Bayesian Analysis (2007) **2**, Number 2, pp. 385–408

Importance Re-sampling MCMC for Cross-Validation in Inverse Problems

S. Bhattacharya* and J. Haslett†

Abstract.

This paper presents a methodology for cross-validation in the context of Bayesian modelling of situations we loosely refer to as 'inverse problems'. It is motivated by an example from palaeoclimatology in which scientists reconstruct past climates from fossils in lake sediment. The inverse problem is to build a model with which to make statements about climate, given sediment. One natural aspect of this is to examine model fit via 'inverse' cross-validation. We discuss the advantages of inverse cross-validation in Bayesian model assessment. In high-dimensional MCMC studies the inverse cross-validation exercise can be computationally burdensome. We propose a fast method involving very many low-dimensional MCMC runs, using Importance Re-sampling to reduce the dimensionality. We demonstrate that, in addition, the method is particularly suitable for exploring multimodal distributions. We illustrate our proposed methodology with simulation studies and the complex, high-dimensional, motivating palaeoclimate problem.

Keywords: Cross-validation, Inverse, Importance Re-sampling, Model fit, Re-use

1 Introduction

Leave-one-out cross-validation is an example of situation, common in much statistical modelling, where it can be necessary to run several similar versions of the same statistical model. When the number of cases n is large and the model is sufficiently complex to require Markov chain Monte Carlo (MCMC), this can be burdensome in the extreme if implemented in 'brute-force' fashion, that is via n very similar 'regular MCMC' runs; by regular MCMC we mean sequential exploration of the parameter space, typically by random walk proposal mechanism. We refer to n repetitions of this as 'n-fold regular MCMC'. Here we propose a generic procedure for leave-one-out cross-validations in situations that we loosely call 'inverse regression'. The procedure uses Importance Resampling (IR) (see Rubin (1988)) (alternatively Sampling/Importance Resampling (SIR)), to reduce, sometimes radically, the high dimensionality. It involves many low-dimensional MCMC runs; but these are fast. The paper explains the trade-off issues involved. We refer to this procedure as IRMCMC; for simplicity we suppress the phrase 'leave-one-out'. We contrast IRMCMC with proposals (Gelfand et al. (1992), Gelfand (1996)) involving importance sampling together with MCMC in 'forward' problems, showing that these are not typically available for inverse problems.

The motivating examples arise in quantitative palaeo-climate reconstruction; see

---

*University of Sheffield, UK, mailto:S.Bhattacharya@sheffield.ac.uk

†Trinity College Dublin, Republic of Ireland, mailto:John.Haslett@tcd.ie

ba0006

Section 6 for details. We present there an example where, in our implementation, n-fold regular MCMC takes 16 hours, but can be accomplished in less than 40 minutes using IRMCMC. In another case, the time is reduced to less than 8 hours from a potential 5 years.

The essential idea of inverse cross-validation can be presented with the following toy problem. Data $(X, Y) = \{(x_i, y_i)\}$ $i = 1, \ldots, n$ are available. We take the y_i as counts, where x_i correspond to a continuous variable. We adopt the Poisson model $P(\theta x_i)$ for the counts y_i, independently of the other cases; θ is an unknown parameter. The objective is to 'predict' x_{n+1} (say) from a future y_{n+1} (say). We describe this as 'inverse regression', to contrast it with the much simpler 'forward' application, the prediction of y_{n+1} for a future x_{n+1}. For the purposes of model validation, we need to contrast each of the n values of x_i with its corresponding leave-one-out posterior distribution $\pi(\cdot|X_{-i}, Y)$; here X_{-i} stands for the data, omitting in each case the corresponding x_i. In all that follow we denote by x and y random variables corresponding to the omitted observed data x_i and y_i respectively. It is to be noted that a prior on x is necessary to obtain the above leave-one-out posterior distribution. For more on this, see Section 2.

Figure 1, which displays artificial data with $n = 10$ cases simulated from the Poisson model, shows that the observed x_8 falls within the support of $\pi(\cdot \mid X_{-8}, Y)$. Given suitable priors the simple Poisson case introduced above may be solved analytically; but in general this is not possible, particularly in the case of high dimensionality. In Vasko et al. (2000), $n = 62$, $dim(y) = 52$, $dim(x) = 1$ and θ involves 3318 parameters. In Bhattacharya and Haslett (2004) (see also Haslett et al. (2006)), $n = 7815$, $dim(y) = 14$, $dim(x) = 2$ and θ is a 9621-dimensional parameter. Both cases use regular MCMC; as the models are large and complex, computational speed is a critical issue.

In brief, our procedure involves first choosing one case, i^*, and investigating $\pi(x, \theta \mid X_{-i^*}, Y)$ by careful implementation of regular MCMC; this generates realisations corresponding to x_{i^*} and of θ. At the second stage, we use IR to resample from these θ values and, for each fixed value of θ, an MCMC run with which to explore $\pi(\cdot \mid y_i, \theta)$, for each of the remaining cases. Thus this involves a large number of MCMC runs. But critically, as x is low-dimensional, these are very fast. The overall procedure is thus regular MCMC (i^*) once, followed by IR and MCMC for all other cases. It is important to note that we re-use (the i^* specific) values of these high-dimensional θ but not those of x. The rationale is that the distributions of θ depend very weakly on i^*. We discuss, in Section 4, the choice of i^*; we will see that we have wide choice; in fact, a random choice of i^* is almost adequate. Previous literature, for forward problems, also involves a single initial regular MCMC, followed by an importance sampling stage for all n cases.

There is a more subtle issue with n-fold regular MCMC for large and complex models. There is always a danger that at least for some cases, the algorithm does not properly explore the parameter space, in this case (x, θ). In other words, the algorithm may not *mix* well. Recommended practice (see, for example, Gilks and Roberts (1996)) is to experiment with a number of MCMC options, including re-parameterisation and the use of several restarts, and to monitor convergence. The modeller will be unlikely to do this in n-fold regular MCMC when n is large. We demonstrate in Section 5 (see also

Section 6) that our procedure has attractive mixing properties.

The examples of inverse problems raise another issue, unrelated to computational speed. Cross-validation is a natural stage in model-fitting. Statistics summarising the observed discrepancies between the y_i and the $\pi(y \mid X, Y_{-i})$ are a natural basis for 'goodness of fit' measures (see, for example, Gelfand et al. (1992), Gelfand and Dey (1994), O'Hagan and Forster (2004)). But when y is high-dimensional such discrepancy measures are difficult to construct. However, if x is low-dimensional, one can easily construct and compute measures from the discrepancies between the x_i and $\pi(x \mid X_{-i}, Y)$. Hence, in such cases, inverse cross-validation may be an easier alternative to forward cross-validation even if the substantive underlying problem is forward. Moreover, Bhattacharya (2006) describes a novel approach based on the construction of reference distributions using data obtained from inverse cross-validation to adequately address the question of this aspect of Bayesian model fit.

The theory of IRMCMC is provided in Section 3. In Section 4, we discuss the important issue of selecting an appropriate i^*, illustrating with the Poisson regression problem. An experiment discussing appropriate run lengths for IRMCMC and illustrating its superior mixing properties compared to regular MCMC is provided in Section 5. In Section 6 we demonstrate the value of IRMCMC in a real example from palaeoclimatology. We discuss further research in Section 7. But before providing the theory, we first discuss forward and inverse problems in more detail.

2 Forward and inverse problems

Note that, assuming conditional independence, the likelihood of Y given X, θ is given by

$$L(Y, X, \theta) = p(y_i \mid x_i, \theta) \prod_{j \neq i; j=1}^{n} p(y_j \mid x_j, \theta) \tag{1}$$

Hence, if the prediction of y_i is of interest, then treating y_i as unknown, in (1) y_i must be replaced with y. One should then compute a posterior predictive distribution of y. This corresponds to the forward problem. Note that the distribution of the unknown y is simply $p(\cdot \mid x_i, \theta)$.

On the other hand, if it is of interest to predict x_i, then it must be treated as unknown, and replaced with x in (1). A posterior predictive distribution of x requires a prior for x. A question that arises now is that how the prior for x should be chosen. Since θ is the model parameter, the prior for θ, denoted by $\pi(\theta)$, should be chosen independently of x. The prior for x may depend upon θ; we denote the joint prior for (x, θ) by $\pi(x, \theta) = \pi(\theta)\pi(x \mid \theta)$. However, in most cases, it will be convenient to use a prior for x, given by $\pi(x)$, independent of θ. An important point to note is that it is important for $\pi(x)$ to include x_i in its support. It is possible to change $\pi(x)$ for each case i, but it is more convenient to use a single prior for x that includes all x_i in its support. But whatever the prior, irrespective of its dependence on θ, our proposed methodology accommodates all. We now provide comparative details of the posterior predictive distributions in the case of forward and inverse problems.

