Contemplating Evidence: properties, extensions of, and alternatives to Nested Sampling

Nicolas Chopin, Christian Robert

Biometrika (2008), **xx**, x, *pp*. 1–18

Printed in Great Britain

Contemplating Evidence: properties, extensions of, and alternatives to Nested Sampling

BY NICOLAS CHOPIN
CREST–ENSAE, F-92245 Malakoff cedex, France
nicolas.chopin@bristol.ac.uk

AND CHRISTIAN P. ROBERT
CEREMADE, Université Paris Dauphine, F-75775 Paris cedex 16, France
xian@ceremade.dauphine.fr

SUMMARY

Nested sampling is a simulation method for approximating marginal likelihoods proposed by Skilling (2006). We establish that nested sampling has an approximation error that vanishes at the standard Monte Carlo rate $\mathrm{O}(N^{-1/2})$, where N is a tuning parameter proportional to the computational effort, and that this error is asymptotically Gaussian. We show that the asymptotic variance of the nested sampling approximation typically grows linearly with the dimension of the parameter. We discuss the applicability and efficiency of nested sampling in realistic problems, and we compare it with two current methods for computing marginal likelihood. We propose an extension that makes it possible to avoid resorting to MCMC to obtain the simulated points.

Some key words: MCMC, Monte Carlo approximation, mixture of distributions, importance sampling.

1. INTRODUCTION

Nested sampling was introduced by Skilling (2006) as a numerical approximation method for integrals of the kind

$$\mathfrak{Z} = \int L(y|\theta)\pi(\theta)\,\mathrm{d}\,\theta\,,$$

when π is the prior distribution and $L(y|\theta)$ is the likelihood. Those integrals are called *evidence* in the above papers and they naturally occur as marginals in Bayesian testing and model choice (Jeffreys, 1939; Robert, 2001, Chapters 5 and 7), even though the pairwise nature of those inferential problems, meaning that $\mathfrak{Z}$ is never computed *per se* but in relation with another marginal $\mathfrak{Z}'$, makes the approximation of the integral ratio (or Bayes factor)

$$\mathfrak{B}_{12} = \int L_1(y|\theta_1)\pi_1(\theta_1)\,\mathrm{d}\,\theta_1 \Big/ \int L_2(y|\theta_2)\pi_2(\theta_2)\,\mathrm{d}\,\theta_2$$

amenable to specific approximations (see, e.g., Chen & Shao, 1997; Gelman & Meng, 1998).

One important aspect of nested sampling is that it resorts to simulating points θ_i from the prior π, constrained to θ_i having a larger likelihood value than some threshold l; the exact principle of nested sampling is described in the next section. In a brief discussion (Chopin & Robert, 2007), we raised concerns about the universality and the formal properties of the method. With respect to the former concern, we pointed out that simulating efficiently from a constrained distribution

may not always be straightforward, even when the MCMC scheme suggested by Skilling (2006) is used. With respect to the latter one, the convergence properties of the method had not been fully established: Evans (2007) showed convergence in probability, but called for further work towards obtaining the rate of convergence and the nature of the limiting distribution.

The purpose of this paper is to investigate both points presented above. Our main contribution is to establish the convergence properties of the nested sampling estimates: the approximation error is dominated by a $\mathrm{O}(N^{-1/2})$ stochastic term, which has a limiting Gaussian distribution, and where N is a tuning parameter proportional to the computational effort. We also investigate the impact of the dimension d of the problem on the performances of the algorithm. In a simple example, we show that the asymptotic variance of nested sampling estimates grows linearly with d; this means that the computational cost is $\mathrm{O}(d^3/e^2)$, where e is the selected error bound.

In a second part, we discuss the difficulty to sample from the constrained prior. Using MCMC, as suggested by Skilling (2006), could incur a curse of dimensionality, although this pitfall seems model-dependent in our simulations. Murray's PhD thesis (2007, University College London) also includes a numerical comparison of nested sampling with other methods for several models.

Since the ability to simulate from the constrained prior is crucial in the applicability of the algorithm, we further propose an extension of nested sampling, based on importance sampling, that introduces enough flexibility so as to perform the constrained simulation without resorting to MCMC. Finally, we examine two alternatives to nested sampling for computing evidence, both based on the output of MCMC algorithms. We do not aim at an exhaustive comparison with all existing methods (see, e.g., Chen et al., 2000, for a broader review), and restrict our attention to methods that share the property with nested sampling that the same algorithm provides approximations of both the posterior distribution and the marginal likelihood, at no extra cost. We provide numerical comparisons between those methods.

2. NESTED SAMPLING: A DESCRIPTION

2·1. *Principle*

We describe briefly here the nested sampling algorithm, as provided in Skilling (2006). We use $L(\theta)$ as a short-hand for the likelihood $L(y|\theta)$, omitting the dependence on y.

Nested sampling is based on the following identity:

$$\mathfrak{Z} = \int_0^1 \varphi(x)\,\mathrm{d}x\,, \tag{1}$$

where φ is the inverse of

$$\varphi^{-1} : l \to P^{\pi}(L(\theta) > l)\,.$$

Thus, φ is the inverse of the survival function of the random variable $L(\theta)$, assuming $\theta \sim \pi$ and φ^{-1} is a decreasing function, which is the case when L is a continuous function and π has a connected support. (The representation $\mathfrak{Z} = \mathbb{E}^{\pi}[L(\theta)]$ holds with no restriction on either L or π.) Formally, this one-dimensional integral could be approximated by standard quadrature methods,

$$\widehat{\mathfrak{Z}} = \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i \tag{2}$$

where $\varphi_i = \varphi(x_i)$, and $0 < x_j < \cdots < x_1 < x_0 = 1$ is an arbitrary grid over $[0, 1]$. (This reduction in the dimension due to a change of measure can be found in the earlier numerical literature,

like Burrows, 1980.) Function φ is intractable in most cases however, so the φ_i's are approximated by an iterative random mechanism:

– Iteration 1: draw independently N points $\theta_{1,i}$ from the prior π, determine $\theta_1 = \arg\min_{1\le i\le N} L(\theta_{1,i})$, and set $\varphi_1 = L(\theta_1)$.
– Iteration 2: obtain the N 'current' values $\theta_{2,i}$, by reproducing the $\theta_{1,i}$'s, except for θ_1 that is replaced by a draw from the prior distribution π conditional upon $L(\theta) \ge \varphi_1$; then select θ_2 as $\theta_2 = \arg\min_{1\le i\le N} L(\theta_{2,i})$, and set $\varphi_2 = L(\theta_2)$.
– Iterate the above step until a given stopping rule is satisfied, for instance observing very small changes in the approximation $\hat{\mathfrak{Z}}$ or reaching the maximal value of $L(\theta)$ when it is known.

In the above, the value $x_i^\star = \varphi^{-1}(\varphi_i)$ that should be used in the quadrature approximation (2) is unknown. An interesting property of the generating process is however that the random variables defined by $t_i = \varphi^{-1}(\varphi_{i+1})/\varphi^{-1}(\varphi_i) = x_{i+1}^\star/x_i^\star$ are independent Beta$(N, 1)$ variates. Skilling (2006) takes advantage of this property by setting $x_i = \exp(-i/N)$, so that $\log x_i$ is the expectation of $\log \varphi^{-1}(\varphi_i)$. Alternatively to this deterministic scheme, Skilling (2006) proposes a *random scheme* where the x_i's are random, by mimicking the law of the t_i's, i.e. $x_{i+1} = x_i \cdot t_i$, where $t_i \sim$ Beta$(N, 1)$. In both cases the relation $\varphi_i = \varphi(x_i)$ does not hold; instead, φ_i should be interpreted as a 'noisy' version of $\varphi(x_i)$.

We focus on the deterministic scheme in this paper. It seems reasonably easy to establish a central limit theorem and other results for the random scheme, but the random scheme always produces less precise estimates, as illustrated by the following example.

Example 1. Consider the artificial case of a posterior distribution equal to $\pi(\theta|y) = \exp\{-\theta\}$ for a specific value of y, derived from the model $\pi(\theta) = \delta \exp\{-\delta\theta\}$ and $L(\theta) = \exp\{-(1-\delta)\theta\}/\delta$, so that $\mathfrak{Z} = 1$ for every $0 < \delta < 1$. Nested sampling can then be implemented with no MCMC approximation, each new θ in the running sample being simulated from an exponential $\mathcal{E}(\delta)$ distribution truncated to $(0, \theta_i)$, θ_i being the point with lowest likelihood excluded from the running sample. A small experiment summarised by Table 1 shows that the random scheme is systematically doing twice as worse than the deterministic scheme, both for the variance and for the mean square error (MSE) $\mathbb{E}[(\hat{\mathfrak{Z}} - \mathfrak{Z})^2]$ criteria. Both quantities decreases in $\mathcal{O}(1/N)$.

