Nested sampling for Bayesian computations: A discussion

Nicolas Chopin and Christian P. Robert
University of Bristol, Université Paris Dauphine and CREST-INSEE

June 21, 2006

The approximation of marginal densities is central to the Bayesian approach to testing of hypotheses since ratios $m_1(x)/m_2(x)$ of those marginals are providing Bayes factors. It is thus of interest to see the emergence of a novel proposal for the approximative computation, although we are less confident than the author about the applicability of nested sampling in realistic Bayesian problems.

Rewriting Z as an integral over $[0,1]$

A first difficulty stems from the convoluted presentation of the equivalence between Z in eqn (1) and z in eqn (4). We do not see why a discretisation and ordering would be necessary at this stage. Indeed,

$$Z = \mathbb{E}^{\pi}[L(\theta)] = \mathbb{E}^{\tilde{\pi}}[L] = \int_0^{\infty} X(\lambda)\,\mathrm{d}\lambda$$

where $\tilde{\pi}$ denotes the distribution of $L(\theta)$, associated with the cdf $(1 - X(\lambda))$. So it is only under the minimal restriction that X is strictly decreasing that we have the representation of eqn (4), since

$$Z = \int_0^{L^{\max}} \ell\,\mathrm{d}X(\ell) = \int_0^1 X^{-1}(x)\,\mathrm{d}x\,.$$

We may add at this point that the use of the same notation for the likelihood L and the inverse of the complementary cdf $X(\lambda)$ is unnecessarily confusing. (Another difficulty that is not alluded to in the paper but that we cannot discuss here is the case of an unbounded likelihood, as for instance in the simple case of a two component Gaussian mixture.)

This representation, while not novel, is quite interesting because it looks at a familiar quantity from an unusual perspective.

Constrained sampling

However, this representation involves the implicit function $X(\lambda)$, which is approximated quite crudely in the remainder of the paper, and the corresponding algorithm requires in addition a constrained sampling that is certainly not *"easier than the traditional Metropolis–Hastings sampling involving likelihood-weighting and detailed-balance"*. We also note that the corpus of work on bridge and path sampling for Bayes factor approximation [Chen et al., 2000] is omitted, as is the possibility of using reversible jump techniques [Green, 1995] to explore simultaneously a series of models.

The second point is that eqn (4) is rather useless in justifying the corresponding nested sampling algorithm since it is simply based on the Riemann decomposition [Robert and Casella, 2004, Chapter

2]

$$Z = \int_0^{L^{\max}} \ell \, \mathrm{d}X(\ell) = \sum_{i=1}^{m} \int_{L_{i-1}}^{L_i} \ell \, \mathrm{d}X(\ell) \approx \sum_{i=1}^{m} L_i (X_{i-1} - X_i)$$

that works no matter what the values L_i are, as long as the differences $(X_{i-1} - X_i)$ all converge to 0 with m going to infinity (which happens to not be the case here, but hopefully the first differences should have a extremely small contribution in terms of L-values.). The choice made in the paper of simulating the θ_i's from $\pi(\theta)$ is thus pertinent but not necessarily optimal. In particular, vague priors are likely to induce a lack of efficiency when the likelihood function is quite concentrated within the support of π. (In importance sampling, simulation from the prior is only done to ensure finiteness of the variance, as in defensive sampling, Robert and Casella, 2004, Chapter 3). And, obviously, it is impossible to simulate from an improper prior. We note however this point is made implicitly in §3.2, where the author suggests to substitute π with a better suited 'base' distribution.

From an algorithmic point of view, the experience of perfect sampling [Mira et al., 2001] shows that simulating from $\pi(\theta)$ restricted under the constraint $L(\theta) \geq L(\theta_i)$ is not always possible. In large dimension spaces, simulating from the prior till the constraint is satisfied is unrealistic: simple calculations involving say Gaussian distributions are enough to show that this becomes exponentially difficult in the dimension of the problem. In that respect, Fig. 1 and Fig. 2 (right side) are quite optimistic, as we would rather expect a very sharp peak on the left, which would fall to zero almost immediately. More fundamentally, sampling from MCMC offers no clear justification for a finite number of simulations. (A single iteration clearly does not work since the chain may remain at the same place and thus repeat the value of $L(\theta_i)$.) Adding a single value to the sample of N points at each iteration of the algorithm also seems quite inefficient, compared with a new generation of a constrained sample of N points, because the chances of getting high values of $L(\theta)$ are then necessarily much lower. We also note that the multimodality issue is not treated in a completely satisfactory manner: if N is too small, the initial sample may miss a narrow but primary mode and it is then quite uncertain whether or not this mode can be recovered at a later stage.

Approximating the X_i's

Our main concern however is with the rather fleeting description of how the X_i's, the X-values attached to the simulated θ_i's, are obtained in §2.1. Sampling X_i (independently?) does not seem to make sense, while approximating it by $\exp(-i/N)$ is rather crude, unless N is very large. And even in that case, to establish if, how quick, and in which sense our approximated integral does converge, in N but also in m, is far from obvious. This is especially true if we follow the author's suggestion to recycle the $N-1$ simulated X_i's that remain after deleting the lowest value.

We propose an alternative description of this particular aspect of the algorithm, which we hope will clarify things: in an initial version, a pair (X_i, L_i) is simulated with respect to an appropriate distribution (constrained prior), in a way that ensures that $L_i = X^{-1}(X_i)$. To improve the algorithm through a form of Rao-Blackwellisation, we replace $\log(X_i)$ by its expectation $-i/N$, but keep L_i as random. This second algorithm is correct provided either (a) the simulated L_i has expectation $\mathbb{E}[L_i] = X^{-1}(\exp(-i/N))$; or (b) it converge to this value in some sense. Condition (a) allows for unbiased estimation (thanks to the linearity of the approximated integral), but should be met only if $X^{-1}(\exp(\cdot))$ is linear, a constrain that never holds. Condition (b) is more reasonable, but should make it difficult to establish convergence results, as the *joint* convergence of all L_i (along iterations) would have to be established.

Moving along X axis at a geometric rate

Because of the curse of dimensionality mentioned above, even geometric steps (along the X-dimension) may not be fast enough to reach the likelihood mode in a reasonable number of iterations. This problem seems to be aggravated by the fact that, in the version of the algorithm with N points, the ratio X_i/X_{i-1} is then the largest of N uniform variates, and therefore should be close to one. This is yet another complication for proving any form of convergence.

More examples

Rather than the toy problems exposed in the paper (which we reproduced to convince ourselves), it would have been nice to have the method illustrated by realistic statistical examples.

References

M.H. Chen, Q.M. Shao, and J.G. Ibrahim. *Monte Carlo Methods in Bayesian Computation.* Springer-Verlag, New York, 2000.

P.J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

A. Mira, J. Møller, and G.O. Roberts. Perfect slice samplers. *J. Royal Statist. Soc. Series B*, 63: 583–606, 2001.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer-Verlag, New York, second edition, 2004.