# Advice on “Computational methods for Bayesian model choice”

John Skilling

*Maximum Entropy Data Consultants Ltd, Killaha East, Kenmare, Kerry, Ireland*

**Abstract.** At last year’s meeting, Robert and Wraith (2009) gave an account of algorithms in current statistical practice. After giving an overview of the principles of computational inference, this paper follows and refers to theirs, though from a more critical and principled viewpoint. Their mistaken account of nested sampling is corrected.

**Keywords:** Bayesian inference, evidence, Bayes factor, importance sampling, bridge sampling, Chib approximation, nested sampling.
**PACS:** 02.50Ng;02.70Uu

## PRINCIPLES

Bayesian inference quantifies a joint distribution, factored on input as prior density $\pi(\theta)$ and likelihood $L(\theta)$, and on output as **evidence** $Z$ and **posterior** density $P(\theta)$.

$$L(\theta)\ \underbrace{\pi(\theta)d\theta}_{\text{prior}} = Z\,\underbrace{P(\theta)d\theta}_{\text{posterior}}$$

Prior and posterior are both measures over parameter $\theta$, and all such measures are related by modulating functions (technically “Radon-Nikodym derivatives”) such as $L$. Toy problems apart, we assume that $L$ can be evaluated anywhere but not everywhere, meaning that it has no analytic integral. Only the prior, and trivial modifications of it, can be integrated analytically. Hence the measures we seek are represented computationally as weighted samples $\{\theta^{(1)},\theta^{(2)},\ldots\}$, and not as analytic expressions.

Averages such as

$$Z = \int L(\theta)\,\pi(\theta)d\theta \equiv \langle L\rangle_\pi$$

can be estimated computationally as sums over random samples. Generally

$$\Big\langle\, f\,\Big\rangle_g \equiv \int f(\theta)g(\theta)d\theta \approx \frac{1}{n}\sum_{i=1}^{n} f(\theta^{(i)}) \quad \text{where} \quad \theta^{(i)} \sim g \text{ with} \int g(\theta)d\theta = 1$$

but there have to be enough samples from $g$ to locate the bulk of $f$. With complicated multi-dimensional likelihood functions, this direct Monte Carlo algorithm for Bayesian inference fails because $L$ confines the posterior to too small a part of the prior.

# Information

The mismatch from source density $q$ to destination density $p$ is quantified by the **information**

$$H(p;q) = \int \log\left(\frac{p(\theta)}{q(\theta)}\right) p(\theta)d\theta \equiv \left\langle \log\frac{p}{q} \right\rangle_p$$

whose exponential $\exp(H)$ represents compression from $q$ to $p$. This form is the unique relationship consistent with the symmetry of independence, that analysing independent problems together yields the same results as analysing them separately.

Information is asymmetric, $H(p;q) \neq H(q;p)$. In particular, $q$ must support $p$, but $p$ need not support $q$. This accords with the requirements of inference, where the prior always supports the posterior (bounded $L$ suffices for this) though the posterior need not support the prior (because $L$ can be and often is 0).

Inference is one-way prior-to-posterior compressive, controlled by information.

# Compression

Only about 1 in $\exp(H)$ samples from $q$ informs about $p$. The others miss the bulk of $p$, so the cost of a single compression is $\exp(H)$. Practical necessity requires $H \leq O(1)$ to avoid exponential inefficiency, and this limits what can be done in one step. Interestingly difficult applications all have extensive data, for which the prior-to-posterior information $H(P;\pi)$ greatly exceeds 1, being thousands or maybe millions in line with the dataset size. Accordingly, practical algorithms must break the problem into steps $k$, repeatedly extracting $\delta H_k \leq O(1)$ bits of information, and then recovering full sampling of the temporary destination before proceeding to the next iteration.

The overall cost is $\sum \exp \delta H_k$ whereas the achieved compression is $\prod \exp \delta H_k$. Computational efficiency demands balanced steps with roughly equal $\delta H_k$. Thus Bayesian computation needs many $H(P;\pi)$ balanced steps of moderate $O(1)$ compression.

# Dimension

Probability is a measure, and measures are commutative $P + Q = Q + P$ so that elements can be arbitrarily re-ordered.

