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

Nicolas Chopin, Christian Robert

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

Nicolas Chopin[1] and Christian P. Robert[1,2]
[1]CREST–ENSAE and [2]Université Paris Dauphine

January 25, 2008

Abstract

Nested sampling is a novel simulation method for approximating marginal likelihoods, proposed by Skilling (2006, 2007). We establish that nested sampling leads to an error that vanishes at the standard Monte Carlo rate $N^{-1/2}$, where N is a tuning parameter that is proportional to the computational effort, and that this error is asymptotically Gaussian. We show that the corresponding asymptotic variance typically grows linearly with the dimension of the parameter. We use these results to discuss the applicability and efficiency of nested sampling in realistic problems, including posterior distributions for mixtures. We propose an extension of nested sampling that makes it possible to avoid resorting to MCMC to obtain the simulated points. We study two alternative methods for computing marginal likelihood, which, in contrast with nested sampling, are based on draws from the posterior distribution and we conduct a comparison with nested sampling on several realistic examples.

Keywords: Convergence rate, MCMC, Monte Carlo approximation, mixtures of distributions, importance sampling, simulation.

1 Introduction

Nested sampling was introduced by Skilling (2007, 2006) as a numerical approximation method for integrals of the kind

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

when π is the prior distribution and $L(\theta|x)$ is the likelihood, called *evidence* in the above papers. These quantities naturally occur as marginals in Bayesian testing theory and Bayesian model choice (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)

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

open specific avenues of approximation (see, e.g., Chen and Shao, 1997; Gelman and Meng, 1998).

One important aspect of nested sampling is that it resorts to simulating iteratively points θ_i from the prior π, constrained to θ_i having a larger likelihood value than some increasing threshold l; the exact principle of nested sampling is described in the next section. In a previous discussion (Chopin and Robert, 2007), we wondered about both the universality and the convergence properties of the method. With respect to the former, we pointed out that simulating from a constrained distribution is not always straightforward. With respect to the latter, we noted that Skilling (2007, 2006) does not provide any formal assessment of the method.

The purpose of this paper is to investigate formally both points presented above. Our main result is to establish formally the convergence properties of nested sampling estimates: we show that the approximation error is dominated by a stochastic term, which is $O(N^{-1/2})$ and has a limiting Gaussian distribution, where N is a tuning parameter, and is proportional to the computational effort. In that respect, nested sampling seems to have properties comparable to most Monte Carlo algorithms.

Then, we show in a simple example that the asymptotic variance of nested sampling estimates typically grows linearly with the dimension d of the problem, and that the overall computational cost is $O(d^3/e^2)$, where e the desired level of accuracy. Note this result assumes that one simulates *exactly* from the constrained prior. Unfortunately, they are numerous cases where such a constrained simulation cannot be performed. Skilling (2007, 2006) advocates using MCMC for obtaining the simulated points. We discuss this approach and provide numerical evidence that the obtained estimates then suffer from a more severe curse of dimensionality.

Since the ability to simulate from the constrained prior seems to be determinant in the applicability of the algorithm, we further propose an extension of nested sampling, based on the principle of importance sampling, that introduces enough flexibility so as to allow for performing the constrained simulation without resorting to MCMC. Finally, we discuss two alternatives to nested sampling for computing evidence, which are both based on the output of MCMC algorithms. These alternatives are quite comparable with nested sampling in terms of convenience: nested sampling, as shown by Skilling (2007, 2006), provides approximations of posterior quantities at no extra cost. Conversely, the methods we propose allow for recycling the MCMC output, primarily used for computing posterior quantities, so as to approximate the evidence. We provide numerical comparisons of those three methods.

The paper is organised as follows. Section 2 describes the nested sampling algorithm. Section 3 gives a formal analysis of the approximation error of nested sampling estimates, and shows that this error is dominated by a stochastic term that has a limiting Gaussian distribution and vanishes at rate $N^{-1/2}$. Section 4 discusses some practical limitations of the algorithm, and shows that the variance of produced estimates tends to grow linearly with the dimension of the problem. Section 5 describes the extension of nested sampling called 'nested importance sampling'. Section 6 describes the two alternative methods based on MCMC mentioned above.

2 Nested sampling: A description

For the sake of completeness, and in order to set notations, we describe briefly the nested sampling algorithm; see Skilling (2006, 2007) for more details. We use $L(\theta)$ as a short-hand for $L(x|\theta)$ from now on, omitting the dependence on x.

2.1 Principle

Nested sampling is based on the following generic if formal representation:

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

where φ is the inverse of

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

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

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

where $\varphi_i = \varphi(x_i)$, and $0 < x_j < \ldots < x_1 < x_0 = 1$ is an arbitrary grid over $[0, 1]$.

In most cases however, including some simple toy examples, the function φ is not tractable. Instead, the values φ_i are approximated by an iterative random mechanism:

- Iteration 1: draw independently N points $\theta_{1,i}$ from π, denote by θ_1 the value such that $L(\theta_{1,i})$ is smallest,
$$\theta_1 = \arg\min_i L(\theta_{1,i}),$$
and set $\varphi_1 = L(\theta_1)$.
- Iteration 2: obtain the N 'current' values $\theta_{2,i}$, by reproducing the N previous points $\theta_{1,i}$, except for θ_1 that is replaced by a draw from the prior distribution π conditional on $L(\theta) > \varphi_1$; then select θ_2 as the point such that $L(\theta_{2,i})$ is smallest, and set $\varphi_2 = L(\theta_2)$.
- Iterate the above step until a given stopping iteration j is reached, for instance observing very small changes in the approximation $\widehat{\mathfrak{Z}}$ or reaching the maximal value of $L(\theta)$ when the likelihood is bounded and its maximum is known.

The output of the algorithm is then the approximation of $\mathfrak{Z}$ by the sum (2), when the x_i's are replaced with $x_i = \exp(-i/N)$, as detailed below. As stressed out in the introduction, a key ingredient of this algorithm is the ability to simulate from the prior distribution π under the constraint $L(\theta) > l$, for $l > 0$.

Let $x_i^{\star} = \varphi^{-1}(\varphi_i)$. An interesting property of this generating process is that the quantities 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, 2007) takes advantage of this property by setting $x_i = \exp(-i/N)$, so that $\log x_i$ is the expectation of $\log \varphi^{-1}(\varphi_i)$ (we will call this approach the deterministic scheme). Alternatively, Skilling (2006) also proposes a *random scheme* where the x_i's are random, by mimicking the law of the t_i's, i.e. $x_{i+1} = x_i * t_i$, where $t_i \sim \mathrm{Beta}(N, 1)$. Note that in both cases the relation $\varphi_i = \varphi(x_i)$ does not hold; at best, φ_i can be interpreted as a 'noisy' version of $\varphi(x_i)$. Our discussion concentrates on the deterministic scheme, since it seems to us that the random scheme only independently creates additional noise and thus does not improve the precision of the approximation of $\mathfrak{Z}$, established in Section 3.

2.2 Variations and posterior simulation

Skilling (2006, 2007) mentions the possibility of replacing (2) with an higher-order quadrature approximation:

$$\widehat{\widehat{\mathfrak{Z}}} = \sum_{i=1}^{j} (x_{i-1} - x_i)(\varphi_i + \varphi_{i-1})/2, \tag{3}$$

assuming $\varphi_0 = 0$, so as to reduce the deterministic error to $O(N^{-2})$ in the quadrature (following results of Yakowitz et al., 1978 and Philippe, 1997b). Since $x_i = \exp(-i/N)$, one has

$$\widehat{\widehat{\mathfrak{Z}}} - \widehat{\mathfrak{Z}} = \left(\frac{e^{1/N} - 1}{2} \right) \widehat{\mathfrak{Z}} - \frac{1}{2} \varphi_j (e^{-j/N} - e^{-(j+1)/N}) = O(N^{-1}).$$

We shall prove in Section 3 than the approximation error of $\widehat{\mathfrak{Z}}$ is dominated by a $O(N^{-1/2})$ stochastic term. Thus, the improvement obtained by replacing $\widehat{\mathfrak{Z}}$ with $\widehat{\widehat{\mathfrak{Z}}}$ is negligible relatively to the approximation error itself.

