Monte Carlo methods for Bayesian palaeoclimate reconstruction

Jonathan Rougier*

Department of Mathematics
University of Bristol

Source: `supranetAddendum.tex`, March 28, 2011

Abstract

In palaeoclimate reconstruction, the natural modelling direction is forwards from climate to sensors to proxy measurements. Statistical methods can be used to invert this direction, making climate inferences from proxy measurements. Among these methods, the Bayesian method would seem to deal best with the substantial epistemic uncertainties about climate, and about its impact on sensors. The main challenge is to perform this inference efficiently within a simulation approach. This paper reviews the Importance Sampling approach to Bayesian palaeoclimate reconstruction, and then goes on to demonstrate the value of recent advances in Markov chain Monte Carlo (MCMC) inference.

KEYWORDS: MCMC, PSEUDO-MARGINAL

1 Brief introduction

One purpose of this note is to provide precision and justification for the simulation-based approach to palaeoclimate reconstruction outlined in SUPRAnet (2011), hereafter SUPRA. An outline of the general Bayesian inference for palaeoclimate reconstruction is given in section 2. Section 3 describes Importance Sampling, which is the estimation technique described in more general terms in SUPR. This technique is chosen because of its simplicity, and because it often corresponds quite closely to informal practice.

Importance Sampling has its limitations, and section 4 explains how these can be somewhat ameliorated with more sophisticated sampling techniques based on Markov chain Monte Carlo (MCMC). This is not a review of standard MCMC material, but presents some recent and powerful theoretical results that will prove

*`j.c.rougier@bristol.ac.uk`

to be extremely useful for palaeoclimate reconstruction. These results are stated as algorithms and proved. The proofs are fairly elementary, and are given here because they are not easily extracted from the original paper (Andrieu *et al.*, 2010). They demonstrate the power of the probability calculus, not just in terms of constructing probabilistic reprentations of complex structured quantities, such as palaeoclimate, but also in terms of deriving algorithms for estimating the properties of such representations. Section 5 contains a brief summary.

2 Outline of the inference

In the following analysis it is very important to distinguish between three ways in which a probability density function may be realised. For the purposes of illustration, consider an arbitrary random quantity X, for which we would like to know $\pi(x)$ for specified x. First, it might be possible to *evaluate* $\pi(x)$ *pointwise*, which is to say that $\pi(x)$ gives a definite computable value, and $\int \pi(x)\,\mathrm{d}x = 1$. Second, it might be possible to evaluate $\pi(x)$ pointwise *up to a constant*, which is to say that we have some other function $\pi'(x)$ which we can evaluate pointwise and for which $\pi'(x) \propto \pi(x)$, but $\int \pi'(x)\,\mathrm{d}x \neq 1$. Third, it might *only be possible to simulate* from $\pi(x)$, written $X \sim \pi(x)$. In the second and third cases there are techniques to construct a density function which we can evaluate pointwise, but if x is high-dimensional then these are infeasible, and approximations must suffice.

Let C denote climate and z^{m} be the measurements from climate proxies derived from sensors. For example, C might be the collection

$$\{(\mathrm{GDD5}_g, \mathrm{MTCO}_g, \mathrm{Precip}_g)\}_{g \in \mathcal{G}}$$

where $\mathcal{G}$ is a collection of gridcells tiling North America, for the mid-Holocene. The sensor might be North American vegetation, and z^{m} might be the collection

$$\{(d_{1s}, \ldots, d_{ks})\}_{s \in \mathcal{S}}$$

where d_{is} is the number of grains of pollen taxa i recorded at site s and dated to the mid-Holocene, and $\mathcal{S}$ is a collection of sites in North America. Within a Bayesian approach, we would like to describe the probability density function $\pi(c \mid z^{\mathrm{m}})$, where '|' denotes 'conditional upon', and adopting the standard conventions that (i) capital letters denote uncertain quantities, and small letters denote typical or specified values, and (ii) π denotes a generic probability density function, indexed by its arguments.[1] Note that 'probability density function' is taken to include the case where the uncertain quantites are discrete, when 'probability mass function' might also be used. The space of possible values of C, 'climate space', is denoted $\mathcal{C}$.

In the Bayesian approach, which supposes that it is possible for us to state our judgements about uncertainty using probability density functions, there is no

[1] Technically, indexed by the capitalisation of its unornamented arguments, so that, for example, $\pi(c \mid z^{\mathrm{m}}) \equiv \pi_{C|Z}(c \mid z^{\mathrm{m}})$, where $\pi_{C|Z}$ is the probability density function of C conditional on Z.

reason in principle why we might not assert $\pi(c \mid z^{\mathrm{m}})$ directly, for all values of $c \in \mathcal{C}$. Usually, however, it is easier and more transparent to build this probability density function from conceptually-simpler ones using the rules of the probability calculus, which is how we proceed here. One simplifying assumption is made up-front, which is that there is no feedback from the climate sensors to climate; this is relaxed in section 4.3.