For example, the 2-dimensional Gaussian likelihood $\exp(-(x^2+y^2)/2)$ on uniform prior can be re-ordered to 1-dimensional exponential $\exp(-t/2)$, again on flat prior, by writing $(x,y)$ in polar coordinates as $(r\cos\theta, r\sin\theta)$ and rastering along a spiral line $r = t^{1/2}$, $\theta = \lambda t$, tightly wound with large $\lambda$.

One can explore $\theta = (x,y)$ with 2-dimensional steps $(\delta x, \delta y)$, or equivalently explore the 1-dimensional spiral with linear steps $\delta t$. Whether planar or linear exploration is more efficient depends on the particular likelihood involved. Either way, equilibrated samples are equivalent, and there's no way of telling which dimension was used.

Dimension is merely a practical tool that may or may not assist the generation of exploratory proposals. Dimension isn't fundamental. Information is.

## Volume

Except in low dimension where smooth interpolation may be allowable, random point samples do not themselves define accompanying volumes.

Problem 1 has Gaussian likelihood $L(\theta) = \exp(-\frac{1}{2}\sum_{i=1}^{1000}\theta_i^2)$ in a moderate thousand dimensions, within an encompassing flat prior. There are $2^{1000}$ orthants and random samples from any feasible ensemble will almost certainly all be in different ones. Problem 2 has the same prior and the same likelihood, *except* that the likelihood restricts the signs of $\theta_{501},\ldots,\theta_{1000}$ to be the same as the signs of $\theta_1,\ldots,\theta_{500}$. Other patterns of signs have $L = 0$. The posterior is thus focussed by a factor $2^{500}$, and $Z$ is decreased by this same substantial factor.

Yet the thousand samples will seem much the same as before, with the same $L$'s and $\pi$'s. Only a particularly astute observer would notice the repetition in

```
++-++---+-+--+--+++++-+-++-+-++-++-+++--+---+-+++++++++---++
+------+-++-++++-++-+-++-+-+++--++++-----+--+-+-----+-+++--+
+++-----+------+-+-----+--+-+---++++++--+++++-+--+---++++--+
--++--+--+----+++---+++-----+-+-+-+---+--+-++-+----+++-++-+-
++-+-++--+-+-+++------++++--+-+++++-+++-+-+-+--+++---++++-+-
++++-+--++--+-+-+-+-+-----+-++--+-+-+-+--+-+-+-++-+-+++-+++-
++--+-+-+-+-++++--+-++-++----+-++-++-++----+--------++-+--++
---+-+----++++++--+-++----+++-+--++++-----+---++--+-+++-++-+
++++---+----+--+--++++-++---+-+--+--+++++-+-++-+-++-++-+++--
+---+-+++++++++---+++------+-++-++++-++-+-++-+-+++--++++----
-+--+-+-----+-+++--++++-----+------+-+-----+--+-+---++++++--
+++++-+--+---++++--+--++--+--+----+++---+++-----+-+-+-+---+-
-+-++-+----+++-++-+-++-+-++--+-+-+++------++++--+-+++++-+++-
+-+-+--+++---++++-+-++++-+--++--+-+-+-+-+-----+-++--+-+-+-+-
-+-+-+-++-+-+++-+++-++--+-+-+-+-++++--+-++-++----+-++-++-++-
---+--------++-+--++---+-+----++++++--+-++----+++-+--++++---
--+---++--+-+++-++-+++++---+----+--+--++
```

and that observer could then take credit for a new model with refined prior that gained a Bayes factor of $2^{500}$. But there are myriad similar constraints and in general there's no way of knowing that $L$ has been hugely confined.

Point samples don't know their associated volume, and it's not generally possible to get normalisation directly from samples.

# IMPORTANCE SAMPLING SCHEMES

The evidence can be written as

$$Z = \int \frac{L(\theta)\pi(\theta)}{\varpi(\theta)}\,\varpi(\theta)\,d\theta \equiv \left\langle \frac{L\pi}{\varpi} \right\rangle_{\varpi}$$

where the arbitrary density $\varpi$ (the "importance function") need no longer coincide with $\pi$. The relevant information, which defines the efficiency of this procedure, is $H(P;\varpi)$, which must not exceed $O(1)$. In practice, it's not possible to reduce the original $H(P;\pi)$ from thousands or millions to $H(P;\varpi) \le O(1)$ by clever selection of $\varpi$. We can't find an analytic $\varpi$ close enough to $P$. Especially so, since experienced users will already have done their best in that direction by choosing suitably sympathetic coordinates and weights, leaving little scope for improvement.

