under model $G_1/I_2$; (c) under model $G_1/I_4$; and (d) model averaged. Solid circles denote observations used for the analysis.*

## 5. Discussion

In this paper we discuss the generic ill-posed inverse problem and the characteristic difficulty of the associated statistical inference. We discuss how the Bayesian approach using MCMC-based tools provides a flexible framework for analysing models of this sort and, in particular, how problems associated with model uncertainty may be overcome. The statistical analysis of ill-posed inverse problems via MCMC pose particularly difficult computational problems stemming from the fact that the associated likelihood tends to have either no convenient closed form and must be estimated or, is simply very expensive to evaluate. This inherent complexity makes standard techniques for improving mixing very difficult to apply, especially in the context of trans-dimensional chains. However, we have shown how tempering-type ideas can be very naturally applied and illustrated how they may

be applied in several general settings. We demonstrate the utility of the methods by analysing data from an intravenous glucose tolerance test which would be impossible to analyse without the use of the simulated tempering scheme to improve, between-model mixing in particular. Our analysis demonstrates that a new alternative to the standard minimal model commonly used to describe data of this sort appears not to describe our own data very well in comparison to the minimal model.

Of course everything we discuss in this paper can be more generically applied beyond the context of ill-posed inverse problems, though many of the advantages of the tempering techniques would not improve upon problem-specific analytic work to derive efficient proposal schemes, especially for the trans-dimensional transitions that may be possible in simpler settings. That being said, when we need to move between different models with the same number of parameters, it is very difficult in almost any setting to construct maps between the parameter spaces under the two models that ensure high acceptance rates. In such cases, as in this paper, the use of simulated tempering schemes that facilitate between-model transitions in higher temperatures are invaluable and, in many cases, provide the only mechanism for ensuring adequate between-model mixing.

## 6. Acknowledgements

The authors gratefully acknowledge the support of Novo Nordisk both for providing the data used in Section 4 and for partially funding the work of KEA and MH. The work of SPB was funded by the UK Engineering and Physical Sciences research Council under grant number AF/000537.

## References

Andersen, K. E., S. P. Brooks and M. B. Hansen (2001), A Bayesian Approach to Crack Detection in Electrically Conducting Media. *Inverse Problems* **17**, 121–136

Andersen, K. E. and M. Højbjerre (2003), A Bayesian Approach to Bergman's Minimal Model. In C. M. Bishop and B. J. Frey (eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pp. 236 – 243

Atkinson, K. (1976), *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*. Philadelphia: SIAM

Baussard, A., D. Premel and O. Venard (2001), A Bayesian Approach for Solving Inverse Scattering from Microwave Laboratory-Controlled Data. *Inverse Problems* **17**, 1659–1669

Bergman, R. N., Y. Z. Ider, C. R. Bowden and C. Cobelli (1979), Quantitative estimation of insulin sensitivity. *American Journal of Physiology* **236**(6), E667 – E677

Brooks, S. P. (1998), Markov Chain Monte Carlo Method and its Application. *The Statistician* **47**, 69–100

Brooks, S. P., P. Giudici and G. O. Roberts (2003), Efficient Construction of Reversible Jump Proposal Distributions (with discussion). *Journal of the Royal Statistical Society, Series B* **65**, 3–55

Bryan, K. and M. Vogelius (1994), A Computational Algorithm to determine Crack Locations from Electrostatic Boundary Measurements. The case of multiple cracks. *International Journal of Engineering Science* **32**, 579–603

Casella, G. and E. I. George (1992), Explaining the Gibbs sampler. *Journal of the American Statistical Association* **46**, 167–174

Clyde, M. A. (1999), Bayesian Model Averaging and Model Search Strategies. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid and J. O. Berger (eds.), *Bayesian Statistics 6*, pp. 157–186, Oxford University Press

de Gaetano, A. and O. Arino (2000), Mathematical modelling of the intravenous glucose tolerance test. *Journal of Mathematical Biology* **40**, 136 – 168

Gelman, A., G. O. Roberts and W. R. Gilks (1996), Efficient Metropolis Jumping Rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 5*, pp. 599–608, New York: Oxford University Press

Geyer, C. J. (1991), Markov Chain Monte Carlo Likelihood. In E. M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface*, pp. 156–163, Interface Foundation

Geyer, C. J. and E. A. Thompson (1995), Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of the American Statistical Association* **90**, 909–920