Skilling (2006, 2007) points out that nested sampling can provide 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 $\omega_i L_i$ (Skilling, 2006)."

(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

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

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 $\mathfrak{Z}$) based on the same random sample. The denominator is estimated by (2), and the numerator is estimated by

$$\sum_{i=1}^{j} (x_{i-1} - x_i) \varphi_i f(\theta_i) \tag{4}$$

which is a noisy version of

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

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

$$\int_0^1 \varphi(x) \widetilde{f}\{\varphi(x)\}\, \mathrm{d}x\,,$$

which is shown to be equal to the numerator in the following lemma.

Lemma 2.1. *Let* $\widetilde{f}(l) = \mathbb{E}^{\pi}[f(\theta)|L(\theta) = l]$ *for* $l > 0$*, then, if* $\widetilde{f}$ *is absolutely continuous,*

$$\int_0^1 \varphi(x)\widetilde{f}\{\varphi(x)\}\, dx = \int \pi(\theta)L(\theta)f(\theta)\, d\theta. \tag{5}$$

A proof is provided in Appendix A. Clearly, the estimate of $\mu(f)$ obtained by dividing (4) with (3) is equal to the estimate obtained by computing the weighted average mentioned above. We do not discuss further the extension of nested sampling to posterior inference, but mention that it seems reasonably easy to extend our convergence results to such approximations of posterior estimators. We also note that, in practice, the distribution of the weights $w_i L_i$ may often be highly asymmetric, with a few weights dominating the others; thus such approximations may often have a large variance.

3 A central limit theorem for nested sampling

Using the (deterministic) nested sampling scheme detailed in the previous section, we now undertake a rigorous study of its convergence properties, with the outcome that nested sampling gives an error that vanishes at rate $N^{-1/2}$ and is asymptotically Gaussian.

To this effect, we decompose the approximation error as follows:

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

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 the first iteration j such that $x_j = e^{-j/N} \leq \varepsilon$, i.e. $j = \lceil(-\log\varepsilon)N\rceil$. (More practical stopping rules will be discussed in Section 7). Assuming φ is bounded from above, or equivalently L is 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 $O(N^{-1})$, since $x_{i-1} - x_i = O(N^{-1})$.

3. The third term is stochastic, and is denoted

$$e_N = \sum_{i=1}^{\lceil(-\log\varepsilon)N\rceil} (x_{i-1} - x_i)\left[\varphi(x_i^{\star}) - \varphi(x_i)\right] .$$

   where the $x_i^{\star}$ 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 given by the following theorem.

Theorem 3.1. *Provided φ is twice continuously-differentiable over $[\varepsilon, 1]$,*

$$N^{1/2} e_N \overset{d}{\rightsquigarrow} \int_{\varepsilon}^{1} s\varphi'(s) B(-\log s)\,\mathrm{d}s$$

where $\overset{d}{\rightsquigarrow}$ denotes convergence in distribution, and $B(\cdot)$ is a standard Brownian motion over the positive real line. The right term above defines a zero-mean Gaussian random variable, with variance:

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

The stochastic error is of order $O_P(N^{-1/2})$ and it dominates both other error terms. The proof of this theorem relies on the functional central limit theorem (also known as Donsker's theorem), and is detailed in Appendix B.

In conclusion, the nested sampling approximation enjoys the same convergence properties as standard Monte Carlo methods, rather than potentially higher convergence rates deduced from the formal application of numerical approximations, because the approximation error due to random sampling is the dominating term.

4 Properties of the nested sampling algorithm

4.1 Simulating from a constrained prior

It seems to us that, despite its satisfactory convergence properties, the main drawback of nested sampling is that it requires simulating θ from the prior distribution $\pi(\theta)$ subject to the constraint $L(\theta) > L(\theta_i)$; this is an intractable problem in many realistic set-ups. It is actually of the same complexity as a one-dimensional slice sampler (see, e.g., Robert and Casella, 2004, Chapter 8), which produces an uniformly ergodic Markov chain when the likelihood L is bounded.

To overcome this difficulty, Skilling (2006, 2007) proposes to sample values of θ by iterating, say, k MCMC steps, using the truncated prior as the invariant distribution and the point selected at the previous iteration as the starting value. While this implementation is formally possible, it raises several concerns. First, this introduces an awkward bias in the procedure, depending on the choice of k, since the k-th iterate of a MCMC chain is not distributed according to the constrained prior. In fact, the convergence result established above cannot be extended in full generality to the case where the algorithm resorts to k MCMC steps for sampling θ. In Section 7.1, we report simulation results that expose this 'MCMC bias'.

Second, there are settings when implementing an MCMC move that leaves the truncated prior invariant is far from straightforward. In that case, one may instead implement a regular MCMC move (e.g., a random walk Metropolis-Hastings proposal) 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 and closer to the highest values of L, given the diminishing weight under the prior π. 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 Section 5, we propose an extension of nested sampling, based on the principle of importance sampling, which makes it easier to avoid MCMC in this hard-constrained prior simulation step.

4.2 Impact of dimensionality

Although nested sampling focusses on the integral (1) that is always unidimensional, we show in this section that its theoretical performance typically depends on the dimension of the problem in the following way: the required number of iterations (for a fixed truncation error), and the asymptotic variance both grow linearly with the dimension d of θ.

A corollary of this result is that, under the assumption that the cost of a single iteration is $O(d)$, which should be the best possible case, the computational cost of nested sampling is $O(d^3/e^2)$, where e denotes a given error level. Note this discussion applies to the *exact* nested algorithm. For MCMC-based nested sampling, performances seem to deteriorate exponentially with the dimension in our simulation experiments; see §7.1.

Consider the following toy example: for $k = 1, \ldots, d$,

$$\theta^{(k)} \sim \mathcal{N}(0, \sigma_0^2) \quad \text{and} \quad y^{(k)}|\theta^{(k)} \sim \mathcal{N}(\theta^{(k)}, \sigma_1^2)\,,$$

independently in both cases. Set $y^{(k)} = 0$ and $\sigma_0^2 = \sigma_1^2 = 1/4\pi$, in order that

$$\mathfrak{Z} = \prod_{i=1}^{d} \varphi(y^{(k)}; 0, \sigma_0^2 + \sigma_1^2) = 1$$

for all values of d. (This simplifies comparisons across dimensions.)

In this toy example, exact simulation from the constrained prior can be performed as follows: simulate r^2 from a $\chi^2(d)$ distribution truncated to $r^2 \leq -\sqrt{2}\log l$ (using, e.g. inverse method or Philippe, 1997a), then simulate $u_1, \ldots, u_d \sim \mathcal{N}(0,1)$, and set $\theta^{(k)} = r\, u_k / \sqrt{u_1^2 + \ldots + u_d^2}$.

Since $\mathfrak{Z} = \int_0^1 \varphi(x)\,\mathrm{d}x = 1$, we assume that the truncation point ε_d is chosen so that $\varphi(0)\varepsilon_d = \tau \ll 1$, $\tau = 10^{-6}$ say, where $\varphi(0) = (2\pi\sigma_1^2)^{-d/2} = 2^{d/2}$ 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:

Lemma 4.1. *For the toy example introduced above, denoting by V_d the asymptotic variance of the nested sampling estimator (with truncation point set to $\varepsilon_d = \tau 2^{-d/2}$), 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 proven in Appendix C. It is easy to generalise these results to any toy example 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 at least linearly with the 'actual' dimensionality of the problem, as measured for instance by the DIC criterion of Spiegelhalter et al. (2002).

5 Nested importance sampling

In this section, we show how to extend nested sampling so as to provide enough flexibility to avoid resorting to MCMC. This extension is very close in spirit to importance sampling, and we therefore call it 'nested importance sampling'. Let $\widetilde{\pi}(\theta)$ denote an instrumental prior, and $\widetilde{L}(\theta)$ an instrumental 'likelihood' (that can in fact be any positive and measurable function). Define the weight function $w(\theta)$ so that:

$$\widetilde{\pi}(\theta)\widetilde{L}(\theta)w(\theta) = \pi(\theta)L(\theta)$$

(We assume that the support of π is included in the support of $\widetilde{\pi}$.)

