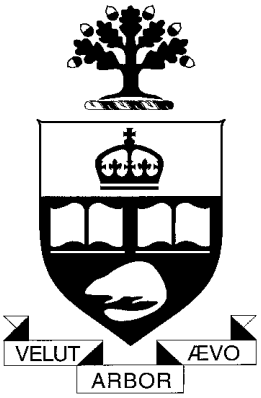

Discussion of Nested Sampling for Bayesian Computations by John Skilling

by

Michael J. Evans
Department of Statistics
University of Toronto

Technical Report No. 0608 June 1, 2006

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Discussion of Nested Sampling for Bayesian Computations by John Skilling

Michael J. Evans
University of Toronto, Canada
mevans@utstat.utoronto.ca

Summary

We comment on several aspects of Skilling's paper presented at Valencia 8. In particular we prove the convergence in probability of the algorithm for a wide class of situations, comment on its potential utility and discuss aspects where further work is needed to assess the approach.

Keywords and Phrases: Bayesian model checking, Integration; quadrature; convergence in probability.

1. INTRODUCTION

The paper contains a number of interesting contributions. We have chosen to comment on two aspects of the paper, namely, the statements concerning the necessity of computing the prior predictive density (the evidence) at the observed data prior to implementing a posterior analysis, and the integration algorithm presented in the paper and referred to as nested sampling.

2. WHAT IS A BAYESIAN ANALYSIS?

The ingredients to a Bayesian analysis comprise the sampling model $\{P_\theta : \theta \in \Omega\}$ for the response X, the prior Π for the parameter θ, and the observed value $X = x$. This is equivalent to specifying the joint model $P_\theta \times \Pi$ for (X, θ) and the observed value $X = x$. In many situations we may also have a loss function but we ignore this here as it is not material to our discussion. We refer to $P_\theta \times \Pi$ as the Bayesian model. The Bayesian model and the data comprise the full information available and we ask how this information is to be used in carrying out a statistical analysis.

The joint model can be factored as $P_\theta \times \Pi = M \times \Pi(\cdot \mid X)$ where M is the marginal prior predictive for X and $\Pi(\cdot \mid X)$ is the posterior for θ. The principle of conditional probability then says that probability statements about θ, that are initially based on the prior Π, should instead be based on the observed posterior $\Pi(\cdot \mid X = x)$. What then, is the role of M in a statistical analysis?

If our goal is inference about θ, then it might seem that we can ignore M and proceed directly to work with $\Pi(\cdot \mid X = x)$. This may seem even preferable if it is

Michael Evans is a Professor of Statistics at the University of Toronto.

possible to compute or sample from $\Pi(\cdot \mid X = x)$ without making any direct reference to M. In fact this is a feature of the MCMC algorithms that have revolutionized Bayesian computing over the past few years. While this may have some appeal, it still leaves one wondering about the role that M might or should have.

A more direct approach to obtaining $\Pi(\cdot \mid X = x)$ proceeds as follows. If we denote the densities of P_θ and Π by f_θ and π respectively, then the density of M at x is given by $m(x) = E_\Pi(f_\theta(x))$ and the density of $\Pi(\cdot \mid X = x)$ is $\pi(\theta \mid x) = f_\theta(x)\pi(\theta)/m(x)$. We can then work with $\pi(\theta \mid x)$ to compute various posterior characteristics. This approach makes some reference to M but only through the value of its density at x and we are still left wondering about any role for the full distribution M in the analysis.

Dr. Skilling makes the point that $m(x)$ (denoted as Z and referred to in the paper as *the evidence*) is something of great importance and perhaps even comes logically before the computation of the posterior. The suggestion is made that this value "assesses the model in question" and, if it leads to doubts about the validity of the model, then there is no real need to proceed to the computation of the posterior.

The paper doesn't make clear how one is supposed to use $m(x)$ in this process, but the suggestion is made that, if we had two Bayesian models, leading to $m_1(x)$ and $m_2(x)$ respectively, then we would decide between them using the ratio $m_1(x)/m_2(x)$. This is the Bayes factor in favour of the first Bayesian model and equals $m_1(x)/m_2(x) = (1-p_1)p_{1|x}/p_1(1-p_{1|x})$ where p_1 is the prior probability of the first Bayesian model and $p_{1|x}$ is the posterior probability of this model. When more than two models are being considered this ratio is a similar function of the conditional prior and posterior probabilities of the first model given that only these two models are possible.

