UNIVERSITY OF TEXAS AT SAN ANTONIO

Nested Sampling: Introduction and Implementation

Liang Jiang

May 2009

1 ABSTRACT

Nested Sampling is a new technique to calculate the evidence, $Z = P(D|M) = \int p(D|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|M)\, \mathrm{d}\boldsymbol{\theta}$ (alternatively the marginal likelihood, marginal density of the data, or the prior predictive, $Z = \int L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$), in a way that uses Monte Carlo methods. These integrals are usually very difficult to calculate for complex models but they play an important role in statistical inference, for example Bayesian model comparison. Though there are already many approaches, both sampling-based and deterministic, have been proposed, nested sampling, first introduced by John Skilling in 2004, has caught a lot of attention because of its robustness, broad applicability, power on dealing with difficult posterior distributions, and little requirement of manual tuning. The key technical requirement of nested sampling is an ability to draw samples uniformly from prior distribution with restriction that the likelihoods of samples need to be larger than certain value. In this paper, the basic idea and algorithm of nested sampling is introduced and a practical implementation in R, including examples and result analysis, is conducted.

2 INTRODUCTION

The primary task is to calculate

$$Z = \text{evidence} = \int L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{2.1}$$

where $L(\boldsymbol{\theta})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is the prior density of the unknown parameter(s). Under background model assumptions $\mathscr{I}$, the probabilistic context of this is usually in the form of Bayes' theorem:

$$\begin{aligned} P(D|\boldsymbol{\theta}, \mathscr{I}) \times P(\boldsymbol{\theta}|\mathscr{I}) &= P(D|\mathscr{I}) \times P(\boldsymbol{\theta}|D, \mathscr{I}) \\ \text{Likelihood} \times \text{Prior} &= \text{Evidence} \times \text{Posterior} \\ L(\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} &= Z \times p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{2.2}$$

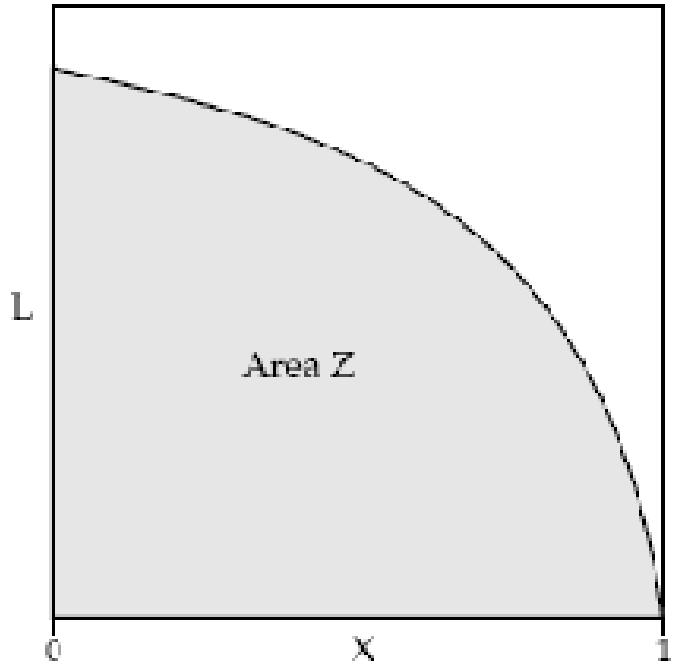
If we define the cumulative prior mass as

$$X(\lambda) = \int_{L(\boldsymbol{\theta})>\lambda} \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}, \tag{2.3}$$

then $X(\lambda)$ will decrease from 1 to 0 as λ increases. And now the integration of evidence is transformed from multiple-dimensional parameter space into one-dimensional axis, see Figure 2.1.