We can then approximate $\mathfrak{Z}$ by applying nested sampling to the pair $(\widetilde{\pi}, \widetilde{L})$, that is, simulating iteratively from $\widetilde{\pi}$ constrained to $\widetilde{L}(\theta) > l$, and computing the following generalised nested sampling estimator, which we introduced in (4):

$$\sum_{i=1}^{j}(x_{i-1} - x_i)\varphi_i w(\theta_i). \tag{6}$$

The advantage of this extension is that one can choose $(\widetilde{\pi}, \widetilde{L})$ in such a way that simulating from $\widetilde{\pi}$ constrained to $\widetilde{L}(\theta) > l$ does not require MCMC steps. As an example, consider the following strategy. Take $\widetilde{\pi}$ as Gaussian, say, $\mathcal{N}_d(\hat{\theta}, \tau^2 I_d)$, where d is the dimension of θ, and take $\widetilde{L}$ as

$$\widetilde{L}(\theta) = \lambda(\|\theta - \hat{\theta}\|^2),$$

where $\hat{\theta}$ is an arbitrary centre and $\lambda(\cdot)$ is a decreasing function, so that the constraint $L(\theta) < l = L(\theta_i)$ defines a ball centred at $\hat{\theta}$. We have already explained in Section 4.2 how to simulate such a constrained Gaussian distribution.

Interestingly, we do not need to specify the exact expression of λ, since the estimator (6) does not depend on this function:

$$\varphi_i w(\theta_i) = \widetilde{L}(\theta_i)w(\theta_i) = \frac{\pi(\theta_i)L(\theta_i)}{\widetilde{\pi}(\theta_i)},$$

and the only tuning parameters of the algorithm are the hyper-parameters $\hat{\theta}$ and τ of the instrumental prior $\widetilde{\pi}$. In our simulations, we found out that if $\hat{\theta}$ was set to the (true) posterior mode, and τ to some large value, the obtained estimator showed good performance, as illustrated in §7.3. Therefore, in situations where one can obtain a good approximation of the (true) posterior mode, but not necessarily of the exact shape of the posterior distribution, the extended algorithm seems an interesting alternative to simple methods like importance sampling, which requires more tuning. Note however that nested importance sampling should suffer from the same curse of dimensionality as standard importance sampling: the weight function $w(\theta)$ should get more and more skewed as the dimension d of θ increases.

6 Alternative algorithms

We have seen in Section 2.2 that the output of nested sampling can be "recycled" so as to provide (at no extra cost and in addition to the evidence) approximations of posterior quantities. The aim of this section is to show that, contrary to common belief, this can also be achieved with

MCMC. More precisely, it is possible to recycle the output of an MCMC algorithm, which was used first for computing posterior quantities, so as to estimate the evidence, with no or little additional programming effort. (Earlier works pointing out this possibility include Gelfand and Dey, 1994 and Chen and Shao, 1997. See also Bartolucci et al., 2006 for a more recent perspective.)

6.1 Approximating $\mathfrak{Z}$ from a posterior sample

A first simple solution is to use a reversible jump MCMC algorithm (Green, 1995). However, it may be argued that this is due to the simultaneous simulation of parameters from several models and thus that the approximation of $\mathfrak{Z}$ uses this extra amount of simulation. But this is not the case: 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 single 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)

$$\varrho_{\mathfrak{M}\to\mathfrak{M}'} = \frac{1/2\varphi(\theta)}{1/2\pi(\theta)L(\theta)} \wedge 1$$

and from $\mathfrak{M}'$ to $\mathfrak{M}$ with probability

$$\varrho_{\mathfrak{M}'\to\mathfrak{M}} = \frac{1/2\pi(\theta)L(\theta)}{1/2\varphi(\theta)} \wedge 1\,,$$

$\varphi(\theta)$ being an arbitrary proposal on θ that corresponds to the move from $\mathfrak{M}'$ to $\mathfrak{M}$. Were we to actually run this reversible jump MCMC algorithm, the frequency of visits to $\mathfrak{M}$ would then be proportional to $\mathfrak{Z}$.

Now, an interesting remark is that we do not need to run this formal reversible sampler to get an estimate of $\mathfrak{Z}$: indeed, if we run a standard MCMC algorithm on θ and compute the probability of moving to $\mathfrak{M}'$, the expectation of the ratio $\varphi(\theta)/\pi(\theta)L(\theta)$ (under stationarity) is equal to the inverse of $\mathfrak{Z}$:

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

no matter what the proposal $\varphi(\theta)$ is, in the spirit of Bartolucci et al. (2006). Obviously, the choice of $\varphi(\theta)$ matters in the precision of the approximation to $\mathfrak{Z}$ and we suggest using a kernel approximation to $\pi(\theta|x)$ based on earlier MCMC simulations. Note that, from an importance sampling point of view, we are faced with a constraint opposite to the usual one, namely that $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi(\theta)L(\theta)$ for the approximation

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

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

In the comparison below (§7.2), we run a comparison of $\widehat{\mathfrak{Z}_1}$ with a more standard importance sampling approximation

$$\widehat{\mathfrak{Z}_2} = \frac{1}{T}\sum_{t=1}^{T} \frac{\pi(\theta^{(t)})L(\theta^{(t)})}{\varphi(\theta^{(t)})}$$

where the $\theta^{(t)}$'s are now generated from the density $\varphi(\theta)$, which can also be a non-parametric approximation of $\pi(\theta|x)$, this time with heavier tails than $\pi(\theta)L(\theta)$.

6.2 Approximating $\mathfrak{Z}$ using a mixture representation

Another (related) approach in the approximation of $\mathfrak{Z}$ is to design a specific mixture for simulation purposes, with density (up to a normalising constant):

$$\omega_1\pi(\theta)L(\theta) + \varphi(\theta)\,,$$

where again $\varphi(\theta)$ is an arbitrary (fully specified) density. Since simulating from this mixture offers the same complexity as simulating from the posterior, an extension of the MCMC code used to simulate from $\pi(\theta|x)$ can be used to simulate from the mixture, based on the introduction of an auxiliary variable δ that indicates whether or not the current simulation is from $\pi(\theta|x)$ or from $\varphi(\theta)$. The basic MCMC steps are as follows:
At iteration t

1. Take $\delta^{(t)} = 1$ with probability

$$\omega_1\pi(\theta^{(t-1)})L(\theta^{(t-1)}) \Big/ \left(\omega_1\pi(\theta^{(t-1)})L(\theta^{(t-1)}) + \varphi(\theta^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;

2. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim \text{MCMC}(\theta^{(t-1)}, \theta^{(t)})$ where $\text{MCMC}(\theta, \theta')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi(\theta|x) \propto \pi(\theta)L(\theta)$;

3. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim \varphi(\theta)$ independently from the previous value $\theta^{(t-1)}$

Note this algorithm is a Gibbs sampler: Step 1 simulates $\delta^{(t)}$ conditional on $\theta^{(t-1)}$, while Steps 2 and 3 simulate $\theta^{(t)}$ conditional on $\delta^{(t)}$. It is immediate to check that the average of the $\delta^{(t)}$'s converges to $\omega_1\mathfrak{Z}/\{\omega_1\mathfrak{Z}+1\}$. A natural Rao-Blackwell improvement is to take the average of the expectations of the $\delta^{(t)}$'s, i.e.

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

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

Note also that, while this is not the primary concern of this study, simulation from the above algorithm induces a natural regeneration scheme that can improve convergence assessment for the overall MCMC scheme (Robert and Casella, 2004).

6.3 Error approximations

Usual confidence intervals can be produced on the empirical averages $1/\widehat{\mathfrak{Z}_1}$, $\widehat{\mathfrak{Z}_2}$ and $\omega_1\widehat{\mathfrak{Z}}_3/\{\omega_1\widehat{\mathfrak{Z}}_3+1\}$. One-to-one transforms then produce confidence intervals on the $\widehat{\mathfrak{Z}_i}$'s and therefore error estimates on the approximations.

