Diffusive Nested Ensemble Sampling

ABSTRACT

This paper contains an exposition and some refinements of diffusive nested sampling as an approach to computing Bayesian evidence of Bayesian model selection. We show that an affine invariant ensemble sampler is effective in some cases. We use the modified algorithm to study multi-companion fits to radial velocity data for stars.

Subject headings: methods: nested sampling — methods: markov chain monte carlo — methods: data analysis — bayesian decision theory

1. Introduction

When presented with competing models and data $\mathcal{D}$, Bayes theorem tells us to compare models as follows:

$$P(\text{Model}_j \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \text{Model}_j)\, P(\text{Model}_j)}{\sum_j P(\mathcal{D} \mid \text{Model}_j)\, P(\text{Model}_j)}. \tag{1}$$

More precisely, suppose model j has parameters $\boldsymbol{\theta}_j = (\theta_1, \ldots, \theta_{n_j})$, where n_j is the dimension of model j. Let $\pi_j(\boldsymbol{\theta}_j)$ be the prior, and let $L_j(\mathcal{D} \mid \boldsymbol{\theta}_j)$ be the likelihood of data $\mathcal{D}$ in model j with parameters $\boldsymbol{\theta}_j$. The probability of data $\mathcal{D}$ given model j in Eqn. (1), which is also the evidence of model j, is

$$Z_j(\mathcal{D}) = P(\mathcal{D} \mid \text{Model}_j) = \int L(\mathcal{D} \mid \boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j \ . \tag{2}$$

Our goal is to estimate Z_j using Markov chain Monte Carlo (MCMC) and diffusive nested sampling.

We have in mind the application to estimating the number of companions about a star based on radial velocity data. Let j represent a model with j companions, and let s represent the number of data sources. The parameters for model j consist of two parameters (velocity offset and jitter) per data source and five orbital parameters per companion. This gives

$$n_j = 2s + 5j$$

in this case.

The *likelihood* in Eqn. (1) is in fact the evidence $Z_{\text{Model }j}$ for model j.

$$Z_{\text{Model }j} \equiv P(\text{Data}|\text{Model}_j) \equiv \int P(\text{Data}|\theta, \text{Model}_j)\, P(\theta|\text{Model}_j)\mathrm{d}\theta, \tag{3}$$

where $P(\text{Data}|\theta, \text{Model}_j)$ is the likelihood of parameter θ for model j and $P(\theta|\text{Model}_j)$ is the prior of parameter θ for model j. To make notations simple, we will use $L(\theta)$ for the likelihood and $\pi(\theta)$

for the prior. We'll also drop the index in $Z_{\mathrm{Model}\,j}$, because the discussion applies to all models. So the evidence can be written as

$$Z = \int L(\theta)\, \pi(\theta)\, \mathrm{d}\theta. \tag{4}$$

The prior $\pi(\theta)$ is normalized in the parameter space, that is

$$\int \pi(\theta)\, \mathrm{d}\theta = 1, \tag{5}$$

while the likelihood $L(\theta)$ is not.

However, evaluating the evidence integral using markov chain Monte Carlo (MCMC) has always been challenging, because the integral Z is the normalization of a probability density. Diffusive Nested Sampling proves to be an efficient and accurate method to evaluate the evidence (Brewer *et al.* 2011). We hope to take advantage of an affine invariant ensemble sampler (Goodman *et al.* 2010; Hou *et al.* 2012; Foreman-Mackey *et al.* 2013) to make diffusive nested sampling even more efficient.

2. Diffusive Nested Sampling

In this section, we explain how diffusive nested sampling works and how to apply the affine invariant sampler to diffusive nested sampling. We first introduce some basic concepts about diffusive nested sampling and describe how diffusive nested sampling works in general. Then, we discuss the algorithm in detail. At last, we try to make the evidence evaluation more accurate and analyze the uncertainty.

We change the variable in the evidence integral from θ to

$$M(L^*) \equiv \int_{L(\theta)>L^*} \pi(\theta)\, \mathrm{d}\theta, \tag{6}$$

which is, in another word, cumulant prior mass covering the area which has likelihood values greater than L^* (Skilling 2006). M is a monotonically decreasing function of L^*, it ranges from 0 to 1. The mapping between M and L^* is a bijection. An infinitesimal increment of M is

$$\mathrm{d}M = \int_{L^*-\mathrm{d}L^*<L(\theta)<L^*} \pi(\theta)\, \mathrm{d}\theta = \pi(\theta) \times \text{volume of}\, \theta, \text{ satisfying}\, L^* - \mathrm{d}L^* < L(\theta) < L^*, \tag{7}$$