Table 1. *Comparison of the deterministic and random schemes in Example 1.* First row: *variance*, second row: *MSE, when using* 10^3 *replications,* $\delta = .1, .5, .9$ (left, centre, right) *and a stopping rule chosen as* $\max(L_i) < 10^{-3}\hat{\mathfrak{Z}}$.

N	Deterministic	Random	N	Deterministic	Random	N	Deterministic	Random
50	325	646	50	46.4	10.5	50	1.81	3.41
	327	646		46.5	10.5		1.82	3.41
100	172	307	100	24.7	49.0	100	0.883	0.176
	175	308		24.9	50.2		0.249	0.176
500	29.2	57.7	500	5.49	10.1	500	0.180	0.387
	29.3	57.7		5.50	11.4		0.181	0.387
10^3	17.6	32.7	10^3	2.47	4.81	10^3	0.090	0.170
	17.6	32.9		2.48	4.83		0.091	0.171

All values are multiplied by 10^{-4}

2·2. *Variations and posterior simulation*

Skilling (2006) points out that nested sampling provides simulations from the posterior distribution at no extra cost: "the existing sequence of points $\theta_1, \theta_2, \theta_3, \cdots$ already gives a set of

posterior representatives, provided the i'th is assigned the appropriate importance weight $\omega_i L_i$". (The weight ω_i is equal to the difference $(x_{i-1} - x_i)$ and L_i is equal to φ_i.) This can be justified as follows. Consider the computation of the posterior expectation of a given function f

$$\mu(f) = \int \pi(\theta)L(\theta)f(\theta)\,\mathrm{d}\theta \Big/ \int \pi(\theta)L(\theta)\,\mathrm{d}\theta\,.$$

One can then use a single run of nested sampling to obtain estimates of both the numerator and the denominator (the latter being the evidence $\mathfrak{Z}$, estimated by (2)). The estimator

$$\sum_{i=1}^{j}(x_{i-1} - x_i)\varphi_i f(\theta_i) \tag{3}$$

of the numerator is a noisy version of

$$\sum_{i=1}^{j}(x_{i-1} - x_i)\varphi_i \widetilde{f}(\varphi_i)\,,$$

where $\widetilde{f}(l) = \mathbb{E}^{\pi}[f(\theta)|L(\theta) = l]$, the (prior) expectation of $f(\theta)$ conditional on $L(\theta) = l$. This Riemann sum is, following the principle of nested sampling, an estimator of the evidence.

LEMMA 1. *Let $\widetilde{f}(l) = \mathbb{E}^{\pi}[f(\theta)|L(\theta) = l]$ for $l > 0$, then, if $\widetilde{f}$ is absolutely continuous,*

$$\int_0^1 \varphi(x)\widetilde{f}\{\varphi(x)\}\,\mathrm{d}x = \int \pi(\theta)L(\theta)f(\theta)\,\mathrm{d}\theta. \tag{4}$$

A proof is provided in Appendix 1. Clearly, the estimate of $\mu(f)$ obtained by dividing (3) by (2) is the estimate obtained by computing the weighted average mentioned above. We do not discuss further this aspect of nested sampling, but our convergence results can be easily extended to such estimates. In many cases, however, the distribution of the weights $w_i L_i$ may be quite skewed, since a certain proportion of points is simulated from the prior constrained by a low likelihood, and such approximations may thus suffer from a large variance.

2·3. *Connection with slice sampling*

In every situation where simulating independently from the constrained prior is feasible, a corresponding slice sampler (e.g., Robert & Casella, 2004, Chapter 8) can be implemented with at most the same computational cost (in the sense that increasing the bound l on the likelihood may induce a diminishing efficiency in computing). Thus, in settings where slice samplers are slow to converge (e.g. Roberts & Rosenthal, 1998), it is likely that nested sampling requires a large computational effort as well. Consider the following example, adapted from Roberts & Rosenthal (1999): $L(\theta) \propto \exp(-||\theta||)$, and $\pi(\theta) \propto ||\theta||^{(1-d^2)/d}\,\mathbb{I}(||\theta|| < 1)$, which is rotation invariant, hence $\mathfrak{Z} = \int_0^1 \exp(-\omega^{1/d})\,\mathrm{d}\omega$. Since the maximum of $\exp(-\omega^{1/d})$ is 1, if we set the stopping rule for the maximum observed likelihood to be at least .99, the number m of uniform simulations that is necessary to get under the limit $\beta_d = (-\log .99)^d \approx 10^{-2d}$ is given by $\mathbb{P}^{\pi}(\min(\theta_1, \ldots, \theta_m) < \beta_d) \approx 0.95$, namely $m \approx 3\,10^{2d}$. Using a sequence of uniforms to reach the maximum of the likelihood is therefore delicate for $d > 3$ and the slice sampler of Roberts & Rosenthal (1999) performs more satisfactorily for such dimensions.

3. A CENTRAL LIMIT THEOREM FOR NESTED SAMPLING

We establish in the section the convergence rate and the limiting distribution of nested sampling estimates. To this effect, we decompose the approximation error as follows:

$$\sum_{i=1}^{j}(x_{i-1}-x_i)\varphi_i - \int_0^1 \varphi(x)\,\mathrm{d}x = -\int_0^\varepsilon \varphi(x)\,\mathrm{d}x + \left[\sum_{i=1}^{j}(x_{i-1}-x_i)\varphi(x_i) - \int_\varepsilon^1 \varphi(x)\,\mathrm{d}x\right] + \sum_{i=1}^{j}(x_{i-1}-x_i)\left\{\varphi_i-\varphi(x_i)\right\}$$

where

1. The first term is a truncation error, resulting from the feature that the algorithm is run for a finite time. For simplicity's sake, we assume that the algorithm is stopped at iteration $j = \lceil(-\log\varepsilon)N\rceil$. so that $x_j = \exp(-j/N) \leq \varepsilon < x_{j-1}$. (More practical stopping rules will be discussed in §7.) Assuming φ, or equivalently L, bounded from above, the error $\int_0^\varepsilon \varphi(x)\,\mathrm{d}x$ is exponentially small with respect to the computational effort.
2. The second term is a (deterministic) numerical integration error, which, provided φ' is bounded over $[\varepsilon,1]$, is of order $\mathrm{O}(N^{-1})$, since $x_{i-1}-x_i = \mathrm{O}(N^{-1})$.
3. The third term is stochastic and is denoted

$$e_N = \sum_{i=1}^{j}(x_{i-1}-x_i)\left[\varphi(x_i^\star)-\varphi(x_i)\right],\ .$$

where the $x_i^\star$'s are such that $\varphi_i = L(\theta_i) = \varphi(x_i^\star)$, i.e. $x_i^\star = \varphi^{-1}(\varphi_i)$.

The asymptotic behaviour of e_N is characterised as follows.

THEOREM 1. *Provided that φ is twice continuously-differentiable over $[\varepsilon,1]$, and that its first and second derivatives are bounded over $[\varepsilon,1]$, $N^{1/2}e_N$ converges in distribution to a Gaussian distribution with mean zero and variance*

$$V = -\int_{s,t\in[\varepsilon,1]} s\varphi'(s)t\varphi'(t)\log(s\vee t)\,\mathrm{d}s\,\mathrm{d}t.$$

The stochastic error is of order $\mathrm{O}_P(N^{-1/2})$ and it dominates both other error terms. The proof of this theorem relies on the functional central limit theorem and is detailed in Appendix 2.

As pointed out by one referee, it usually is more relevant in practice to consider the log-scale error, $\log\widehat{\mathfrak{Z}} - \log\mathfrak{Z}$. A straightforward application of the delta-method shows that the log-error has the same asymptotic behaviour as above, but with asymptotic variance $V/\mathfrak{Z}^2$.