The prediction of y_i involves computing the posterior

$$\begin{aligned}\pi(y \mid X_{-i}, Y_{-i}, x_i) &= \int p(y \mid x_i, \theta)\pi(\theta \mid X_{-i}, Y_{-i})d\theta \\ &\propto \int \pi(\theta)p(y \mid x_i, \theta) \prod_{j\neq i; j=1}^{n} p(y_j \mid x_j, \theta)d\theta \qquad (2)\end{aligned}$$

which is equivalent to the expectation of $p(y \mid x_i, \theta)$ with respect to the posterior $\pi(\theta \mid X_{-i}, Y_{-i})$. For the Poisson regression problem, using a flat prior for θ it follows from (2) that,

$$\pi(y \mid X_{-i}, Y_{-i}, x_i) \propto \frac{x_i^y}{y!}\frac{\Gamma(y + \sum_{j\neq i} y_j + 1)}{(\sum_{j=1}^n x_j)^{(y+\sum_{j\neq i} y_j + 1)}}$$

But our interest is in the ‘inverse’ of the above problem. In other words, we are interested in learning about $\pi(x \mid X_{-i}, Y_{-i}, y_i)$ in each case. This is given by

$$\begin{aligned}\pi(x \mid X_{-i}, Y_{-i}, y_i) &= \int \pi(x \mid y_i, \theta)\pi(\theta \mid X_{-i}, Y)d\theta \\ &\propto \int \pi(\theta)\pi(x \mid \theta)p(y_i \mid x, \theta) \prod_{j\neq i, j=1}^{n} p(y_j \mid x_j, \theta)d\theta \qquad (3)\end{aligned}$$

Observe that $\pi(x \mid X_{-i}, Y_{-i}, y_i)$ is the expectation of $\pi(x \mid y_i, \theta)$ with respect to the posterior $\pi(\theta \mid X_{-i}, Y)$. For the Poisson regression problem, using flat priors on both x and θ it follows that

$$\pi(x \mid X_{-i}, Y_{-i}, y_i) \propto \frac{x^{y_i}}{(x + \sum_{j\neq i} x_j)^{(\sum_{j=1}^n y_j + 1)}}$$

Note that, except in simple cases, as the Poisson regression example, simple analytic solutions to equations (2) and (3) are not available.

In such cases n-fold regular MCMC is a naturally available methodology for cross-validation. But clearly, as in Vasko et al. (2000) and Bhattacharya and Haslett (2004), where the parameters have very high dimension and the number of cases is large, repeating the regular MCMC procedure n times can be burdensome in the extreme.

There are particular problems with n-fold regular MCMC. In the real problem addressed by Vasko et al. (2000), regular MCMC with a fixed proposal mechanism is not only slow but completely fails to explore the bimodal solution in one case. Since n-fold regular MCMC is the cause of both problems described above, one way to avoid the problems is to replace regular MCMC by importance sampling (see, for example, Geweke (1989), Robert and Casella (1999), Geyer (1991)). The idea can be explained as follows.

Suppose that interest lies in estimation of the expectation of a function $h(\theta)$ with respect to a distribution $f(\theta)$. Suppose further that a sample $\theta^{(1)}, \cdots, \theta^{(N)}$ is available from another distribution $g(\theta)$. Then the expected value of $h(\theta)$ may be estimated as

$$\hat{E}_N(h(\theta)) = \frac{\sum_{\ell=1}^{N} h(\theta^{(\ell)}) w(\theta^{(\ell)})}{\sum_{\ell=1}^{N} w(\theta^{(\ell)})} \tag{4}$$

where $w(\theta) \propto f(\theta)/g(\theta)$ is the importance weight of θ. The estimate thus obtained is called the importance sampling estimate and the density $g(\theta)$ is known as the importance sampling density. The key observation is that, for implementation, the ratio $f(\theta)/g(\theta)$ need only be known up to a proportionality constant. For the conditions of the convergence of the importance sampling estimate see Geweke (1989). The quality of the estimate depends heavily on the variability of the importance weights, which depends on how similar $f(\theta)$ and $g(\theta)$ are (Robert and Casella (1999)). An important condition is that the support of the importance sampling density $g(\theta)$ should not be included in that of the density of interest, $f(\theta)$.

In forward cross-validation problems, $\pi(\theta \mid X, Y)$ may be used as the importance sampling density; we refer to this as the saturated posterior since it involves the complete available dataset. The idea was proposed by Gelfand et al. (1992) and Gelfand (1996). For each cross-validation, the sample available from the saturated posterior can be used to estimate $\pi(y \mid X, Y_{-i})$ as in (4). The weights, which are simply available, are given by

$$w_i(\theta) = \frac{\pi(\theta \mid X, Y_{-i})}{\pi(\theta \mid X, Y)} \propto \frac{1}{p(y_i \mid x_i, \theta)}. \tag{5}$$

In the Poisson case, the weights are proportional to $\exp(\theta x_i)(\theta x_i)^{-y_i}$.

Typically, however, importance weights with respect to the saturated posterior density are not in general available in inverse problems. This is because the weight function in this case, given by

$$w_i(\theta) = \frac{\pi(\theta \mid X_{-i}, Y)}{\pi(\theta \mid X, Y)} = \frac{\int \pi(x, \theta \mid X_{-i}, Y)dx}{\pi(\theta \mid X, Y)}, \tag{6}$$

may not be available if the integration on the right hand side of the above expression is not tractable analytically. Another difficulty with the importance sampling approach described above is that the normalizing constant of $\pi(x \mid y_i, \theta)$ in (3) may be unknown. This means that $h(\theta)$ of (4) is not known completely. Thus the importance sampling estimate given by (4) can not be evaluated. Hence the leave-one-out posterior distribution in inverse cross-validation problems may not be accessed by importance sampling. To overcome such difficulties we propose to combine very fast and easily implementable MCMC runs with IR, outlined below. One key contribution in this paper is the proposal of a novel importance sampling density that completely avoids the problem of analytic integration.

3 Importance Resampling MCMC

Our proposed procedure can be stated in the following manner.

1. Choose an initial case i^*. Use $\pi(x, \theta \mid X_{-i^*}, Y)$ as the importance sampling density.

2. From this density, sample values $(x^{(\ell)}, \theta^{(\ell)}); \ell = 1, \cdots, N$, for large N. Typically, regular MCMC will be used for sampling.

3. For $i \in \{1, \cdots, i^* - 1, i^* + 1, \cdots, n\}$ do

   a. For each sample value $(x^{(\ell)}, \theta^{(\ell)})$, compute importance weights $w_{i^*,i}^{(\ell)} = w_{i^*,i}(x^{(\ell)}, \theta^{(\ell)})$, where the importance weight function is given by

   $$w_{i^*,i}(x,\theta) = \frac{\pi(x,\theta \mid X_{-i}, Y)}{\pi(x,\theta \mid X_{-i^*}, Y)} \propto \frac{L(Y, X_{-i}, x, \theta)}{L(Y, X_{-i^*}, x, \theta)} = \frac{p(y_{i^*} \mid x_{i^*}, \theta) p(y_i \mid x, \theta)}{p(y_{i^*} \mid x, \theta) p(y_i \mid x_i, \theta)}. \tag{7}$$

   Thus, for the Poisson regression problem, the weights are given by

   $$w_{i^*,i}(x,\theta) \propto x^{y_i - y_{i^*}} \exp\{\theta(x_i - x_{i^*})\}. \tag{8}$$

   b. For $k \in \{1, \cdots, K\}$

      (i) Sample $\tilde{\theta}^{(k)}$ from $\theta^{(1)}, \cdots, \theta^{(N)}$ where the probability of sampling $\theta^{(\ell)}$ is proportional to $w_{i^*,i}^{(\ell)}$.