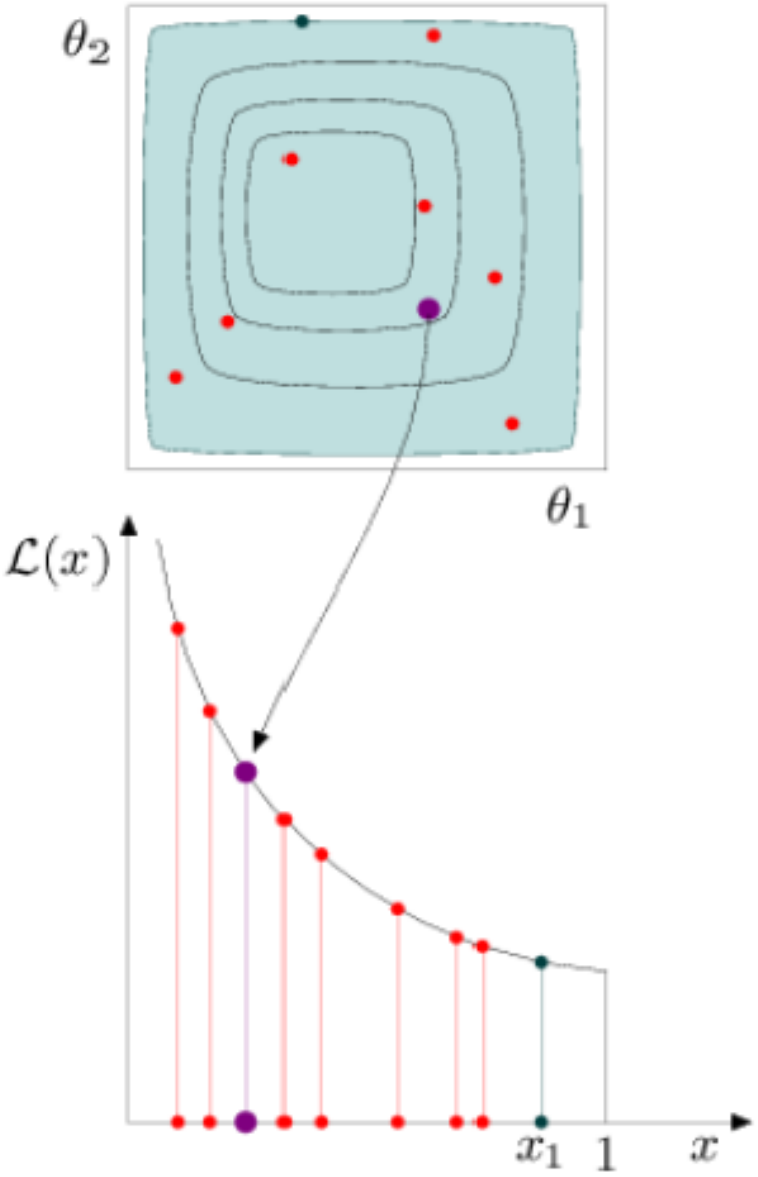
$$Z = \int_0^1 L(X)\,\mathrm{d}X \tag{2.4}$$

Figure 2.1: The integration of evidence: X is the cumulative prior mass



An example for two-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is illustrated in Figure 2.2. In two-dimensional parameter plane, each point is mapped to a point on X-axis within interval [0,1]. And each point on $L(X)$-curve represents the corresponding likelihood contour on $\theta_1 - \theta_2$ plane. As the contour area shrinks to maximum of likelihood, the point on X -axis moves to 0, because $X(L_{max}) = \int_{L(\boldsymbol{\theta})>L_{max}} \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = 0$.

Figure 2.2: Transformation from two-dimensional parameter plane to one-dimension



3 METHODOLOGY

3.1 COMPUTATION WITH MONTE CARLO METHOD

For equation (2.4), one-dimensional integration can be easily approximated by

$$Z \approx \sum_{i=1}^{m} w_i L_i \tag{3.1}$$

where $w_i = X_i - X_{i+1}$. So the Monte Carlo method to approximate the value of Z is

1. uniformly generate a sequence of samples for X, say $X^{(1)}, ..., X^{(m)}$;
2. map each $X^{(i)}$ to the corresponding $\boldsymbol{\theta}^{(i)}$ and calculate $L^{(i)} = L(X^{(i)}) = L(\boldsymbol{\theta}^{(i)})$;
3. then equation (3.1) can used to approximate Z.

However, we can also work in the opposite direction. Let $X_1, ..., X_m$ be the order sample such that $0 < X_m < ... < X_1 < 1$ and the corresponding likelihood and parameters are $L_i, \boldsymbol{\theta}_i$ where $L_i = L(X_i) = L(\boldsymbol{\theta}_i)$. It is more natural to directly sample the i^{th} point $\boldsymbol{\theta}_i$ from prior distribution $\pi(\boldsymbol{\theta})$, but with the restriction that $L(\boldsymbol{\theta}_i) > L_{i-1}$.