where signs have been ignored for clarity. Multiplying both sides by L^* and integrating, we get

$$\int_0^1 L^*\, \mathrm{d}M = \int L(\theta)\, \pi(\theta)\, \mathrm{d}\theta. \tag{8}$$

so the evidence Z can be expressed as

$$Z = \int_0^1 L^*(M)\, \mathrm{d}M. \tag{9}$$

In most cases, it is impossible to know the function $L^*(M)$ analytically. Based on the definition of M, Eqn. (6), if we generate N samples from the prior $\pi(\theta)$, $M(L^*)$ is the proportion of samples which have likelihood larger than L^*. Nested sampling takes advantage of this to find $L^*(M)$ statistically.

2.1. Level and Constrained Prior

In nested sampling, we first try to find several points, $\{(M_0, L_0^*), (M_1, L_1^*), \ldots\}$, on the $L^*(M)$ curve. We call these points 'levels'. The L_j^* is called a level's likelihood threshold or simply threshold. Each level defines a constrained prior,

$$p_j(\theta) = \frac{\pi(\theta)}{M_j}\, \mathbb{1}_{L(\theta)>L_j^*}, \tag{10}$$

where

$$\mathbb{1}_{L(\theta)>L_j^*} = \begin{cases} 1 & \text{if } L(\theta) > L_j^*, \\ 0 & \text{otherwise.} \end{cases}$$

Normalized by M_j, $p_j(\theta)$ is a properly defined probability density function. We can also define a mixture of these constrained priors,

$$p(\theta) = \sum_j w_j\, p_j(\theta),$$

where w_j are weights assigned to each $p_j(\theta)$ and w_j's should sum up to 1. The mixture of constrained priors will be discussed in more detail later.

2.2. Setting Level Thresholds

Our nested sampler first attempts to make the levels' prior masses $M_j = e^{-j}$ and estimate the corresponding likelihood thresholds L_j^*. The algorithm to achieve this is described in Section (2.3). Then we keep L_j^* unchanged and look for a more accurate prior mass $\widehat{M_j}$ that corresponds to L_j^*. We call this procedure prior mass refinement. This is described in Section (2.4). The zeroth level has $M_0 = \widehat{M_0} = 1$ and $L_0^* = 0$.

To estimate L_1^*, we generate N samples from the prior density $\pi(\theta)$, θ_1, ..., θ_N. We choose L_1^* so that the number of k with $L(\theta_k) > L_1^*$ is N/e. This may be done with the *quick find* algorithm that is part of the standard template library (STL) of C++. The actual prior mass corresponding to L_1^* is subject to round-off error. Luckily, we are able to find N's that make M_1 extremely close to e^{-1}.[1] We treat M_j and e^j as synonyms in this paper.

[1] For example, $N = 1084483$ and N/e is rounded to be 398959. $\log(398959/1084483) = -0.999999999999823$.

To estimate the next level (M_2, L_2^*), we need N samples with likelihood larger than L_1^* from the prior. There are many different ways to do this. One is sampling the constrained density $p_1(\theta)$ defined in Eqn. (10). Another would be sampling a mixture of $p_1(\theta)$ and prior $\pi(\theta)$. Sampling the mixture is a better method because the area covered by level 1 may be disconnected in parameter space and only sampling the constrained prior $p_1(\theta)$ may get us stuck in only one or few of those disconnected areas. In order to balance efficiency and the need to circumvent discontinuity, we give the latest level more weight. For example, we can use $w_1/w_0 = e$. We keep sampling until we have a chain of N likelihoods which are all larger than L_1^*, rank these likelihoods in descending order and find the N/e-th likelihood, which we call level 2. L_2^* is the likelihood threshold of level 2. $M_2 = 1/e^2$ is the prior mass that level 2 covers.

Continuing with the method described above, suppose we now have levels (M_0, L_0^*), (M_1, L_1^*), (M_2, L_2^*), ... All these levels are in fact estimations. This can be seen from two aspects. One aspect is that the L_j^*'s are the estimations of the true likelihood thresholds corresponding to prior masses $M_j = e^{-j}$. The other aspect is that the M_j's are the estimations of the true prior masses covered by L_j^*. We choose the 2nd aspect and will refine the prior masses M's in Section (2.4).

