Stat Comput (2015) 25:37–43
DOI 10.1007/s11222-014-9503-z

Cuts in Bayesian graphical models

Martyn Plummer

Received: 18 March 2014 / Accepted: 4 August 2014 / Published online: 27 November 2014
© Springer Science+Business Media New York 2014

Abstract The cut function defined by the OpenBUGS software is described as a "valve" that prevents feedback in Bayesian graphical models. It is shown that the MCMC algorithm applied by OpenBUGS in the presence of a cut function does not converge to a well-defined limiting distribution. However, it may be improved by using tempered transitions. The cut algorithm is compared with multiple imputation as a gold standard in a simple example.

Keywords Bayesian inference · Cutting feedback · Multiple imputation

1 Introduction

Cut models are an alternative to Bayesian full probability models that are used to modulate the flow of information from data to parameters. Cut models arise in applications with multiple data sources that provide information about different parameters in the model. A simplified graphical representation of a typical cut model is shown in Fig. 1. Interest lies in parameters $\boldsymbol{\theta}$, which are informed by data $\mathbf{Y}$. The likelihood for $\boldsymbol{\theta}$ includes nuisance parameters $\boldsymbol{\varphi}$ that are estimated using auxiliary data $\mathbf{Z}$. In a full probability model, information from $\mathbf{Y}$ "feeds back" through the graph to influence the posterior distribution of $\boldsymbol{\varphi}$. There are circumstances in which this feedback may be considered unhelpful:

1. If the dose-response relationship that describes the relationship between $\boldsymbol{\theta}$ and $\mathbf{Y}$ is speculative, we may prefer to use only information from $\mathbf{Z}$ to estimate $\boldsymbol{\varphi}$.
2. If there is conflict between the different data sources such that $p(\boldsymbol{\varphi} \mid \mathbf{Y}, \mathbf{Z})$ is very different from $p(\boldsymbol{\varphi} \mid \mathbf{Z})$, we may consider $\mathbf{Z}$ to be a more reliable source of information about $\boldsymbol{\varphi}$ and so down-weight the influence of $\mathbf{Y}$. For example, in measurement error models, $\mathbf{Z}$ represents surrogate exposure data, and the study design typically includes a validation or reliability sub-study, which is directly informative about the error properties of the surrogate. The more complex graphical models associated with such designs are explored by Richardson and Gilks (1993).
3. If there are computational problems with Markov Chain Monte Carlo (MCMC), convergence and mixing may be improved in a model in which $\boldsymbol{\varphi}$ is estimated only from $\mathbf{Z}$.

Cuts have most extensively been investigated in population pharmacokinetic / pharmacodynamic (PK/PD) models. In such models, separate data sources provide information on the PK and the PD parameters. When both data sources are analyzed using a single probability model, feedback occurs from the PD data into the PK model. Sequential analysis, in which the PK data are analysed first, and then estimates are plugged into the model for the PD data, may be both faster and more robust (Bennett and Wakefield 2001; Zhang et al. 2003a, b). Sequential analysis is also the basis of regression calibration (Carroll et al. 2007), a frequentist method used to correct for measurement error in generalized linear models.

A disadvantage of Bayesian sequential analysis is that uncertainty in the imputed values in the first phase is not carried forward to the second phase. A potential solution to this problem is provided by the WinBUGS and OpenBUGS soft-

Electronic supplementary material The online version of this article (doi:10.1007/s11222-014-9503-z) contains supplementary material, which is available to authorized users.

M. Plummer (✉)
International Agency for Research on Cancer, Lyon, France
e-mail: plummerm@iarc.fr

ware packages, which implement a modified MCMC algorithm through the use of a "cut" function. The cut function modifies a key step in MCMC on a graph – the construction of the full conditional distribution. This is the distribution of a node given the values of all other nodes in the graph. In practice, the full conditional distribution depends only on local nodes (the Markov blanket) so that the full conditional density can typically be expressed as the product of few factors, even in a large graphical model. The cut function simplifies the full conditional density further by ignoring some terms. In Fig. 1, the graph is divided by a cut into two subgraphs G_1 and G_2. When constructing the full conditional distributions for parameters in G_1, likelihood terms involving random variables in G_2 are ignored. Hence $\boldsymbol{\varphi}$ is sampled using only $\mathbf{Z}$ and not $\mathbf{Y}$, despite the dependence of $\mathbf{Y}$ on $\boldsymbol{\varphi}$. Conversely when parameters in G_2 are sampled, prior terms involving random variables in G_1 are included as normal in the construction of the full conditional density.