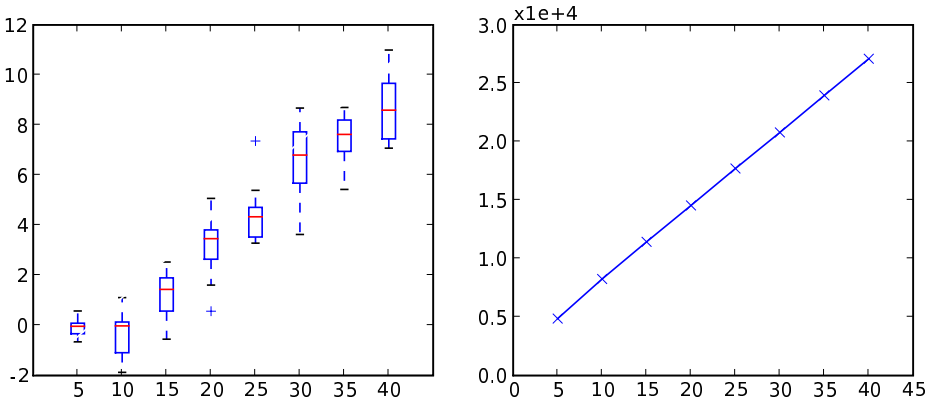

Figure 1: **Box-plots of the log-relative error for different dimensions d (left), and scatter-plot of average number of iterations versus dimension (right), for the decentred Gaussian example.**

7 Numerical experiments

7.1 A decentred Gaussian example

We slightly modify the Gaussian toy example presented in §4.2: $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(d)})$, 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 we now set all the y_k's equal to 3. To simulate from the prior truncated to $L(\boldsymbol{\theta}) > l = L(\boldsymbol{\theta}_0)$, we perform $T = 10$ iterations of a Gibbs sampler with respect to this truncated distribution: the full conditional distribution of $\theta^{(k)}$, given $\theta^{(j)}$, $j \neq k$, is an univariate $\mathcal{N}(0,1)$ distribution that is truncated to the interval $[y^{(k)} - \delta, y^{(k)} + \delta]$ with

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

(Note that the difference inside the squared root is always positive, due to the successive conditional simulations.)

The nested sampling algorithm is run 20 times for $d = 5$, 10, ..., 50, and $N = 100$. 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. Figure 1 reports, for different values of d, the numbers of iterations and the box-plots corresponding to the log-relative error, i.e. $\log \widehat{\mathfrak{Z}} - \log \mathfrak{Z}$, where $\widehat{\mathfrak{Z}}$ is the estimate produced by the algorithm (since $\mathfrak{Z}$ is known in this case). Note that the variability of the number of iterations for a given d is very small, so we only report averages over the 20 runs.

The box-plots exhibit evidence of the bias introduced by MCMC sampling, which seems to increase exponentially with the dimension. Thus, for large dimensions, the efficiency of nested sampling seems to critically depend on the forgetting properties of its MCMC updating strategy. In practice, a way to detect such a bias would be to implement trial-and-error runs, increasing progressively the value of k, the number of MCMC steps performed at each iteration, but this may quickly prove cumbersome.

7.2 A mixture example

The study of the posterior distribution on (μ, σ) associated with the mixture

$$p\mathcal{N}(0,1) + (1-p)\mathcal{N}(\mu, \sigma)\,, \tag{7}$$

when p is known, has several distinctive features that make this example worthwhile considering for a comparison of nested sampling with the more conventional alternatives we recalled above:

- this is a moderately complex if realistic model in that the posterior distribution is not available for computing Bayes estimates but the likelihood associated with a sample $x_1, \ldots, x_n$ from (7) and a value (μ, σ) of the parameter can be computed in linear (i.e. O(n)) time;
- the likelihood is unbounded: when σ converges to 0 and μ converges to any of the x_i's ($1 \leq i \leq n$), the likelihood diverges. This thus represents a challenging problem for exploratory schemes and in particular for nested sampling;
- efficient MCMC strategies have been developed and tested for mixture models since the early 1990's (Diebolt and Robert, 1990, 1994; Richardson and Green, 1997; Celeux et al., 2000; Marin et al., 2004), but Bayes factors are notoriously difficult to approximate in this setting and their dependence on the prior modelling also is noteworthy;
- the two-dimensional nature of the parameter space allows for graphical representations of the posterior surface and for numerical approximations of the marginal $\mathfrak{Z}$, as described below.

We thus designed a numerical experiment where we simulated n observations from (7) with $\mu = -2$ and $\sigma = 3/2$, and then computed the various estimates of $\mathfrak{Z}$ described in the previous sections. Our prior distribution was a uniform both on $(-2, 6)$ for μ and on $(.001, 16)$ for σ^2. (The prior square is chosen arbitrarily to allow all possible values and still retain a compact parameter space.) As described on Figure 3, the likelihood/posterior surface shows tiny bursts in the vicinity of each one of the n observations when σ goes to 0. These unbounded modes are attractors for a method like nested sampling which considers increasing values of the likelihood, but they do not necessarily have a large posterior mass. (The more observations the less relevant those modes are, as shown by Figure 2.) The difference between the MCMC and the nested sampling coverages of the likelihood is clear when comparing Figures 6 and 7, obtained for $n = 6$. The MCMC sample does not visit the six boundary modes corresponding to the six observations in the sample, while the nested sampling sequence accumulates in one of those modes.

As pointed out above, 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 850×950 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. The MCMC algorithm we use is the standard completion of Diebolt and Robert (1994) and it does not suffer from the usual label switching deficiency because (7) is an identifiable model. As shown by the MCMC sample displayed on Figure 3, the exploration of the modal regions by the MCMC chain is satisfactory. We also use this MCMC sampler to simulate from the prior truncated by the likelihood level within the nested sampling algorithm: starting at the current value of (μ, σ), we then ran 50 MCMC steps above the corresponding likelihood value and then weighted those 50 points by the inverse of their likelihood to compensate for a simulation from the posterior (rather than from the prior). While this stage