There is a simple stopping criterion to tell how many levels are enough, assuming we have solved the optimization problem to find L_{max}. Suppose we already have j new levels besides level 0. The evidence integral is

$$Z = \int_{M_j}^{1} L^*(M)\,\mathrm{d}M + \int_0^{M_j} L^*(M)\,\mathrm{d}M = Z_j + \int_0^{M_j} L^*(M)\,\mathrm{d}M.$$

Because $L^*(M) < L_{max}$ always, the 2nd term cannot be larger than $L_{max}\,M_j$. We choose a stopping point J so that $L_{max}\,M_J \le \epsilon Z_J$. We usually choose $\epsilon = 10^{-6}$. Z_J can be roughly estimated from all the levels already built. We do not simply throw away the integration from 0 to M_J. We just do not build new levels in that interval.

With total J levels, the mixture of constrained priors can be defined as

$$p(\theta) = \sum_{j=0}^{J} w_j\, p_j(\theta), \tag{11}$$

where $p_j(\theta)$ is the constrained prior defined in Eqn. (10) and w_j are the weights of each level which sum up to 1,

$$\sum_{j=0}^{J} w_j = 1. \tag{12}$$

The choice of weight may change according to different purposes. For example, when we are building a new level, we can use ‘exponential’ weight

$$w_j \propto \exp\left(\frac{j-J}{\lambda}\right),$$

where J is the latest level index and λ is some constant. (Brewer *et al.* 2011) But when we refine the prior masses, we need to sample all the levels with equal weight.

Nested sampling can be better understood in the context of importance sampling. The goal is to evaluate the integral Z, defined in Eqn. (8). By performing nested sampling, we are in fact trying to find a probability density similar in shape with the integrand $L(\theta)\,\pi(\theta)$. The mixture of constrained priors $p(\theta)$ is that probability density function we are looking for. However, because both M_j and L_j^* are estimations, $p(\theta)$ is not a well-defined probability density function since normalization depends on both M_j and L_j^*. This is why we will have to refine the levels.

2.3. Nested Sampling by Stretch Move

We define the joint probability density of θ and j as

$$p(\theta, j) = w_j\, p_j(\theta), \tag{13}$$

so $p(\theta)$ defined in Eqn. (11) can be seen as the marginal density of $p(\theta, j)$ summed over j. We can sample $p(\theta, j)$ by partial re-sampling. That is to say, we first sample $p(\theta|j)$ with j fixed and then sample $p(j|\theta)$ with θ fixed. The two conditional probability densities can be expressed more explicitly as

$$p(\theta|j) = \frac{p(\theta, j)}{\int p(\theta, j)\,\mathrm{d}\theta} = p_j(\theta), \tag{14}$$

and

$$p(j|\theta) = \frac{p(\theta, j)}{\sum_j p(\theta, j)} = C\,\frac{w_j}{M_j}\,\mathbb{1}_{L(\theta)>L_j^*} = C\,e^j\,w_j\,\mathbb{1}_{L(\theta)>L_j^*}, \tag{15}$$

where C is the normalization and a function of θ but not of j.

We use stretch move to sample $p(\theta|j)$. Stretch move has the feature of affine invariance and is a very efficient ensemble sampler with low auto-correlation time and few tuning parameters. (Goodman *et al.* 2010) In order to apply stretch move to nested sampling, we assign a level index to every walker in the ensemble. The likelihood threshold of the level assigned to a walker must be smaller than the likelihood of that walker. So a walker with level index j can be seen as a sample from $p_j(\theta)$, a.k.a. $p(\theta|j)$. The ensemble size has to be larger than both the dimension of θ and the total number of j in order not to get stuck in a subspace. The stretch move can be described in pseudo-code as:

- to propose a new location for walker X, randomly choose a helping walker Y in the ensemble different from X.
- propose a new location with stretch move: $X_{new} = Y + z\,(X - Y)$, where z is a random variable from some distribution (Goodman *et al.* 2010).
- accept the proposed X_{new} with probability: $\max\left(z^{\mathrm{d}-1}\,\frac{p_j(X_{new})}{p_j(X_{old})}, 1\right)$.

There are many ways to sample $p(j|\theta)$ defined in Eqn. (15). Simple Metropolis-Hastings would suffice. (Brewer *et al.* 2011) If w_j's are simple functions like constant or exponential, we can also use direct sampling. For example, when w_j are the same for all the levels, the sampling can be summarized as:

- find the largest possible level index j_{max} for walker X.
- the new index is $j_{new} = [j_{max} + \log(U)]$, where U is a uniform random variable and the square bracket means truncating. If $j_{new} < 0$, $j_{new} = 0$.

This is actually slightly biased towards moving to smaller j's but the bias is almost negligible. However, whichever method we use, we in fact have assumed that, if a walker is in level j, the probability that the walker can be assigned to any higher level $k > j$ is proportional to $\exp(j - k)$. This assumption cannot be guaranteed true because samples from $p(\theta|j)$ may not be independent, which means the new samples may be too close to the original ones.