More fundamentally, importance sampling is based on analytic input $\pi$, so it can only be set up once at the beginning. It can only be iterated by having a sequence of analytically integrable weights that will only be available in toy problems.

## Bridge sampling

Bridge sampling relates two models, 1 and 2, with common parameters $\theta$ and reasonably close posteriors. Initially, $P_2$ is used as the sampling distribution $\varpi$ to get the Bayes factor

$$B_{12} = \frac{Z_1}{Z_2} = \frac{\int L(\theta)\pi_1(\theta)d\theta}{Z_2} \equiv \frac{1}{Z_2}\left\langle \frac{L(\theta)\pi_1(\theta)}{P_2} \right\rangle_{P_2} = \left\langle \frac{\pi_1(\theta)}{\pi_2(\theta)} \right\rangle_{P_2}$$

This compressive (from 2 to 1) estimator requires $P_1$ to be supported by $P_2$ with $H(P_1;P_2) \leq O(1)$. Otherwise, if $P_1$ leaks into regions of small $P_2$, it will be sampled there at negligible frequency $P_2$ but huge weight $1/P_2$, which destroys the estimate by mixing it with $0/0$.

R&W then consider use of $|\pi_1(\theta) - \pi_2(\theta)|$ as the sampling density $\varpi$. But they correctly dismiss the suggestion because users would not be able to extract the necessary normalisation $\int \varpi(\theta)d\theta$, and because the difference could vanish over important regions which would become unsampled. The Bayes factor could then be altered by changing the models within an unsampled region, so that the estimate (being unchanged) would necessarily be unreliable.

R&W next write the Bayes factor as

$$B_{12} = \frac{Z_1}{Z_2} \frac{\int \alpha(\theta)\, L(\theta)\pi_1(\theta)\, L(\theta)\pi_2(\theta)\, d\theta}{\int \alpha(\theta)\, L(\theta)\pi_1(\theta)\, L(\theta)\pi_2(\theta)\, d\theta}$$

(the latter fraction is 1) with arbitrary distribution $\alpha$, which can be rewritten as

$$B_{12} = \frac{\int \alpha(\theta)\, L(\theta)\pi_1(\theta)\, P_2(\theta)\, d\theta}{\int \alpha(\theta)\, L(\theta)\pi_2(\theta)\, P_1(\theta)\, d\theta} \equiv \frac{\langle \alpha L \pi_1 \rangle_{P_2}}{\langle \alpha L \pi_2 \rangle_{P_1}}$$

The conditions for this estimator are rather less restrictive, because it's only the crossover distribution $\alpha P_1 P_2$ that needs to be adequately supported (both by $P_1$ and by $P_2$).

Of course, Bayes factor estimates don't yield the evidence relative to a prior that's exp(thousands or millions) distant from these posteriors. Also, bridge sampling is diffusive, so inefficient for large-scale compression.

## Mixture bridge sampling

Mixture bridge sampling is another method requiring a tractable density $\varphi$. It uses a particle labelled by parameter $\theta$ and switch $\delta$ subject to likelihood defined by

$$\widetilde{L}(\theta,\delta) = \begin{cases} wL(\theta)\,\pi(\theta) & \text{if } \delta = 1, \\ \varphi(\theta) & \text{if } \delta = 2. \end{cases}$$

which is equilibrated by Gibbs sampling. The equilibrium occupancy ratio is

$$\frac{\Pr(\delta = 1)}{\Pr(\delta = 2)} = \frac{w \int L(\theta)\,\pi(\theta)\,d\theta}{\int \varphi(\theta)\,d\theta} = wZ$$

and this yields $Z$ provided the weight $w$ is tuned to a reasonably appropriate value. At each exploratory step, a fresh estimate of $\Pr(\delta = 1)$ is available as $wL\pi/(wL\pi + \varphi)$, and it's more accurate to average these than to rely on cruder "1 versus 2" occupancy counts. This proper Bayesian use of the relevant data is called Rao-Blackwellisation.

When $w$ takes its optimal value $1/Z$, the occupancy ratio is 1:1, and equilibration occurs at its maximum rate

$$\zeta = \int P(\theta)\varphi(\theta)d\theta$$

For this to be acceptably speedy, it must not be exponentially small. The crossover density $\rho(\theta) = P(\theta)\varphi(\theta)/\zeta$ must be accessible from both $P$ and $\varphi$.