      (ii) For *fixed* $\theta = \tilde{\theta}^{(k)}$, draw M times from $\pi(x \mid y_i, \tilde{\theta}^{(k)})$. Thus, for the Poisson regression case, with flat prior on x,

      $$\pi(x \mid y_i, \theta) \propto \exp(-\theta x) x^{y_i}, \tag{9}$$

      which we recognise as the Gamma distribution. Note that in general it is not easy to sample from $\pi(x \mid y_i, \theta)$, even when x is univariate, and we recommend MCMC for generality. For example, for the Poisson regression case, if the prior on x is given by a Cauchy distribution, truncated on $(0, \infty)$, then

      $$\pi(x \mid y_i, \theta) \propto \frac{1}{1 + x^2} \exp(-\theta x) x^{y_i}. \tag{10}$$

      To generate samples from (10), MCMC seems to be the simplest methodology.

   c. Store the $K \times M$ draws of x as the posterior for x_i as $\hat{x}_i^{(1)}, \cdots, \hat{x}_i^{(KM)}$.

The key idea in the above proposal is the use of $\pi(x, \theta \mid X_{-i^*}, Y)$ as the importance sampling density, for some particular i^*. Recall that in forward problems importance weights given by (5) are easily computable but those in inverse problems, given by (6), require analytic integration and so may not be available. Note that, unlike (6), importance weights (7) do not require integration for tractability; hence they are generally easily computable. Thus steps 1, 2 and 3a eliminate the problem of analytic integration in (6). Observe that the importance weights are independent of the prior on (x, θ).

An important technical question is whether IR whould be used with or without replacement. Indeed, most of the references to IR in the literature use sampling with replacement. See, for example, Gelfand et al. (1992), Newton and Raftery (1994), O'Hagan and Forster (2004). However, Gelman et al. (1995), Stern and Cressie (2000) recommend IR without replacement. They argue that sampling without replacement can provide protection against highly variable importance weights. Recently, Skare et al. (2003) formally prove a theorem that IR without replacement is better than IR with replacement, with respect to the total variation norm. Thus, in our proposal we recommend the former. Bhattacharya (2004) provides further details in this context including a comparison of IR with/without replacement.

4 Selection of appropriate importance sampling density

It follows from Section 2 that for IR to be most effective, it is desirable that the importance sampling density resembles the target density as closely as possible. It is shown in Bhattacharya (2004) that the total variation distance between $\pi(x,\theta \mid X_{-i^*},Y)$ and $\pi(x,\theta \mid X_{-i},Y)$ tends to zero as the data size n increases indefinitely. However, it is of interest to examine the situation for relatively small data size. We recognize that the task is to choose that $\pi(x,\theta \mid X_{-i^*},Y)$ which is 'closest overall' to the remaining $\pi(x,\theta \mid X_{-i},Y)$. We thus consider methods of choosing a 'central density'. We do this by consideration of pairwise 'distance measures' $d(i,j)$ between the posterior densities corresponding to cases i and j; given such distance measures, we select case i^* where $i^* = \arg\min_j \sum_i d(i,j)$. We propose two definitions of distance.

4.1 A KL motivation for the selection of i^*

Denoting the expectation with respect to $\pi(x,\theta \mid X_{-j},Y)$ by E_j, we note that,

$$E_j[w_{j,i}(x,\theta)] = \int \frac{\pi(x,\theta \mid X_{-i},Y)}{\pi(x,\theta \mid X_{-j},Y)}\pi(x,\theta \mid X_{-j},Y)dxd\theta = 1$$

Observe that $E_j[\log\{w_{j,i}(x,\theta)\}] = \log\{E_j[w_{j,i}(x,\theta)]\} = 0$ if and only if the weights are equal. We describe $d_{KL}(i,j) = |E_j[\log\{w_{j,i}(x,\theta)\}]|$ as a measure of difficulty in using $\pi(x,\theta \mid X_{-j},Y)$ as a basis for estimating $\pi(x,\theta \mid X_{-i},Y)$. Note that if $w_{j,i}(x,\theta)$ includes normalizing constants of the posteriors $\pi(x,\theta \mid X_{-i},Y)$ and $\pi(x,\theta \mid X_{-j},Y)$, then $-E_j[\log\{w_{j,i}(x,\theta)\}]$ is in fact the KL distance between these posteriors. We note here that Bradlow and Zaslavsky (1997) have used KL distance to measure synergy between pairs of cases in the forward context of Bayesian hierarchical models.

We point out in our context that if the normalising constants of the posteriors are not included, then $-E_j[\log\{w_{j,i}(x,\theta)\}]$ may not be non-negative (an example being that $w_{ji}(x,\theta) = c$, a constant greater than one, independent of (x,θ)). Hence, for the sake of generality, we use $|E_j[\log\{w_{j,i}(x,\theta)\}]|$. Also, since $d_{KL}(i,j)$ is not symmetric we modify it to $\hat{d}_{KL}(i,j) = d_{KL}(i,j) + d_{KL}(j,i)$.

For the Poisson regression case, it follows from (8) that

$$d_{KL}(i,j) = \mid (y_i - y_j)E_j(\log(x)) + E_j(\theta)(x_i - x_j) \mid . \tag{11}$$

Using the above equation, $\hat{d}_{KL}(i,j)$ can easily be calculated.

It is clear from (11) that the above theory can be used formally only when the posteriors $\pi(x,\theta \mid X_{-i},Y)$ are known or samples already available from them. However, this is not the case in reality. One approximation is to use the absolute value of

$$E_j[\log\{w_{j,i}(x,\theta)\}] \approx \log\{w_{j,i}\left(E_j(x),E_j(\theta)\right)\} \tag{12}$$

A natural alternative for θ is to use expected values from the saturated posterior. This we denote by $E_{sat}(\theta)$. For x, we propose to approximate $E_j(x)$ by x_j. Thus it follows from (7) that,

$$E_j[\log\{w_{j,i}(x,\theta)\}] \approx \log\frac{p(y_i|x_j,\theta)}{p(y_i|x_i,\theta)}, \tag{13}$$

which equals zero when $x_i = x_j$. Note that this is close in spirit to the KL-distance between $p(y_i \mid x_i,\theta)$ and $p(y_i \mid x_j,\theta)$. Thus, for the Poisson regression, (11) may be modified to

$$d_{KL}(i,j) = |(y_i - y_j)\log(x_j) + E_{sat}(\theta)(x_i - x_j)|, \tag{14}$$

Thus it is simple to compute $\hat{d}_{KL}(i,j)$ for all j, given $E_{sat}(\theta)$. Note however, that $E_{sat}(\theta)$ is not simply available. A separate regular MCMC, in addition to the regular MCMC run for $\pi(x,\theta \mid X_{-i^*},Y)$, is necessary for its computation. This extra computing effort is not negligible when θ is of very high dimensionality (for example, in Vasko et al. (2000), Bhattacharya and Haslett (2004)). It is not worth the effort if there exist simpler methods that perform as well or better. Moreover, it will be demonstrated in Section 4.3 that the effects of approximations used to compute $\hat{d}_{KL}$ may not be negligible and may adversely affect performance.

4.2 Other measures of centrality to determine i^*

An adequate approximation to equality of weights is to choose i^* such that $L(Y,X_{-i^*},x,\theta)$ is roughly 'central' in the set of $L(Y,X_{-i},x,\theta)$. We elaborate with the Poisson example for which the importance weights are given by (8). We propose two distance measures whose minimisation offers different versions of centrality :

$$\text{(a)}\quad d_1(i) = \sum_{j=1}^{n}\left(\frac{|\,x_j - x_i\,|}{S_X} + \frac{|\,y_j - y_i\,|}{S_Y}\right) \tag{15}$$

$$\text{(b)}\quad d_2(i) = \sqrt{\sum_{j=1}^{n}\left(\frac{(x_j - x_i)^2}{S_X^2} + \frac{(y_j - y_i)^2}{S_Y^2}\right)} \tag{16}$$

where S_X and S_Y denote sample standard deviations of the X column and the Y column respectively. Note that the above measures can be easily extended to situations where x_i and y_i are multivariate. In the case where $x_i = (x_{i1},\cdots,x_{ip})$ and $y_i = (y_{i1},\cdots,y_{iq})$, the distance measures are defined by

$$(c)\quad d_1(i) = \sum_{j=1}^{n}\left(\sum_{k=1}^{p}\frac{\mid x_{jk}-x_{ik}\mid}{S_{X_k}}+\sum_{k=1}^{q}\frac{\mid y_{jk}-y_{ik}\mid}{S_{Y_k}}\right) \tag{17}$$

$$(d)\quad d_2(i) = \sqrt{\sum_{j=1}^{n}\left(\sum_{k=1}^{p}\frac{(x_{jk}-x_{ik})^2}{S^2_{X_k}}+\sum_{k=1}^{q}\frac{(y_{jk}-y_{ik})^2}{S^2_{Y_k}}\right)} \tag{18}$$

In the above S_{X_k} and S_{Y_k} denote sample standard deviations of the k^{th} column of X and Y respectively.