First, let us specify 'prior' judgements about C in the form of a probability density function $\pi(c)$. 'Prior' is used here to denote 'prior to adjusting our judgements about C by the numerical values z^{m}'. The climate prior is likely to be structured in space and time, and also across types of climate variable. In the illustration given above a particular time was specified (the mid-Holocene), and so the indexing would be by location and type. The most satisfactory way to induce a physically-consistent structure in $\pi(c)$ is to use a climate simulator. In the simplest approach, the true climate C might be modelled as 'climate simulator output plus noise', where the 'noise' is the structural uncertainty of the simulator. If the noise is, say, multivariate Gaussian, then $\pi(c)$ can be evaluated pointwise, providing that the climate simulator has been evaluated. This situation is explored in Rougier (2007). In the more general case, the climate simulator itself might be stochastic, or the climate prior might be represented by a stochastic process, in which case it is likely that it can *only* be sampled from, i.e. we can sample $C \sim \pi(c)$, but *not* evaluate $\pi(c)$ or some function that is proportional to it.

Second, let Y denote the sensor, defined in such a way that the difference between Y and z^{m} is predominantly measurement uncertainty, of a type which can be straightforwardly specified in terms of a probability density function $\pi(z \mid y, c)$. The measurement uncertainty model is expected to be straightforward to specify, as it captures mainly aleatory uncertainty from the recording process. So, for example, a Multinomial distribution for counts, a Poisson distribution for numbers of events, an Exponential distribution for waiting times, a Gaussian distribution for measurements on a continuous scale; and their various generalisations. Therefore $\pi(z \mid y, c)$ will be treated as though it can be evaluated pointwise up to a constant. Note that while the distinction between the sensor Y and measurement z^{m} is not artificial, it has been introduced here mainly to simplify the exposition and the statistical modelling. There are closely-related inferential approaches that do not require a 'measurement uncertainty' probability density function that can be evaluated pointwise, i.e. approaches based on sampling all the way to Z (see, e.g., Beaumont *et al.*, 2002; Toni *et al.*, 2009). Note also that the measurement uncertainty probability density $\pi(z \mid y, c)$ would usually be invariant to the value c, but c is included as an argument for completeness and clarity.

Third, let us specify the conditional density function $\pi(y \mid c)$, which describes our judgements about Y conditional on knowledge of the climate. This distribution represents the 'forward model' from climate to sensor. The distribution $\pi(y \mid c)$ represents all of the uncertainty that sits between knowing the climate and knowing the response of the sensor. This would comprise mainly epistemic uncertainty—uncertainty induced by our lack of knowledge. In SUPR this would be represented by a cascade of uncertainties through a linked series of forward models. The use

of a probability distribution does not rule out a deterministic function; formally such a function $y = f(c)$ is represented probabilistically by a Dirac delta function, $\pi(y \mid c) = \delta(\|y - f(c)\|)$.[2] For the greatest possible generality, we will assume only that it is possibly to simulate $Y \sim \pi(y \mid c)$.

The 'no feedback' assumption made above ensures that it is reasonable to conceive of the joint distribution of climate and sensor in terms of the product of distributions

$$\pi(c, y) = \pi(c)\,\pi(y \mid c);$$

this factorisation is always possible (it is a core result in the probability calculus), but it is not always practicable. As Pearl (2000, chapter 1) notes, conditional probabilities are much easier to specify when they respect causality; without feedback, the direction $C \to Y$ is causal, and $\pi(y \mid c)$ is a natural conditional distribution to specify.

Table 1 summarises the probability density functions that need to be specified in this Addendum.

Once these three probability density functions have been specified we can construct our target probability density function:

$$\pi(c \mid z^{\mathrm{m}}) = (1/\mu)L(c)\,\pi(c) \propto L(c)\,\pi(c) \tag{1a}$$

according to Bayes's Theorem, where $L(c)$ is the *likelihood function* of C, and μ is the *marginal likelihood*,

$$L(c) := \pi(z^{\mathrm{m}} \mid c) \tag{1b}$$

$$\mu := \pi(z^{\mathrm{m}}) = \int_{\mathcal{C}} L(c)\,\pi(c)\,\mathrm{d}c, \tag{1c}$$

where ':=' denotes 'defined to be'. To evaluate $L(c)$ requires the Law of Total Probability:

$$L(c) = \int_{\mathcal{Y}} \pi(z^{\mathrm{m}} \mid y, c)\,\pi(y \mid c)\,\mathrm{d}y, \tag{1d}$$

where we introduce and then integrate out Y over the sensor-space $\mathcal{Y}$. The computational challenge is to sample from or otherwise summarise $\pi(c \mid z^{\mathrm{m}})$ according to the relationships in (1), but under the constraints that $\pi(c)$ and $\pi(y \mid c)$ can be sampled from but not evaluated pointwise. It is also interesting to consider the additional advantages that accrue if $\pi(c)$ can be evaluated pointwise up to a constant.