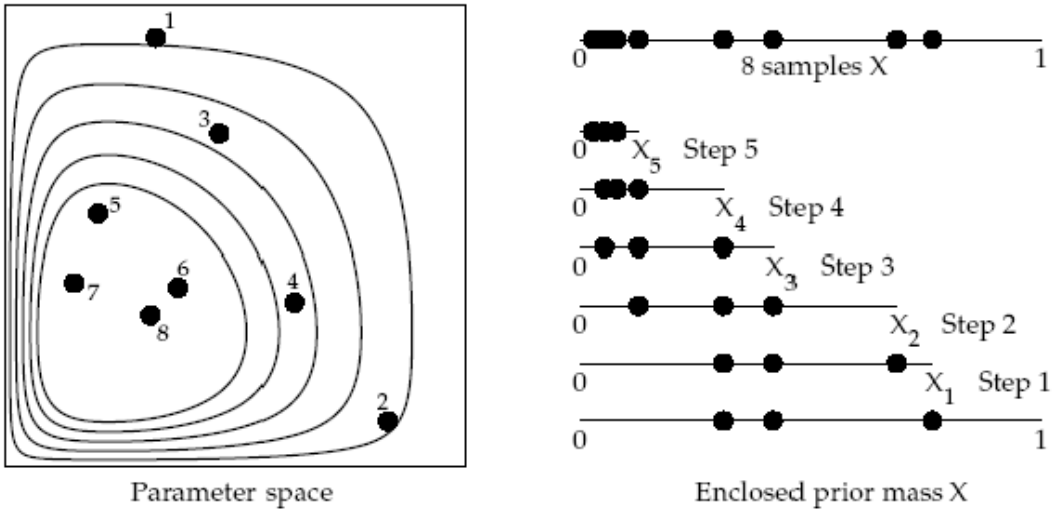
More generally, to obtain X_i, we can sample N points in parameter space which all satisfy $L(\boldsymbol{\theta}_i) > L_{i-1}$ and select the point with lowest L (highest X) as the i^{th} point. This procedure provides $X_i = t_i X_{i-1}$ with $P(t_i) = N t_i^{N-1}$ where $t_i \in (0,1)$. For this distribution, we know $E(\log t) = -1/N$ and $Var(\log t = 1/N^2)$. Since individual $\log t$'s are independent, after i steps X_i is expected to shrink to $\log X_i = -(i \pm \sqrt{i})/N$. Thus, in case of a crude implementation, X_i can be treated as a known value, $\log X_i = -i/N$.

Another trick here is we don't need to sample N new points at each step. Instead, we only need to sample one new point, because $N-1$ points, except the point with lowest L (highest X), from the previous step still satisfy the restriction condition and can be used in current step.

This method is illustrated in Figure 3.1 for the case $N = 3$. At first step, three points are generated, labeled as 1, 3, 4. Point 1 has highest likelihood, so set $L_1 = L(\boldsymbol{\theta}_1)$. Then sample another point under restriction $L(\boldsymbol{\theta}) > L_1$ and the generated point is point 2. Replace point 1 with point 2 and find

the point with highest likelihood, which is point 2, from current N-point sequence 2, 3, 4. So set $L_2 = L(\boldsymbol{\theta}_2)$. In next step, point 5 is generated under restriction $L(\boldsymbol{\theta}) > L_2$ and it replaces point 2. After five steps, the five discarded points $(1,2,3,4,5)$ are augmented with the final three survivors $(6,7,8)$ to approximate the evidence by using equation (3.1).

Figure 3.1: The nested sampling procedure for $N = 3$ case



3.2 ALGORITHM

The algorithm based on the method discussed in previous section is described in below.

1. First: sample N points in parameter-space $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N$ from prior $\pi(\boldsymbol{\theta})$, and set initial values $Z = 0, X_0 = 1$.