With the above proposition, we then have $i^* = \arg\min\{d_k(i); 1 \leq i \leq n\}$, for $k = 1, 2$. Note that unlike measures based on d_{KL}, no knowledge is required of any quantity to be estimated and thus seems far more reasonable and simpler to compute. We next demonstrate with the Poisson regression example that the measures d_1 and d_2 may outperform the procedure motivated by KL distance. In fact, we show that even a random choice of i^* from $\{1, \cdots, n\}$ may perform more adequately than the latter.

4.3 Comparison between methods of choosing an appropriate i^*

In this section we use the Kolmogorov-Smirnov (KS) measure to evaluate the performance of IRMCMC with respect to different choice of i^* in the case of the Poisson problem. The KS measure is defined by

$$\sup_{z\in R} \mid G_n(z) - G(z) \mid \tag{19}$$

In (19), G is the true distribution function of the marginal posterior of x corresponding to case i has been omitted and G_n denotes the empirical distribution function of the marginal posterior of x defined as

$$G_n(x) = \frac{1}{N}\sum_{\ell=1}^{N}\delta_{(-\infty,x]}(\hat{x}_i^{(\ell)})$$

where δ denotes the indicator function. For details and related issues see Lehmann (1986), Billingsley (1995), Rao (1965). Recall that the true distribution is easily available in this toy problem.

For a fixed value of θ, we simulate 500 replicates of (X, Y) such that, for $i = 1, \cdots, 10$, $y_i \sim P(\theta x_i)$. For each of the 500 replications, the P-values associated with the observed KS-measure are computed for the 10 cases. This has been repeated for different ways of selecting i^*. For the procedure motivated by the KL distance we can envisage four versions, each version shedding different light on the basic issue of selecting an appropriate i^*.

(1) KL-1: Approximate version of d_{KL} given by (14) is used. Implementation of this version seems to be feasible and sensible in practice.

(2) KL-2: Exact version of d_{KL} is used with normalising constants, making d_{KL} a KL-distance, is used. This is given by

$$d_{KL}(i,j) = |(y_i - y_j)E_j(\log(x)) + (x_i - x_j)E_j(\theta)| + \log(c_i) - \log(c_j),$$

where c_i is the normalising constant of $\pi(x,\theta \mid X_{-i}, Y)$, given by

$$c_i = \frac{(\sum_{k\neq i} x_k)^{(\sum_{k\neq i} y_k)}}{\Gamma(\sum_{k\neq i} y_k)\Gamma(y_i+1)}.$$

$E_j(\theta)$ is given by $\sum_{k\neq j} y_k / \sum_{k\neq j} x_k$ and $E_j(\log(x))$ has been evaluated numerically. Note that in this simple Poisson regression case, where analytical solution is available, such form of d_{KL} might be used. But in reality such analytical solutions may not be available. However, this simple problem where analytical solutions are available will help us expose the fact that the approximation (14), although seems realistic and easy to compute, may not be sufficiently accurate and hence the performance of IRMCMC may be affected in that case.

(3) KL-3: In this case, the distance d_{KL} has been made independent of x by integrating it out. Here

$$d_{KL}(i,j) = |(y_j - y_i)E_j(\log(\theta)) - (y_j - y_i)E_j(\theta)|.$$

In the above, it has been assumed that $E_j(\log(\theta)) \approx E_{sat}(\log(\theta))$ and $E_j(\theta) \approx E_{sat}(\theta)$. Also note that the normalizing constants corresponding to cases i and j are not considered.

This version involves the assumption that x can be integrated out analytically, which is unrealistic, but is helpful in demonstrating that the random variable x involved in the measure d_{KL} can adversely affect selection of an appropriate i^*.

(4) KL-4: This is similar to KL-3 in essence but uses exact values of $E_j(\log(\theta))$ and $E_j(\theta)$ instead of approximations and normalizing constants (here $c_i = (\sum_{k\neq i} x_k)^{(\sum_{k\neq i} y_k)}/\Gamma(\sum_{k\neq i} y_k))$ taken into account making d_{KL} a KL-distance. It will be demonstrated that this is the best version; however, this will be unavailable in practice since analytical solutions are needed.

Apart from the performances of the above four versions of the procedure motivated by the KL distance, the performances of d_1, d_2 and the method of simple random selection are also considered and compared. It will be demonstrated that even with very small samples there is very little difference to choose between all the candidates for i^*, but that very simple measures seem to offer a choice that is easy to compute. We remark that both d_1 and d_2 seemed to exhibit similar performances; indeed in the realistic applications described in Chapters 6, 7 and 9 of Bhattacharya (2004) both yielded same results.

We say that IRMCMC corresponding to a given value of i^* satisfactorily approximates the target distribution of x at case i if the P-value for that case is greater than

0.05. In each replication, the number of P-values (note that there are ten P-values in each replication) exceeding 0.05 is noted. Let us denote this number corresponding to the r^{th} replication by $\mathcal{N}_r$.

The proportion of times $\mathcal{N}_r = 10$ with respect to the proposed measures, may be used to compare the performances of the measures. This is given by

$$\hat{P}(\mathcal{N}_r = 10) = \sum_{r=1}^{500} \delta_{\{\mathcal{N}_r=10\}}(\mathcal{N}_r)/500,$$

where δ denotes the indicator function. A high value of $P(\mathcal{N}_r = 10)$ indicates satisfactory performance of IRMCMC, given a particular distance measure.

Using the above criterion, Table 1 compares the performances of different ways of selecting i^* for different values of θ.

Observe that all seven proposals perform quite adequately, the proportons $\hat{P}(\mathcal{N}_r = 10)$ being high. That this holds despite the fact that the variabilities of the simulated data sets change as θ changes demonstrates the considerable robustness of IRMCMC. It is particularly satisfying to note that even a randomly selected i^* performs very adequately. .

However, compared to other procedures, the performance of KL-1 is the poorest. The exact version of KL-1, denoted by KL-2, performs better than KL-1, indicating that crude approximations involved in KL-1 might have adversely affected its performance.

The version KL-3, which corresponds to approximation after integrating out the random variable x, perform better than both KL-1 and KL-2. This is not unexpected, since in both KL-1 and KL-2 the random variable x, which can be regarded as a nuisance parameter while resampling θ only, is retained. This causes loss of efficiency of the procedure. Since only posteriors of θ, not x, are of interest, while resampling θ, and since an appropriate choice of i^* is needed only to ensure efficiency of the resampling procedure, the choice of i^* should not explicitly depend on x. Since KL-3 avoids this problem, it performs much better than both KL-1 and KL-2.

The version KL-4 is the exact version of KL-3, and hence outperforms KL-3. The fact that x has been integrated out and that exact solutions have been used help KL-4 perform excellently. Clearly, this is the best performer among all seven proposals.

Note that the distance measures d_1 and d_2 perform better than all procedures other than KL-4. This is because the measures use information from the data only for their determination of i^* and involve no unknown parameters and consequently no approximations. This makes them safe from unreliable approximations that could have made them inefficient. Neither do they involve the undesirable random variable x in the distance calculation. Thus they perform better than KL-1, KL-2, KL-3. Since d_1 and d_2 use information from the data and the proposal of the random choice of i^* use absolutely no information, they also perform better than the random choice proposal. On the other hand, since they do not use information on the posterior of θ, which is of interest, they perform less efficently than KL-4, which rightly use information on θ. Thus d_1 and d_2

seem to be compromise between the most desirable and less desirable characteristics. However, since KL-4 is unrealistic in practice, we recommend d_1 and d_2 as reliable and realistic measures of determining i^* appropriately. Since both perform adequately, we arbitrarily recommend d_1.

4.4 Extreme observation

Figure 2 presents a data set with an influential observation at case 2. Using measure d^*, $i^* = 5$ is obtained as the minimiser. However, $\pi(\theta \mid X_{-5}, Y)$ approximates $\pi(\theta \mid X_{-2}, Y)$ very poorly; see Figure 3. This is because (see Section 2) the support of the posterior of θ at the extreme observation case includes that of case 5 and all other cases; see Figure 4. The smaller support of case 5 does not allow adequate represention of the parameter space of θ at case 2 which in turn causes poor approximation of the posterior of x at case 2. Thus we see that IRMCMC can fail, if there are unsuspected extreme cases or cases that are influential in some unsuspected way.