The OpenBUGS manual describes the cut function informally as follows: "The cut function acts as a kind of valve in the graph: prior information is allowed to flow downwards through the cut, but likelihood information is prevented from flowing upwards" (Spiegelhalter et al. 2004). The cut function has been used in a variety of applications in different fields (Haining et al. 2007; Carrigan et al. 2007; Scollnick 2004; Mwalili et al. 2005; Jackson et al. 2008; Choi et al. 2009). Its use in PK/PD models is reviewed by Lunn et al. (2009). The same modified MCMC sampling technique has also been applied in custom software (Rougier 2008; He and Zaslavsky 2009). In the sequel, this algorithm will be referred to as the "naive cut algorithm".

Liu et al. (2009) describe the cut function as an example of "modularization"—the division of a large complex model into smaller modules that interact more weakly than in a full Bayesian analysis. They note that attempts to modify the flow of information in Bayesian models can also be found in inconsistent dependency networks (Heckerman et al. 2000) and inconsistent Gibbs (Raghunathan et al. 2001).

The naive cut algorithm has been used either as a useful computational technique to increase speed and improve convergence of MCMC, or informally introduced as an approach to model robustness. However, as noted by Liu et al. (2009), modularization occurs in the context of settings that are too complex for formal analysis. Thus it can be difficult to determine the validity of cut models. The purpose of the current paper is to provide a simple illustration of cutting feedback with a toy example drawn from epidemiology, to demonstrate the lack of convergence of the naive cut algorithm and to provide an alternative approximate solution.

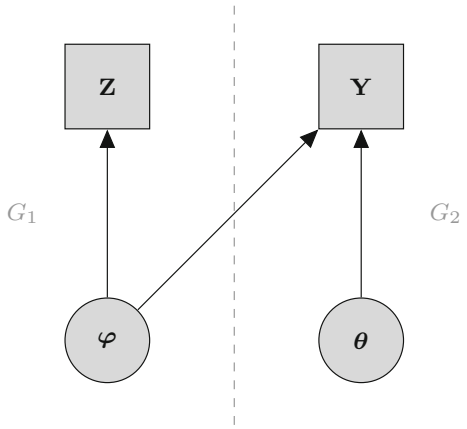

Fig. 1 Graphical representation of a cut model

2 Target distribution for cut models

The naive cut algorithm is defined operationally, in terms of a modified MCMC update scheme, rather than in terms of simulating from a target distribution. In order to evaluate the algorithm it is helpful to define this target distribution.

By assumption, the variables represented by nodes in Fig. 1 have a joint distribution that factorizes on the graph as:

$$p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}) = p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\varphi}) p(\mathbf{Z} \mid \boldsymbol{\varphi}) p(\boldsymbol{\theta}) p(\boldsymbol{\varphi}) \tag{1}$$

We assume that the factors on the right-hand side of (1) can be expressed in closed form, or otherwise can be easily computed.

In the simplest form of the naive cut algorithm, it is possible to sample directly from the conditional distributions $p(\boldsymbol{\varphi} \mid \mathbf{Z})$ and $p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\varphi})$, where $p(\boldsymbol{\varphi} \mid \mathbf{Z}) \propto p(\boldsymbol{\varphi}) p(\mathbf{Z} \mid \boldsymbol{\varphi})$ is the posterior distribution of $\boldsymbol{\varphi}$ when $\mathbf{Z}$ is observed but $\mathbf{Y}$ is not. If direct sampling is possible, then the resulting sequence of samples $(\boldsymbol{\varphi}^t, \boldsymbol{\theta}^t)$ for $t = 1, 2, \ldots$ are independent draws from the joint distribution

$$p^*(\boldsymbol{\theta}, \boldsymbol{\varphi}) = p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi} \mid \mathbf{Z}) \tag{2}$$

This differs from the full Bayesian posterior

$$p(\boldsymbol{\theta}, \boldsymbol{\varphi} \mid \mathbf{Y}, \mathbf{Z}) = p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi} \mid \mathbf{Y}, \mathbf{Z}) \tag{3}$$

In the Bayesian posterior (3), the second factor depends on $\mathbf{Y}$. In the cut distribution (2) it does not. Hence p^*, while being a valid probability distribution, is not a true posterior.

In general, it is not possible to sample directly from the conditional distributions $p(\boldsymbol{\varphi} \mid \mathbf{Z})$ and $p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\varphi})$. If (2) can be expressed in closed form (up to a multiplicative constant) then it may be used as a target for a standard MCMC algorithm. However, this is rarely possible. The target may

be rewritten as

$$p^*(\boldsymbol{\theta}, \boldsymbol{\varphi}) \propto \frac{p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi})}{p(\mathbf{Y} \mid \boldsymbol{\varphi})}$$

The numerator can be expressed in closed form via the factorization (1), but the marginal likelihood in the denominator is an integral

$$p(\mathbf{Y} \mid \boldsymbol{\varphi}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\varphi}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

which can only be expressed in closed form in very simple models.