This can **only** work for close distributions, and it doesn't directly relate $P$ to a distant prior $\pi$. Also, the method is one-way analytic-to-ensemble, so not suitable for large-scale compression.

## Harmonic means

Just as the simplest expression for the evidence is $Z = \langle L \rangle_\pi$, so the reciprocal gives the "harmonic mean" formula

$$\frac{1}{Z} = \frac{\int \pi(\theta)d\theta}{Z} = \int \frac{P(\theta)}{L(\theta)}d\theta = \left\langle \frac{1}{L} \right\rangle_P$$

But, just as numerical evaluation of $\langle L \rangle_\pi$ fails because samples from the prior fail to find the posterior, so does numerical evaluation of $\langle 1/L \rangle_P$ fail because samples from the posterior miss almost all of the prior.

In fact, the situation is even worse. The prior necessarily contains the posterior somewhere, but the posterior is not required to contain the prior. Model 1 has likelihood $L$ confined within some domain supported by the prior $\pi_1$. Model 2 has the same likelihood $L$, but prior $\pi_2$ is diluted by a factor of two because its support has expanded to twice the size. Clearly $Z_2 = Z_1/2 \neq Z_1$. Nevertheless, the posterior is unchanged, and so are the samples, so the harmonic mean estimates will also be unchanged, $\widehat{Z}_2 = \widehat{Z}_1$. This is a definitive counter-example from which the harmonic mean cannot recover.

What's gone wrong is that the integrand is dominated by huge values of $1/L$ where $L$ is small, whereas those places of small $L$ are sampled at negligible frequency. Thus the formula is merely an expensive way of calculating $0/0$. It's not surprising that evaluations notoriously have unbounded variance, which should be interpreted as a symptom warning of incurable failure. More fundamentally, in spaces of moderate or large dimension, the weights $1/L$ soon become dominated by just one or a very few particularly small $L$, so that effective sampling is hopelessly inadequate.

Nevertheless, the reciprocal can be written as

$$\frac{1}{Z} = \frac{\int \varphi(\theta)d\theta}{Z} = \left\langle \frac{\varphi}{L\pi} \right\rangle_P$$

for arbitrary density $\varphi$. R&W uncritically suggest exploiting this by making $\varphi$ narrower than $P$. The summation average is then stabilised with finite variance, which would be fine *if* this were feasible.

The problem is that most samples from $P$ miss $\varphi$ and don't contribute, unless $\varphi$ is only just narrower than $P$. The method needs $H(\varphi;P) \leq O(1)$. In a realistically complicated application with $P$ exponentially narrower than $\pi$ and defined only by a set of random samples (quite likely fewer than the dimension), how could such a closely under-fitting analytic density $\varphi$ be found? And, if such a close density could be found, why not make it cover $P$ properly and avoid the reciprocal by using direct importance sampling?

R&W's toy example in two dimensions defines $\varphi$ as a uniform ellipse covering the top 10% "high posterior density" (HPD) domain. But such a domain rapidly becomes impossible to locate in spaces of merely moderate dimension. How could one know that it was fully but only just inside $P$? The fact that weighted harmonic mean doesn't fail in a tiny 2-dimensional example does *not* mean that the method will be usable for more difficult problems. It won't be.

## CHIB APPROXIMATION

Chib [2] observed that we can get $Z$ merely by evaluating

$$Z = L(\theta)\,\pi(\theta)\,/\,P(\theta)$$

at any convenient $\theta^*$, provided we can obtain an estimate $\widehat{P}(\theta^*)$ of the posterior there (we already know $L$ and $\pi$). But point samples do not define their normalisation. Some sort of iteration between prior and posterior is needed. Chib proposed iterating on dimension (or blocks of a few dimensions) as follows.