2. Loop: for $i = 1, 2, ..., m$

   a) find the point $\boldsymbol{\theta}_l$ with lowest likelihood from current N-point sequence $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N$ and set $L_i = L(\boldsymbol{\theta}_l)$;

b) set $X_i = \exp(-i/N)$ (or sample t_i with $P(t_i) = Nt_i^{N-1}$ where $t_i \in (0,1)$ and set $X_i = t_i X_{i-1}$);

c) set $w_i = X_{i-1} - X_i$ (or $w_i = \frac{X_{i-1} - X_{i+1}}{2}$);

d) update Z by following $Z_i = Z_{i-1} + w_i L_i$;

e) sample one point $\boldsymbol{\theta}_k$ from $\pi(\boldsymbol{\theta})$ with restriction $L(\boldsymbol{\theta}_k) > L_i$ and then replace $\boldsymbol{\theta}_l$ with $\boldsymbol{\theta}_k$ to obtain the new N-point sequence.

3. Last: update Z with addition of $X_m(L(\boldsymbol{\theta}_1) + \ldots + L(\boldsymbol{\theta}_N))/N$.

The last step is based on consideration of boundary effect (more details can be found in John Skilling's paper).

3.3 Termination

The simplest and straightforward way to set termination condition for main loop is to examine the value of current addition $\Delta Z_i = w_i L_i$. If it is very small comparing to current evidence Z_{i-1}, which means the following iterations are not likely to contribute significantly to the accumulation of Z, then terminate the loop. Or if an upper bound $L \leq L_{max}$ can be found, the similar termination condition is to terminate when $\Delta Z_i^* = w_i L_{max} \ll Z_{i-1}$.

Another termination condition discussed in Skilling's paper is: "continue iterating until the count i significantly exceeds NH" where $H = \int \log(\mathrm{d}P/\mathrm{d}X)\,\mathrm{d}P$ represents the information that how much of the prior mass the bulk of the posterior mass contains.

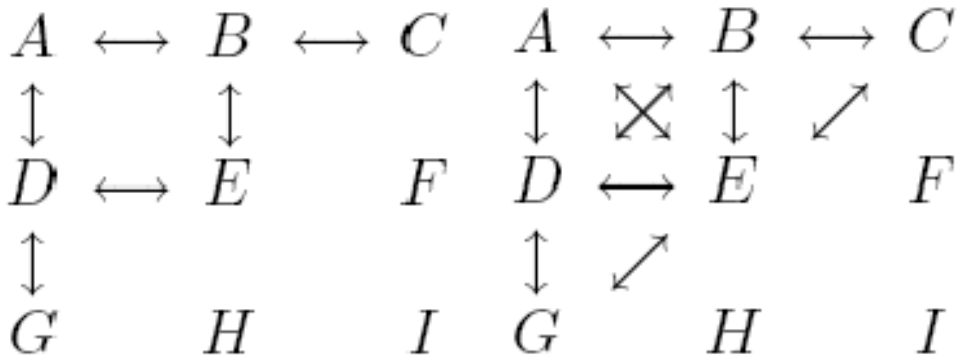
3.4 Difficulty

To sample one point $\boldsymbol{\theta}_k$ from $\pi(\boldsymbol{\theta})$ with restriction $L(\boldsymbol{\theta}_k) > L_i$ is not always possible. Even when it is possible, after certain steps L_i becomes very close to L_max and the potential region of $\boldsymbol{\theta}_k$ in parameter-space that satisfies $L(\boldsymbol{\theta}_k) > L_i$ will shrink to very small region. In this case, if we sample $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta})$ in whole parameter-space, it usually takes too long to get a point that falls in the potential region.

One solution for this difficulty is transforming prior distribution to uniform prior and use MCMC algorithm to explore new point. The procedure is: randomly choose one point from previous $N-1$ survivors and random

walk in parameter-space by certain small distance; use wrap-around or reflection technique when walking out of the restricted region. In Figure 3.2, assume A, B, C, D, E, G are states in restriction region and F, H, I fall outside. So if random walk starts from some state inside of restriction region, the transitions to outside can be blocked by using *wrap-around* technique (when next random walk state is outside we redirect it to certain inside state that is on boundary of the restriction region) or *reflection* technique (redirect it to previous state). Let random walk last long enough, eventually any specified state in the region will be visited with uniform probability.