3 Importance sampling

Importance Sampling is a stochastic integration technique based on the Weak Law of Large Numbers (WLLN). The WLLN concerns the distribution of the sample

[2] The Dirac delta function is a spike at zero, defined as

$$\delta(x) := \lim_{\sigma \to 0} (2\pi\sigma^2)^{-1/2} \exp\{-x^2/(2\sigma^2)\}.$$

Table 1: Probability density functions that need to be specified.

Symbol	Meaning	Nature
$\pi(c)$	'Prior' distribution for climate	Simulate from, or evaluate pointwise up to a constant
$\pi(y \mid c)$	Conditional distribution of sensor given climate	Simulate from
$\pi(z \mid y, c)$	Measurement uncertainty: conditional distribution of proxy measurements given sensor and climate	Evaluate pointwise up to a constant

average of a collection of independent and identically distributed random variables. Suppose that $X^1, \ldots, X^m \overset{\text{iid}}{\sim} \pi(x)$, where '$\overset{\text{iid}}{\sim}$' denotes 'independent and identically distributed', and in this paper superscripts are used to denote replications. Then the WLLN states that the sample average $\bar{X}_m := (1/m) \sum_{i=1}^{m} X^i$ is a *consistent* estimator of $E(X)$, providing that this expectation exists; from now on, assume that all of the expectations that we want to compute actually exist. Consistency means, informally, that in the limit as the sample size becomes large, the probability that the sample average will diverge from the actual value goes to zero. The necessary theory is covered in Davison (2003, sec. 2.2). As a general property, referred to again below, the sample average is also an unbiased estimator of $E(X)$; i.e. $E(\bar{X}_m) = E(X)$ for all $\pi(x)$.

Importance sampling is a simulation method which is used to approximate integrals of functions of C. These integrals are summary statistics of the distribution of C, and, were it possible to compute enough of them, then the distribution of C would be known to arbitrary accuracy. In general, write $h(c)$ for an arbitrary specified function of c, such that we would like to compute

$$E(h(C) \mid z^{\mathrm{m}}) = \int_{\mathcal{C}} h(c)\, \pi(c \mid z^{\mathrm{m}})\, \mathrm{d}c. \tag{2}$$

If we wanted to compute the mean of $C \mid z^{\mathrm{m}}$, then we would set $h(c) = c$. Another choice for h might be the indicator function $h(c) = 1(c \leq v)$, for which (2) computes the probability that C is no greater than v. Usually, however, we would only consider low-order moments like the mean and the variance, because experience suggests that higher-order moments and tail probabilities tend to be less well estimated by simulation methods, and to be sensitive to aspects of our distributional choices about which we do not feel well-informed (Rougier, 2006; Goldstein and Wooff, 2007).

Suppose initially that $\pi(c \mid z^{\mathrm{m}})$ can be evaluated pointwise, and consider the introduction of a 'proposal' distribution $q(c)$ which can be evaluated pointwise,

and sampled from. Then (2) could be written

$$\begin{aligned} E[h(C) \mid z^{\mathrm{m}}] &= \int_{\mathcal{C}} h(c)\, \pi(c \mid z^{\mathrm{m}})\, \frac{q(c)}{q(c)}\, \mathrm{d}c \\ &= \int_{\mathcal{C}} h(c)\, \frac{\pi(c \mid z^{\mathrm{m}})}{q(c)}\, q(c)\, \mathrm{d}c \\ &\approx \sum_{i=1}^{m} h(C^i) W^i \qquad \text{for } C^i \overset{\text{iid}}{\sim} q(c) \end{aligned} \tag{3}$$

where $W^i := (1/m)\pi(C^i \mid z^{\mathrm{m}})/q(C^i)$. Provided only that $q(c) > 0$ whenever $\pi(c \mid z^{\mathrm{m}}) > 0$, this approximation is unbiased, and a consistent estimator of $E[h(C) \mid z^{\mathrm{m}}]$ according to the WLLN. Usually the idea is to make an efficient choice for q (one for which $W^i \approx 1/m$), but in our case this is not feasible if we cannot evaluate $\pi(c \mid z^{\mathrm{m}})$ pointwise. This difficulty is circumvented by setting $q(c) = \pi(c)$. This gives

$$E[h(C) \mid z^{\mathrm{m}}] \approx \sum_{i=1}^{m} h(C^i) W^i \qquad \text{where } C \overset{\text{iid}}{\sim} \pi(c), \tag{4}$$

and where $W^i := (1/m\mu)L(C^i)$, and $\pi(c \mid z^{\mathrm{m}})$ has been written as $(1/\mu)L(c)\pi(c)$. The calculation is still not finalised, because $L(c)$ cannot be evaluated pointwise, and μ is not known. However, both of these terms represent integrals, as shown in (1c) and (1d). Therefore both of these integrals can also be estimated by Importance Sampling.