We remark, however, that in the case of such potential problems, the distance d^* takes extreme values, thus pointing to such influential cases. In this particular Poisson example, the value of $d^*(i, 2)$ for each $i \neq 2$ (and hence $\sum_{i=1}^{10} d^*(i, 2)$) was extremely large; see Bhattacharya (2004). Once the problem is diagnosed, carefully designed regular MCMC may be employed for that case. For more discussion on extreme cases, see Section 7.

5 Mixing

In our experience, sensible choice of K and M is not difficult for adequate performance of IRMCMC. In fact, any sensible choice that makes $K \times M$ sufficiently large so that the Monte Carlo error (see, for example, Jones and Hobert (2001)) of x falls below a certain pre-specified level, seems to be acceptable. Bhattacharya (2004) proposed a 'quick and generic' method of determining K and M. However, we demonstrate below that moderate values of K and M lead to superior mixing properties of IRMCMC, compared to regular MCMC.

Clearly, since sampling θ by IR is computationally very much cheaper than using regular MCMC, IRMCMC requires much less time than regular MCMC to generate samples of a given size, that is, to perform a fixed number of iterations. We demonstrate with the Poisson problem that even in a fixed number of iterations, IRMCMC explores the target posterior as well as, or even better, than regular MCMC for moderate choices of K and M. Thus IRMCMC mixes better than regular MCMC both with respect to computational time and per iteration. To compare the performances of IRMCMC (here $i^* = 8$) and regular MCMC we use the KS measure.

Table 2 displays $R_{K,M}$, the ratio of the KS measure corresponding to IRMCMC and regular MCMC respectively, each having a run of length $K \times M$. A small (< 1) value of $R_{K,M}$ indicates that IRMCMC is more accurate than regular MCMC given respective values of K and M. On the other hand, a ratio greater than one indicates that regular MCMC is better. We observe that for $K = 1$, IRMCMC is generally more accurate than regular MCMC when M is relatively small but, for large M, regular MCMC seems better. This is because the posterior correlation between x and θ makes the autocorrelation in the regular MCMC samples of x higher than in the IRMCMC case for each fixed θ and makes small sample sizes less adequate in the former case. This has been supported by our experiments (not shown). The last column, giving the ratios when $K = 50$ and $M = 100$ shows that IRMCMC and regular MCMC produce quite similar results in some cases, but that the former is significantly better than the latter in other cases. This is in accord with the previous columns and suggests that it is worth taking K and M of moderate value. For the computationally challenging problem of Section 6 we have used $K = 50$ and $M = 100$.

This provides insight into the real advantage of IRMCMC over regular MCMC. For even where x and θ are low dimensional, given moderate K and M, IRMCMC mixes as fast as, and sometimes faster than, regular MCMC. But unlike in regular MCMC, the computational cost of IRMCMC does not increase with the dimensionality of θ. Thus in situations where θ is very high-dimensional, simulating just a few values of θ by IR and running an MCMC chain for the low-dimensional variable x is clearly very much less expensive than simulating a large number of them by regular MCMC methods.

We have thus argued that a two-stage procedure such as IRMCMC has particular advantages over n-fold regular MCMC in terms of both computation and mixing. We have argued - both by general consideration of cases with large n and by detailed examination of a simple case with very small n - that the choice of i^* is not difficult.

6 Application of IRMCMC to the motivating palaeoclimate problem

Vasko et al. (2000) reported a n-fold regular MCMC cross-validation exercise for a data set comprising multivariate counts y_i on $m = 52$ species of chironomid at $n = 62$ lakes (sites) in Finland. The unidimensional x_i denote mean July air temperature. As species respond differently to summer temperature, the variation in the composition provides the analyst with information on summer temperatures. This information is exploited to reconstruct past temperatures from count data derived from fossils in the lake sediment; see Korhola et al. (2002). Thus, counts y_i are simply related to x_i in a forward sense; but interest lies in temperature x_i, the 'inverse' of the relationship.

The inverse cross-validation exercise was computationally challenging and IRMCMC led to very considerable computational savings. In particular, our implementation of regular MCMC on a modern personal computer took 16 hours. In contrast, the IRMCMC implementation took 16 minutes for the initial run and 20 minutes for the remaining 61. Additionally, IRMCMC drew attention to the bimodality of one of the posteriors, a point completely missed by the regular MCMC implementation. We provide details of this below. First we explain the high dimensionality of θ and our implementation of cross-validation by IRMCMC.

In Vasko et al. (2000), the vector y_i of counts at site i followed the multinomial distribution,

$$(y_i \mid y_{i+}, p_i) \sim Multinomial(y_{i+}, p_i). \tag{20}$$

Here $y_i = (y_{i1}, \cdots, y_{im})$, $y_{i+} = \sum_{k=1}^{m} y_{ik}$ and p_i is an (unobserved) vector of relative abundances $(p_{i1}, \cdots, p_{im})$, of length $m = 52$. The unobserved $\{p_i; i = 1, \cdots, n\}$ thus provide 62×51 parameters, even before temperature x_i is related to the relative abundances. The Dirichlet distribution, because of its conjugacy with the multinomial model, is a convenient way to relate these. In particular, Vasko et al. (2000) suppose,

$$(p_i \mid x_i) \sim Dirichlet(\Lambda_i). \tag{21}$$

In (21), the kth component λ_{ik} of Λ_i was modelled as $\lambda_{ik} = \lambda(x_i, \Psi_k)$, for a simple function λ of x_i and of $\Psi_k = (\alpha_k, \beta_k, \gamma_k)$, a 3-component parameter vector associated with the kth species. Vasko et al. (2000) chose a simple unimodal function of these species specific parameters, given by $\lambda(x_i, \Psi_k) = \alpha_k \exp\{(\beta_k - x_i)/\gamma_k\}^2$. The mode, β_k represents the value of temperature at which species k is most abundant. Tolerance of the species is denoted by γ_k and α_k is a scaling factor. There are thus an additional 3×52 parameters, yielding 3318 in total. For further detail, and choice of priors, see Vasko et al. (2000).

6.1 Cross validation of the model using IRMCMC

The importance weight function leaving out site i^* is given by

$$w_{i^*,i}(x, \theta) = \frac{\Gamma(\sum_{k=1}^{52} \lambda(x_{i^*}, \Psi_k))}{\Gamma(\sum_{k=1}^{52} \lambda(x_i, \Psi_k))} \prod_{k=1}^{52} \frac{\Gamma(\lambda(x_i, \Psi_k))}{\Gamma(\lambda(x_{i^*}, \Psi_k))} \cdot \frac{p_{ik}^{\lambda(x,\Psi_k) - \lambda(x_i,\Psi_k)}}{p_{i^*k}^{\lambda(x,\Psi_k) - \lambda(x_{i^*},\Psi_k)}} \tag{22}$$

Observe that weights (22) are independent of the count data y_i. Hence, it follows from the definition of d^* that $i^* = \{i : x_i = median(X)\}$. Among the two medians in this case (since $n = 62$ is even), we arbitrarily choose $i^* = 38$.

In this case the density of p_i depends on x; hence simulation from $\pi(x \mid y_i, \theta)$ is done by generating samples of both x and p_i (and finally ignoring realisations of p_i) from the joint density of (x, p_i), given y_i, θ. The latter is given by

$$\pi(x, p_i \mid y_i, \theta) \propto \pi(x) \frac{\Gamma(\sum_{k=1}^{52} \lambda(x, \Psi_k))}{\prod_{k=1}^{52} \Gamma(\lambda(x, \Psi_k))} \prod_{k=1}^{52} p_{ik}^{\lambda(x,\Psi_k) + y_{ik} - 1} \tag{23}$$

In (23), $\pi(x)$ is a normal density with specified mean and variance. The above density (available up to the normalising constant) is non-standard, highly complicated and MCMC seems to be the only feasible methodology for generating samples from it. We recall from (10) that simple methods of simulation from $\pi(x \mid y_i, \theta)$ may not be available even for very simple problems and in general MCMC is necessary.

For the implementation of IRMCMC, we choose $N = 5000, K = 50, M = 100$. Results of cross-validation obtained by IRMCMC have been compared with 62-fold regular MCMC. Certainly due to lack of space, we omit details here. However, all 62 posteriors obtained by both IRMCMC and regular MCMC are available in Bhattacharya (2004). Both the methods agreed with each other at all sites except site 6.