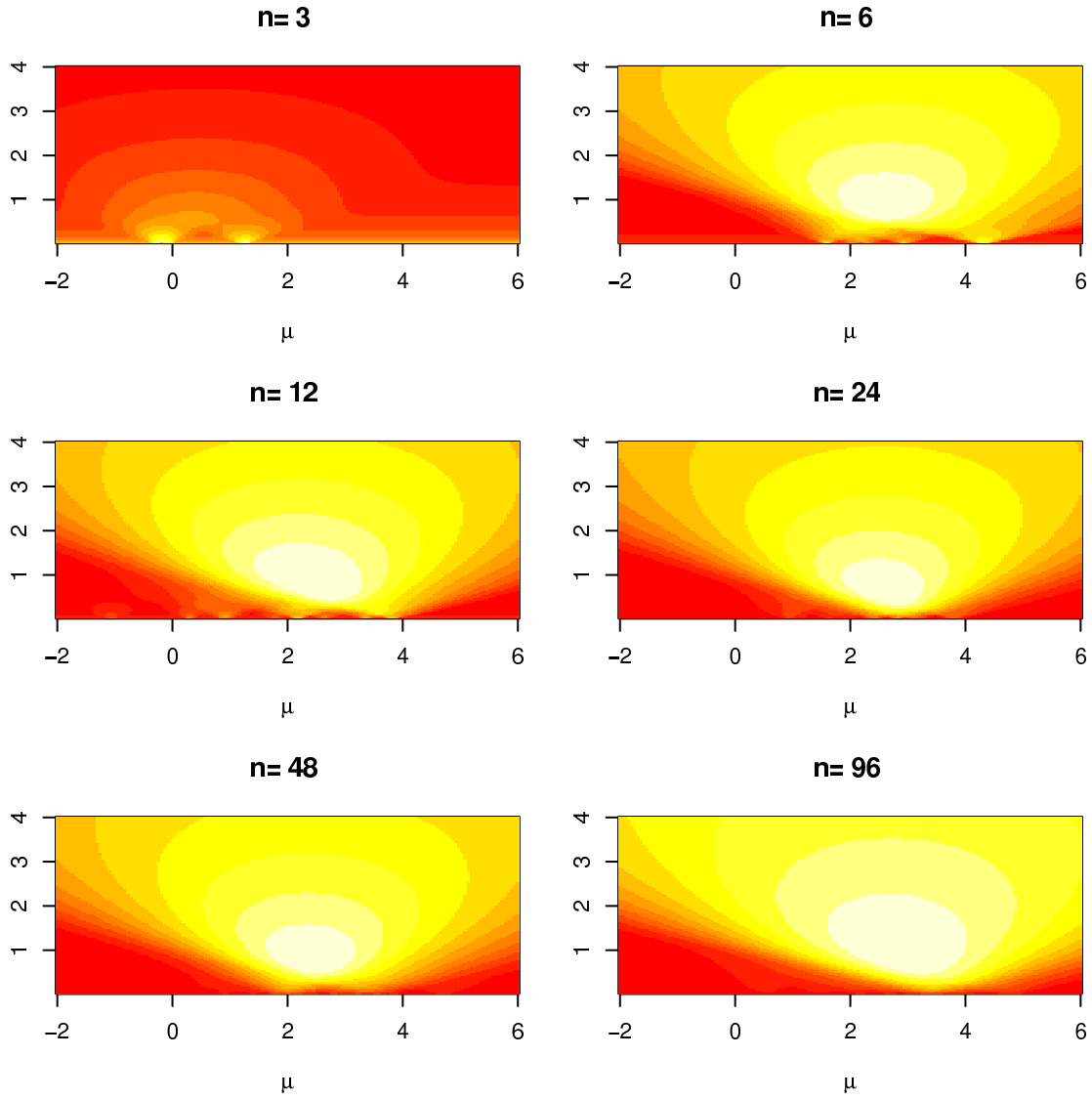

Figure 2: **Comparison of log-likelihood surfaces for different values of the sample size n, for the mixture example.**

Table 1: **Comparison of five approximations of $\mathfrak{Z}$ for a sample of 3 observations from the normal mixture (7) simulated with $\mu = -2$ and $\sigma = 3/2$. All Monte Carlo and MCMC algorithms are based on $T = 104$ simulations, while the numerical integration is based on a 850×950 grid in the (μ, σ) parameter space that is also used for the graphical representations, and the nested sampling approximation is based on a starting sample of $M = 1000$ points followed by at least 103 further simulations from the constrained prior and a stopping rule at 95% of the observed maximum likelihood. The constrained prior simulation is based on 50 values simulated from the MCMC kernel by starting from the current value of (μ, σ) and accepting only MCMC steps that lead to a likelihood higher than the bound; the 50 values are then weighted by the inverse likelihood to compensate for the simulation from the posterior. The error evaluations are given between parentheses and are described above, with the nested sampling error being derived from the approximate standard deviation $\sqrt{H/M}$ on $\log \mathfrak{Z}$ and H being estimated the same way as $\mathfrak{Z}$.**

Experiment	Numerical	Nested	MCMC	MC	Mixed
1	0.02638185	0.7994272	0.02631638	0.06540345	0.02868031
		(0.009371732)	(0.0260898)	(0.02863879)	(0.0004610243)
2	0.002720217	0.02069206	0.002714722	0.002990156	0.002809786
		(0.00393269)	(0.00264507)	$(9.172314e-05)$	$(3.857435e-05)$
3	0.005917445	0.07945572	0.0060799	0.006434044	0.006052107
		(0.002041785)	(0.006063258)	(0.0001077968)	$(2.834609e-05)$
4	0.002649868	0.09996364	0.002575131	0.003753396	0.002926395
		(0.002708107)	(0.002552974)	(0.0003436786)	$(6.014119e-05)$
5	0.01604548	0.7359017	0.01561999	0.03075327	0.01739375
		(0.01846696)	(0.01548540)	(0.01243047)	(0.0003401821)

(moderately) increased the computing time for the nested sampler, it seemed to us this was the most feasible approach. We then ran the nested sampling algorithm till the current value of (μ, σ) neither contributed significantly to the approximation of $\mathfrak{Z}$, not was far from the maximum value of the likelihood observed during the first MCMC round (see Figure 5).

The analysis of this experiment in Tables 1–3 first shows that nested sampling often gives very different evaluations of $\mathfrak{Z}$ when compared with the four other approaches. The most reliable approach—besides the numerical evaluation which cannot be used in general settings—is the mixture approach of Section 6.2, which provides estimates of $\mathfrak{Z}$ that are quite similar to the numerical evaluation on a stable basis. The nested sampling evaluation is on many occurrences quite above the numerical value and this may be attributed to a fatal attraction of the tiny modes corresponding to the mean close to one of the observations and the variance close to zero. Note also that the Monte Carlo method leading to $\widehat{\mathfrak{Z}}_2$ is second in producing poor approximations to $\mathfrak{Z}$. (The kernel ϕ used in $\widehat{\mathfrak{Z}}_2$ is a t non-parametric kernel estimate with standard bandwidth estimation.)

7.3 A probit example for nested importance sampling

To illustrate the extended algorithm described in Section 5 we consider the arsenic dataset and one of the probit models studied in Chapter 5 of Gelman and Hill (2006). The observations are independent Bernoulli variables y_i such that $\Pr(y_i = 1) = \Phi(x_i^T \theta)$, where x_i is a vector of d covariates

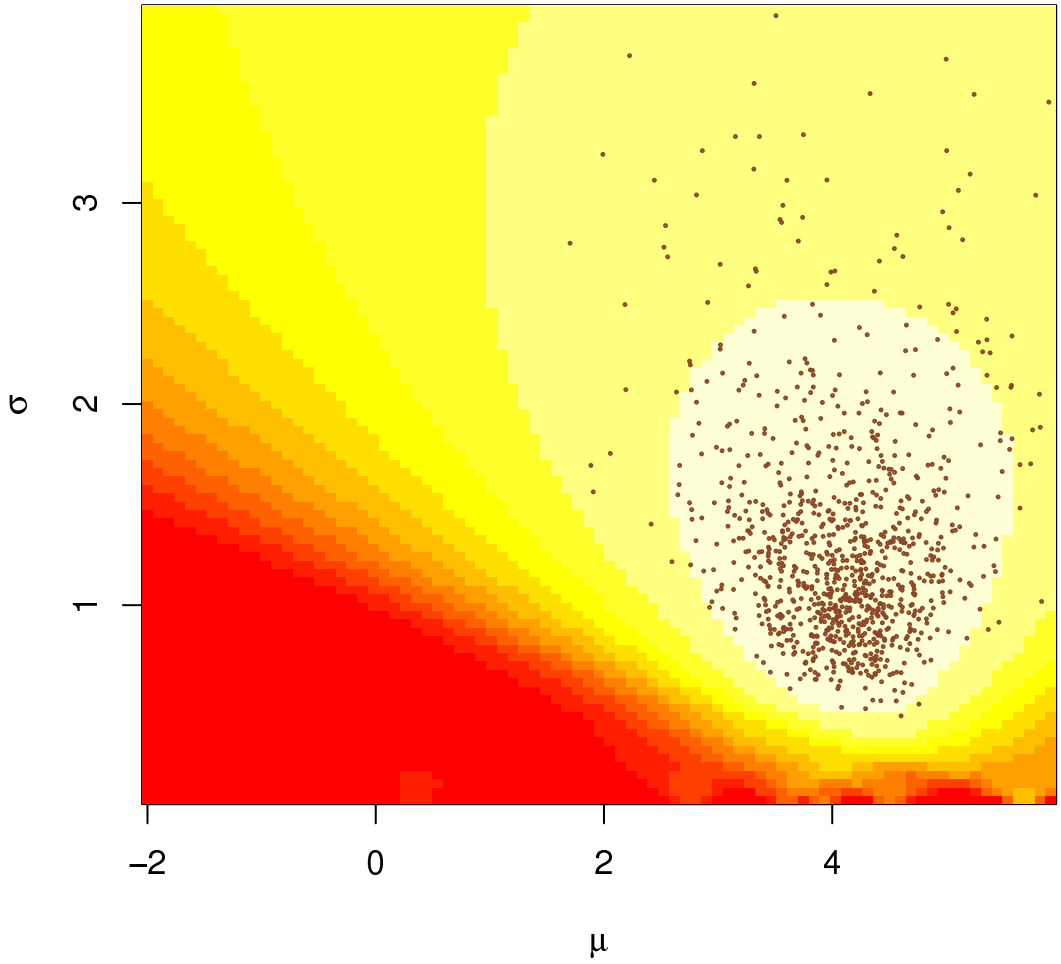

Figure 3: **MCMC sample of** 104 **simulations plotted on the log-likelihood surface in the** (μ, σ) **space for** $n = 63$ **observations from** (7)**, for the mixture example.**