4. PROPERTIES OF THE NESTED SAMPLING ALGORITHM

4·1. *Simulating from a constrained prior*

The main difficulty of nested sampling is to simulate θ from the prior distribution π subject to the constraint $L(\theta) > L(\theta_i)$; exact simulation from this distribution is an intractable problem in many realistic set-ups. As noted in § 2·3, it is at least of the same complexity as a one-dimensional slice sampler, which produces an uniformly ergodic Markov chain when the likelihood L is bounded but may be slow to converge in other settings (Roberts & Rosenthal, 1999).

Skilling (2006) proposes to sample values of θ by iterating M MCMC steps, using the truncated prior as the invariant distribution, and a point chosen at random among the $N-1$ survivors as the starting point. Since the starting value is already distributed from the invariant distribution, a finite number M of iterations produces an outcome that is marginally distributed from the correct distribution. This however introduces correlations between simulated points. Our central limit theorem applies no longer and it is unclear whether a nested sampling estimate based on MCMC converges as $N \to +\infty$, for a fixed M, or if it should merely be interpreted as an approximation of an ideal nested sampling output based on independent samples. A reason why such a theoretical result seems difficult to establish is that each iteration involves both a different MCMC kernel and a different invariant distribution.

In addition, there are settings when implementing an MCMC move that leaves the truncated prior invariant is not straightforward. In those cases, one may instead implement an MCMC move (e.g., random walk Metropolis-Hastings) with respect to the unconstrained prior, and subsample only values that satisfy the constraint $L(\theta) > L(\theta_i)$, but this scheme gets increasingly inefficient as the constraint moves closer to the highest values of L. Obviously, more advanced sampling schemes can be devised that overcome this difficulty, as for instance the use of a diminishing variance factor in the random walk, with the drawback that this adaptive scheme requires more programming effort, when compared with the basic nested sampling algorithm.

In §5, we propose an extension of nested sampling based on importance sampling. In some settings, this may facilitate the design of efficient MCMC steps, or even allow for sampling independently from the (instrumental) constrained prior.

4·2. *Impact of dimensionality*

Although nested sampling is based on the unidimensional integral (1), this section shows that its theoretical performance typically depends on the dimension d of the problem in that the required number of iterations (for a fixed truncation error) and the asymptotic variance both grow linearly with d. A corollary of this result is that, under the assumption that the cost of a single iteration is $\mathrm{O}(d)$, the computational cost of nested sampling is $\mathrm{O}(d^3/e^2)$, where e denotes a given error level, as also stated in Murray's PhD thesis, using a more heuristic argument. This result applies to the *exact* nested algorithm. Resorting to MCMC usually entails some additional curse of dimensionality, although simulation studies in §7 indicate that the severity of this problem is strongly model-dependent.

Example 2. Consider the case where, for $k = 1, \ldots, d$, $\theta^{(k)} \sim \mathcal{N}(0, \sigma_0^2)$, and $y^{(k)}|\theta^{(k)} \sim \mathcal{N}(\theta^{(k)}, \sigma_1^2)$, independently in both cases. Set $y^{(k)} = 0$ and $\sigma_0^2 = \sigma_1^2 = 1/4\pi$, so that $\mathfrak{Z} = 1$ for all d's. Exact simulation from the constrained prior can be performed as follows: simulate $r^2 \leq -\sqrt{2}\log l$ from a truncated $\chi^2(d)$ distribution and $u_1, \ldots, u_d \sim \mathcal{N}(0,1)$, then set $\theta^{(k)} = r\, u_k / \sqrt{u_1^2 + \ldots + u_d^2}$.

Since $\mathfrak{Z} = 1$, we assume that the truncation point ε_d is chosen so that $\varphi(0)\varepsilon_d = \tau \ll 1$, $\tau = 10^{-6}$ say, where $\varphi(0) = 2^{d/2}$ is the maximum likelihood value. Therefore, $\varepsilon_d = \tau 2^{-d/2}$ and the number of iterations required to produce a given truncation error, i.e. $j = \lceil (-\log \epsilon) N \rceil$, grows linearly in d. To assess the dependence of the asymptotic variance with respect to d, we state the following lemma, established in Appendix 3:

LEMMA 2. *In the setting of Example 2, if V_d is the asymptotic variance of the nested sampling estimator with truncation point ε_d, there exist constants c_1, c_2 such that $V_d/d \leq c_1$ for all $d \geq 1$, and $\liminf_{d\to+\infty} V_d/d \geq c_2$.*

This lemma is easily generalised to setups where the prior is such that the components are independent and identically distributed, and the likelihood factorises as $L(\theta) = \prod_{k=1}^{d} L(\theta^{(k)})$. We conjecture that V_d/d converges to a finite value in all these situations and that, for more general models, the variance grows linearly with the 'actual' dimensionality of the problem, as measured for instance in Spiegelhalter et al. (2002).

5. NESTED IMPORTANCE SAMPLING

We introduce an extension of nested sampling based on importance sampling. Let $\widetilde{\pi}(\theta)$ an instrumental prior with the support of π included in the support of $\widetilde{\pi}$, and let $\widetilde{L}(\theta)$ an instrumental likelihood, namely a positive measurable function. We define an importance weight function $w(\theta)$ such that $\widetilde{\pi}(\theta)\widetilde{L}(\theta)w(\theta) = \pi(\theta)L(\theta)$. We can approximate $\mathfrak{Z}$ by nested sampling for the pair $(\widetilde{\pi}, \widetilde{L})$, that is, by simulating iteratively from $\widetilde{\pi}$ constrained to $\widetilde{L}(\theta) > l$, and by computing the generalised nested sampling estimator

$$\sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i w(\theta_i). \tag{5}$$

The advantage of this extension is that one can choose $(\widetilde{\pi}, \widetilde{L})$ so that simulating from $\widetilde{\pi}$ under the constraint $\widetilde{L}(\theta) > l$ is easier than simulating from π under the constraint $L(\theta) > l$. For instance, one may choose an instrumental prior $\widetilde{\pi}$ such that MCMC steps wr.t. the instrumental constrained prior are easier to implement than w.r.t. the actual constrained prior, as illustrated in §7·2. In a similar vein, nested importance sampling facilitates contemplating several priors at once, as one may compute the evidence for each prior by producing the same nested sequence (based on the same pair $(\widetilde{\pi}, \widetilde{L})$) and by simply modifying the weight function.

Ultimately, one may choose $(\widetilde{\pi}, \widetilde{L})$ so that the constrained simulation is performed exactly. For instance, if $\widetilde{\pi}$ is a Gaussian $\mathcal{N}_d(\hat{\theta}, \hat{\Sigma})$ distribution with arbitrary hyper-parameters, take

$$\widetilde{L}(\theta) = \lambda\left((\theta - \hat{\theta})^T \hat{\Sigma}^{-1} (\theta - \hat{\theta})\right),$$

where λ is an arbitrary decreasing function. Then

$$\varphi_i w(\theta_i) = \widetilde{L}(\theta_i) w(\theta_i) = \pi(\theta_i) L(\theta_i) / \widetilde{\pi}(\theta_i).$$

In this case, the x_i's in (2) are error-free: at iteration i, θ_i is sampled uniformly over the ellipsoid that contains exactly $\exp(-i/N)$ prior mass as $\theta_i = q_i C v / \|v\|_2^{1/2}$, where C is the Cholesky lower triangle of $\hat{\Sigma}$, $v \sim N_d(0, I_d)$, and q_i is the $\exp(-i/N)$ quantile of a $\chi^2(d)$ distribution.

The nested ellipsoid strategy seems useful in two scenarios. First, assume both the posterior mode and the Hessian at the mode are available numerically and tune $\hat{\theta}$ and $\hat{\Sigma}$. In this case, this strategy should outperform standard importance sampling based on the optimal Gaussian proposal, because the nested ellipsoid strategy uses a $O(N^{-1})$ quadrature rule on the radial axis, along which the weight function varies the most; see §7·4 for an illustration. Second, assume only the posterior mode is available, so one may set $\hat{\theta}$ to the posterior mode, and set $\hat{\Sigma} = \tau I_d$, where τ is an arbitrary, large value. §7·4 indicates that the nested ellipsoid strategy may still perform reasonably in such a scenario. Models such that the Hessian at the mode is tedious to compute include in particular Gaussian state space models with missing observations (Brockwell & Davis, 1996), Markov modulated Poisson processes (Rydén, 1994), or, more generally, models where the EM algorithm (see, e.g. MacLachlan & Krishnan, 1997) is the easiest way to compute

the posterior mode (although one may use Louis' 1982 method for computing the information matrix from the EM output).