Exploration of the posterior yields an ensemble $\{\theta^{(1)},\ldots,\theta^{(N)}\}$. Putting aside the earlier coordinates in $n$-dimensional $\theta = (\theta_1,\theta_2,\ldots,\theta_n)$, the $N$ values $\{\theta_n^{(1)},\ldots,\theta_n^{(N)}\}$ represent the marginal density $P(\theta_n)$ of $\theta_n$ alone. Then (the dangerous step) this marginal is approximated by an interpolating trapezoid or other analytic form $\widehat{P}(\theta_n)$ which can be evaluated as $\widehat{P}(\theta_n^*)$ at any chosen $\theta_n^*$. With $\theta_n^*$ now an assigned constant, the procedure is repeated with dimension reduced by 1, iterating until dimensionality is exhausted.

| Iterate number | Marginal as ensemble | Marginal as interpolant | Assigned coordinate | Estimated marginal value |
|---|---|---|---|---|
| 1 | $P(\theta_4)$ | $\widehat{P}(\theta_4)$ | $\theta_4^*$ | $\widehat{P}(\theta_4^*)$ |
| 2 | $P(\theta_3 \mid \theta_4^*)$ | $\widehat{P}(\theta_3 \mid \theta_4^*)$ | $\theta_3^*$ | $\widehat{P}(\theta_3^* \mid \theta_4^*)$ |
| 3 | $P(\theta_2 \mid \theta_3^*,\theta_4^*)$ | $\widehat{P}(\theta_2 \mid \theta_3^*,\theta_4^*)$ | $\theta_2^*$ | $\widehat{P}(\theta_2^* \mid \theta_3^*,\theta_4^*)$ |
| 4 | $P(\theta_1 \mid \theta_2^*,\theta_3^*,\theta_4^*)$ | $\widehat{P}(\theta_1 \mid \theta_2^*,\theta_3^*,\theta_4^*)$ | $\theta_1^*$ | $\widehat{P}(\theta_1^* \mid \theta_2^*,\theta_3^*,\theta_4^*)$ |
| | | | | $\widehat{P}(\theta_1^*,\theta_2^*,\theta_3^*,\theta_4^*)$ |

On completion, the required $\widehat{P}(\theta^*)$ is available as the product of the estimated marginals, yielding an estimate

$$\widehat{Z} = L(\theta^*)\,\pi(\theta^*)\,/\,\widehat{P}(\theta^*)$$

This is a compressive algorithm (which is good), and it's ensemble-to-ensemble (which is essential), though it does rely on dimension (which isn't fundamental). The danger point is the diversion through an analytic interpolant. Consider this ensemble of a dozen uniform-looking 3-digit integers:

$$171,\ 435,\ 849,\ 675,\ 261,\ 321,\ 159,\ 447,\ 282,\ 876,\ 930,\ 555.$$

They were actually generated as integers from

$$\pi = \texttt{Uniform}[100, 999], \qquad L = \begin{cases} 1 \text{ if divisible by 3,} \\ 0 \text{ otherwise,} \end{cases}$$

which has $Z = 1/3$. But a user would need to be astute to notice the divisibility property of the posterior samples, and brave to remove those integers not divisible by 3 from the support of the estimated interpolant $\widehat{P}$. Most users would interpolate the ensemble with flattish posterior and reach the wrong estimate $\widehat{Z} \approx 1$. The Chib approximation is effectively blind to microscopic structure.

Of course, there may be no dimension to start with. In the extreme, if the hypothesis space is a set of equivalent locations (such as the vertices of a simplex), then there are no prior adjacency relationships from which a dimension might plausibly be constructed. In that case, the only plausible adjacency derives from likelihood values, suggesting that the locations be sorted by likelihood. That idea leads naturally to ....

## NESTED SAMPLING

Nested sampling [3, 4, 5] is based on the one-dimensional representation

$$Z = \int_0^1 L(X)\,dX$$

of the evidence, where

$$X(L^*) = \int_{L(\theta)>L^*} \pi(\theta)\,d\theta$$

is the prior mass enclosed by likelihood contour $L^*$, and $L(X)$ is the inverse which labels the contour that encloses mass $X$. The evidence is estimated from likelihood-ordered samples by quadrature

$$\widehat{Z} = \sum_{i=1}^{\nu} L_i\,\delta X_i\,, \qquad \delta X_i = X_{i-1} - X_i$$

where the increasing $L$'s are known (they are the likelihood values of $\nu$ known sample points $\theta^{(i)}$), but the decreasing $X$'s are unknown in detail (except by impossibly expensive multidimensional integration) and are estimated statistically.

Iterate $i$ enters with $N$ live points (allowably as few as 1) constrained within $L(\theta) > L_{i-1}$ but otherwise sampled uniformly over the prior (and hence over the measure $X_{i-1}$).