The two procedures described above can happen in any order. In our code, as well as in (Brewer *et al.* 2011), half the times we sample the parameter space first and the level indices second and the other way around for the other half.

2.4. Refining Level Masses

Assume we have constructed J levels following previous sections, $\{(M_1, L_1^*), (M_2, L_2^*), \ldots, (M_J, L_J^*)\}$. We sample $p(\theta, j)$, defined in Eqn. (13), to obtain a long enough chain of both the visited level indices and the likelihoods of the walkers during those visits. In practice, we keep the samples from visiting different levels in different chains. If we have J levels besides level 0, we would have $J + 1$ chains. The length of each of these chains should be at least a few times larger than N which is the length we use to build each level in Section (2.2). For the j-th chain ($0 \le j < J$), we can define an indicator function

$$\mathbb{1}_j(\theta) = \begin{cases} 1 & \text{if } L(\theta) > L_{j+1}^*, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

In another word, $\mathbb{1}_j$ indicates whether the walker is within level $j + 1$ during a visit to level j. Let n_j be the length of the j-th chain. Let n_j^{j+1} be the number of times when the likelihood exceeds level $j + 1$'s threshold L_{j+1}^* during those n_j visits to level j. n_j^{j+1} can be expressed as

$$n_j^{j+1} = \sum_{k=1}^{n_j} \mathbb{1}_j(\theta_k), \tag{17}$$

where k is just a label for samples. The refined prior mass $\widehat{M_j}$ is then defined as (Brewer *et al.* 2011)

$$\widehat{M_j} = \widehat{M_{j-1}} \, \frac{n_{j-1}^j + C \, M_j/M_{j-1}}{n_{j-1} + C} = \widehat{M_{j-1}} \, \frac{n_{j-1}^j + C \, e^{-1}}{n_{j-1} + C}, \tag{18}$$

where C is a constant that reflects one's confidence in the accuracy of original levels. If n_{j-1} and n_{j-1}^j are large enough, C will not be very important. As mentioned before, $\widehat{M_0} = M_0 = 1$ and we start with $\widehat{M_1}$, so $\widehat{M_{j-1}}$ will be known when we arrive at $\widehat{M_j}$.

In fact, if both n_j and n_j^{j+1} are large enough and C is neglectible, $\widehat{M_{j+1}}/\widehat{M_j}$ can be expressed as

$$\frac{\widehat{M_{j+1}}}{\widehat{M_j}} \approx \frac{n_j^{j+1}}{n_j} = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathbb{1}_j(\theta_k). \tag{19}$$

So the expectation of $\widehat{M_{j+1}}/\widehat{M_j}$ is approximately the same as the expectation of $\mathbb{1}_j$. And $\widehat{M_j}$ is

$$\widehat{M_j} \approx \prod_{l=0}^{j-1} \left(\frac{1}{n_l} \sum_{k=1}^{n_l} \mathbb{1}_l(\theta_k) \right). \tag{20}$$

The variance of $\mathbb{1}_j$ is

$$\mathrm{Var}(\mathbb{1}_j) = \mathrm{E}(\mathbb{1}_j) \left(1 - \mathrm{E}(\mathbb{1}_j)\right) \approx \frac{n_j^{j+1}}{n_j} \left(1 - \frac{n_j^{j+1}}{n_j} \right). \tag{21}$$

By central limit theorem, taking the variance of both sides of Eqn. (19), we get

$$\mathrm{Var}\left(\frac{\widehat{M_{j+1}}}{\widehat{M_j}} \right) \approx \mathrm{Var}\left(\frac{1}{n_j} \sum_{k=1}^{n_j} \mathbb{1}_j(\theta_k) \right) = \frac{\mathrm{E}(\mathbb{1}_j) \left(1 - \mathrm{E}(\mathbb{1}_j)\right)}{n_j/\tau} \approx \frac{n_j^{j+1}/n_j \left(1 - n_j^{j+1}/n_j \right)}{n_j/\tau}, \tag{22}$$

where τ is the auto-correlation time of the chain of $\mathbb{1}_j(\theta_k)$, $k = 1, 2, \ldots, n_j$. And we can get the variance of $\widehat{M_j}$ using the following relationship

$$\mathrm{Var}\left(\widehat{M_j}\right) = \mathrm{Var}\left(\frac{\widehat{M_j}}{\widehat{M_{j-1}}} \right) \mathrm{Var}\left(\widehat{M_{j-1}}\right) + \mathrm{Var}\left(\frac{\widehat{M_j}}{\widehat{M_{j-1}}} \right) \mathrm{E}\left(\widehat{M_{j-1}}\right)^2 + \mathrm{E}\left(\frac{\widehat{M_j}}{\widehat{M_{j-1}}} \right)^2 \mathrm{Var}\left(\widehat{M_{j-1}}\right), \tag{23}$$