Since $\boldsymbol{\varphi}$ is a nuisance parameter, interest may lie only in the marginal distribution of $\boldsymbol{\theta}$:

$$p^*(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi} \mid \mathbf{Z}) d\boldsymbol{\varphi} \tag{4}$$

The target distribution $p^*(\theta)$ can always be estimated by multiple imputation (MI), *i.e.* generating a sequence of samples $\boldsymbol{\varphi}^1, \ldots \boldsymbol{\varphi}^T$ and then fitting T separate models for $\boldsymbol{\theta}$ such that, under model t, $\boldsymbol{\varphi}$ is considered to be observed at $\boldsymbol{\varphi} = \boldsymbol{\varphi}^t$. Pooled MCMC samples from all T models can be used to estimate the marginal density $p^*(\boldsymbol{\theta})$. Moreover, if $p^*(\theta)$ is approximately normal and the $\boldsymbol{\varphi}^t$ are independent, the multiple imputation combining rules of Little (1992) may be used to provide an approximation to $p^*(\theta)$ and in this case relatively few imputations may be required.

3 Example: ecological study of HPV and cervical cancer

To illustrate cuts, we consider a toy example that illustrates many of the important features of cut models while being simple and reproducible. The example is derived from a real epidemiological study.

The motivation for the example is an investigation of the international correlation between human papillomavirus (HPV) prevalence and cervical cancer incidence (Maucort-Boulch et al. 2008). Cervical cancer is known to be caused by around 20 types of HPV, which are known collectively as "high-risk" types. In this study, high-risk HPV prevalence data came from a series of surveys in 13 different countries. Incidence data come from cancer registries in the same populations. We consider the oldest age group (55–64 years) in the survey, which demonstrated the strongest correlation between HPV and cervical cancer.

In population i, the outcome Y_i is the number of cancer cases arising from T_i woman-years of follow-up; Z_i is the number of women infected with high-risk HPV in a sample of size N_i from the same population. We assume a Poisson and a Binomial distribution for these data respectively:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i) \\ \mu_i &= \lambda_i T_i \\ Z_i &\sim \text{Binomial}(N_i, \varphi_i) \end{aligned}$$

We postulate a log linear relationship between high-risk HPV prevalence φ_i and incidence λ_i in the same population:

$$\log(\lambda_i) = \theta_1 + \theta_2 \varphi_i \tag{5}$$

and focus on the slope parameter θ_2.

Since the dose-response relationship (5) is speculative, it falls into the first case considered in Sect. 1 motivating the use of a cut model to prevent feedback from the incidence data when estimating the prevalence parameters $\boldsymbol{\varphi}$.

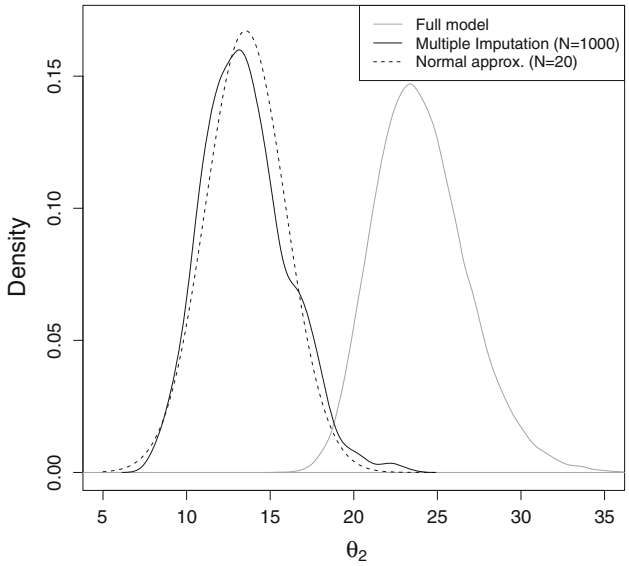

Fig. 2 Comparison of the posterior for θ_2 under the full probability model and under the cut model estimated by multiple imputation

3.1 Multiple imputation

Figure 2 shows a comparison of the posterior distribution of θ_2 for the full probability model and for the cut model using multiple imputation. Two results for multiple imputation are shown: an approximation using the combining rules of Little (1992) for 20 imputations, and a full density estimate based on 1000 imputations.

There is little overlap between the supported regions of the two models, showing the strong influence of feedback in the probability model. The approximate posterior based on the multiple imputation combining rules captures the location and scale of the cut model. This approximation is less accurate in the tails, as may be expected.