Figure 3.2: Wrap-around and redirect on the boundary during random walk



4 Implementation and Results

4.1 Case I

The model and the prior are

$$Y_i \sim N(\mu, \sigma^2) \tag{4.1}$$

$$\mu \sim N(0, 100) \tag{4.2}$$

Two exploration methods are tested:

Table 4.1: Performance for exploration method (1)

Step Number	CPU Time (second)	Estimate of Z
10	0.11	5.920883e-05
15	0.21	0.0002560032
20	1.01	0.003112191
25	326.97	0.02808788

Table 4.2: Performance for exploration method (2)

Step Number	CPU Time (second)	Estimate of Z
10	0.18	1.536947e-05
15	2.31	0.000693845
20	237.89	0.01182036

(1) searching in whole parameter space without using any other technique;

(2) start searching from a random survivor point and use Metropolis-Hasting algorithm with fixed successful accept number 20.

The performance results are summarized in Table 4.1 and 4.2. We can see that the estimation of Z converges to the true value quicker when using exploration method (2) than using (1) but in the mean time it takes more time to obtain the point in each step in method (2). A possible reason for this phenomena is that: since the number of successful acceptance in M-H is fixed to 20, the sampling may not be good enough to be proportional to prior density, which could lead to the fact that the generated point tends to be "stuck" around where maximum likelihood is. We can also notice that the searching time increases dramatically after certain steps in method (1) due to the quick shrinkage of restricted region.

Furthermore, we fix the parameter-space in a rectangle,

$$\mu \sim \text{Unif}(-5,5), \sigma^2 \sim \text{Unif}(0,5),$$

and the results shown in Table 4.3 and Figure 4.1 suggest that after 20 steps the estimation almost converges to true value.

Table 4.3: Rectangle parameter region

Step Number	CPU Time (second)	Estimate of Z
10	0.11	0.02711679
15	0.15	0.09266391
20	10.51	0.1534983
25	90.76	0.1537126

Figure 4.1: The trace plot of X and L

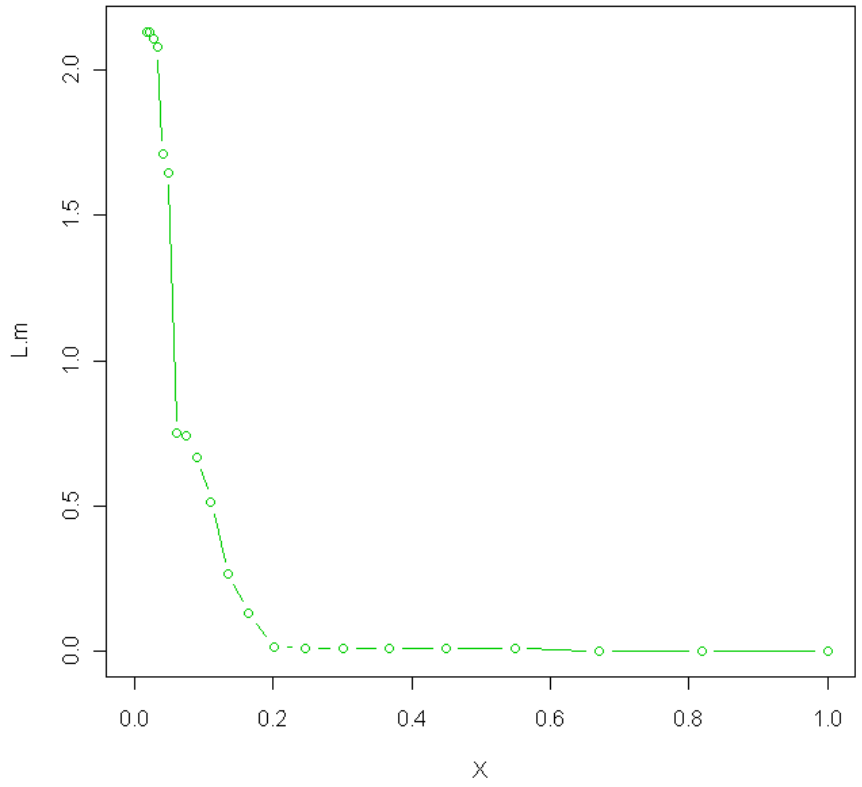


Table 4.4: Exponential distribution

Step Number	CPU Time (second)	Estimate of Z
10	0.05	0.01057942
15	0.42	0.01068455
20	71.29	0.01068455

4.2 Case II

The model and the prior are

$$Y_i \sim \text{Exp}(\lambda) \tag{4.3}$$

$$\lambda \sim \text{Unif}(a, b) \tag{4.4}$$

The results, shown in Table 4.4 and Figure 4.2 and 4.3, suggests a fast converge for exponential distribution.