Therefore in order to estimate (2) subject to our constraints on sampling, we need two nested sets of samples:

1. Sample $C^{1:m} \overset{\text{iid}}{\sim} \pi(c)$,
2. For each C^i, sample $Y_i^{1:n} \overset{\text{iid}}{\sim} \pi(y \mid C^i)$,

where $C^{1:m} := (C^1, \ldots, C^m)$, and $Y_i^{1:n} := (Y_i^1, \ldots, Y_i^n)$. The second set of samples is used to construct the estimators

$$\hat{L}_n(C^i) := (1/n) \sum_{j=1}^{n} \pi(z^{\mathrm{m}} \mid Y_i^j, C^i) \qquad i = 1, \ldots, m, \tag{5}$$

which are unbiased, and consistent estimators of $L(C^i)$ according to the WLLN, using (1d). Hats are used to denote simulation-based estimators, with the subscript showing the size of the sample. Note that the unbiasedness of $\hat{L}_n$ will be an important feature of section 4. These estimators are used to construct the estimator

$$\hat{\mu}_{mn} := (1/m) \sum_{i=1}^{m} \hat{L}_n(C^i) \tag{6}$$

which is unbiased, and a consistent estimator of p according to the WLLN, using (1c). Finally, these are combined to construct the estimator

$$\hat{E}_{mn} := \sum_{i=1}^{m} h(C^i)\, \hat{W}^i \tag{7}$$

where now $\hat{W}^i := (1/m\hat{\mu}_{mn})\hat{L}_n(C^i)$. This should be compared with (4), to see the way in which approximate estimated quantities have been substituted for the unkowns. $\hat{E}_{mn}$ is a biased estimator of $E[h(C)\,|\,z^{\mathrm{m}}]$, but still a consistent estimator according to the WLLN, plus additional results on probabilistic convergence. The bias comes from taking the ratio in which there are stochastic quantities in the denominator—an instance of the result that $E(1/X) \neq 1/\,E(X)$ if X is not certain.

Note that the same nested sets of samples can be used for different choices of h. Therefore the collection $\{C^{1:m}, \hat{W}^{1:m}\}$ stands in for the probability density function $\pi(c\,|\,z^{\mathrm{m}})$, for all inferences that we desire to make about C. The informal description of probabilistic inference in SUPR corresponds to sampling $C^{1:m}$ and then, through sampling $Y_i^{1:n}$ for each C^i, computing $\hat{W}^{1:m}$. Note that $\hat{\mu}_{mn}$ does not have to be computed explicitly, because $\hat{W}^i \propto \hat{L}_n(C^i)$, such that $\sum_{i=1}^{m} W^i = 1$. In other words, once the $\hat{L}_n(C^i)$ values have been computed, they are turned into weights by scaling them to sum to one.

An introduction to Importance Sampling is given in, e.g., Davison (2003, sec. 11.3.2). There are many refinements; see, e.g., Evans and Swartz (2000), Liu (2001), and Robert and Casella (2004). Improvements in efficiency are possible if $\pi(c)$ can be evaluated pointwise up to a constant, rather than simply sampled from. In this case the random sample for C can be from a proposal $q(c)$, which is tuned to $\pi(c\,|\,z^{\mathrm{m}})$ in some way. Here we note the obvious drawback of Importance Sampling, which is that the result is only an approximation, and may not be a very good one if m or n is small. Happily, recent developments in Markov chain Monte Carlo sampling offer a partial solution.

4 Markov chain sampling

Markov chain Monte Carlo (MCMC) has as its objective the construction of a Markov chain for which the stationary distribution is the target distribution one wants to sample from. In our case the target distribution is

$$\pi(c\,|\,z^{\mathrm{m}}) \propto L(c)\,\pi(c)$$

from (1a), where the constant of proportionality $(1/\mu)$ has been dropped. For the moment, suppose that both $L(c)$ and $\pi(c)$ can be evaluated pointwise up to a constant. The most general MCMC implementation, the Metropolis-Hastings algorithm, also requires a proposal distribution for moving around in $\mathcal{C}$, denoted $q(c'\,|\,c)$, which we can both sample from and evaluate pointwise up to a constant.