Table 2: **Same caption as Table 1 for** $n = 6$ **observations (**$T = 105$ **simulations).**

Experiment	Numerical	Nested	MCMC	MC	Mixed
1	0.003000692	0.005226228	0.003113305	0.003065143	0.002994274
		(0.001367028)	(0.003104752)	$(8.89131e-05)$	$(1.241009e-05)$
2	$2.311333e-05$	$1.784006e-05$	$2.37307e-05$	$2.366789e-05$	$2.342896e-05$
		$(1.052050e-05)$	$(2.366576e-05)$	$(1.236103e-07)$	$(7.743941e-08)$
3	0.0001575844	0.0001809647	0.0001590132	0.0001577876	0.0001576939
		$(8.071656e-0)$	(0.0001585772)	$(7.622973e-07)$	$(5.967563e-07)$

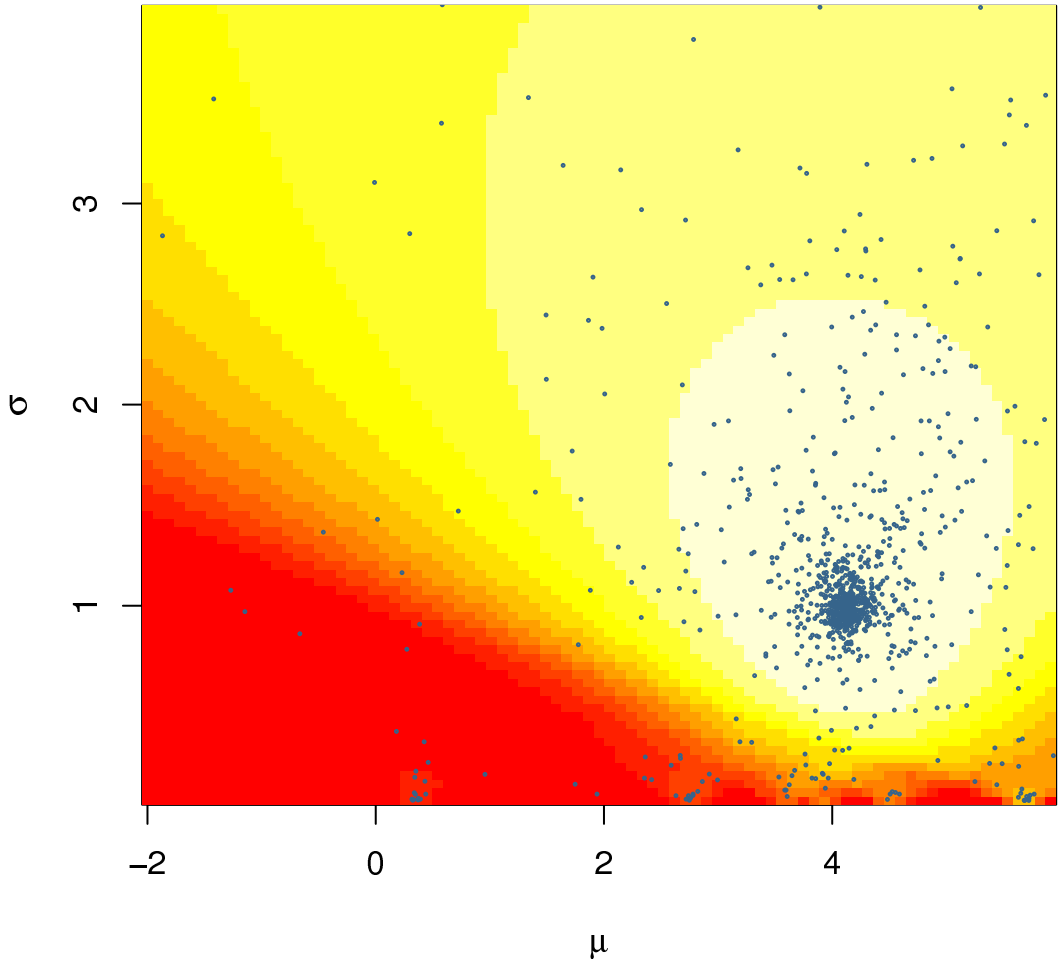

Figure 4: **Nested sampling sequence based on $M = 1000$ starting points for the same dataset as Figure 3, for the mixture example.**

Table 3: **Same caption as Table 1 for $n = 12$ observations ($T = 105$ simulations).**

Experiment	Numerical	Nested	MCMC	MC	Mixed
1	$5.06238e-08$	$4.275614e-08$	$5.056614e-08$	$5.066061e-08$	$5.07019e-08$
		$(7.697119e-08)$	$(5.042801e-08)$	$(1.433309e-10)$	$(1.321545e-10)$
2	$1.917207e-08$	$4.759043e-07$	$1.925406e-08$	$1.916241e-08$	$1.917270e-08$
		$(3.735215e-08)$	$(1.920146e-08)$	$(5.052397e-11)$	$(5.060781e-11)$

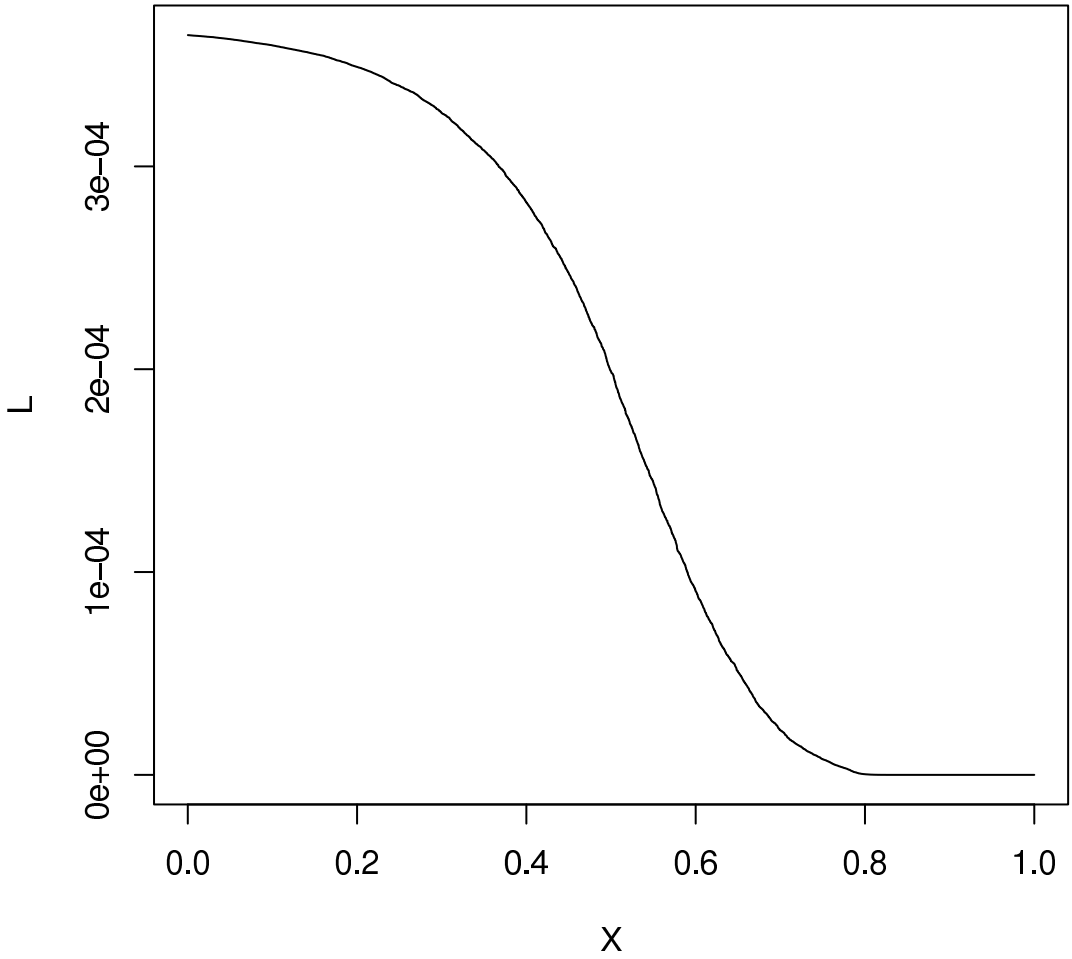

Figure 5: **Likelihood evaluations of the sequence plotted on Fig. 4, for the mixture example.**