6. ALTERNATIVE ALGORITHMS

6·1. *Approximating $\mathfrak{Z}$ from a posterior sample*

As shown in §2·2, the output of nested sampling can be "recycled" so as to provide approximations of posterior quantities. From the opposite perspective, we can recycle the output of an MCMC algorithm so as to estimate the evidence, with no or little additional programming effort. Several solutions are available in the literature, including Gelfand & Dey (1994), Meng & Wong (1996), and Chen & Shao (1997). We describe below those solutions used in the subsequent comparison with nested sampling, but first we stress that we do not pretend at an exhaustive coverage of those techniques (see Chen et al., 2000 or Han & Carlin, 2001 for deeper coverage) nor at using the most efficient approach (see, e.g., Meng & Schilling, 2002). In her evaluation of Chib's (1995) method, Frühwirth-Schnatter (2004) used the solutions we present below.

6·2. *Approximating $\mathfrak{Z}$ by a formal reversible jump*

We first recover Gelfand and Dey's (1994) solution of reverse importance sampling by an integrated reversible jump, as a natural approach to compute a marginal likelihood is to use a reversible jump MCMC algorithm (Green, 1995). However, this may seem wasteful as it involves simulating from several models, while only one is of interest. But we can in theory contemplate a single model $\mathfrak{M}$ and still implement reversible jump in the following way. Consider a formal alternative model $\mathfrak{M}'$, for instance a fixed distribution like the $\mathcal{N}(0,1)$ distribution, with prior weight $1/2$ and build a proposal from $\mathfrak{M}$ to $\mathfrak{M}'$ that moves to $\mathfrak{M}'$ with probability (Green, 1995) $\varrho_{\mathfrak{M}\to\mathfrak{M}'} = \{1/2g(\theta)\}\big/\{1/2\pi(\theta)L(\theta)\} \wedge 1$ and from $\mathfrak{M}'$ to $\mathfrak{M}$ with probability $\varrho_{\mathfrak{M}'\to\mathfrak{M}} = \{1/2\pi(\theta)L(\theta)\}\big/\{1/2g(\theta)\} \wedge 1$, $g(\theta)$ being an arbitrary proposal on θ. Were we to actually run this reversible jump MCMC algorithm, the frequency of visits to $\mathfrak{M}$ would then converge to $\mathfrak{Z}$.

However, the reversible sampler is not needed since, if we run a standard MCMC algorithm on θ and compute the probability of moving to $\mathfrak{M}'$, the expectation of the ratio $g(\theta)/\pi(\theta)L(\theta)$ (under stationarity) is equal to the inverse of $\mathfrak{Z}$:

$$\mathbb{E}\left[g(\theta)\big/\pi(\theta)L(\theta)\right] = \int \frac{g(\theta)}{\pi(\theta)L(\theta)}\,\frac{\pi(\theta)L(\theta)}{\mathfrak{Z}}\,\mathrm{d}\theta = 1\big/\mathfrak{Z}\,,$$

no matter what $g(\theta)$ is, in the spirit of both Gelfand & Dey (1994) and Bartolucci et al. (2006).

Obviously, the choice of $g(\theta)$ impacts on the precision of the approximated $\mathfrak{Z}$. When using a kernel approximation to $\pi(\theta|y)$ based on earlier MCMC simulations and considering the variance of the resulting estimator, the constraint is opposite to the one found in importance sampling, namely that $g(\theta)$ must have lighter (not fatter) tails than $\pi(\theta)L(\theta)$ for the approximation

$$\widehat{\mathfrak{Z}_1} = 1\Big/\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)})\big/\pi(\theta^{(t)})L(\theta^{(t)})$$

to have a finite variance. This means that light tails or finite support kernels (like an Epanechnikov kernel) are to be preferred to fatter tails kernels (like the t kernel).

In the comparison in §7·3, we compare $\widehat{\mathfrak{Z}_1}$ with a standard importance sampling approximation

$$\widehat{\mathfrak{Z}_2} = \frac{1}{T}\sum_{t=1}^{T} \pi(\theta^{(t)})L(\theta^{(t)})\big/g(\theta^{(t)})\,, \qquad \theta^{(t)} \sim g(\theta)\,,$$

where g can also be a non-parametric approximation of $\pi(\theta|y)$, this time with heavier tails than $\pi(\theta)L(\theta)$. Frühwirth-Schnatter (2004) uses the same importance function g in both $\widehat{\mathfrak{Z}_1}$ and $\widehat{\mathfrak{Z}_2}$, and obtain similar results that $\widehat{\mathfrak{Z}_2}$ performs better than $\widehat{\mathfrak{Z}_1}$.

6·3. *Approximating $\mathfrak{Z}$ using a mixture representation*

Another approach in the approximation of $\mathfrak{Z}$ is to design a specific mixture for simulation purposes, with density proportional to $\omega_1\pi(\theta)L(\theta) + g(\theta)$ ($\omega_1 > 0$), where $g(\theta)$ is an arbitrary (fully specified) density. Simulating from this mixture offers the same complexity as simulating from the posterior, the MCMC code used to simulate from $\pi(\theta|y)$ can be easily extended by introducing an auxiliary variable δ that indicates whether or not the current simulation is from $\pi(\theta|y)$ or from $g(\theta)$. The t-th iteration of this extension is as follows, where MCMC(θ, θ') denotes an arbitrary MCMC kernel associated with the posterior $\pi(\theta|y) \propto \pi(\theta)L(\theta)$:

1. Take $\delta^{(t)} = 1$ (and $\delta^{(t)} = 2$ otherwise) with probability
$$\omega_1\pi(\theta^{(t-1)})L(\theta^{(t-1)}) \Big/ \left\{\omega_1\pi(\theta^{(t-1)})L(\theta^{(t-1)}) + g(\theta^{(t-1)})\right\} ;$$
2. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim$ MCMC$(\theta^{(t-1)}, \theta^{(t)})$, else generate $\theta^{(t)} \sim g(\theta)$ independently from the previous value $\theta^{(t-1)}$.

This algorithm is a Gibbs sampler: Step 1 simulates $\delta^{(t)}$ conditional on $\theta^{(t-1)}$, while Step 2 simulates $\theta^{(t)}$ conditional on $\delta^{(t)}$. While the average of the $\delta^{(t)}$'s converges to $\omega_1\mathfrak{Z}/\{\omega_1\mathfrak{Z} + 1\}$, a natural Rao-Blackwellisation is to take the average of the expectations of the $\delta^{(t)}$'s,

$$\hat{\xi} = \frac{1}{T}\sum_{t=1}^{T} \omega_1\pi(\theta^{(t)})L(\theta^{(t)}) \Big/ \left\{\omega_1\pi(\theta^{(t)})L(\theta^{(t)}) + g(\theta^{(t)})\right\} ,$$

since its variance should be smaller. A third estimate is then deduced from this approximation by solving $\omega_1\hat{\mathfrak{Z}}_3/\{\omega_1\hat{\mathfrak{Z}}_3 + 1\} = \hat{\xi}$.

The use of mixtures in importance sampling in order to improve the stability of the estimators dates back at least to Hesterberg (1998) but, as it occurs, this particular mixture estimator happens to be almost identical to the bridge sampling estimator of Meng & Wong (1996). In fact,

$$\hat{\mathfrak{Z}}_3 = \frac{1}{\omega_1}\sum_{t=1}^{T} \frac{\omega_1\pi(\theta^{(t)})L(\theta^{(t)})}{\omega_1\pi(\theta^{(t)})L(\theta^{(t)}) + g(\theta^{(t)})} \Bigg/ \sum_{t=1}^{T} \frac{g(\theta^{(t)})}{\omega_1\pi(\theta^{(t)})L(\theta^{(t)}) + g(\theta^{(t)})}$$

is the Monte Carlo approximation to the ratio $\mathbb{E}_\varphi[\alpha(\theta)\pi(\theta)L(y|\theta)]/\mathbb{E}_{\pi(\cdot|y)}[\alpha(\theta)g(\theta)]$ when using the optimal function $\alpha(\theta) = 1\big/\omega_1\pi(\theta)L(\theta) + g(\theta)$. The only difference with Meng & Wong (1996) is that, since $\theta^{(t)}$'s are simulated from the mixture, they can be recycled for both sums.