6.2 IRMCMC and exploration of bimodal posterior

The densities at site 6 corresponding to regular MCMC and IRMCMC are shown in Figure 5. We observe that regular MCMC explores a unimodal posterior but IRMCMC explores a distribution with two modes, one being a minor mode, which in fact explains the observed datum as indicated by the vertical line.

Indeed, a bimodal posterior is not unexpected at that site since the species composition there is dominated by the presence of a very large count of a particular species. The situation may be interpreted as the abundant species and the remaining species having preference for disjoint regions of the climate space. Thus they send conflicting signals for these two regions, resulting in bimodality. Simulation studies confirm bimodality under these circumstances; see Bhattacharya (2004). For more on bimodality in the context of palaeoclimate problems, see Bhattacharya and Haslett (2004) and Haslett et al. (2006).

We remark that with *a priori* knowledge of bimodality, a carefully designed regular MCMC algorithm could have been adopted to explore the leave-one-out posterior distribution at site 6. But the problem was unsuspected and only a later examination of the data revealed it.

7 Conclusions and future work

This paper introduces IRMCMC as an effective methodology for cross-validation in inverse problems. The usefulness of our proposal has been illustrated with two examples; a toy Poisson regression problem and a real palaeoclimatological problem. We have also demonstrated its superiority over the default methodology n-fold regular MCMC. Guidelines on selecting an importance sampling density appropriately have been provided in Section 4 and their validity demonstrated.

The MCMC runs needed in Step 3b(ii) are easy to implement since the dimensionality of x is low. It is also useful to mention that in cases where it is easy to sample directly from $\pi(x \mid y_i, \tilde{\theta}^{(k)})$ (if the density is a standard one) one can realise an almost *iid* sample from the true posterior of x at the cost of no burn-in. In the Poisson example (but not in the case of Vasko et al. (2000)) such direct simulation is possible. However, we have demonstrated that even for very low dimensional cases, no generic method of simulation from $\pi(x \mid y_i, \tilde{\theta}^{(k)})$ is available. Thus we continue to recommend MCMC for its generality.

We have demonstrated that it does not matter much which i^* is chosen, even when the data size, n, is very small, providing that sensible steps are taken to avoid using extrema. At a very extreme case k, $\pi(\theta \mid X_{-k}, Y)$ will have much wider support compared to $\pi(\theta \mid X_{-j}, Y); j \neq k$, thus making the importance sampling density a poor approximation to the target density. In fact, all approximate methods for computing 'deletion diagnostics' (see, for example Haslett and Dillane (2004)) will fail in extreme cases and common sense needs to be used; this is same in the forward use of IR as well. We have argued that the distance measure d^* can help diagnose such problems. We remark that the example of the extreme case provided in this paper is very extreme indeed. We have conducted simulation studies with much less extreme observations (not reported) and in such cases any i^* seems quite satisfactory. When n is large, such problem of extreme case has very little effect. Since for small n there is no pressing need for IRMCMC, it can be argued that extreme cases are not, in practice, a problem for IRMCMC. However, in the presence of extreme observations it will be of interest to study the properties of an importance sampling density which is a mixture of extreme as well as central densities. This we reserve for future work.

In principle, IRMCMC can also be applied to general problems not related to cross-validation. Observe that the joint distribution of a set of at least two random variables can be factorized in the form $P(x, \theta) = P(x \mid \theta)P(\theta)$. IR may be used to sample from $P(\theta)$ and given θ, x may be sampled from $P(x \mid \theta)$ using MCMC.

Bhattacharya (2004) indicated that the output obtained by IRMCMC in a cross-validation exercise can be used to construct useful reference distributions of omnibus measures of model fit. He further showed that IRMCMC may also be very usefully employed for sensitivity analysis. These ideas will be communicated elsewhere.

Appendix

IRMCMC is MCMC with a special proposal mechanism

We demonstrate in this section that IRMCMC is really a version of MCMC with a special proposal mechanism.

For notational convenience in this section we use the shorthand notation $\pi_i(\cdot)$ for $\pi(\cdot \mid X_{-i}, Y)$. For the IRMCMC methodology, the proposal kernels of θ and x are given by

$$\begin{aligned} Q_{1N}(\theta^{(t+1)} \mid x^{(t)}, \theta^{(t)}) &= P(t : x^{(t)} \in \mathcal{S})\pi_{iN}(\theta^{(t+1)} \mid x^{(t)}) \\ &+ P(t : x^{(t)} \notin \mathcal{S})\delta_{\{\theta^{(t)}\}}(\theta^{(t+1)}) \end{aligned} \tag{24}$$

$$\begin{aligned} Q_2(x^{(t+1)} \mid \theta^{(t+1)}, x^{(t)}) &= P(t : x^{(t)} \in \mathcal{S})\pi_i(x^{(t+1)} \mid \theta^{(t+1)}) \\ &+ P(t : x^{(t)} \notin \mathcal{S})q(x^{(t+1)} \mid x^{(t)}) \end{aligned} \tag{25}$$

In the above, δ is the indicator function. $P(t : x^{(t)} \in \mathcal{S})$ denotes the probability that t is a stopping time. In other words, t is a stopping time if $x^{(t)}$ takes value in the set $\mathcal{S}$. $\pi_{iN}(\theta)$ is the empirical distribution of $\pi_i(\theta)$. Observe that the empirical distribution function in this case is given by

$$F_{i,N}(\theta) = \frac{\sum_{\ell=1}^{N} \delta_{(-\infty,\theta]}(\theta^{(\ell)}) w_i(\theta^{(\ell)})}{\sum_{\ell=1}^{N} w_i(\theta^{(\ell)})}$$

Clearly, by the ergodic theorem the above empirical distribution function converges almost surely to the true distribution function as N is made infinitely large. q is a distribution that may or may not depend on values $x^{(t)}$ and $\theta^{(t+1)}$. Note that we suppress $\theta^{(t+1)}$ in the notation. This is because we generally choose the distribution to be independent of $\theta^{(t+1)}$ (normally, this is a random walk).

We now have a closer look at the proposal kernels Q_1 and Q_2. Q_1 says that if t is a stopping time propose a new value, $\theta^{(t+1)}$ from the empirical distribution $\pi_{iN}(\theta^{(t+1)})$; if not then set $\theta^{(t+1)} = \theta^{(t)}$. A new value $\theta^{(t+1)}$ will be proposed by IR without replacement. The interpretation of Q_2 is similar to Q_{1N}. It says that if t is a stopping time propose a new value $x^{(t+1)}$ from the distribution $\pi_i(x \mid \theta^{(t+1)})$ (which will be typically done by MCMC) and if not then propose $x^{(t+1)}$ from any distribution q that may (or may not) depend on the current value $x^{(t)}$. Observe that in our case τ is a stopping time if $\tau \in \{M, 2M, \ldots, KM\}$; that is $P(\tau \in \{M, 2M, \ldots, KM\}) = 1$ and $P(\tau \notin \{M, 2M, \ldots, KM\}) = 0$. Note that our proposal implies keeping θ fixed for M consecutive realisations of x. This is a deterministic definition; however, randomness may be introduced by agreeing to stop the chain when the Monte Carlo error falls below a certain level.

Denoting the acceptance probability of $\theta^{(t+1)}$ given $\theta^{(t)}$ and $x^{(t)}$ by $\alpha_N(\theta^{(t+1)} \mid x^{(t)}, \theta^{(t)})$, and using (24), we observe that $\alpha_N(\theta^{(t+1)} \mid x^{(t)}, \theta^{(t)}) \to 1$ as $N \to \infty$, for $\pi_i(\cdot)$-almost all $\big(\theta^{(t)}, x^{(t)}\big)$.