Consider the situation at iteration t, where t indexes the Markov chain—this is conventional notation. There may well be time-indexed quantities in C, but these

are suppressed at our level of abstraction. Instead, t here corresponds roughly to computing time. At iteration t, the Markov chain has value C_t. Then the Metropolis-Hastings algorithm has three steps:

1. Sample $C' \sim q(c' \mid C_t)$;
2. Compute
$$\alpha_t := \frac{L(C')\,\pi(C')}{L(C_t)\,\pi(C_t)} \times \frac{q(C_t \mid C')}{q(C' \mid C_t)}; \tag{8}$$
3. Sample $U_t \sim \text{Uniform}(0,1)$, and then
$$\text{if } \begin{cases} U_t \leq \alpha_t & \text{set } C_{t+1} = C', \\ U_t > \alpha_t & \text{set } C_{t+1} = C_t. \end{cases}$$

Note that by taking ratios, all unknown constants cancel. In general, the target distribution for the Metropolis-Hastings algorithm can be identified as that part of the numerator of the acceptance ratio α_t which is not the 'reverse' of the proposal. This characterisation will be used below.

This Metropolis-Hastings algorithm requires $\pi(c)$ to be evaluated pointwise up to a constant in order to compute α_t. Where this is not possible, i.e. where we can only sample from $\pi(c)$, then the simplest adjustment is to set the proposal $q(c' \mid c)$ equal to $\pi(c')$, in which case the $\pi(c)$ terms in the target cancel with the proposal, and $\alpha_t = L(C')/L(C_t)$. However, for clarity in what follows, the distinction between $q(c' \mid c)$ and $\pi(c')$ will be preserved.

Typically, one chooses a reasonable starting-point for C_0, runs the chain for a while, and then, after checking that it has attained its stationary distribution in some fashion, takes a subset of values from the chain as an approximately random sample from $\pi(c \mid z^{\text{m}})$. There has been a huge amount written about MCMC in the last twenty years. For an introduction, see, e.g., Davison (2003, sec. 11.3.3). For more detail see, e.g., Robert and Casella (2004), or Gelman *et al.* (2003).

The reason for proposing MCMC as an alternative to Importance Sampling is the surprising but gratifying fact that it is possible to replace the true but unknown value $L(c)$ with the estimator $\hat{L}_n(c)$ in (8) *and still end up sampling from* $\pi(c \mid z^m)$, after an appropriate modification. In other words, the stationary distribution of the modified chain is the correct $\pi(c \mid z^{\text{m}})$, even though the value used for $L(c)$ is only an approximation, a result which holds for all $n \geq 1$. This is *not* the case with Importance Sampling, where if n is finite then the resulting expectation is only approximate. This result was first noted by Beaumont (2003), and developed by Andrieu and Roberts (2009) and Andrieu *et al.* (2010). The next two subsections re-present some of the results from Andrieu *et al.* (2010), because this gripping but highly technical paper is expressed at a greater level of generality than required here. Subsection 4.3 relaxes the no-feedback assumption.

4.1 Exact approximation

Andrieu *et al.* (2010) refer to their algorithms as 'exact approximations', in the sense that although $\hat{L}_n(c)$ is an approximation for $L(c)$, the stationary distribution of the Markov chain is exactly $\pi(c \,|\, z^{\mathrm{m}})$. We write $\boldsymbol{Y} := Y^{1:n}$ for compactness and clarity, and $\hat{L}_n(c)$ as $\hat{L}_n(c, \boldsymbol{Y})$ to emphasise the role of $\boldsymbol{Y}$, as expressed in (5). The crucial property is that $\hat{L}_n(c, \boldsymbol{Y})$ is an unbiased positive estimator of $L(c)$, written as

$$E[\hat{L}_n(c, \boldsymbol{Y}) \mid c] = L(c)$$

for all $c \in \mathcal{C}$.

Consider the following Metropolis-Hastings algorithm, defined on the extended set of variables $\{C, \boldsymbol{Y}\}$:

1. Sample $C' \sim q(c' \mid C_t)$, and then sample $\boldsymbol{Y}' \overset{\text{iid}}{\sim} \pi(y \mid C')$;
2. Compute
$$\alpha_t := \frac{\hat{L}_n(C', \boldsymbol{Y}')\,\pi(C')}{\hat{L}_n(C_t, \boldsymbol{Y}_t)\,\pi(C_t)} \times \frac{q(C_t \mid C')}{q(C' \mid C_t)}; \tag{9}$$
3. Sample $U_t \sim \text{Uniform}(0, 1)$, and then
$$\text{if } \begin{cases} U_t \leq \alpha_t & \text{set } C_{t+1} = C' \text{ and } \boldsymbol{Y}_{t+1} = \boldsymbol{Y}', \\ U_t > \alpha_t & \text{set } C_{t+1} = C_t \text{ and } \boldsymbol{Y}_{t+1} = \boldsymbol{Y}_t. \end{cases}$$

This is the same algorithm as before, except that $\hat{L}_n$ is used instead of L, and $\boldsymbol{Y}$ is propagated in the Markov chain, as well as C. Of course $\boldsymbol{Y}_t$ can be discarded once $\boldsymbol{Y}_{t+1}$ has been sampled (for now: see section 4.2).