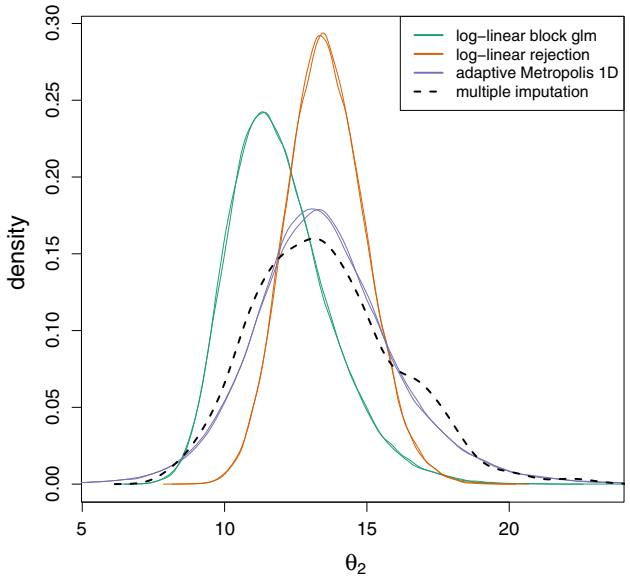

Fig. 3 Posterior density of θ_2 using the naive cut algorithm with three different update methods, and a comparison with multiple imputation

A large number of imputations are required for the density estimate due to the fact that the variance of $\boldsymbol{\theta}$ between imputations is much higher than the variance within imputations. This can be summarized as a rate of missing information (Little 1992) of 0.98 in this example.

Although multiple imputation with a large number of replicates is feasible in this small example, it would be useful to have a way to sample from the target distribution in a single MCMC chain.

3.2 Naive cut algorithm

Figure 3 shows posterior estimates of θ_2 using the naive cut algorithm implemented in OpenBUGS 3.2.2. OpenBUGS is designed to be a "black box" that automatically chooses appropriate sampling methods for the user. Nevertheless the user has partial control over the sampling methods-referred to as "updaters" in OpenBUGS-which allows us to fit the same model with different sampling methods (See appendix for details). Figure 3 shows results for three updaters, labelled "log-linear block glm", "log-linear rejection" and "adaptive Metropolis 1D". Updaters are not documented by OpenBUGS, so a full understanding of the sample methods in use in Fig. 3 can only be obtained by inspecting the source code. Each updater defines a different transition kernel for the Markov chain. In standard MCMC, all of these transition kernels converge to the same posterior distribution, although they may differ in their efficiency.

Each updater was run in two parallel chains with initial values drawn from the prior. The density estimates for the two parallel chains coincide, demonstrating convergence of all three updaters. However, the three transition kernels converge to different limiting distributions. Of the three updaters, the "adaptive Metropolis 1D" comes closest to the result from multiple imputation (Fig. 3). This is not the default updater chosen by OpenBUGS, and in the absence of the multiple imputation estimate, we would not know which of the sampling methods produced the best approximation to the target density.

4 Convergence of the naive cut algorithm

In general, MCMC methods do not sample directly from the target density, but supply a sequence of reversible transitions in detailed balance with the full conditional distribution. For notational simplicity, we consider only block sampling in which $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are updated in separate blocks. Hence, under the naive cut algorithm, $\boldsymbol{\varphi}$ is updated in detailed balance with $p(\boldsymbol{\varphi} \mid \mathbf{Z})$ and $\boldsymbol{\theta}$ is updated in detailed balance with $p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\varphi})$.

The transition kernel of $\boldsymbol{\varphi}$ is defined by the probability of moving from the old value $\boldsymbol{\varphi}^{t-1}$ to the new value $\boldsymbol{\varphi}^{t}$ at iteration t. We represent the probability density of this move with the notation $p(\boldsymbol{\varphi}^{t-1} \rightarrow \boldsymbol{\varphi}^{t})$. Unlike the posterior density $p(\boldsymbol{\varphi}|\mathbf{Z})$, which is determined by the model, we have free choice in defining the transition kernel, but it must satisfy the detailed balance relation:

$$p(\boldsymbol{\varphi}^{t-1} \mid \mathbf{Z})p(\boldsymbol{\varphi}^{t-1} \rightarrow \boldsymbol{\varphi}^{t}) = p(\boldsymbol{\varphi}^{t} \mid \mathbf{Z})p(\boldsymbol{\varphi}^{t} \rightarrow \boldsymbol{\varphi}^{t-1}) \quad (6)$$

This ensures that if $\boldsymbol{\varphi}^{t-1}$ is a random sample from the posterior $p(\boldsymbol{\varphi}|\mathrm{Z})$ then so is $\boldsymbol{\varphi}^{t}$ (and vice versa if one is observing the process in reverse time).