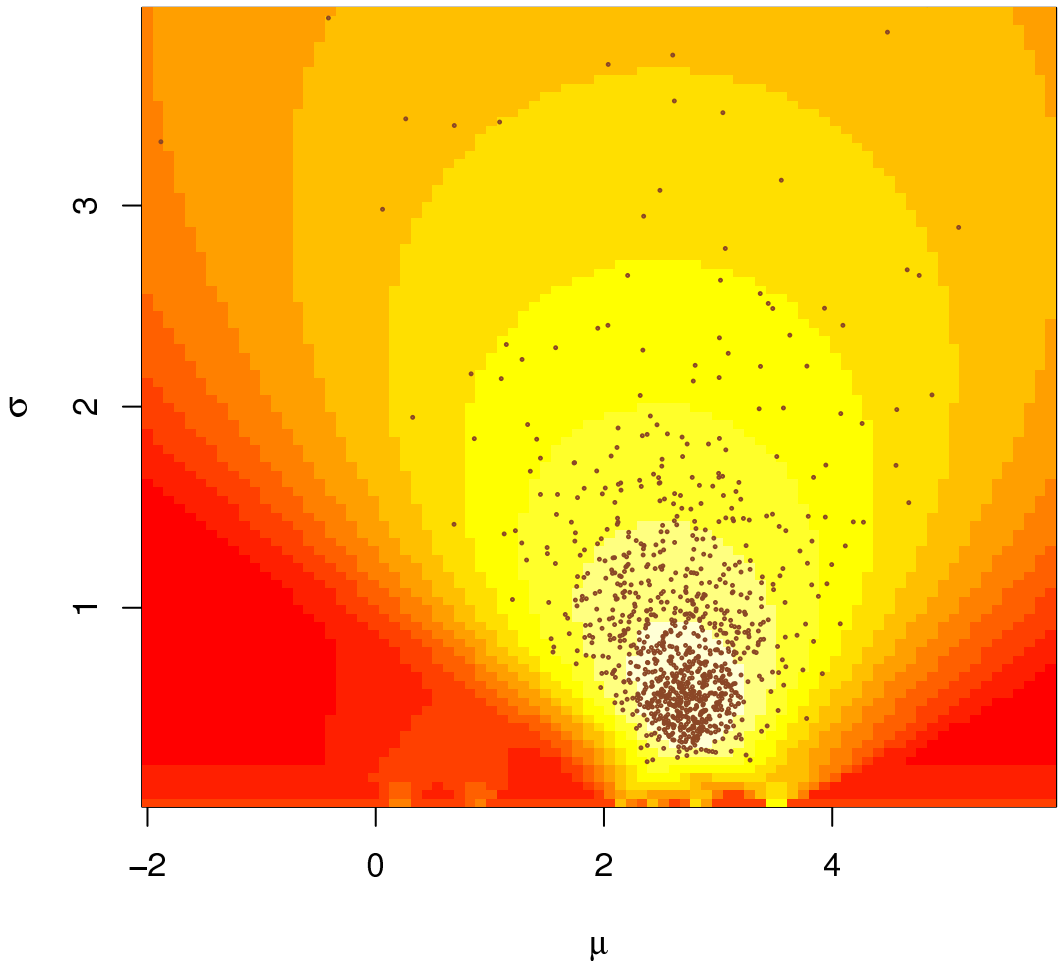

Figure 6: **MCMC sample of** 104 **simulations plotted on the log-likelihood surface in the** (μ, σ) **space for** $n = 6$ **observations from the mixture** (7)**.**

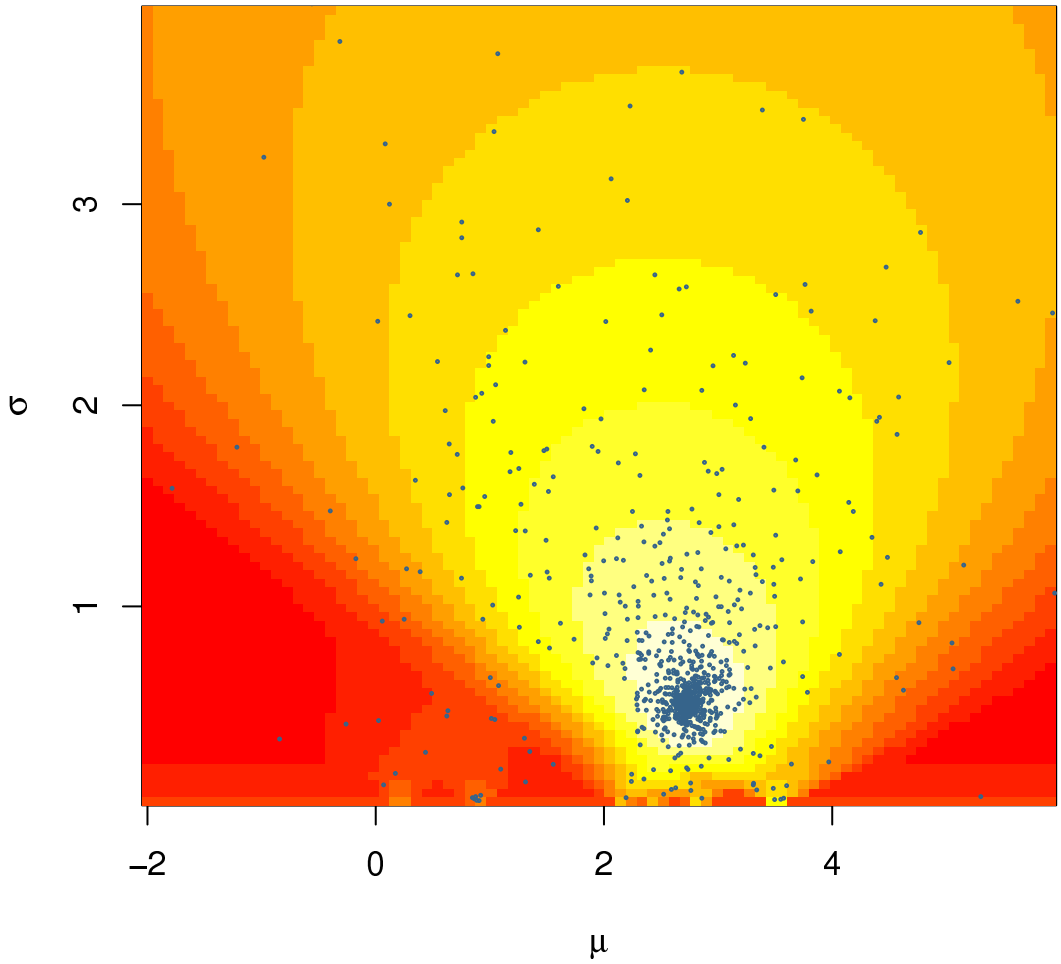

Figure 7: **Nested sampling sequence based on $M = 1000$ starting points for the same dataset as Figure 6.**

associated with y_i, 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, at `http://www.stat.columbia.edu/~gelman/arm/examples/arsenic`.

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 he or she drinks from in the past three years, and the seven covariates are an intercept, distance to nearest safe well (in 100 meters unit), education level, log of arsenic level, and cross-effects for these three variables.

The likelihood reads:

$$L(\theta) = \prod_{i=1}^{n} = \Phi(x_i^T\theta)^{y_i}\{1 - \Phi(x_i^T\theta)\}^{1-y_i}$$

and, for illustration's purpose, we assign $N_d(0, 10^2 I_d)$ as prior on θ.