It is now easily proved that the c-margin of the stationary distribution of this Markov chain is $\pi(c \mid z^{\mathrm{m}})$, which is our target distribution. First, note that the proposal on the extended set of variables is

$$q(c' \mid c_t)\,\pi(\boldsymbol{y}' \mid c'),$$

writing $\pi(\boldsymbol{y} \mid c) := \prod_{i=1}^{n} \pi(y^i \mid c)$. The acceptance ratio (9) can be written

$$\alpha_t := \frac{\hat{L}_n(C', \boldsymbol{Y}')\,\pi(C')\,\pi(\boldsymbol{Y}' \mid C')}{\hat{L}_n(C_t, \boldsymbol{Y}_t)\,\pi(C_t)\,\pi(\boldsymbol{Y}_t \mid C_t)} \times \frac{q(C_t \mid C')\,\pi(\boldsymbol{Y}_t \mid C_t)}{q(C' \mid C_t)\,\pi(\boldsymbol{Y}' \mid C')} \tag{10}$$

which has simply introduced a factor of one. But referring back to the original Metropolis-Hastings algorithm, notably (8) and the subsequent comment, this shows that the target distribution on the extended space is proportional to

$$\hat{L}_n(c, \boldsymbol{y})\,\pi(c)\,\pi(\boldsymbol{y} \mid c),$$

the non-proposal part of the numerator α_t in (10). Integrating over $\boldsymbol{y}$ to find the c-margin of the target distribution:

$$\int \cdots \int_{y^n} \hat{L}_n(c, \boldsymbol{y})\,\pi(c)\,\pi(\boldsymbol{y} \mid c)\,\mathrm{d}\boldsymbol{y} = \pi(c)\,E[\hat{L}_n(c, \boldsymbol{Y}) \mid c] = \pi(c)\,L(c)$$

for all $c \in \mathcal{C}$, according to the unbiasedness of $\hat{L}_n$. This is proportional to $\pi(c \,|\, z^{\mathrm{m}})$, by Bayes's Theorem, as was to be shown.

As noted above, in practice we replace $q(c' \,|\, c)$ with $\pi(c')$ in the algorithm if we can only sample from $\pi(c)$. In this case $\alpha_t = \hat{L}_n(C', \boldsymbol{Y}')/\hat{L}_n(C_t, \boldsymbol{Y}_t)$ in (9).

4.2 Sampling the sensor as well

Andrieu *et al.* (2010) show that it is *also* possible to adapt the Metropolis-Hastings algorithm so that the target distribution is exactly $\pi(c, y \,|\, z^{\mathrm{m}})$. In other words, we can make inferences simultaneously about both the climate and the sensor, based on the proxy measurements. The proof is an extension of that in section 4.1.

The set of variables is extended again by one additional variable, $K \in \{1, \ldots, n\}$. This K will then be used to select one value from $\boldsymbol{Y}$, as explained after the algorithm:

1. Sample $C' \sim q(c' \,|\, C_t)$, and then sample $\boldsymbol{Y}' \overset{\text{iid}}{\sim} \pi(y \,|\, C')$;
2. Sample K from the set $\{1, \ldots, n\}$ according to probabilities
$$\frac{\pi(z^{\mathrm{m}} \,|\, (Y^j)', C')}{n\hat{L}_n(C', \boldsymbol{Y}')} \qquad j = 1, \ldots, n.$$
3. Compute
$$\alpha_t := \frac{\hat{L}_n(C', \boldsymbol{Y}')\,\pi(C')}{\hat{L}_n(C_t, \boldsymbol{Y}_t)\,\pi(C_t)} \times \frac{q(C_t \,|\, C')}{q(C' \,|\, C_t)}; \tag{11}$$
4. Sample $U_t \sim \mathrm{Uniform}(0, 1)$, and then
$$\text{if } \begin{cases} U_t \le \alpha_t & \text{set } C_{t+1} = C',\ K_{t+1} = K, \text{ and } \boldsymbol{Y}_{t+1} = \boldsymbol{Y}', \\ U_t > \alpha_t & \text{set } C_{t+1} = C_t,\ K_{t+1} = K_t, \text{ and } \boldsymbol{Y}_{t+1} = \boldsymbol{Y}_t. \end{cases}$$

This is the same algorithm as in section 4.1, except now a K_t is sampled, and also saved at each step of the chain. As before, $\alpha_t = \hat{L}_n(C', \boldsymbol{Y}')/\hat{L}_n(C_t, \boldsymbol{Y}_t)$ if $q(c' \,|\, c)$ is replaced with $\pi(c')$.