Similarly we represent the probability density of the move from $\boldsymbol{\theta}^{t-1}$ to $\boldsymbol{\theta}^{t}$ at iteration t as $p(\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^{t}|\boldsymbol{\varphi}^{t})$. This transition kernel is more complex, as it may depend on the current value of $\boldsymbol{\varphi}$. The transition kernel must satisfy the detailed balance relation:

$$\begin{aligned} &p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t})p(\boldsymbol{\theta}^{t-1} \rightarrow \boldsymbol{\theta}^{t} \mid \boldsymbol{\varphi}^{t}) \\ &\quad = p(\boldsymbol{\theta}^{t} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t})p(\boldsymbol{\theta}^{t} \rightarrow \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t}) \end{aligned} \quad (7)$$

So that if $\boldsymbol{\theta}^{t-1}$ is a random sample from the conditional posterior $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\varphi}^{t})$ then so is $\boldsymbol{\theta}^{t}$ (and vice versa).

Suppose that $(\boldsymbol{\varphi}^{t-1}, \boldsymbol{\theta}^{t-1})$ is a random sample from $p^{*}(\boldsymbol{\varphi}, \boldsymbol{\theta})$ and that first $\boldsymbol{\varphi}$ is updated according to (6) then $\boldsymbol{\theta}$ is updated according to (7). Then $(\boldsymbol{\varphi}^{t}, \boldsymbol{\theta}^{t})$ is a random sample from the weighted density $p^{*}(\boldsymbol{\varphi}, \boldsymbol{\theta})w(\boldsymbol{\varphi}, \boldsymbol{\theta})$ where the weight function is given by

$$w(\boldsymbol{\varphi}, \boldsymbol{\theta}) = \int \frac{p(\boldsymbol{\theta}' \mid \mathbf{Y}, \boldsymbol{\varphi}')}{p(\boldsymbol{\theta}' \mid \mathbf{Y}, \boldsymbol{\varphi})} p(\boldsymbol{\varphi} \rightarrow \boldsymbol{\varphi}')p(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}' \mid \boldsymbol{\varphi})d\boldsymbol{\varphi}'d\boldsymbol{\theta}' \quad (8)$$

In a standard MCMC update, the first factor in the integrand cancels out with likelihood terms from the full conditional distribution of $\boldsymbol{\varphi}$. However, with the naive cut algorithm, these terms are ignored when $\boldsymbol{\varphi}$ is updated, and cancellation does not take place. Hence p^* is not the stationary distribution of the Markov chain under the naive cut algorithm. Note that the weight w is a function of the transition kernels $p(\boldsymbol{\varphi} \to \boldsymbol{\varphi}')$ and $p(\boldsymbol{\theta} \to \boldsymbol{\theta}' \mid \boldsymbol{\varphi})$. This explains why different transition kernels converge to different limiting distributions.

For the naive cut algorithm to draw approximate samples from (2), w should be as close as possible to 1, at least over the range of $(\boldsymbol{\varphi}, \boldsymbol{\theta})$ with posterior support. This may occur in two limiting situations

- The probability of the transition $\boldsymbol{\theta} \to \boldsymbol{\theta}'$ does not depend on $\boldsymbol{\theta}$. In this limit

$$\frac{p(\boldsymbol{\theta} \to \boldsymbol{\theta}' \mid \boldsymbol{\varphi})}{p(\boldsymbol{\theta}' \mid \mathbf{Y}, \boldsymbol{\varphi})} \to 1$$

- The transition $\boldsymbol{\varphi} \to \boldsymbol{\varphi}'$ only permits very small steps leading to slow mixing of the Markov chain for $\boldsymbol{\varphi}$. In this limit

$$\frac{p(\boldsymbol{\theta}' \mid \mathbf{Y}, \boldsymbol{\varphi}')}{p(\boldsymbol{\theta}' \mid \mathbf{Y}, \boldsymbol{\varphi})} \to 1$$

In the example of Sect. 3 the $\boldsymbol{\varphi}$ parameters have a conjugate beta distribution and are sampled directly from $p(\boldsymbol{\varphi} \mid \mathbf{Z})$. This leads to large jumps in $\boldsymbol{\varphi}$ between iterations and is as far as possible from the ideal situation for the good behaviour of the naive cut algorithm

Working backwards from (8) it is possible determine what conditions on the transitions are necessary in order to have p^* as the stationary distribution. The conditions on the transitions for $\boldsymbol{\varphi}$ are unchanged. However, for $\boldsymbol{\theta}$, a different balance equation is required:

$$\begin{aligned} &p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1}) p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^{t-1}, \boldsymbol{\varphi}^t) \\ &\quad = p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^t) p(\boldsymbol{\theta}^t \to \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^t, \boldsymbol{\varphi}^{t-1}) \end{aligned} \tag{9}$$

This balance relation uses both the current and previous values of $\boldsymbol{\varphi}$.