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996), *Markov Chain Monte Carlo in Practice*. Chapman and Hall

Glad, I. K. and G. Sebastiani (1995), A Bayesian Approach to Synthetic Magnetic Resonance Imaging. *Biometrika* **82**, 237–250

Green, P. J. (1995), Reversible Jump MCMC Computation and Bayesian Model determination. *Biometrika* **82**, 711–732

Green, P. J. (2000), Efficient Construction of Reversible Jump MCMC Proposal Distributions. In *Highly Structured Stochastic Systems Volume*, Oxford University Press, To appear

Higdon, D., H. Lee and Z. Bi (2001), A Bayesian Approach to Characterizing Uncertainty in Inverse Problems Using Coarse and Fine Scale Information. Technical report, Duke University, Institute of Statistics and Decision Sciences

Hukushima, K. and K. Nemoto (1996), Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan* **65**, 1604–1608

Kaipio, J. P., V. Kolehmainen, M. Vauhkonen and E. Somersalo (1999), Inverse problems with structural prior information. *Inverse Problems* **15**, 713 – 729

Kass, R. E. and A. E. Raftery (1995), Bayes Factors. *Journal of the American Statistical Association* **90**, 773–795

King, R. and S. Brooks (2001), Model Selection for Integrated Recovery/Recapture Data. Technical report, University of Cambridge

Madigan, D. M., A. E. Raftery, C. Volinsky and J. Hoeting (1996), Bayesian Model Averaging. In P. Chan, S. Stolofo and D. Wolpert (eds.), *Integrating Multiple Learned Models (IMLM-96)*, pp. 77–83

Marinari, E. and G. Parisi (1992), Simulated Tempering: A New Monte Carlo Scheme. *Europhysics letters* **19**, 451–458

Pacini, G. and R. N. Bergman (1986), MINMOD: a computer program to calculate insulin sensitivity and pancreatic responsivity from the frequently sampled intravenous glucose tolerance test. *Computer Methods and Programs in Biomedicine* **23**, 113 – 122

Pillonetto, G., G. Sparacino, P. Magni, R. Bellazzi and C. Cobelli (2002), Minimal model $S_I = 0$ problem in NIDDM subjects: nonzero Bayesian estimates with credible intervals. *American Journal of Physiology - Endocrinology and Metabolism* **282**, E565 – E573

Richardson, S. and P. J. Green (1997), On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society, Series B* **59**, 731–792

Roberts, G. O. and S. K. Sahu (1997), Updating Schemes, Covariance Structure, Blocking and Parameterisation for the Gibbs Sampler. *Journal of the Royal Statistical Society, Series B* **59**, 291–318

Stephens, M. (2000), Bayesian Analysis of Mixture Models with an Unknown Number of Components: An Alternative to Reversible Jump Methods. *Annals of Statistics* **28**, 40–74

Toffolo, G., D. R.N. Bergman, C. R. Finegood, Bowden and C. Cobelli (1980), Quantitative estimation of beta cell sensitivity to glucose in the intact organism: a minimal model of insulin kinetics in the dog. *Diabetes* **29**, 979–990

## A. The Interpretation of Model Parameters

The interpretation of the parameters within the minimal model are as follows (units are given in brackets.)

$G_b$: basal pre-injection level of glucose [mmol/l].

$I_b$: basal pre-injection level of insulin [pmol/l].

$S_G$: insulin-independent rate constant of glucose uptake in muscles, liver and adipose tissue [$\text{min}^{-1}$].

$p_2$: rate for decrease in tissue glucose uptake ability [$\text{min}^{-1}$].

$S_I$: insulin-dependent increase in glucose uptake ability in tissue per unit of insulin concentration above $I_b$ [$\text{min}^{-1}(\text{pmol/l})^{-1}$].

$n$: first order decay rate for insulin in plasma [$\text{min}^{-1}$].

$h$: threshold value of glucose [mmol/l] above which the pancreatic $\beta$-cells release insulin.

$\gamma$: rate of the pancreatic $\beta$-cells' release of insulin after the glucose injection and with glucose concentration above $h$ [$\text{pmol/l min}^{-2}(\text{mmol/l})^{-1}$].

$G_0$: theoretical glucose concentration in plasma [mmol/l] extrapolated at time 0.

$I_0$: theoretical insulin concentration in plasma [pmol/l] extrapolated at time 0.

Under the delay-differential model of de Gaetano and Arino (2000) the interpretation of the additional parameter is as follows.