It is now easily-proved that the (c, y^k)-margin of the stationary distribution of the Markov chain is $\pi(c, y \,|\, z^{\mathrm{m}})$. Just as before, the full proposal over the extended set of variables is found to be

$$q(c' \,|\, c)\,\pi(\boldsymbol{y}' \,|\, c')\,\frac{\pi(z^{\mathrm{m}} \,|\, (y^k)', c')}{n\hat{L}(c', \boldsymbol{y}')},$$

where the final term is the probability of selecting $K = k$. Again, as before, this shows that the stationary distribution is proportional to

$$\hat{L}_n(c, \boldsymbol{y})\,\pi(c)\,\pi(\boldsymbol{y} \,|\, c)\,\frac{\pi(z^{\mathrm{m}} \,|\, y^k, c)}{n\hat{L}_n(c, \boldsymbol{y})} \propto \pi(c)\,\pi(\boldsymbol{y} \,|\, c)\,\pi(z^{\mathrm{m}} \,|\, y^k, c).$$

Integrating out all $\boldsymbol{y}$ except y^k gives

$$\pi(c)\,\pi(y^k \mid c)\,\pi(z^{\mathrm{m}} \mid y^k, c) = \pi(z^{\mathrm{m}} \mid y^k, c)\,\pi(y^k, c).$$

This is proportional to $\pi(c, y^k \mid z^{\mathrm{m}})$ by Bayes's Theorem, as was to be shown.

Unlike $\boldsymbol{Y}'$, K is not needed for the progression of the chain, but only for the selection of the chain value for Y. In practice, at each iteration of the Markov chain, if the chain moves, then K is sampled, and Y_{t+1} is recorded as the Kth element of $\boldsymbol{Y}_{t+1}$. It is very intuitive that K is sampled according to probabilities that are proportional to $\pi(z^{\mathrm{m}}|y^j, c)$, because these probabilities are larger for those y^j that provide a better explanation for the measurements z^{m}.

Why not set n = 1? The exact approximation works even in the case $n = 1$. In this case a new pair $\{C, Y\}$ is sampled at each step. In some ways this seems natural—one wants to make an inference about $\{C, Y\}$, and so one samples them repeatedly, every new candidate for climate getting its own candidate for the sensor. In the case of Importance Sampling this is clearly a bad idea. The Importance Sampling estimator is biased but consistent, which indicates that larger n means smaller bias. Bias in the estimator does not translate automatically into inaccuracy in the estimate, but a simple way to reduce bias, such as increasing n, would be foolish to reject, if resources allowed.

The MCMC approach has an analogous weakness. $\hat{L}_n(c)$ is an estimator for $L(c)$, and when n is small this estimator has a large standard error, which is to say that $\hat{L}_n(c)$ can be substantially smaller or larger than $L(c)$. Occasionally, then, the estimate will substantially overestimate. From the mechanics of MCMC, this means that the chain will get stuck, because proposals away from this point, even though they might be in the direction of increasing probability, cannot offset the 'luck' of the current point in the chain. In the end, of course, the chain will move, because it is just a matter of time before a really good proposal or an even luckier estimate comes along. But this chain will move very slowly—in the jargon it will be sticky. Therefor huge numbers of iterations will be necessary to ensure that it has attained the target distribution, and, once there, huge numbers of iterations will be required to draw an approximately independent sample.

The real benefit of exact approximation is that it allows us to treat n as a tuning parameter. In general, MCMC is highly tunable, through the effectively limitless range of possibilities for the proposal q. Selecting q can even be automated, and allowed to adapt to the chain's behaviour (see, e.g., Andrieu and Thoms, 2008). But where we can sample from $\pi(c)$ but not evaluate it pointwise up to a constant, selecting q is not an option: it must be set to $\pi(c)$. In this situation, n becomes the main tuning parameter. The analyst knows that all choices of n are valid, and can therefore choose one which balances the stickiness of the chain against the computing resources available. Therefore control over n increases the accuracy of the inference through improving its efficiency. A choice such as $n = 1$, while not invalid, is likely to be highly inefficient, and, consequently, when resources are limited, highly inaccurate as well.

4.3 Incorporating feedback

We have assumed that there is no feedback from the sensor to climate, in order that the two-stage sampling strategy $\pi(c, y) = \pi(c)\,\pi(y \mid c)$ is well-specified. Sometimes it will be necessary to incorporate feedback from the sensor to climate, in which case it makes no sense to first simulate climate, and then simulate the sensor given climate. We now suppose that we can sample directly from $\pi(c, y)$. How does this change the inference? In principle, hardly at all. All that is needed is a 'rebadging', where C becomes $\{C, Y\}$ and Y is suppressed. However, this is now an $n = 1$ calculation, and has the potential to be inaccurate, as described immediately above. Possibly it is better to replace feedback from the sensor to climate with an additional source of uncertainty, and preserve the computationally more attractive simulation of $\pi(c)$ and then $\pi(y \mid c)$.