We ran 10 times the nested importance sampling algorithm, based on the nested ball strategy described in §5, for $N = 10$, 100, 1000. The tuning parameters $\hat{\theta}$, τ are respectively set to the posterior mode (as obtained numerically), and 100. Figure 8 compares box-plots of the obtained estimates to the true value of the evidence on the logarithmic scale. The average number of simulated points was respectively, 645, 6628 and 65,926, for $N = 10$, 100, 1000. The precision of estimates seems quite satisfactory, given the small computational effort, and the relatively straightforward tuning (based on the posterior mode only). We note however a small bias, which may due to some numerical artifact. As a matter of comparison, a basic importance sampling estimator, based on only 645 simulated points, and a Gaussian proposal fitted with the same hyper-parameters (mean set to posterior mode, variance set to 100 times the identity matrix), has a variance which is larger by several orders of magnitude.

Thus, this example shows that nested importance sampling is generally more convenient that standard importance sampling for evaluating the evidence, in that it exhibits good performance even if it is supplied only with an approximation of the posterior mode (at least in a problem of reasonable dimension). This may be particularly helpful in situations where the shape of the posterior support is difficult to determine: say when the posterior mode (or maximum likelihood estimator) must be obtained by the EM algorithm, which does not allow for calculating the Hessian of the posterior log density at the mode.

Acknowledgements

The authors are grateful to Omiros Papaspiliopoulos for his helpful comments.

References

Bartolucci, F., Scaccia, L., and Mira, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, 93:41–52.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixtures posterior distribution. *J. American Statist. Assoc.*, 95(3):957–979.

Chen, M. and Shao, Q. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, 25:1563–1594.

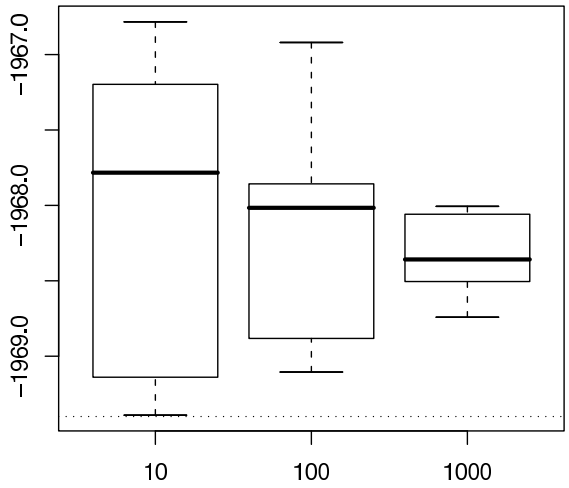

Figure 8: **Box-plots of nested importance sampling estimates for** $N = 10$, 100, 1000 **compared to true value of evidence (dotted line), for the probit example [log scale].**

Diebolt, J. and Robert, C. (1990). Bayesian estimation of finite mixture distributions, Part i: Theoretical aspects. Technical Report 110, LSTA, Université Paris VI, Paris.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Soc. Series B*, 56:363–375.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. (Ser. B)*, 56:501–514.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science*, 13:163–185.

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

Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer-Verlag, New York.

Marin, J., Mengersen, K., and Robert, C. (2004). Bayesian modelling and inference on mixtures of distributions. In Rao, C. and Dey, D., editors, *Handbook of Statistics*, volume 25 (to appear). Springer-Verlag, New York.

Philippe, A. (1997a). Simulation of right and left truncated Gamma distributions by mixtures. *Stat. Comput.*, 7:173–181.

Philippe, A. (1997b). Simulation output by Riemann sums. *J. Statist. Comput. Simul.*, 59(4):295–314.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Soc. Series B*, 59:731–792.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

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

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–860.

Skilling, J. (2007). Nested sampling for general Bayesian computation. In Bernardo, J., Bayarri, M., Berger, J., David, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 8*. (to appear).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Soc. Series B*, 64(2):583–639.

Yakowitz, S., Krimmel, J., and Szidarovszky, F. (1978). Weighted Monte Carlo integration. *SIAM J. Numer. Anal.*, 15(6):1289–1300.

A Proof of Lemma 2.1

We first note that 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 (5) 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:

$$\begin{aligned}
\int \pi(\theta)L(\theta)f(\theta)\,d\theta &= \mathbb{E}^{\pi}[\psi\{L(\theta)\}] \\
&= \int_0^{+\infty} P^{\pi}(\psi\{L(\theta)\} > l)\,dl \\
&= \int_0^{+\infty} \varphi^{-1}\{\psi^{-1}(l)\}\,dl \\
&= \int_0^{1} \psi\{\varphi(x)\}\,dx.
\end{aligned}$$

which concludes the proof.

B Proof of Theorem 1

Let $t_i = x^{\star}_{i+1}/x^{\star}_i$, for $i = 1, 2, \ldots$ We first prove the following lemma. (This result is mentioned by Skilling (2006, 2007).)

Lemma B.1. *The t_i's are independent Beta$(n, 1)$ variates.*

Proof. For $i = 1, \ldots, N$, let $\hat{t}_i = \varphi^{-1}(L(\theta_{1,i}))$, where the $\theta_{1,i}$'s denote the N initial draws from the prior, and let $\hat{t}_{(i)}$ denote the elements of the corresponding ordered vector. The $\hat{t}_i$ are i.i.d. $[0,1]$ uniform variates, so $t_1 = \hat{t}_{(N)}$ is Beta$(n,1)$. The density of the $\hat{t}_i$'s for $i = 1, \ldots, N-1$, conditional on $\hat{t}_{(N)}$ is computed as

$$p(\hat{t}_{(1)}, \ldots, \hat{t}_{(N-1)} | \hat{t}_{(N)}) = \frac{(N-1)!}{\hat{t}_{(N)}^{N-1}} I[\hat{t}_{(1)} < \ldots < \hat{t}_{(N)}] \,.$$

Hence, by a simple symmetry argument, if the largest element $t_1 = \hat{t}_{(N)}$ is removed from the vector $(\hat{t}_1, \ldots, \hat{t}_N)$, the remaining components divided by t_1 are independent $[0,1]$ uniform variates. The result follows by induction. □

A direct consequence of the above lemma is that $u_i = t_i^N$ defines a sequence of independent uniform $[0,1]$ variates. A Taylor expansion of e_N then leads to:

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

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

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

is a sum of independent, standard variables, as $\mathbb{E}[\log u_i] = -1$ and $\mathrm{var}[\log u_i] = 1$. Thus, $(\log x_i - \log x_i^\star) = O_P(N^{-1/2})$ and

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

since $\psi'(t) = \psi'(i/N) + O(N^{-1})$ for $t \in [(i-1)/N, i/N]$, and provided $B_N(t)$ is defined as:

$$B_N(t) = (cN)^{-1/2} S_{\lfloor cNt \rfloor}$$

for $t \in [0,1]$. According to Donsker's theorem (e.g. Kallenberg, 2002, p.275), B_N converges to a Brownian motion B on $[0,1]$, in the sense that $f(B_N) \overset{d}{\rightsquigarrow} f(B)$ for any measurable and a.s. continuous function f. Thus

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

which has the same distribution as

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

that is, a zero-mean Gaussian variate.

C Proof of Lemma 4.1

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

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

for $d\geq 1$, since $-\int_{\varepsilon_d}^1 s\varphi_d'(s)\,\mathrm{d}s \leq -\int_0^1 s\varphi_d'(s)\,\mathrm{d}s = \int_0^1 \varphi_d(s)\,\mathrm{d}s = 1$. This gives the first result, with $c_1 = \log(\sqrt{2}/\tau)$.

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

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

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. Take α so that the right-hand side is negative, i.e. $\alpha > \sqrt{2}\exp(-1/2)$. Using large deviations calculations, e.g. (Kallenberg, 2002, Chap. 27), one gets that $-\log(s_d)/d$ converges to some $\gamma>0$ as $d\to+\infty$. Then

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

as $\int_0^1 \varphi_d(s)\,ds = 1$. As $d\to+\infty$, $-\log(s_d)/d\to\gamma$, $s_d\to 0$, $\varphi_d(s_d)=\alpha^d\to 0$, $\int_{s_d}^1 \varphi_d(s)\,ds \leq \varphi_d(s_d)(1-s_d)\to 0$, and

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

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