A point that can be made here, is that calibrating the value of $m_1(x)/m_2(x)$, i.e., saying when this quantity is small or large, depends intrinsically on the assignments of the prior probabilities. When these are known, then the size of the Bayes factor, i.e., is it strong evidence in favour of the first model or otherwise, is interpreted in terms of the posterior probabilities. It may be argued that the Bayes factor does not formally require the specification of the prior probabilities for the models and that this is a virtue of this approach. But then we think it is reasonable to question how we should interpret the value of $m_1(x)/m_2(x)$. So one could argue that in model selection problems the issues are a bit more involved than just the computation of the $m_i(x)$ values. Of course, the computation of the ratios of these values is absolutely necessary.

More generally, the value of $m(x)$ is involved in model criticism. If the value of $m(x)$ is a surprising value, then we have evidence that the Bayesian model is incorrect in some sense. In such circumstances, proceeding to inferences about θ seems at least of questionable validity. It seems essential that checks be made on the ingredients put into an analysis to ensure that these make sense in light of the data obtained. In Box (1980) it was proposed that $m(x)$ be compared with its distribution induced by M to determine if $m(x)$ is surprising and, if it was, conclude that we had evidence against the validity of the Bayesian model.

Actually there are two ways in which a Bayesian model can fail. The sampling model can fail by the data being surprising for every distribution in $\{P_\theta : \theta \in \Omega\}$ or, if the sampling model is correct, the prior may place its mass primarily on values of θ for which the data are surprising. We refer to this second failure as prior-data conflict. The consequences of these two failures are somewhat different. With large amounts of data prior-data conflict can be ignored, as the data swamp the prior,

while no amount of data can fix a bad sampling model. Accordingly, as argued in Evans and Moshonov (2006a), it seems sensible to check for these errors separately. First we check the sampling model, if no problems are detected we then check for prior-data conflict, and then depending on this assessment, possibly proceed to inference about θ.

In Evans and Moshonov (2006a) it is argued that the appropriate approach for checking for prior-data conflict is based on comparing the observed value $T(x)$, of a minimal sufficient statistic T, with its conditional prior-predictive distribution $M_T(\cdot \mid U(T(x))$ given an ancillary $U(T)$. This leads to the factorization $M = P(\cdot \mid T(x)) \times P_{U(T)} \times M_T(\cdot \mid U(T(x))$ where $P(\cdot \mid T(x))$ is the conditional distribution of the data given $T(x)$ and $P_{U(T)}$ is the marginal distribution of $U(T(X))$. Both $P(\cdot \mid T(x))$ and $P_{U(T)}$ are independent of the prior and so are available for checking the sampling model. Interestingly the lack of a unique maximal ancillary does not pose a problem in this context as we simply get different checks with different maximal ancillaries.

We see that this approach leads to using the full information available to the statistician in carrying out a Bayesian analysis. Each step in the analysis is based on a component of a factorization of the joint model. A further factorization of $M_T(\cdot \mid U(T(x))$ is sometimes available, as discussed in Evans and Moshonov (2006b), when we want to check for prior-data conflict with specific components of the prior. We also note that this approach avoids the double use of the data problems inherent in posterior model checks, as discussed in Bayarri and Berger (2000).

So we are in agreement with Dr. Skilling's assertion concerning the importance of looking at Z before proceeding to inference about θ and that logically this step cannot be avoided. As just discussed, however, we would go much further than just computing Z.

3. NESTED SAMPLING

Nested sampling appears to be a new integration method that is applicable to the evaluation of any integral. We will first focus on its application to the evaluation of $Z = E_{\Pi}(f_\theta(x))$, as is done in the paper, although there does not appear to be anything that makes this problem, as opposed to a general integration problem, particularly suitable for nested sampling. We modify Dr. Skilling's presentation slightly to bring the notation and terminology more in line with what is customary in probability and statistics. One simple approach to estimating Z is to generate a sample $\theta_1, \ldots, \theta_N$ from the prior Π and compute $N^{-1} \sum_{i=1}^{N} f_{\theta_i}(x)$. We refer to this as naive importance sampling and note that it is widely recognized as being inadequate for this problem.