One possibility to correct the transition probabilities is to consider a transition generated by (6) and (7) and add a Metropolis acceptance step. The acceptance probability would be $\min(1, R)$ where

$$R = \frac{p(\boldsymbol{\theta}^t \mid \mathbf{Y}, \boldsymbol{\varphi}^t)}{p(\boldsymbol{\theta}^{t-1} \mid \mathbf{Y}, \boldsymbol{\varphi}^{t-1})} \frac{p(\boldsymbol{\theta}^t \to \boldsymbol{\theta}^{t-1} \mid \boldsymbol{\varphi}^{t-1})}{p(\boldsymbol{\theta}^{t-1} \to \boldsymbol{\theta}^t \mid \boldsymbol{\varphi}^t)} \tag{10}$$

In general, R will be hard to evaluate. The second factor requires the transition probabilities to be available in closed form. However, some sampling methods such as slice sampling and Hamiltonian Monte Carlo are designed to generate reversible transitions without providing an explicit formula for the transition probabilities. This problem could be overcome by restricting sampling methods to those that generate explicit transition probabilities, such as Metropolis-Hastings. A second problem that is not so easily overcome, is that the first factor (which also appears inside the integrand of the weight function in (8)) is the ratio of two normalized probability distributions. To make this more explicit, it can be further factorized as

$$\frac{p(\boldsymbol{\theta}^t, \mathbf{Y} \mid \boldsymbol{\varphi}^t)}{p(\boldsymbol{\theta}^{t-1}, \mathbf{Y} \mid \boldsymbol{\varphi}^{t-1})} \frac{p(\mathbf{Y} \mid \boldsymbol{\varphi}^{t-1})}{p(\mathbf{Y} \mid \boldsymbol{\varphi}^t)}$$

Hence calculation of R requires a formula for the marginal likelihood $p(\mathbf{Y} \mid \boldsymbol{\varphi})$. Thus attempting to rescue the naive cut algorithm with a Metropolis step brings us back to the problem of the intractable marginal likelihood.

5 Tempered cut algorithm

The two limiting cases that allow the weight function (8) to approach the value 1 suggest some approximate methods that may improve on the naive cut function. One possibility is to run multiple updates of $\boldsymbol{\theta}$ for each update of $\boldsymbol{\varphi}$. The new value $\boldsymbol{\theta}^t$ would be retained after discarding a number of "burn-in" iterations of $\boldsymbol{\theta}$. This is effectively an MCMC implementation of multiple imputation. If the burn-in is sufficiently long the new sample $\boldsymbol{\theta}^t$ will depend weakly on the previous value $\boldsymbol{\theta}^{t-1}$ and the weight function w should be close to 1. Note however that if the jump $\boldsymbol{\varphi}^{t-1} \to \boldsymbol{\varphi}^t$ is large, the burn-in period may have to be correspondingly long. This suggests a second modification. Instead of moving directly from $\boldsymbol{\varphi}^{t-1}$ to $\boldsymbol{\varphi}^t$, we move along a linear path

$$\boldsymbol{\varphi}(c) = c\boldsymbol{\varphi}^t + (1-c)\boldsymbol{\varphi}^{t-1}$$

in a sequence $c^1, \ldots c^m$ with $c^i = i/m$. At each step we draw a new sample of $\boldsymbol{\theta}$ but keep only the sample at the last step m, which becomes $\boldsymbol{\theta}^t$. The linear path provides a tempered transition between the probability densities at $\boldsymbol{\varphi}^{t-1}$ and $\boldsymbol{\varphi}^t$. If $\boldsymbol{\varphi}$ is discrete-valued then a parametric path is not possible, but an alternate path through distribution space can be constructed, e.g.

$$\left\{p(\mathbf{Y}, \boldsymbol{\theta} \mid \boldsymbol{\varphi}^{t-1})\right\}^{1-c} \left\{p(\mathbf{Y}, \boldsymbol{\theta} \mid \boldsymbol{\varphi}^t)\right\}^c$$

As the number of steps m increases, the sampled value $\boldsymbol{\theta}^t$ becomes less dependent on the value $\boldsymbol{\theta}^{t-1}$ at the previous iteration, and the changes in $\boldsymbol{\varphi}(c)$ between steps get smaller. Both of these approach the asymptotic conditions for good behaviour of the naive cut algorithm. For sufficiently large m, the tempered cut algorithm should therefore provide approx-