6·4. *Error approximations*

Usual confidence intervals can be produced on the averages $1/\widehat{\mathfrak{Z}_1}$, $\widehat{\mathfrak{Z}_2}$ and $\omega_1\hat{\mathfrak{Z}}_3/\{\omega_1\hat{\mathfrak{Z}}_3 + 1\}$, from which confidence intervals on the $\widehat{\mathfrak{Z}_i}$'s and error estimates are easily deduced.

7. NUMERICAL EXPERIMENTS

7·1. *A decentred Gaussian example*

We modify the Gaussian toy example presented in §4·2: $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$, where the $\theta^{(k)}$'s are i.i.d. $\mathcal{N}(0, 1)$ and $y_k|\theta^{(k)} \sim \mathcal{N}(\theta^{(k)}, 1)$ independently, but setting all the y_k's to 3. To simulate

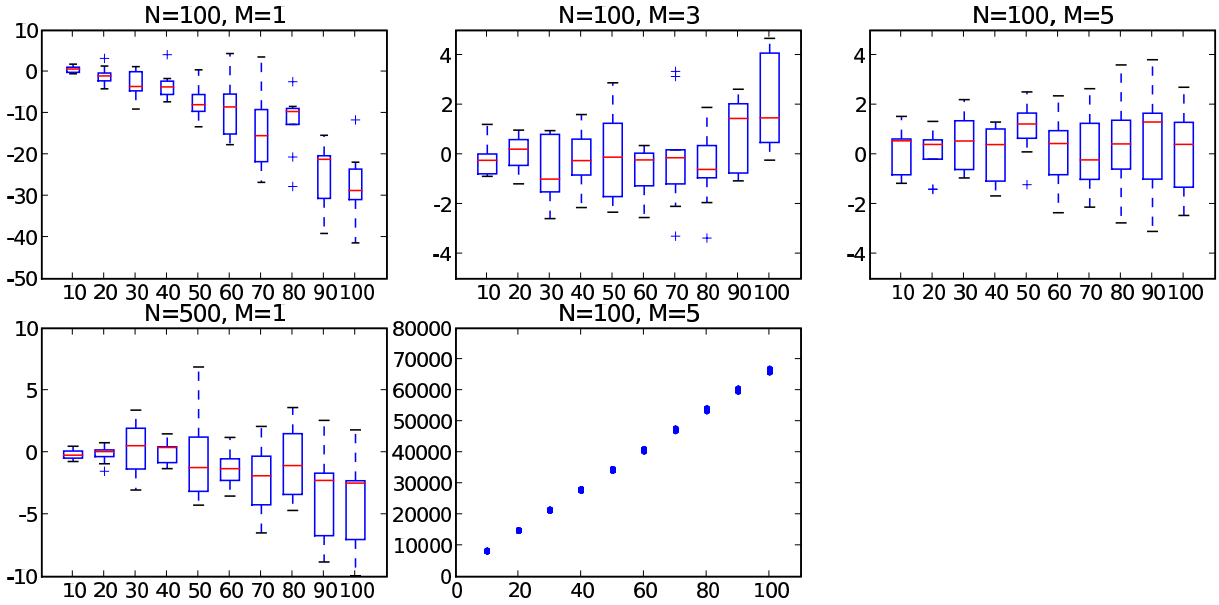

Fig. 1. Decentred Gaussian example: Box-plots of the log-relative error $\log \widehat{\mathfrak{Z}} - \log \mathfrak{Z}$ versus dimension d for several values of (N, M), and total number of iterations vs dimension for $(N, M) = (100, 5)$

from the prior truncated to $L(\theta) > L(\theta_0)$, we perform M Gibbs iterations with respect to this truncated distribution, with $M = 1$, 3 or 5: the full conditional distribution of $\theta^{(k)}$, given $\theta^{(j)}$, $j \neq k$, is a $\mathcal{N}(0, 1)$ distribution that is truncated to the interval $[y^{(k)} - \delta, y^{(k)} + \delta]$ with

$$\delta^2 = \sum_j (y_j - \theta_0^{(j)})^2 - \sum_{j \neq k} (y_j - \theta^{(j)})^2$$

The nested sampling algorithm is run 20 times for $d = 10$, 20, ..., 100, and several combinations of (N, M): $(100, 1)$, $(100, 3)$, $(100, 5)$, and $(500, 1)$. The algorithm is stopped when a new contribution $(x_{i-1} - x_i)\varphi_i$ to (2) becomes smaller than 10^{-8} times the current estimate. Focussing first on $N = 100$, Figure 1 exposes the impact of the mixing properties of the MCMC step: for $M = 1$, the bias sharply increases with respect to the dimension, while, for $M = 3$, it remains small for most dimensions. Results for $M = 3$ and $M = 5$ are quite similar, except perhaps for $d = 100$. Using $M = 3$ Gibbs steps seems to be sufficient to produce a good approximation of an *ideal* nested sampling algorithm, where points would be simulated independently. Interestingly, if N increases to 500, while keeping $M = 1$, then larger errors occur for the same computational effort. Thus, a good strategy in this case is to increase first M until the distribution of the error stabilises, then to increase N to reduce the Monte Carlo error. As expected, the number of iterations linearly increases with the dimension.

While artificial, this example shows that nested sampling performs quite well even in large dimension problems, provided both prior and likelihood are close to Gaussianity.

7·2. *A stochastic volatility example*

We consider a simplified stochastic volatility model $(t = 1, \ldots, T)$:

$$h_0 = 0\,, \quad h_t = \rho h_{t-1} + \sigma \varepsilon_t\,, \quad \varepsilon_t \sim \mathcal{N}(0, 1)\,, \quad y_t | h_t \sim \mathcal{N}\{0, \exp(h_t)\}\,,$$

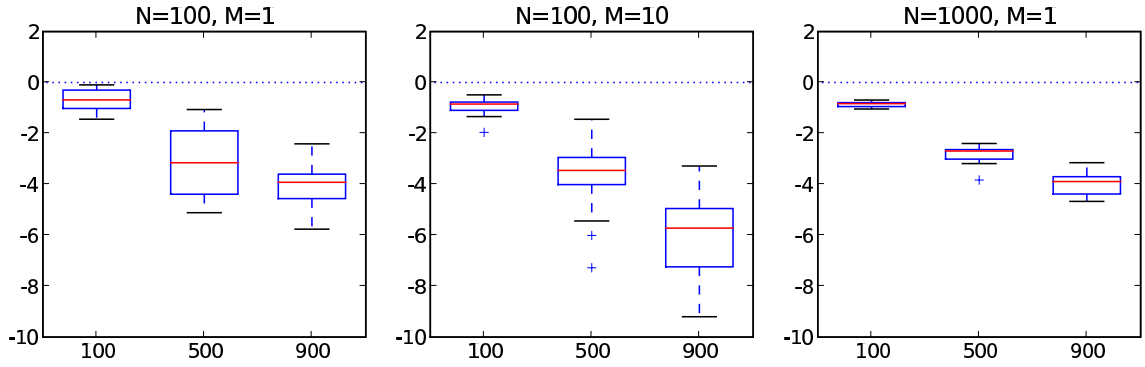

Fig. 2. Stochastic volatility example: box-plots of log-errors for different values of T (sample size), N and M

with a prior $\rho \sim \mathcal{U}([-1,1])$, $\sigma^{-2} \sim \mathcal{G}(1/2, (0.05)^2/2)$ on the remaining components of the parameter $\theta = (\rho, \sigma, h_1, \ldots, h_T)$. The data is simulated, using $\rho = 0.9$ and $\sigma = 0.05$. We implemented a MCMC strategy where realisations from the prior were generated using M steps of a fully conditional Gibbs sampler targeted at the constrained prior, the full conditionals being reasonably easy to simulate.

Figure 2 shows that, in contrast to the previous example, one gets better results with $(N, M) = (1000, 1)$ than with $(N, M) = (100, 10)$, although both scenarios cost the same. However, when we tried to increase N further to 10^5, with $M = 1$, we obtained sensibly the same biases as for $(N, M) = (1000, 1)$ (results not shown). So this may a case where nested sampling based on MCMC should be interpreted as a possibly good, but non necessarily convergent, approximation of the ideal nested sampling algorithm based on independent samples. On the other hand, stochastic volatility models are notoriously difficult to estimate, see e.g. Kim et al. (1998), in particular because Gibbs samplers tend to converge slowly; this difficulty may be the best explanation for this observed bias. For $T = 900$ a bias of order -4 may be small enough for model comparison purposes. (The actual log evidence is -1297.06.)