Figure 4.2: The trace plot of X and L (exponential distribution)

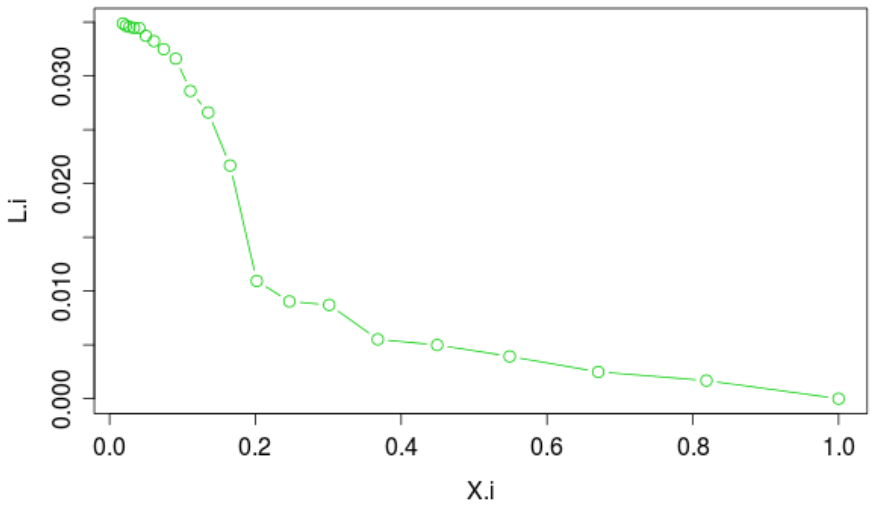
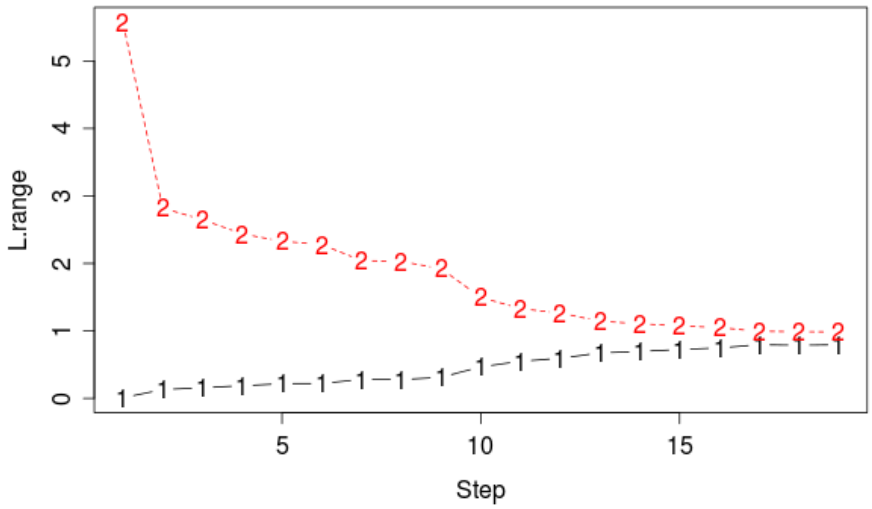


Figure 4.3: The shrinkage of likelihood (exponential distribution)



5 Discussion

As a new computation method, nested sampling is attractive in many aspects which includes,

- applicable generally;
- capable to deal with difficult likelihood function and prior densities;
- calculate the evidence (marginal likelihood) and can be extended to use the posterior samples;
- require very little manual tuning.

However, it still suffer several problems and needs to be improved:

- the searching for potential point in parameter-space could be very slow in some cases;
- the likelihood function with multiple local modes needs to be taken care of to make sure all the restricted sub-regions are included during searching.

REFERENCE

1. John Skilling, Nested Sampling for General Bayesian Computation, Bayesian Analysis (2006) 1, Number 4, pp. 833-860.

2. Nicolas Chopin and Christian P. Robert, Nested sampling for Bayesian computations: A discussion, June 21, 2006.

3. Iain Murray, David J.C. MacKay, Zoubin Ghahramani, and John Skilling, Nested sampling for Potts models.

4. D.S. Sivia and J. Skilling, Data Analysis: A Bayesian Tutorial.