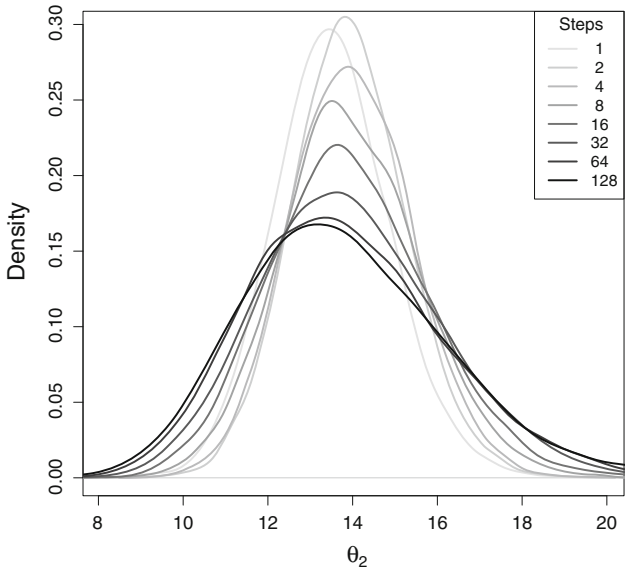

Fig. 4 Posterior density of θ_2 using the tempered cut algorithm with an increasing number of steps

imate samples from the target distribution p^*. The number of steps m required must be determined empirically.

In the cervical cancer example, each prevalence parameter φ_i has a conjugate beta distribution and can be sampled directly from its posterior distribution. For sampling $\boldsymbol{\theta}$ we include an inner loop at each iteration in which $\boldsymbol{\varphi}$ is moved from its value at the previous iteration (denoted `phi.old` below) to the newly sampled value (`phi.new`) in m steps. This may be represent in pseudocode:

```
For j = 1 to m
   For k = 1 to LENGTH(m)
      Phi [k] = (j * phi.new[k] + (m-j) * phi.old[k]) / m
   ENDFOR
   FOR k = 1 to LENGTH(theta)
      Sample new theta [k]
   ENDFOR
ENDFOR
```

The R implementation of this algorithm is provided in the appendix. Elements of $\boldsymbol{\theta}$ were sampled using a random walk Metropolis algorithm with separate step sizes tuned to give acceptance probabilities in the range $0.26 - 0.45$.

Figure 4 shows the results of applying the tempered cut algorithm to the example with the number of steps increasing in powers of two between $m = 1$ (corresponding to the naive cut algorithm) and $m = 128$. Between $m = 64$ and $m = 128$ the posterior distribution shows evidence of convergence.

6 Discussion

Cut models are widely used, due largely to the implementation of the naive cut algorithm in OpenBUGS but also to the intuitive attractiveness of the "modularization" that cut models appear to offer. Unfortunately the naive cut algorithm does not converge to a well-defined distribution. The present article should serve as a warning and avoid further applications of the naive cut algorithm without due caution (e.g. sensitivity analyses using different sampling methods).

I have proposed a modified algorithm based on tempered transitions. However, this offers only an approximate solution. Moreover, it requires an additional convergence check (of the number of tempering steps required) in addition to the usual evaluation of MCMC convergence. It should also be noted that the tempered algorithm has only a heuristic justification. It should not be confused with the well-founded tempered transition algorithm for multi-model posterior distributions developed by Neal (1996). Exact MCMC sampling from a cut model seems to require evaluating the marginal likelihood $p(Y \mid \boldsymbol{\varphi})$ which is computationally expensive (Gelman and Meng 1998). It might be possible, as suggested by a reviewer, that auxiliary variable methods such as the one proposed by Møller et al. (2006) may allow this evaluation to be skipped. Otherwise, it appears unlikely that any MCMC method can sample exactly from the target density $p^*(\boldsymbol{\theta}, \boldsymbol{\varphi})$.

Multiple imputation can always be used for approximate inference in cut models, and is an attractive approach in an increasingly parallel computing environment. The combining rules of Little (1992) can be applied with few imputations, but depend crucially on an assumption of normality. This may not be valid in the complex, nonlinear models to which the cut algorithm is typically applied, such as PK/PD models.

This article has concentrated on the computational aspects of cut models. Clearly more work needs to be done to see if more efficient MCMC approximations are available. It is not clear if this goal can be realised, and even if it can then a more in-depth statistical critique of this methodology is required.