$b_5$: length of the past period in which plasma glucose concentrations influence the current pancreatic insulin secretion [min].

## B. Reversible Jump Moves for the Glucose-Insulin Models

This appendix provides a more detailed description on the updating mechanisms required within the MCMC simulations.

### *B.1. Within-model updates*

We consider here two distinct updating mechanisms for the observational error precisions $\nu_{g^o}$ and $\nu_{i^o}$ and the remaining parameters in $\boldsymbol{\theta}_m$. Assume that in model $m$ at temperature $\tau$ we propose a within-model move in which any one of the remaining parameters in $\boldsymbol{\theta}_m$ is updated by sampling a value uniformly in a symmetric interval around the current position with width $2l$, $l > 0$. In proposing a change to any one of these parameters, we propose a move from $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}'_m$ according to $q(\boldsymbol{\theta}_m, \boldsymbol{\theta}'_m) = 1/2l$. Since this proposal is symmetric, the acceptance probability becomes $\alpha(\boldsymbol{\theta}_m, \boldsymbol{\theta}'_m \,|\, \tau) = \min(1, A(\boldsymbol{\theta}_m, \boldsymbol{\theta}'_m \,|\, \tau))$ with

$$
\begin{aligned}
A(\boldsymbol{\theta}_m, \boldsymbol{\theta}'_m \,|\, \tau) &= \frac{\left[f_{m,\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}'_m) f_{m,\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m)\right]^{g(\tau)} p_m(\boldsymbol{\theta}'_m) p(m) q(\boldsymbol{\theta}'_m, \boldsymbol{\theta}_m)}{\left[f_{m,\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_m) f_{m,\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m)\right]^{g(\tau)} p_m(\boldsymbol{\theta}_m) p(m) q(\boldsymbol{\theta}_m, \boldsymbol{\theta}'_m)} \\
&= \frac{\left[\nu^{3|\Lambda|/2} \exp(-V_\tau(\boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m))(\nu_{g^o}\nu_{i^o})^{|\mathcal{T}|/2} \exp(-W_\tau(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m))\right]^{g(\tau)} p_m(\boldsymbol{\theta}'_m)}{\left[\nu^{3|\Lambda|/2} \exp(-V_\tau(\boldsymbol{\Phi}^s, \boldsymbol{\theta}_m))(\nu_{g^o}\nu_{i^o})^{|\mathcal{T}|/2} \exp(-W_\tau(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m))\right]^{g(\tau)} p_m(\boldsymbol{\theta}_m)} \\
&= \exp\Big(\Big[V_\tau(\boldsymbol{\Phi}^s, \boldsymbol{\theta}_m) - V_\tau(\boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m) + W_\tau(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m) - W_\tau(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m)\Big] g(\tau)\Big) \frac{p_m(\boldsymbol{\theta}'_m)}{p_m(\boldsymbol{\theta}_m)},
\end{aligned}
$$

where the posterior potentials $V$ and $W$ now depend upon $\tau$. Once the first of the parameters in $\boldsymbol{\theta}_m$ has had a new value proposed and subsequently either accepted or rejected, we move on to the next. This continues until all parameters have undergone this procedure.

The error precisions $\nu_{g^o}$ and $\nu_{i^o}$ are updated via a simple Gibbs sampler update.

### *B.2. Between-model updates*

Of course, jumps between some models (e.g., updating the insulin model from $I_1$ to $I_t, t = 2, 3, 4$, say) do not involve changing the number of parameters. For moves of this kind, we simply compare the posterior distribution under the current and proposed new model with the current parameter values and adopt the usual acceptance ratio. For moves that involve adding or deleting a parameter (e.g., when we update the glucose model) we require a reversible jump MCMC update which can be implemented as follows.