To develop nested sampling we put $\psi = \Psi(l) = \Pi(f_\theta(x) > l)$ so that Ψ is the complementary cdf of the random variable $f_\theta(x)$ when $\theta \sim \Pi$ and x is fixed. Now let $\Psi^{-1}(\psi) = \sup\{l : \Psi(l) > \psi\}$. Then, if $\psi \sim U(0,1)$, we have that $\Psi^{-1}(\psi) \sim f_\theta(x)$ and note that this is true whether the distribution of $f_\theta(x)$ is discrete or continuous. Therefore we can write

$$Z = \int_0^1 \Psi^{-1}(\psi)\, d\psi \tag{1}$$

and note that the integrand Ψ^{-1} decreasing. We believe that the representation (1) summarizes what the paper means by "sorting the likelihood".

Since the integral in (1) is 1-dimensional, we can approximate this by a Riemann sum $Z \approx \sum_{i=1}^{m} \Psi^{-1}(\psi_i)\,(\psi_i - \psi_{i+1})$ where $0 = \psi_{m+1} < \psi_m < \cdots < \psi_0 = 1$. The problem with this approximation is the need to calculate $\Psi^{-1}(\psi_i)$, which is the $(1-\psi_i)$-th quantile of the distribution of $f_\theta(x)$ when $\theta \sim \Pi$. Obviously, evaluating this is generally at least as hard as evaluating the original integral.

The paper proposes a novel approach that avoids the need to directly evaluate $\Psi^{-1}(\psi_i)$. There are a number of variants discussed in the paper and we focus on one of these but feel our comments apply generally. The first step in this is to suppose that the ψ_i are randomly generated. These quantities are defined iteratively via $\psi_i = t_1 t_2 \cdots t_i$ where t_i is distributed as the largest order statistic in a sample of N from the $U(0,1)$ distribution. It is then easily shown that $E(\psi_i - \psi_{i+1}) \sim e^{-i/N} - e^{-(i+1)/N}$ as $N \to \infty$. Accordingly we could use instead the approximation $Z \approx \sum_{i=1}^{m} \Psi^{-1}(\psi_i)(e^{-i/N} - e^{-(i+1)/N})$. This still seems to require the evaluation of $\Psi^{-1}(\psi_i)$, but now suppose we generate $\theta_{i1}, \ldots, \theta_{iN}$ from the conditional distribution $\theta \,|\, f_\theta(x) > f_{\theta_{i-1}}(x)$ and let $\theta_i \in \{\theta_{i1}, \ldots, \theta_{iN}\}$ be such that $f_{\theta_i}(x) = \min\{f_{\theta_{i1}}(x), \ldots, f_{\theta_{iN}}(x)\}$. We have to be able to implement this sampling but, when we can, then $\Psi^{-1}(\psi_i)$ has the same distribution as $f_{\theta_i}(x)$ and we can approximate (1) via the randomized estimator

$$Z_{m,N} = \sum_{i=1}^{m} f_{\theta_i}(x)\left(e^{-i/N} - e^{-(i+1)/N}\right). \tag{2}$$

It is of course necessary to show that (2) converges to (1). For this we note that we can study the more general problem concerning the convergence of $\sum_{i=1}^{m} g(\psi_i)$ $(\exp\{-i/N\} - \exp\{-(i+1)/N\})$ for a Riemann integrable $g : [0,1] \to R^1$. For the case of a continuous g we have the following result.

Theorem 1 *For continuous $g : [0,1] \to R^1$, and with the ψ_i generated as described, then $\sum_{i=1}^{m} g(\psi_i)(e^{-i/N} - e^{-(i+1)/N}) \to \int_0^1 g(\psi)\, d\psi$ in probability as $N \to \infty$ and $m/N \to \infty$.*