5 Summary

Whatever method we choose, Bayesian inference about palaeoclimate requires an integration jointly over climate-space and sensor-space. Monte Carlo methods offer the best value in terms of programming effort and probability of coding error, given the power of modern computers. Both Importance Sampling and MCMC are open-ended Mote Carlo methods. One can continue to add to the simulation until resources are exhausted, or until diagnostics indicate that the results have stabilised. If the climate probability density function can be evaluated pointwise up to a constant, then both methods can be tuned, e.g. based on a pilot study, to improve their efficiency. This pilot study would usually be on a reduced set of measurements, for greater speed. Among statisticians, random walk Metropolis-Hastings is usually favoured over non-sequential methods such as Importance Sampling, due to the tendency for the chain to head automatically for a region of concentration of probability, and due to the ease with which the proposal distribution can be implemented, tuned, and adapted on-line.

This Addendum has shown that there is an added advantage to Metropolis-Hastings over Importance Sampling when large numbers of candidate climates can be simulated, which is that long chains can compensate for small numbers of sensor simulations for each candidate climate. Put differently, the analyst has the option of selecting an appropriate number of sensor simulations at each step, confident in the knowledge that the stationary distribution is preserved, although the chain's mixing will be affected. Furthermore, Metropolis-Hastings can be adapted at no additional cost to perform inferences jointly on both climate and the sensor. By-and-large, if an Importance Sampling approach with large numbers of candidate climates is feasible, then one would imagine that a Metropolis-Hastings approach along the lines suggested in section 4.2 would be better.

Finally, though, a word of caution. Anyone who has attempted these types of calculations will have a horror story about how long it took the Markov chain to move into an area of high concentration of probability, or, to put it another way, how sticky the Markov chain got in the tails. A well-specified Markov chain has

to visit the tails occasionally, and so this is bound to happen sooner or later in the simulation. Usually, it happens at the start, which tends to be at a climate which has low posterior probability density. While the MCMC approach described here should work better than Importance Sampling, especially given the opportunity to tune the latter by the choice of the number of sensor simulations per climate, one cannot assert with confidence that either of them will work efficiently. There is, however, a simple solution, which is to increase the amount of computing power available. For each simulated climate, the simulations of the sensor are independent, and can be farmed out across a computer cluster.

It is important to appreciate that the inference outlined here is not a hard calculation, just a long one. The hard part—also the expensive part—is collecting the proxy measurements, and constructing the climate simulator and the sensor simulator. After that, the statistical analysis for the reconstruction of climate on the basis of the measurements and these simulators is somewhat routine, if it follows the outline given here. Having said that, my personal experience of making mistakes in routine calculations suggests that the involvement of a statistician would be recommended at this stage. On balance, though, it would seem more critical to involve a statistician at the previous stages where the stochastic description of climate and of the sensor are developed, and where the experiment to collect proxy measurements is designed.

Acknowledgements

My substantial debt to my colleague Prof. Christophe Andrieu should be obvious to all readers; this Addendum has also benefited from the suggestions of Joe Cainey. I would also like to thank the members of the SUPRAnet project, headed by Profs Caitlin Buck and John Haslett, for many fruitful discussions. And the Isaac Newton Institute, Cambridge, for hosting a programme on the mathematics and statistics of climate (CLP), August to December 2010.

References

C. Andrieu, A. Doucet, and R. Holenstein, 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **72**(3), 269–342. With discussion.

C. Andrieu and G.O. Roberts, 2009. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, **37**(2), 697–725.

C. Andrieu and J. Thoms, 2008. A tutorial on adaptive MCMC. *Statistics and Computing*, **18**, 343–373.

M.A. Beaumont, 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

M.A. Beaumont, W. Zhang, and D.J. Balding, 2002. Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

A.C. Davison, 2003. *Statistical Models*. Cambridge: Cambridge University Press.

M. Evans and T. Swartz, 2000. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, 2003. *Bayesian Data Analysis*. Chapman & Hall, 2nd edition.

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. Chichester, England: John Wiley & Sons.

J.S. Liu, 2001. *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

J. Pearl, 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

C.P. Robert and G. Casella, 2004. *Monte Carlo Statistical Methods*. New York: Springer, 2nd edition.

J.C. Rougier, 2006. Comment on the paper by Haslett *et al. Journal of the Royal Statistical Society, Series A*, **169**(3), 432–433.

J.C. Rougier, 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264.

SUPRAnet, 2011. Studying uncertainty in palaeoclimate reconstruction: A framework for research. In preparation.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M.P.H. Stumpf, 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.