One fundamental issue that must be addressed is whether cut models are admissible from a Bayesian viewpoint. Since $p^*(\boldsymbol{\theta}, \boldsymbol{\varphi})$ is not a standard Bayesian posterior, it cannot represent the coherent belief of any individual. However, it may represent the consensus belief of two individuals who are observing different data sets and communicating with each other via summary statistics. Since this reflects the reality of scientific communication, it may be useful to develop a theoretical and computational framework for understanding this process.

Acknowledgments I started thinking about the cut problem after a presentation by Nicky Best at the IceBUGS meeting in 2006 (http://mathstat.helsinki.fi/openbugs/IceBUGS/Presentations/BestIceBUGS). Over the years, I have had many useful discussions with Nicky Best, Dave Lunn, David Spiegelhalter and Jon Wakefield. I would also like to thank Sylvia Richardson and Nicky Best for inviting me to give a talk on this topic at MCMSki IV.

References

Bennett, J., Wakefield, J.: Errors-in-variables in joint population pharmacokinetic/pharmacodynamic. Biometrics **57**, 803–812 (2001)

Carrigan, G., Barnett, A., Dobson, A., Mishra, G.: Compensating for missing data from longitudinal studies using WinBUGS. J. Stat. Softw. **19**(7), 1–17 (2007)

Carroll, R., Ruppert, D., Stefanski, L.: Measurement error in nonlinear models. Chapman and Hall, New York (2007)

Choi, J., Fuentes, M., Reich, B.J.: Spatial-temporal association between fine particulate matter and daily mortality. J. Comput. Gr. Stat. **53**, 2989–3000 (2009)

Gelman, A., Meng, X.: Simulating normalizing constants: from importance sampling to bridge samplingn to path sampling. Stat. Sci. **13**(2), 163–185 (1998)

Haining, R., Law, J., Maheswaran, R., Pearson, T., Brindley, P.: Bayesian modelling of environmental risk: example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen. Stoch. Environ. Res. Risk Assess. **21**, 501–509 (2007)

He, Y., Zaslavsky, A.: Combining information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies. Biometrics **65**, 946–52 (2009)

Heckerman, D., Chickering, D.M., Meek, C., Rounthwaite, R., Kadie, C.M.: Dependency networks for inference, collaborative filtering, and data visualization. J. Mach Learn. Res. **1**, 49–75 (2000)

Jackson, C., Best, N., Richardson, S.: Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. J. R. Stat. Soc. Ser A **171**(1), 159–178 (2008)

Little, R.: Regression with missing x's: a review. J. Am. Stat. Assoc. **87**, 1227–1237 (1992)

Liu, F., Bayarri, M.J., Berger, J.O.: Modularization in Bayesian analysis, with emphasis on analysis of computer models. Bayesian Anal **4**(1), 119–150 (2009)

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., Neuenschwander, B.: Combining MCMC with 'sequential' PKPD modelling. J. Pharmacokinet Pharmacodyn. (January 2009). doi:10.1007/s10928-008-9109-1

Maucort-Boulch, D., Franceschi, S., Plummer, M.: International correlation between human papillomavirus prevalence and cervical cancer incidence. Cancer Epidemiol. Biomark. Prev. **17**, 717–720 (2008)

Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika **93**(2), 451–458 (2006)

Mwalili, S., Lesaffre, E., Declerck, D.: A Bayesian ordinal logistic regression model to correct for inter-observer measurement error in a geographical oral health study. J. R. Stat. Soc. Ser. C **54**(1), 77–93 (2005)

Neal, R.: Sampling from multimodal distributions using tempered transitions. Stat. Comput. **4**, 353–366 (1996)

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodol. **27**, 85–89 (2001)

Richardson, S., Gilks, W.: Conditional independence models for epidemiological studies with covariate measurement error. Stat. Med. **12**, 1703–1722 (1993)

Rougier, J.: Comment on paper by Sansó, et al. Bayesian Anal. **3**(1), 45–56 (2008)

Scollnick, D.: Bayesian reserving models inspired by chain ladder methods and implemented using WinBUGS. Actuar. Res. Clear. House **2014**(2), (2004). http://www.soa.org/news-and-publications/publications/proceedings/arch/pub-arch-detail.aspx

Spiegelhalter, D.J., Thomas, A., Best, N., Lunn, D.: WinBUGS user manual, version 2.0 (2004)

Zhang, L., Beal, S., Sheiner, L.: Simultaneous vs. sequential analysis for population PK/PD data i: best-case performance. J. Pharmacokinet Pharmacodyn. **30**, 387–404 (2003a)

Zhang, L., Beal, S., Sheiner, L.: Simultaneous vs. sequential analysis for population PK/PD data ii. J. Pharmacokinet Pharmacodyn. **30**, 405–416 (2003b)