The outermost (smallest $L$) of these points is taken as the $i$'th sample $\theta^{(i)}$ with likelihood $L_i$. It is archived and replaced with a new point within $L(\theta) > L_i$ to replenish the ensemble. Meanwhile, the $N-1$ survivors, already uniformly distributed, can be recycled without change. Adjacent ratios $t_i = X_i/X_{i-1}$ are each distributed as

$$\Pr(t) = N t^{N-1}$$

because $X_i$ is the outermost of $N$ samples from $\texttt{Uniform}(0, X_{i-1})$.

| Iterate<br>number | Lower<br>bound | Ensemble<br>ordered by $L$ | Likelihood<br>(increasing) | Enclosed<br>prior mass |
|---|---|---|---|---|
| | | | $L_0 = 0$ | $X_0 = 1$ |
| 1 | $L_0$ | $\{\theta^{(1)}, \bullet, \ldots, \bullet\}$ | $L_1 = L(\theta^{(1)})$ | $X_1 = t_1$ |
| 2 | $L_1$ | $\{\theta^{(2)}, \bullet, \ldots, \bullet\}$ | $L_2 = L(\theta^{(2)})$ | $X_2 = t_1 t_2$ |
| 3 | $L_2$ | $\{\theta^{(3)}, \bullet, \ldots, \bullet\}$ | $L_3 = L(\theta^{(3)})$ | $X_3 = t_1 t_2 t_3$ |
| 4 | $L_3$ | $\{\theta^{(4)}, \bullet, \ldots, \bullet\}$ | $L_4 = L(\theta^{(4)})$ | $X_4 = t_1 t_2 t_3 t_4$ |
| | | | | $\widehat{Z} = \sum L \delta X$ |

On termination, the estimate $\widehat{Z}$ is available as a summation. The plausible range of values of $\log \widehat{Z}$ can be found by taking a few dozen simulations of the $t$'s.

Compression per step is about 1 part in $N$ (technically, $\log t = (-1 \pm 1)/N$), so it takes about $NH$ (technically, $NH \pm \sqrt{NH}$) iterates to compress $e^{-H}$ to the typical posterior, and about the same again to confirm adequate exploration of the interior. The variability of iterate count induces a rms uncertainty $\pm\sqrt{H/N}$ in $\log Z$. Full analysis is in [6].

In statistical physics, it has been known since the 19th century that all macroscopic equilibrium properties of a system can be derived (at any temperature) from the density of states $dX/dL$. Correspondingly, the relation between $X$ and $L$ is an appropriate target for computational inference. Nested sampling aims directly at this fundamental target. It makes no ancillary assumptions about dimensionality or continuity. If the user explores correctly, nested sampling's results are statistically guaranteed.

The Wang-Landau method [7] also aims at the density of states, but through a preassigned ladder of likelihood values, on which ratios $\delta X/\delta L$ are estimated by up-and-down diffusion. By contrast, nested sampling already knows the connection between adjacent levels (it's the compression $t$), so diffusion is not needed. Also, nested sampling's schedule is automatic (1 part in $N$ per step), whereas Wang-Landau requires that likelihoods be set close enough that $\delta \log X \leq O(1)$, otherwise the diffusive transitions are too slow. That becomes impossible at a first-order phase transition. Nested sampling's iterates automatically give the balanced compressions of prior mass anticipated at the outset.

## *Robert and Wraith's account*

R&W begin by referring to [8] (Burrows 1980) as "an earlier related method", though that paper used conventional quadrature — an approach necessarily limited to low dimension. There was no hint of nested sampling's statistical methodology or dimensional invariance.

More controversially, R&W then claimed that "a full convergence assessment" of nested sampling had been given in Chopin and Robert [9]. This claim refers to a partial commentary that

(a) failed to treat the correct statistics of nested sampling,
(b) imposed boundedness and continuity conditions that are not required,
(c) used rough order-of-magnitude analysis,
(d) wrongly claimed dependence on dimension.

Nested sampling is invariant to continuity and dimension, so cannot in logic depend upon such conditions. In fact, the only conditions needed (and the only ones that *could* be needed with such wide invariance) are that $Z$ and $H(P;\pi)$ be bounded. Skilling [6] shows this by using the correct statistics, and properly treating termination error together with statistical uncertainty to get explicit numerical upper bounds (depending only on $H$) for worst cases. R&W omit this reference although it was given at the same meeting.