Proof. Let $\epsilon > 0$ and note that we can find a polynomial h such that $|g(\psi) - h(\psi)| < \epsilon$ for all $\psi \in [0,1]$. Then

$$\begin{aligned} & \left|\sum_{i=1}^{m} g(\psi_i)\left(e^{-i/N} - e^{-(i+1)/N}\right) - \sum_{i=1}^{m} h(\psi_i)\left(e^{-i/N} - e^{-(i+1)/N}\right)\right| \\ \leq & \sum_{i=1}^{m} |g(\psi_i) - h(\psi_i)|\left(e^{-i/N} - e^{-(i+1)/N}\right) \leq \epsilon \sum_{i=1}^{m}\left(e^{-i/N} - e^{-(i+1)/N}\right) \\ = & \epsilon\left(e^{-1/N} - e^{-(m+1)/N}\right) \to 0 \end{aligned}$$

as $N \to \infty$ and $m/N \to \infty$. Therefore, we can prove the result for g a polynomial. Further if we prove the result for all powers $g(\psi) = \psi^k$ with $k > 0$, then the result is established.

For $g(\psi) = \psi^k$ with $k > 0$ we proceed as follows. First we have that, under the sampling scheme specified for the ψ_i, then $E(\psi_i^k) = (E(t_1^k))^i = (1 + k/N)^{-i}$. This leads to

$$\begin{aligned} & E\left(\sum_{i=1}^{m} \psi_i^k\left(e^{-i/N} - e^{-(i+1)/N}\right)\right) \\ = & \frac{(1 - e^{-1/N})(1 + k/N)^{-1} e^{-1/N}}{1 - (1 + k/N)^{-1} e^{-1/N}}\left(1 - ((1 + k/N)^{-1} e^{-1/N})^m\right) \end{aligned} \tag{3}$$

and, when $N \to \infty$ and $m/N \to \infty$, this converges to $E(\psi^k) = 1/(k+1)$. Reasoning in a similar fashion we can prove that

$$Var\left(\sum_{i=1}^{m} \psi_i^k \left(e^{-i/N} - e^{-(i+1)/N}\right)\right) \to 0$$

as $N \to \infty$ and $m/N \to \infty$. Then, using Markov's inequality, we have that

$$\begin{aligned} & P\left(\left|\sum_{i=1}^{m} \psi_i^k \left(e^{-i/N} - e^{-(i+1)/N}\right) - 1/(k+1)|\right| > \eta\right) \\ \leq \quad & \eta^{-2}\left\{E\left(\sum_{i=1}^{m} \psi_i^k \left(e^{-i/N} - e^{-(i+1)/N}\right)\right) - 1/\left(k+1\right)\right\}^2 + \\ & \eta^{-2} Var\left(\sum_{i=1}^{m} \psi_i^k \left(e^{-i/N} - e^{-(i+1)/N}\right)\right) \end{aligned}$$

and the right-hand side converges to 0 establishing the result. □

While the proof is for continuous g, it seems that it can be generalized to arbitrary Riemann integrable functions and in particular for step functions. The function Ψ^{-1} will be a step function when $f_\theta(x)$ has a discrete distribution under Π. Applying the theorem when Ψ^{-1} is continuous, establishes that $Z_{m,N} \to Z$ in probability as $N \to \infty$ and $m/N \to \infty$. This also establishes that $\ln Z_{m,N} \to \ln Z$ in probability but doesn't tell us the rate of convergence. For various reasons we may be more interested in $\ln Z$ than in Z. If we can prove that $r_{m,N}(Z_{m,N} - Z) \to V$ in distribution, where V has mean 0 and variance σ^2, then the delta theorem tells us that $r_{m,N}(\ln Z_{m,N} - \ln Z)$ converges in distribution to V/Z and so the rate of convergence does not change by taking the logarithm. The asymptotic coefficient of variation of $Z_{m,N}$ is $\sigma/(r_{m,N} Z)$ while for $\ln Z_{m,N}$ it is $\sigma/(r_{m,N} Z \ln Z)$. Since Z will often be very large, the estimator $\ln Z_{m,N}$ will be a more accurate estimator of $\ln Z$ than $Z_{m,N}$ is of Z. Still, if $Z_{m,N}$ is a terrible estimator of Z, as will be reflected in an extremely large value of σ, then the improvement in accuracy obtained by using $\ln Z_{m,N}$ will be immaterial. To obtain a better understanding of nested sampling, and so determine if it can be used to obtain reliable results, we need to obtain the distribution of V, the rate $r_{m,N}$, and the dependence of σ^2 on Ψ^{-1}.