Kim et al. (1998) propose a Beta prior as a more sensible choice for ρ. The full conditional distribution of ρ under the constraint is difficult to simulate, requiring an extra Hastings-Metropolis step. A convenient alternative is to use nested importance sampling, with $\widetilde{\pi}(\theta)$ set to $\mathcal{U}[-1,1]$, and $\widetilde{L} = L$, the actual likelihood, in order to recycle the above algorithm, including the MCMC strategy, but with the weight function $w(\theta) = \pi(\theta)$ in the estimate of $\mathfrak{Z}$.

7·3. *A mixture example*

Following Frühwirth-Schnatter (2004)'study of several marginal likelihood estimates, a benchmark example is the posterior distribution on (μ, σ) associated with the normal mixture

$$y_1, \ldots, y_n \sim p\mathcal{N}(0,1) + (1-p)\mathcal{N}(\mu, \sigma)\,, \tag{6}$$

when p is known, for several compelling reasons:

1. Both the posterior distribution and the marginal likelihood are unavailable (unless n is small).
2. When σ converges to 0 and μ is equal to any of the x_i's $(1 \leq i \leq n)$, the likelihood diverges, as illustrated on Figure 3 by the tiny bursts in the vicinity of each observation when σ goes to 0. This represents a challenging problem for exploratory schemes such as nested sampling.

3. Efficient MCMC strategies have been developed and tested for mixture models since the early 1990's (Diebolt & Robert, 1994; Richardson & Green, 1997; Celeux et al., 2000), but Bayes factors are notoriously difficult to approximate in this setting.

We designed a Monte Carlo experiment where we simulated n observations from a $\mathcal{N}(2, (3/2)^2)$ distribution, and then computed the estimates of $\mathfrak{Z}$ introduced above for the model (6). The prior distribution was a uniform both on $(-2, 6)$ for μ and on $(.001, 16)$ for $\log \sigma^2$. (The prior square is chosen arbitrarily to allow all possible values and still retain a compact parameter space. Furthermore, a flat prior allows for an easy implementation of nested sampling since the constrained simulation can be implemented via a random walk move.)

The two-dimensional nature of the parameter space allows for a numerical integration of $L(\theta)$, based on a Riemann approximation and a grid of 800×500 points in the $(-2, 6) \times (.001, 16)$ square. This approach leads to a stable evaluation of $\mathfrak{Z}$ that can be taken as the reference against which we can test the various methods. (An additional evaluation based on a crude Monte Carlo integration using 10^6 terms produced essentially the same numerical values.) The MCMC algorithm implemented here is the standard completion of Diebolt & Robert (1994) and it does not suffer from the usual label switching deficiency (Jasra et al., 2005) because (6) is identifiable. As shown by the MCMC sample of size $N = 10^4$ displayed on the lhs of Fig. 3, the exploration of the modal region by the MCMC chain is satisfactory. This MCMC sample is used to compute the non-parametric approximations g that appear in the three alternatives of §6. For the reverse importance sampling estimate $\mathfrak{Z}_1$, g is a product of two Gaussian kernels with a bandwidth equal to half the default bandwidth of the R function density(), while, for both $\mathfrak{Z}_2$ and $\mathfrak{Z}_3$, g is a product of two t kernels with a bandwidth equal to twice the default Gaussian bandwidth.

We ran the nested sampling algorithm, with $N = 10^3$, reproducing the implementation of Skilling (2006), namely using 10 steps of a random walk in $(\mu, \log \sigma)$ constrained by the likelihood boundary. based on the contribution of the current value of (μ, σ) to the approximation of $\mathfrak{Z}$. The overall number of points produced by nested sampling at stopping time is on average close to 10^4, which justifies using the same number of points for the MCMC algorithm. As shown on the rhs of Fig. 3, the nested sampling sequence visits the minor modes of the likelihood surface but it ends up in the same central mode as the MCMC sequence. All points visited by nested sampling are represented without reweighting, which explains for a larger density of points outside the central modal region.

The analysis of this Monte Carlo experiment in Figure 4 first shows that nested sampling gives approximately the same numerical value when compared with the three other approaches, exhibiting a slight upward bias, but that its variability is much higher. The most reliable approach, besides the numerical and raw Monte Carlo evaluations which cannot be used in general settings, is the importance sampling solution, followed very closely by the mixture approach of §6·3. The reverse importance sampling naturally shows a slight upward bias for the smaller values of n and a variability that is very close to both other alternatives, especially for larger values of n.

7·4. *A probit example for nested importance sampling*

To implement the nested importance sampling algorithm based on nested ellipsoids, we consider the arsenic dataset and a probit model studied in Chapter 5 of Gelman & Hill (2006). The observations are independent Bernoulli variables y_i such that $\Pr(y_i = 1|x_i) = \Phi(x_i^T\theta)$, where x_i is a vector of d covariates, θ is a vector parameter of size d, and Φ denotes the standard normal distribution function. In this particular example, $d = 7$; more details on the data and the covariates are available on the book's web-page (`http://www.stat.columbia.edu/~gelman/arm/examples/arsenic`).

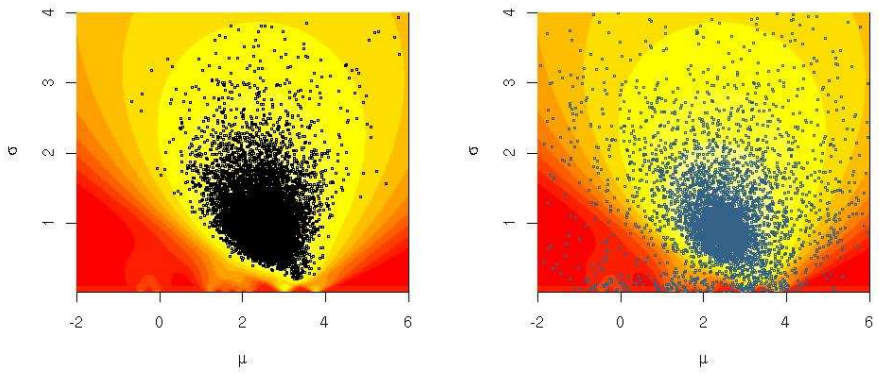

Fig. 3. Mixture example: *(left)* MCMC sample plotted on the log-likelihood surface in the (μ, σ) space for $n = 10$ observations from (6) *(right)* nested sampling sequence based on $N = 10^3$ starting points for the same dataset

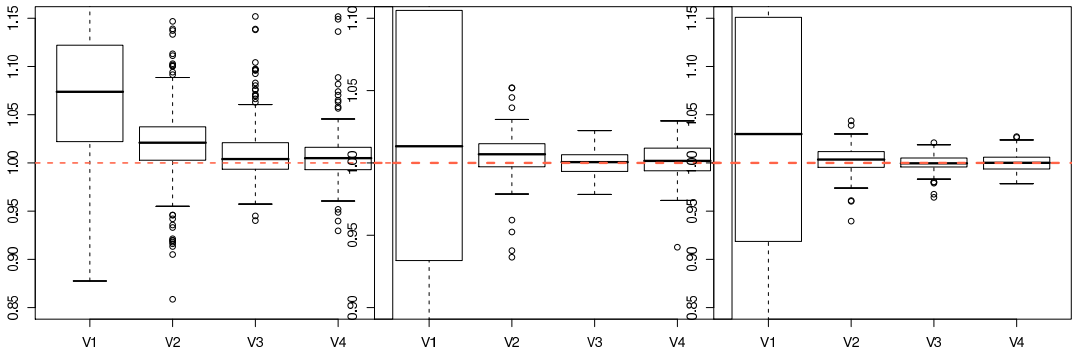

Fig. 4. Mixture model: comparison of the variations of nested sampling (V1), reverse importance sampling (V2), importance sampling (V3) and mixture sampling (V4), relative to a numerical approximation of 3 (dotted line), based on 150 samples of size $n = 10, 50, 100$

The probit model we use is model 9a in the R program available at this address: the dependent variable indicates whether or not the surveyed individual changed the well she drinks from over the past three years, and the seven covariates are an intercept, distance to the nearest safe well (in 100 meters unit), education level, log of arsenic level, and cross-effects for these three variables. We assign $\mathcal{N}_d(0, 10^2 I_d)$ as our prior on θ, and denote θ_m the posterior mode, and Σ_m the inverse of minus twice the Hessian at the mode; both quantities are obtained numerically beforehand.