R&W suggest that the $X$'s "are either deterministic, e.g. $X_i = e^{-i/N}$, or random" (generated by $\Pr(t)$ above). These aren't alternatives. The "deterministic" option is just an approximate simplification which ignores the statistical variability of the compression ratios $t$ in the interests of convenience.

R&W then write misleadingly that "the nested sampling algorithm relies on an estimate of $L(X_i)$ or, equivalently, $X_i$". In fact, the $L$'s are known from the sample likelihoods $L(\theta)$. Only the $X$'s are to be estimated, by sampling the clearly-stated distribution $\Pr(t)$ of ratios.

R&W proceed carelessly to claim a approximation

$$X_i/X_{i-1} = (N-1)/N$$

which is badly wrong for small $N$, and obviously wrong for $N = 1$. If, for simplicity, variability is to be ignored, $X_i/X_{i-1}$ should be $e^{-1/N}$ from the mean logarithmic compression — there's no "e.g." alternative.

## *Numerical example*

R&W furnish a two-dimensional numerical example with flat prior

$$\pi(\theta_1,\theta_2) = 1/6400 \quad \text{on} \quad \theta_1 \in (-40,40),\, \theta_2 \in (-40,40)$$

and likelihood

$$L = \frac{1}{2\pi\sigma}\exp\left(-\frac{\theta_1^2}{2\sigma^2} - \frac{(\theta_2+\beta(\theta_1^2-\sigma^2))^2}{2}\right) \quad \text{where} \quad \sigma = 10,\ \beta = 0.03$$

plotted in their Figure 3 (where their outer contour is at 99% not 99.9%). Direct summation over the prior domain yields

$$\log Z = -8.7642\,, \qquad H(P;\pi) = 3.6249\,.$$

With the prior-to-posterior $H$ already $O(1)$, the efficiency $e^{-H}$ for elementary Monte Carlo integration $\langle L \rangle_\pi$ is as high as $1/38$, making this example a uselessly forgiving testbed.

## *Example by PMC*

R&W subject the example to a "population Monte Carlo (PMC) mixture importance sampler", in which the "importance function to be optimised consists of a mixture of 9 multivariate Student $t$'s with 9 degrees of freedom for each component". With $9 \times 9 = 81$ parameters, such a model $\varpi$ will surely be capable of almost perfect fit to this simple and smooth likelihood in just two dimensions. Eighty-one parameters may not be enough for a practical problem but they surely suffice for this toy.

Unsurprisingly, when 50000 points are taken from this best-fit importance function, direct MC accumulates the evidence $Z = \langle L\pi/\varpi \rangle_\varpi$ to accuracy almost 1 in $\sqrt{50000}$, leading to a deviation of $\pm 0.0045$ in $\log Z$. That agrees satisfactorily with the bulk of the range plotted in R&W's Figure 4 (left), except that the plot shows an upward excess of about 6 outliers out of 100. Similar outliers appear on the location parameters means $\bar{\theta}_1$ and $\bar{\theta}_2$ shown in their Figure 5, and presumably they reflect sporadic failure to optimise the importance function.

R&W do not comment on this repeated danger sign. But they do compare their PMC analysis *which effectively knew the posterior in advance* with nested sampling *which had to discover that for itself*. In this unfair comparison, it's not surprising that PMC calculated $Z$ more accurately.

## *Example by nested sampling*

For nested sampling, R&W use an ensemble of $N = 1000$ live points. According to standard nested sampling theory [3, 4, 5], this ought to be sufficient to yield $\log Z$ accurate to $\pm\sqrt{H/N} = \pm 0.0602$, provided that the number of iterates comfortably exceeds the $NH = 3625$ steps that are needed to reach the bulk of the posterior before continuing further to make sure of the interior.

R&W ignore the theory and treat the number of iterates (they used 10000) as merely empirical. They make no mention of $H$ despite its availability and potential value to the user, and indeed to the programmer seeking sanity checks on reasonable results.

Correct evaluation requires adequate exploration. To re-equilibrate a sample point, R&W take 50 steps of a random walk of step length 0.1 (they refer to 0.1 as a variance, but apparently intend standard deviation). Yet 50 steps of length 0.1 can only diffuse a distance of $0.1\sqrt{50} = 0.7$ which is far short of the $O(20)$ distance that their Figure

5 indicates as apparently necessary. Inadequate exploration would, of course, bias $Z$ downwards by failing to find the desired regions of high likelihood.