The theorem requires both $N \to \infty$ and $m/N \to \infty$ and it might be wondered if both conditions are necessary. Consideration of (3) for $g(\psi) = \psi^k$ shows that both of these conditions are necessary even to obtain asymptotic unbiasedness. Restricting nested sampling to decreasing functions, such as Ψ^{-1}, does not remove the need for both conditions, as consideration of the functions $g(\psi) = 1 - \psi^k$ for $k > 1$ shows. There is a variant of nested sampling where N uniformly distributed points are added in the final interval $[0, \exp\{-(m+1)/N\}]$. For this variant, if we fix m and let $N \to \infty$, then we will have $Z_{m,N} \to Z$ in probability. But note that $\exp\{-(m+1)/N\} \to 1$ and so this algorithm is really nothing more than naive importance sampling. When $m/N \to \infty$, then the contribution in $[0, \exp\{-(m+1)/N\}]$ is of no importance as the mass in this interval goes to 0. If we fix N and let $m \to \infty$, then the estimator will be asymptotically biased as (3) shows with $g(\psi) = \psi^k$ for $k > 0$. For example, with $k = 1, N = 1$ the asymptotic bias is 28% of the true value. Basically it seems

that we have to choose N large enough so that the bias is small and then choose m, larger than N, to make the variance small. Replicating a biased estimator and averaging, as the paper seems to suggest, will not get rid of the bias. If the bias is large, nothing is really accomplished by this.

We might also consider studying the related quadrature rule $\sum_{i=1}^{m} g(\exp\{-i/N\})$ $(\exp\{-i/N\} - \exp\{-(i+1)/N\})$. This can also be shown to converge to $\int_0^1 g(x)\,dx$ as $N \to \infty$ and $m/N \to \infty$ for continuous g. We might ask for what functions g is this quadrature rule particularly effective and see if these correspond to the kinds of Ψ^{-1} functions encountered in practice.

As previously noted, nested sampling can be applied to any integration problem. When w is a probability density with respect to a support measure μ, we have that

$$\int_{\mathcal{X}} f(x)\,\mu\,(dx) = \int_{\mathcal{X}} \frac{f(x)}{w(x)} w(x)\,\mu\,(dx) = E_w\left(\frac{f(X)}{w(X)}\right)$$

and we can apply the nested sampling approach to $f(X)/w(X)$ with $X \sim w$. From this point-of-view nested sampling reminds us somewhat of randomized quadrature rules as discussed, for example, in Evans and Swartz (2000). Like randomized quadrature, nested sampling could be seen as a potentially useful variance reduction technique in the context of importance sampling.

In a specific problem it seems difficult to determine how to choose m and N, but perhaps study of the asymptotic distribution and the rate of convergence will provide guidance. Certainly it doesn't seem likely that naive choices of these characteristics will result in successful approximations. It seems plausible to us that the curse of dimensionality, that applies to high-dimensional integration generally, is hidden in the values of m and N that are required to produce useful approximations. Many integration algorithms that seem to avoid the curse of dimensionality really do not. For example, importance sampling and Metropolis-Hastings require the choice of good samplers to be effective and there is no easy way to find these for arbitrary high-dimensional integration problems.

Overall we feel that Dr. Skilling has produced an interesting new approach to approximating integrals. Transforming an integration problem to a 1-dimensional problem and then using the trick for avoiding the evaluation of the function Ψ^{-1} at a specific point seems particularly ingenious. Only more study of the algorithm, however, will reveal the extent to which it can be effectively used.

REFERENCES

Bayarri, M.J. and Berger, J.O. (2000). P values for composite null models. *J. Amer. Statist. Assoc.* **95:452**, 1127-1142.

Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. A* **143**, 383-430.

Evans, M. and Moshonov, H. (2006a). Checking for prior-data conflict. To appear in *Bayesian Analysis* **1:4**.

Evans, M. and Moshonov, H. (2006b). Checking for prior-data conflict with hierarchically specified priors. Tech. Rep. No. 0503, June 12, 2005, Dept. of Statistics, U. of Toronto and to appear in *Proceedings of the International Workshop/Conference on Bayesian Statistics and its Applications*, Dept. of Statistics, Banaras Hindu U., Varanasi, India.

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