We run the nested ellipsoid algorithm 50 times, for $N = 2$, 8, 32, 128, and for two sets of hyper-parameters corresponding to the two scenarios described in §5. In the first scenario, we set $(\hat{\theta}, \hat{\Sigma}) = (\theta_m, 2\Sigma_m)$. The bottom row of Fig. 5 compares log-errors produced by our method (left), with those of importance sampling based on the optimal Gaussian proposal (with mean θ_m, variance Σ_m), and the same number of likelihood evaluations (as reported on the x-axis of the right plot). In the second scenario, we set $(\hat{\theta}, \hat{\Sigma}) = (\theta_m, 100\, I_d)$. The top row compares log-errors produced by our method (left) with those of importance sampling, based again on the optimal proposal, and the same number of likelihood evaluations. The variance of importance

sampling estimates based on a Gaussian proposal with hyper-parameters $\hat{\theta}$ and $\hat{\Sigma} = 100 I_d$ is higher by several order of magnitudes, and is not reported in the plots.

As expected, the first strategy outperforms standard importance sampling, when both methods are supplied with the same information (mode, Hessian), and the second strategy still does reasonably well compared to importance sampling based on the optimal Gaussian proposal, although only provided with the mode. For too small values of N, however, nested importance sampling is slightly biased.

As pointed out by one referee, results are sufficiently precise that one can afford to compute the evidence for the 2^7 possible models: the most likely model, with posterior probability 0.81, includes the intercept, the three variables mentioned above (distance, arsenic, education) and one cross-effect between distance and education level, and the second most likely model, with posterior probability 0.18, is the same model but without the cross-effect.

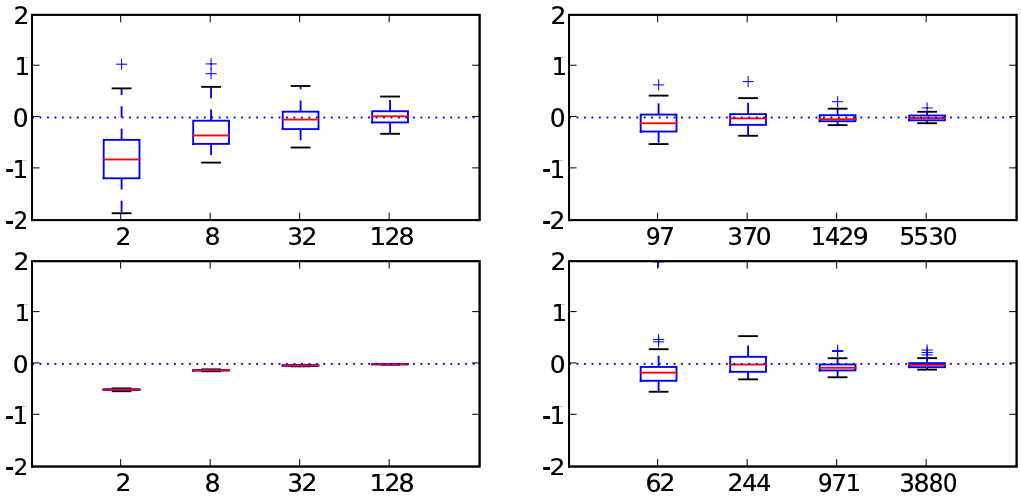

Fig. 5. Probit example: Box-plots of *(left column)* log-errors of nested importance sampling estimates, for $N = 2, 8, 32, 128$, compared with the log-error of importance sampling estimates *(right column)* based on the optimal Gaussian proposal, and the same number of likelihood evaluation (reported on the x axis of the right column plots). Bottom row corresponds to the first strategy (both mode and Hessian available), top row corresponds to the second strategy (only mode available).

8. Conclusion

We have shown that nested sampling is a valid Monte Carlo method, with convergence rate $\mathrm{O}(N^{-1/2})$, which enjoys good performance in some applications, for example those where the posterior is approximately Gaussian, but which may also provide less satisfactory results in some difficult situations. Further work on the formal and practical assessment of nested sampling convergence would be welcomed. The convergence properties of MCMC-based nested sampling are unknown and technically challenging. Methodologically, efforts are required to design efficient MCMC moves with respect to the constrained prior. In that and other respects, nested importance sampling may be a useful extension. Ultimately, our comparison between nested sampling and alternatives should be extended to many more examples, to get a clearer idea of when nested

673 sampling should be the method of choice and when it should not. All the programs implemented
674 for this paper are available from the authors.
675
676
677 ## ACKNOWLEDGEMENTS

678 The authors are grateful to R. Denny, A. Doucet, T. Loredo, O. Papaspiliopoulos, and G.
679 Roberts, for helpful comments and discussions. Discussions with J. Skilling also helped in clari-
680 fying our understanding of the implementation of the method. The second author was supported
681 by the 2005 project ANR-05-BLAN-0299 Adap'MC.
682
683
684 ## REFERENCES
685 ## REFERENCES

686 BARTOLUCCI, F., SCACCIA, L. & MIRA, A. (2006). Efficient Bayes factor estimation from the reversible jump
687 output. *Biometrika* **93**, 41–52.
688 BROCKWELL, P. & DAVIS, P. (1996). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics.
Springer-Verlag, New York.
689 BURROWS, B. L. (1980). A new approach to numerical integration. *IMA J. Appl. Math.* **26**, 151–173.
690 CELEUX, G., HURN, M. & ROBERT, C. (2000). Computational and inferential difficulties with mixtures posterior
691 distribution. *J. American Statist. Assoc.* **95(3)**, 957–979.
692 CHEN, M. & SHAO, Q. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*
25, 1563–1594.
693 CHEN, M., SHAO, Q. & IBRAHIM, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag,
694 New York.
695 CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.* **90**, 1313–1321.
696 CHOPIN, N. & ROBERT, C. (2007). Comments on 'Nested Sampling' by John Skilling. In *Bayesian Statistics 8*,
O. U. P. Bernardo, J. M. et al. (eds), ed.
697 DIEBOLT, J. & ROBERT, C. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist.*
698 *Society Series B* **56**, 363–375.
699 EVANS, M. (2007). Discussion of nested sampling for Bayesian computations by John Skilling. In *Bayesian Statistics*
8, J. Bernardo, M. Bayarri, J. Berger, A. David, D. Heckerman, A. Smith & M. West, eds. Oxford University Press,
700 pp. 491–524.
701 FRÜHWIRTH-SCHNATTER, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models
using bridge sampling techniques. *The Econometrics Journal* **7**, 143–167.
702 GELFAND, A. & DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Royal Statist.*
703 *Society Series B* **56**, 501–514.
704 GELMAN, A. & HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge,
UK: Cambridge University Press.
705 GELMAN, A. & MENG, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling
706 to path sampling. *Statist. Science* **13**, 163–185.
707 GREEN, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–
732.
708 HAN, C. & CARLIN, B. (2001). MCMC methods for computing Bayes factors: a comparative review. *J. American*
709 *Statist. Assoc.* **96**, 1122–1132.
710 HESTERBERG, T. (1998). Weighted average importance sampling and defensive mixture distributions. *Technometrics*
711 **37**, 185–194.
712 JASRA, A., HOLMES, C. & STEPHENS, D. (2005). Markov Chain Monte Carlo methods and the label switching
problem in Bayesian mixture modeling. *Statist. Sci.* **20**, 50–67.
713 JEFFREYS, H. (1939). *Theory of Probability*. Oxford: The Clarendon Press, 1st ed.
714 KALLENBERG, O. (2002). *Foundations of Modern Probability*. Springer-Verlag, New York.
715 KIM, S., SHEPHARD, N. & CHIB, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH
models. *Rev. Econom. Stud.* **65**, 361–393.
716 LOUIS, T. (1982). Finding the observed information matrix when using the EM algorithm. *J. Royal Statist. Society*
717 *Series B* **44**, 226–233.
718 MACLACHLAN, G. & KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley.
719 MENG, X. & SCHILLING, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.* **11**, 552–586.
720 MENG, X. & WONG, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical
exploration. *Statist. Sinica* **6**, 831–860.