R&W instead report (Figure 4) that 100 runs yield deviations $0.17 \pm 0.12$ *above* the true $\log Z$ (mean and standard deviation inferred from the plotted quartiles). They claim that the "results suggest that nested sampling exhibits a *slight* upward bias for the evaluation of the evidence" (my italics). Yet the "90 positive versus 10 negative" bias in their plot is visually obvious to any informed reader. It has chance probability $2^{-100}\,100!/90!10! = 1.4 \times 10^{-17}$, which is hugely significant. Far from being "slight", its presence (if correct) would clearly demonstrate a disastrous flaw. R&W offer no explanation, and invite the reader to believe that nested sampling can't even get this trivial problem right.

Actually, nested sampling cannot possibly behave like this. Downward bias points to inadequate exploration, whereas upward bias can only be mistaken programming (earlier private correspondence to author). In R&W's case, the variability is a factor of two above the predicted $\pm 0.06$, confirming erroneous implementation.

My own nested sampling computation with 1000 runs using adequate exploration led to deviations from the true $\log Z$ of $0.0022 \pm 0.0614$, with quartiles at $-0.0398, 0.0037, 0.0462$. This is in proper statistical agreement with 1000 samples from the theoretically predicted $0 \pm 0.0602$.

*Contrary to R&W's presentation and claim, nested sampling gives no detected bias and the variability is as predicted.*

## COMMENTARY

A principled approach to inference shows that computation of large-scale problems requires systematic iterative compression from prior to posterior. Dimension doesn't matter: information $H$ does. Point samples do not yield their distribution's normalisation, except in dimensions so low that smooth interpolation may be allowed.

Distributions can either be coded analytically (and thereby evaluated anywhere), or as an ensemble of live points. The power of an analytical approach rapidly fades in the presence of a complicated likelihood function, so iterations need to process ensembles. Analytic forms only appear at the beginning, where the prior is analytic, and at the end, where an approximate analytic summary may be passed to the user for undefined future use.

Many algorithms require symmetrical detailed balance, but that's not necessary and indeed is inefficient for compression. There is a place for back-and-forth diffusive methods such as bridge sampling when models are close enough to communicate when programmed together, but the main compression should be systematic. Here, nested sampling, for which compression factors are directly available by construction, seems the natural approach. The alternative Chib approximation relies on analytic interpolation which blinds it to small-scale structure.

When choosing an algorithm, keep basic principles in mind. Do not naïvely believe the literature (particularly the mistaken accounts of nested sampling in [1, 9, 10]). Do not choose an algorithm just for what it's claimed to do, but sceptically seek what it can *not* do.

# REFERENCES

1. Robert, C.P., and D. Wraith. 2009. Computational methods for Bayesian model choice. *AIP Proceedings* (ed. P. Goggans and C.-Y. Chan) 1193: 251–262.
2. Chib, S. 2009. Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* 90: 1313-1321.
3. Skilling, J. 2006. Nested Sampling for general Bayesian computation. *J. Bayesian Analysis* 1: 833–860.
4. Sivia, D. S. and J. Skilling. 2006. *Data analysis; a Bayesian tutorial (2nd ed.)*, chap. 9. Oxford Univ. Press. ISBN 0-19-856831-2.
5. Skilling, J. 2007. Nested sampling for Bayesian computations. In *Bayesian statistics 8*, ed. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckermann, A.F.M. Smith and M. West, Oxford Univ. Press, pp 491–524.
6. Skilling, J. 2009. Nested sampling's convergence. *AIP Proceedings* (ed. P. Goggans and C.-Y. Chan) 1193: 277–291.
7. Wang, F, and D.P. Landau 2001. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E* 64: 056101.
8. Burrows, B. L. 1980. A new approach to numerical integration. *J. Inst. Maths Applics* 26: 151–173.
9. Chopin, N. and C.P. Robert. 2007. Contemplating evidence: properties, extensions of, and alternatives to nested sampling. In *Bayesian statistics 8*, ed. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckermann, A.F.M. Smith and M. West, Oxford Univ. Press, pp 513–515.
10. Chopin, N. and C.P. Robert. 2009. Properties of nested sampling, arXiv:0801.3887 and *Biometrika* 2010 in press.