On the other hand, the acceptance probability of $x^{(t+1)}$ given $\theta^{(t+1)}$ and $x^{(t)}$ depends on whether on not t is a stopping time. If t is a stopping time, then the acceptance probability, $\alpha(x^{(t+1)} \mid \theta^{(t)}, x^{(t)}) = 1$, and $\beta(x^{(t+1)} \mid \theta^{(t)}, x^{(t)})$ otherwise, where

$$\begin{aligned} &\beta(x^{(t+1)} \mid \theta^{(t)}, x^{(t)}) \\ &= \begin{cases} \min\left\{\frac{\pi_i(x^{(t+1)}|\theta^{(t+1)})}{\pi_i(x^{(t)}|\theta^{(t+1)})}\frac{q(x^{(t)}|x^{(t+1)})}{q(x^{(t+1)}|x^{(t)})}, 1\right\} & : \; \pi_i(x^{(t)} \mid \theta^{(t+1)}) q(x^{(t+1)} \mid x^{(t)}) > 0 \\ 1 & : \; \pi_i(x^{(t)} \mid \theta^{(t+1)}) q(x^{(t+1)} \mid x^{(t)}) = 0 \end{cases} \end{aligned} \tag{26}$$

The above facts show that IRMCMC is indeed a version of MCMC with special proposal kernels. The Markov chain can be written as

$$\begin{array}{rcc} \{(x^{(1)}, \theta^{(i_1)}), (x^{(2)}, \theta^{(i_1)}), & \cdots, & (x^{(M)}, \theta^{(i_1)}), \\ (x^{(1+M)}, \theta^{(i_2)}), (x^{(2+M)}, \theta^{(i_2)}), & \cdots, & (x^{(2M)}, \theta^{(i_2)}), \\ & \cdots & \\ (x^{(K+M)}, \theta^{(i_K)}), (x^{(K+M)}, \theta^{(i_K)}), & \cdots, & (x^{(KM)}, \theta^{(i_K)}), \cdots\}. \end{array}$$

Note that an initial burn-in is essential if MCMC is used to draw $x^{(t)}$ from $\pi_i(\cdot \mid \theta^{(1)} = \tilde{\theta}^{(1)})$. Corresponding to $\theta^{(k)}$, for $k > 1$ the last realization of x corresponding

Table 1: Assessment of the performances of methods of selecting an appropriate i^*.

$\hat{P}(\mathcal{N}_r = 10)$							
θ	KL-1	KL-2	KL-3	KL-4	d_1	d_2	Random
0.5	0.974	0.980	0.934	0.980	0.882	0.894	0.908
1.0	0.978	0.972	0.932	0.986	0.970	0.932	0.938
3.0	0.948	0.964	0.976	0.992	0.984	0.968	0.944
5.0	0.912	0.924	0.968	0.998	0.966	0.974	0.946
7.0	0.934	0.948	0.966	0.988	0.976	0.960	0.934
9.0	0.934	0.938	0.964	0.992	0.986	0.980	0.936
11.0	0.900	0.938	0.970	0.994	0.984	0.984	0.964
13.0	0.892	0.904	0.938	0.994	0.994	0.982	0.970
15.0	0.910	0.926	0.968	1.000	0.986	0.962	0.948

to $\theta^{(k-1)}$ could be used as the initial value and hence no burn-in is needed. Note also that typically MCMC (and burn-in) is needed to construct the empirical distribution function F_{iN} but observe that we do not need to evaluate the distribution function at any point but we only draw samples from F_{iN} without replacement.

References

Bhattacharya, S. (2004). "Importance resampling MCMC: a methodology for cross-validation in inverse problems and its applications in model assessment." Doctoral thesis, Department of Statistics, Trinity College Dublin. Available at http://www.tcd.ie/Statistics/JHpersonal/thesis.pdf.

— (2006). "Model assessment using inverse reference distribution approach." Technical report, Institute of Statistics and Decision Sciences, Duke University. Submitted.

Bhattacharya, S. and Haslett, J. (2004). "Fast cross-validation of a palaeoclimate model using IRMCMC." Technical report, Trinity College, Dublin, Ireland. Presented at TIES 2004 conference held at Maine. Available at http://www.tcd.ie/Statistics/JHpersonal/research.htm.

Billingsley, P. (1995). *Probability and measure*. New York: John Wiley and Sons.

Bradlow, E. T. and Zaslavsky, A. M. (1997). "Case influence analysis in Bayesian inference." *Journal of Computational and Graphical Statistics*, 6(3): 314–331.

Gelfand, A. E. (1996). "Model determination using sampling-based methods." In Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.), *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, 145–162. London: Chapman and Hall.

Table 2: Ratio of the KS measure corresponding to IRMCMC and regular MCMC.

Case	$R_{1,1000}$	$R_{1,2000}$	$R_{1,3000}$	$R_{1,4000}$	$R_{1,5000}$	$R_{50,100}$
1	0.337	0.739	1.670	1.807	3.560	1.028
2	0.391	0.317	1.674	0.500	1.243	1.741
3	1.431	1.289	2.299	0.706	1.240	0.799
4	0.712	1.331	2.216	1.055	0.932	0.393
5	0.754	0.625	0.824	2.511	2.684	0.326
6	0.626	0.243	1.838	2.616	2.314	1.209
7	0.905	0.661	0.521	4.181	0.512	1.083
8	–	–	–	–	–	–
9	0.395	2.976	1.115	1.157	1.729	1.160
10	1.509	1.512	2.195	1.200	0.817	0.894

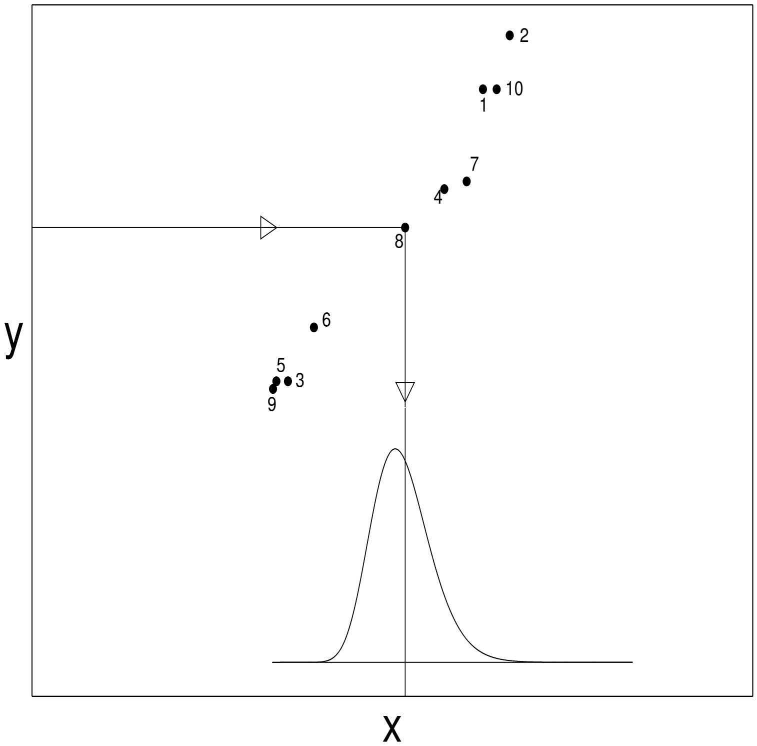

Figure 1: Artificial Poisson data for 10 cases; the case number is displayed alongside the datum. Case 8 illustrates cross-validation with Poisson regression.

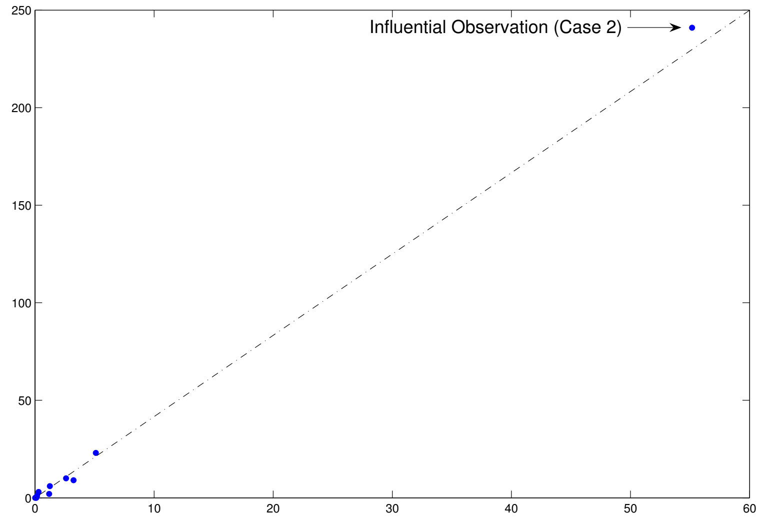

Figure 2: Influential data.