RICHARDSON, S. & GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B* **59**, 731–792.
ROBERT, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, 2nd ed.
ROBERT, C. & CASELLA, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 2nd ed.
ROBERTS, G. & ROSENTHAL, J. (1998). Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Canadian J. Statist.* **26**, 5–32.
ROBERTS, G. & ROSENTHAL, J. (1999). Convergence of slice sampler Markov chains. *J. Royal Statist. Society Series B* **61**, 643–660.
RYDÉN, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Stochastic Models* **10**, 795–829.
SKILLING, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis* **1(4)**, 833–860.
SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Society Series B* **64**, 583–639.

APPENDIX 1

Proof of Lemma 1

It is sufficient to prove this result for functions $\widetilde{f}$ that are real-valued, positive and increasing. First, the extension to vector-valued functions is trivial, so $\widetilde{f}$ is assumed to be real-valued from now on. Second, the class of functions that satisfy property (4) is clearly stable through addition. Since $\widetilde{f}$ is absolutely continuous, there exist functions f^+ and f^-, such that f^+ is increasing, f^- is decreasing, and $\widetilde{f} = f^+ + f^-$, so we can restrict our attention to increasing functions. Third, absolute continuity implies bounded variation, so it always possible to add an arbitrary constant to $\widetilde{f}$ to transform it into a positive function.

Let $\psi : l \to l\widetilde{f}(l)$, which is a positive, increasing function and denote its inverse by ψ^{-1}. One has:

$$\mathbb{E}^\pi[\psi\{L(\theta)\}] = \int_0^{+\infty} P^\pi(\psi\{L(\theta)\} > l)\,\mathrm{d}l = \int_0^{+\infty} \varphi^{-1}\{\psi^{-1}(l)\}\,\mathrm{d}l = \int_0^1 \psi\{\varphi(x)\}\,\mathrm{d}x\,,$$

which concludes the proof.

APPENDIX 2

Proof of Theorem 1

Let $t_i = x^\star_{i+1}/x^\star_i$, for $i = 0, 1, \ldots$. As mentioned by Skilling (2006), the t_i's are independent Beta$(N, 1)$ variates. Thus, $u_i = t_i^N$ defines a sequence of independent uniform $[0, 1]$ variates. A Taylor expansion of e_N gives:

$$\begin{aligned} e_N &= \sum_{i=1}^{\lceil cN \rceil} (x_{i-1} - x_i)\,[\varphi(x^\star_i) - \varphi(x_i)] \\ &= \sum_{i=1}^{\lceil cN \rceil} (x_{i-1} - x_i)\left[\psi'(-\log x_i)\,(\log x_i - \log x^\star_i) + O\,(\log x_i - \log x^\star_i)^2\right] \end{aligned}$$

where $c = -\log\varepsilon$, and $\psi(y) = \varphi(e^{-y})$. Note that

$$S_i = N\,(\log x_i - \log x^\star_i) = \sum_{k=0}^{i-1}(-1 - \log u_k)$$

is a sum of independent, standard variables, as $\mathbb{E}[\log u_i] = -1$ and $\mathrm{var}[\log u_i] = 1$. Thus, $(\log x_i - \log x_i^\star) = \mathrm{O}_P(N^{-1/2})$, where the implicit constant in $\mathrm{O}_P(N^{-1/2})$ does not depend on i, and

$$N^{1/2} e_N = N^{-1/2} \sum_{i=1}^{\lceil cN \rceil} (e^{-(i-1)/N} - e^{-i/N}) S_i \left[\psi'(\frac{i}{N}) + O_P(N^{-1/2}) \right]$$
$$= c^{1/2} \sum_{i=1}^{\lceil cN \rceil} \int_{(i-1)/N}^{i/N} e^{-t} \psi'(t) B_N(\frac{t}{c}) \, \mathrm{d}t \left[1 + O_P(N^{-1/2}) \right]$$

since $\psi'(t) = \psi'(i/N) + \mathrm{O}(N^{-1})$ for $t \in [(i-1)/N, i/N]$, where, again, the implicit constant in $\mathrm{O}(N^{-1})$ can be the same for all i, as ψ'' is bounded, and provided $B_N(t)$ is defined as $B_N(t) = (cN)^{-1/2} S_{\lceil cNt \rceil}$ for $t \in [0,1]$. According to Donsker's theorem (Kallenberg, 2002, p.275), B_N converges to a Brownian motion B on $[0,1]$, in the sense that $f(B_N)$ converges in distribution to $f(B)$ for any measurable and a.s. continuous function f. Thus

$$N^{1/2} e_N = c^{1/2} \int_0^{\lceil cN \rceil / N} e^{-t} \psi'(t) B_N(\frac{t}{c}) \, \mathrm{d}t + O_P(N^{-1/2}) \overset{d}{\rightsquigarrow} c^{1/2} \int_0^c e^{-t} \psi'(t) B(\frac{t}{c}) \, \mathrm{d}t \,,$$

which has the same distribution as the following zero-mean Gaussian variate:

$$\int_0^c e^{-t} \psi'(t) B(t) \, \mathrm{d}t = \int_\varepsilon^1 s \varphi'(s) B(-\log s) \, \mathrm{d}s.$$

APPENDIX 3

Proof of Lemma 2

For the sake of clarity, we make dependencies on d explicit in this section, e.g. φ_d for φ, ε_d for ε, etc. We will use repeatedly the facts that φ is nonincreasing and that φ' is nonnegative. One has:

$$- \int_{s,t \in [\varepsilon_d, 1]} s \varphi_d'(s) t \varphi_d'(t) \log(s \vee t) \, \mathrm{d}t \leq - \log \varepsilon_d \left(\int_{\varepsilon_d}^1 s \varphi_d'(s) \, \mathrm{d}s \right)^2 \leq d \log(\sqrt{2}/\tau)$$

for $d \geq 1$, since $- \int_{\varepsilon_d}^1 s \varphi_d'(s) \, \mathrm{d}s \leq - \int_0^1 s \varphi_d'(s) \, \mathrm{d}s = 1$. This gives the first result.

Let $s_d = \varphi_d^{-1}(\alpha^d)$, for $0 < \alpha < 1$; s_d is the probability that

$$(4\pi/d) \sum_{i=1}^d \theta_i^2 - 1 \leq -2 \log(\alpha/\sqrt{2}) - 1$$

assuming that the θ_i's are i.i.d $\mathcal{N}(0, 1/4\pi)$ variates. The left-hand side is an empirical average of i.i.d. zero-mean variables. We take α so that the right-hand side is negative, i.e. $\alpha > \sqrt{2} \exp(-1/2)$. Using large deviations (Kallenberg, 2002, Chapter 27), one has $-\log(s_d)/d \to \gamma > 0$ as $d \to +\infty$, and

$$\frac{1}{d} V_d = -\frac{1}{d} \int_{s,t \in [\varepsilon_d, 1]} s \varphi_d'(s) t \varphi_d'(t) \log(s \vee t) \, \mathrm{d}s \mathrm{d}t \geq \left(\frac{-\log s_d}{d} \right) \left(\int_{\varepsilon_d}^{s_d} s \varphi_d'(s) \, \mathrm{d}s \right)^2$$
$$\geq \left(\frac{-\log s_d}{d} \right) \left(\int_{\varepsilon_d}^{s_d} \varphi_d(s) \, \mathrm{d}s + \varepsilon_d \varphi_d(\varepsilon_d) - s_d \varphi_d(s_d) \right)^2$$
$$\geq \left(\frac{-\log s_d}{d} \right) \left(1 - \int_0^{\varepsilon_d} \varphi_d(s) \, \mathrm{d}s - \int_{s_d}^1 \varphi_d(s) \, \mathrm{d}s + \varepsilon_d \varphi_d(\varepsilon_d) - s_d \varphi_d(s_d) \right)^2 .$$

As $d \to +\infty$, $-\log(s_d)/d \to \gamma$, $s_d \to 0$, $\varphi_d(s_d) = \alpha^d \to 0$, $\int_{s_d}^1 \varphi_d(s) \, \mathrm{d}s \leq \varphi_d(s_d)(1 - s_d) \to 0$, and

$$0 \leq \int_0^{\varepsilon_d} \varphi_d(s) \, \mathrm{d}s - \varepsilon_d \varphi_d(\varepsilon_d) \leq \varepsilon_d [\varphi_d(0) - \varphi_d(\varepsilon_d)] \leq \tau < 1,$$

by the definition of ε_d, and the squared factor is in the limit greater than or equal to $(1-\tau)^2$.