Suppose that the Markov chain currently is visiting model $m$ at temperature $\tau$ and is in state $\boldsymbol{\theta}_m$. Then a new model $m' \in \mathcal{M} \setminus \{m\}$ is picked uniformly among the remaining models in $\mathcal{M}$, i.e. the acceptance probability for this between-model move becomes $\alpha(m, m' \,|\, \tau) = \min(1, A(m, m' \,|\, \tau))$ with

$$A(m, m' \,|\, \tau) = \frac{\Big[f_{m',\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_{m'}) f_{m',\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_{m'})\Big]^{g(\tau)} p_{m'}(\boldsymbol{\theta}_{m'}) p(m') p_{m'}(r') q_{r',m',m}(\boldsymbol{u}')}{\Big[f_{m,\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_m) f_{m,\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m)\Big]^{g(\tau)} p_m(\boldsymbol{\theta}_m) p(m) p_m(r) q_{r,m,m'}(\boldsymbol{u})} \left| \frac{\partial g(\boldsymbol{\theta}_m, \boldsymbol{u})}{\partial(\boldsymbol{\theta}_m, \boldsymbol{u})} \right|,$$

where $p_{m'}(r') = p_m(r)$ and $p(m') = p(m)$ for any pair $m, m' \in \mathcal{M}$ as described in Section 4. For illustration, we give the details for computing the acceptance probability for moving from the standard minimal model $m$ to the model $m'$ proposed by de Gaetano and Arino (2000). For this particular between-model update, we let $\boldsymbol{u} = b_5$ be generated from a uniform proposal distribution so that $q_{r,m,m'}(b_5) = 1/l_{b_5}$. For the reverse move, however, we need sample $\boldsymbol{u}' = (p_2, \gamma, h)$ uniformly from $q_{r',m',m}(p_2, \gamma, h) = 1/l_{p_2} l_\gamma l_{p_h}$. Thus the acceptance probability for this move becomes

$$\alpha(m, m' \,|\, \tau) = \min\left\{1, \left[\frac{f_{m',\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_{m'}) f_{m',\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_{m'})}{f_{m,\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_m) f_{m,\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m)}\right]^{g(\tau)} \frac{p_{m'}(\boldsymbol{\theta}_{m'})}{p_m(\boldsymbol{\theta}_m)} \frac{l_{b_5}}{l_{p_2} l_\gamma l_{p_h}}\right\},$$

as the Jacobian is simply one. The acceptance probabilities for the remaining between-model moves are derived in a similar fashion.

### *B.3. Simulated tempering updates*

Suppose the Markov chain is in model $m$ at temperature $\tau$, and a new temperature $\tau'$ is proposed according to the directions given in Section 2.3. Then the acceptance probability is given as $\alpha(\tau, \tau' \,|\, \boldsymbol{\theta}_m) = \min(1, A(\tau, \tau' | \boldsymbol{\theta}_m))$ where

$$\begin{aligned}
A(\tau, \tau' \,|\, \boldsymbol{\theta}_m) &= \frac{\Big[f_{m,\tau'}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_m) f_{m,\tau'}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m)\Big]^{g(\tau')} p_m(\boldsymbol{\theta}_m) p(m) c_{\tau'} p_{\tau' \to \tau}}{\Big[f_{m,\tau}(\boldsymbol{\Phi}^s \,|\, \boldsymbol{\theta}_m) f_{m,\tau}(\boldsymbol{\Phi}^o \,|\, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m)\Big]^{g(\tau)} p_m(\boldsymbol{\theta}_m) p(m) c_\tau p_{\tau \to \tau'}} \\
&= \frac{\Big[\nu^{3|\Lambda|/2} \exp(-V_{\tau'}(\boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m)) (\nu_{g^o} \nu_{i^o})^{|\mathcal{T}|/2} \exp(-W_{\tau'}(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m))\Big]^{g(\tau')} c_{\tau'} p_{\tau' \to \tau}}{\Big[\nu^{3|\Lambda|/2} \exp(-V_\tau(\boldsymbol{\Phi}^s, \boldsymbol{\theta}_m)) (\nu_{g^o} \nu_{i^o})^{|\mathcal{T}|/2} \exp(-W_\tau(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}_m))\Big]^{g(\tau)} c_\tau p_{\tau \to \tau'}} \\
&= \Big[\nu^{3|\Lambda|/2} (\nu_{g^o} \nu_{i^o})^{|\mathcal{T}|/2}\Big]^{g(\tau') - g(\tau)} \Big[\exp\big(g(\tau) V_\tau(\boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m) - g(\tau') V_{\tau'}(\boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m) \\
&\qquad + g(\tau) W_\tau(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m) - g(\tau') W_{\tau'}(\boldsymbol{\Phi}^o, \boldsymbol{\Phi}^s, \boldsymbol{\theta}'_m)\big)\Big] \frac{c_{\tau'} p_{\tau' \to \tau}}{c_\tau p_{\tau \to \tau'}}.
\end{aligned}$$