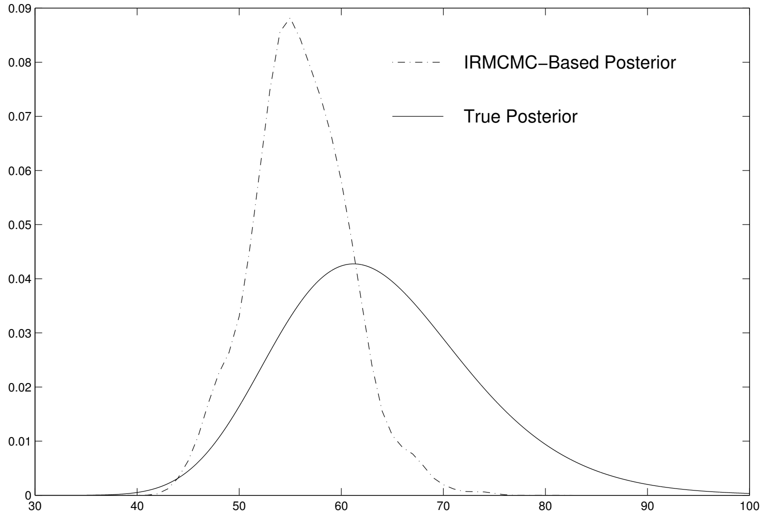

Figure 3: The true posterior $\pi(x \mid X_{-2}, Y)$ and the IRMCMC-approximated posterior with $i^* = 5$ are shown. IRMCMC does not adequately represent the parameter space in this case.

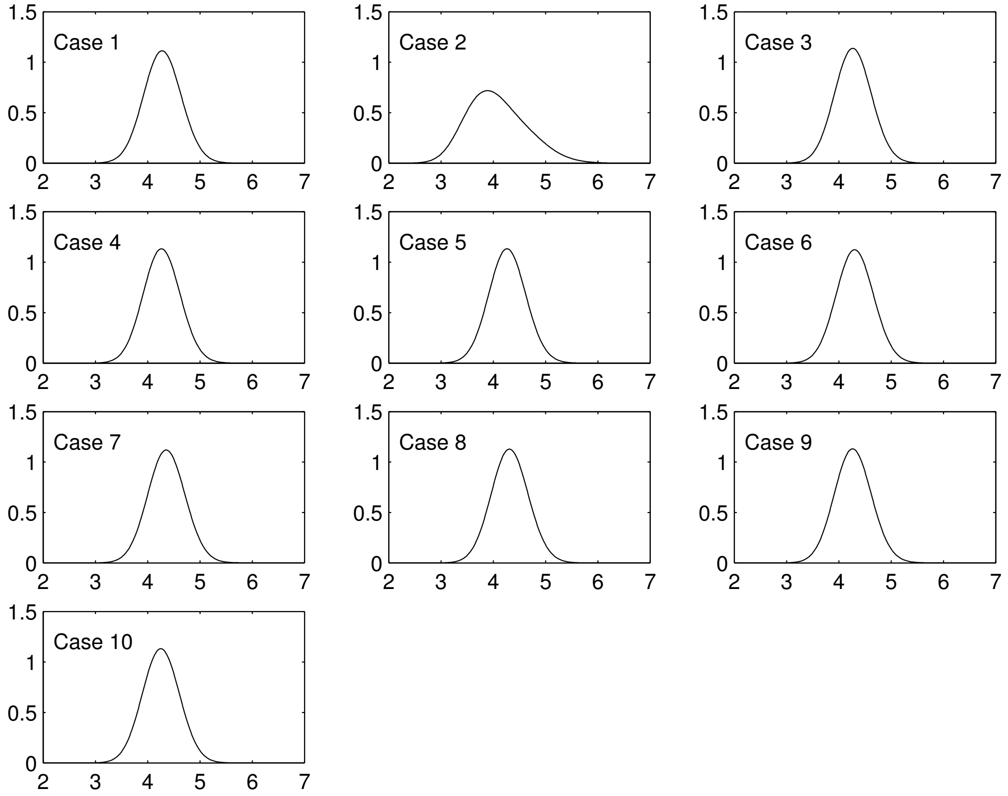

Figure 4: Posteriors $\pi(\theta \mid X_{-i}, Y); i = 1, \cdots, 10$. The support of $\pi(\theta \mid X_{-2}, Y)$ includes those of $\pi(\theta \mid X_{-j}, Y); j \neq 2$.

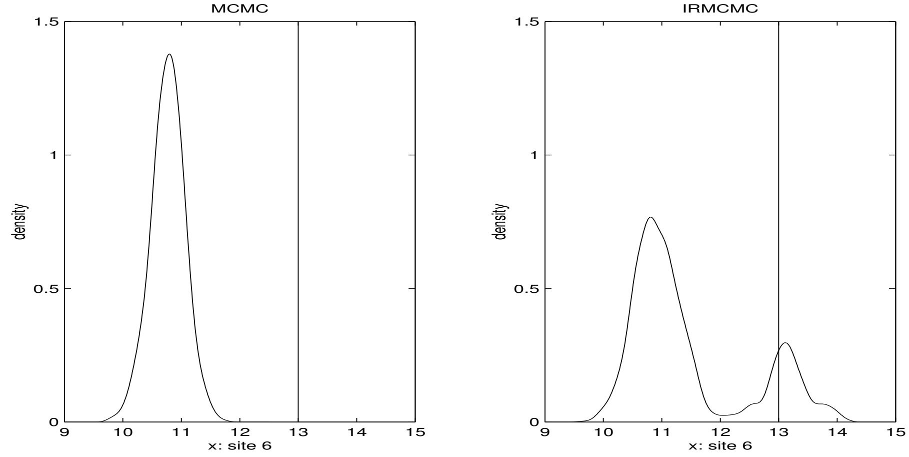

Figure 5: $\pi(x \mid X_{-6}, Y)$ obtained by MCMC and IRMCMC; the vertical line denotes the observed temperature.

Gelfand, A. E. and Dey, D. K. (1994). “Bayesian model choice: asymptotics and exact calculations.” *Journal of the Royal Statistical Society B*, 56(3): 501–514.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). “Model determination using predictive distributions with implementation via sampling methods (with discussion).” In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 147–167. Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. R. (1995). *Bayesian data analysis*. Chapman and Hall. Second Edition.

Geweke, J. (1989). “Bayesian inference in econometric models using Monte Carlo integration.” *Econometrica*, 57(6): 1317–1339.

Geyer, C. J. (1991). “Reweighting Monte Carlo mixtures.” Technical report, University of Minnesota, School of Statistics.

Gilks, W. R. and Roberts, G. O. (1996). “Strategies for improving MCMC.” In Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.), *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, 89–114. London: Chapman and Hall.

Haslett, J. and Dillane, D. (2004). “Application of ‘delete=replace’ to deletion diagnostics for variance component estimation in the linear mixed model.” *Journal of the Royal Statistical Society. Series B*, 66: 131–143.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B., and Mitchell, F. J. G. (2006). “Bayesian palaeoclimate reconstruction (with discussion).” *Journal of the Royal Statistical Society. Series A*, 129(3): 395–438.

Jones, G. L. and Hobert, J. P. (2001). “Honest exploration of intractable probability distributions via Markov chain Monte Carlo.” *Statistical Science*, 16(4): 312–334.

Korhola, A., Vasko, K., Toivonen, H. T. T., and Olander, H. (2002). “Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling.” *Quaternary Science Reviews*, 21: 1841–1860.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. New York, Inc: Springer-Verlag.

Newton, M. A. and Raftery, A. E. (1994). “Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion).” *Journal of the Royal Statistical Society. Series B*, 56(1): 3–48.

O’Hagan, A. and Forster, J. (2004). *Kendall’s advanced theory of statistics, Bayesian inference (Vol 2B)*. London: Arnold.

Rao, C. R. (1965). *Linear statistical inference and its applications*. New York: John Wiley and Sons.

Robert, C. and Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.

Rubin, D. (1988). “Using the SIR algorithm to simulate posterior distributions.” In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics 3*, 395–402. Oxford: New York.

Skare, Ø., Bølviken, E., and Holden, L. (2003). “Improved sampling-importance resampling and reduced bias importance sampling.” *Scandivanian Journal of Statistics*, 30: 719–737.

Stern, H. S. and Cressie, N. (2000). “Posterior predictive model checks for disease mapping models.” *Statistics in Medicine*, 19: 2377–2397.

Vasko, K., Toivonen, H. T., and Korhola, A. (2000). “A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction.” *Journal of Paleolimnology*, 24: 243–250.

Acknowledgments

The authors are grateful to to the editorial board of Bayesian Analysis whose suggestions helped improve the quality and presentation of the paper.

SB has been supported by Enterprise Ireland 14 grant SC/2001/171.