and starting from $\mathrm{Var}\left(\widehat{M_0}\right) = 0$. The uncertainty of interval $\widehat{M_{j-1}} - \widehat{M_j}$ can be similarly found,

$$\begin{aligned} \mathrm{Var}\left(\widehat{M_{j-1}} - \widehat{M_j}\right) =& \mathrm{Var}\left(\frac{\widehat{M_j}}{\widehat{M_{j-1}}} \right) \mathrm{Var}\left(\widehat{M_{j-1}}\right) + \mathrm{Var}\left(\frac{\widehat{M_j}}{\widehat{M_{j-1}}} \right) \left(1 - \mathrm{E}\left(\widehat{M_{j-1}}\right)\right)^2 \\ &+ \mathrm{E}\left(\frac{\widehat{M_j}}{\widehat{M_{j-1}}} \right)^2 \mathrm{Var}\left(\widehat{M_{j-1}}\right). \end{aligned} \tag{24}$$

The only difference form $\mathrm{Var}\left(\widehat{M_j}\right)$ is in the 2nd term.

While sampling the mixture of constrained priors Eqn. (11), it is possible that the weights w_j are not sampled as desired. One cause of this is that the levels are not constructed accurately, which means the prior mass that L^*_j covers is not exactly e^{-1} of the prior mass that L^*_{j-1} covers. Another cause is that the samples are not independent, which is talked about in Section (2.3). As a result, some of the n_j's and n_j^{j+1}'s are too small to make a meaningful refinement. In such cases, we can use the number of visits to each level n_j to enforce that the weights w_j be sampled as desired (Brewer *et al.* 2011). Although such enforcement would violate the Markov property, the violation only happens in sampling the indices and does not happen in the sampling of the parameter space. So the estimation of Z should not be affected.

2.5. Computing Evidence

We take the mean of likelihoods sandwiched between two levels,

$$\bar{L}_j = \frac{1}{l_j} \sum_{L^*_j \le L(\theta) < L^*_{j+1}} L(\theta), \tag{25}$$

where l_j is the number of samples sandwiched between level j and level $j+1$. For the sake of notation simplicity, we can add one extra level whose likelihood threshold L^*_{J+1} is the optimum likelihood and whose prior mass M_{J+1} is 0. So the estimation of the evidence can be expressed as

$$\widehat{Z} = \sum_{j=0}^{J} \bar{L}_j \left(\widehat{M_j} - \widehat{M_{j+1}}\right). \tag{26}$$

The variance of the evidence Z can be estimated via

$$\mathrm{Var}\left(\widehat{Z}\right) \approx \sum_{j=0}^{J} \bar{L}_j^2 \,\mathrm{Var}\left(\widehat{M_j} - \widehat{M_{j+1}}\right) + \mathrm{Var}\left(\bar{L}_j\right) \left(\widehat{M_j} - \widehat{M_{j+1}}\right)^2, \tag{27}$$

where $\mathrm{Var}\left(\widehat{M_j} - \widehat{M_{j+1}}\right)$ can be calculated using Eqn. (24) and

$$\mathrm{Var}\left(\bar{L}_j\right) \approx \frac{1}{{l_j}^2/\tau_j} \sum_{L^*_j \le L(\theta) < L^*_{j+1}} \left(L(\theta) - \bar{L}_j\right)^2, \tag{28}$$

where τ_j is the auto-correlation time for $L(\theta)$ chain that satisfies $L^*_j \le L(\theta) < L^*_{j+1}$. In Eqn. (27), cross terms are ignored because $\widehat{M_j} - \widehat{M_{j+1}}$ and $\bar{L}_j$ are independent. Also ignored is the the product of the variance of both $\widehat{M_j} - \widehat{M_{j+1}}$ and $\bar{L}_j$ because its contribution is of smaller order.

REFERENCES

Brewer, B. J., Pártay, L. B. & Csányi, G, 2011, Statistics and Computing, 21, 649

Foreman-Mackey, D., Hogg, D. W., Lang, D., Goodman, J., http://arxiv.org/abs/1202.3665

Goodman, J., Weare, J., 2010, Comm. App. Math. and Comp. Sci., 5, 65

Hou, F., Goodman, J., Hogg, D. W., Weare, J., Schwab, C., 2012, ApJ, 745, 198

Skilling, J., 2006, Bayesian Analysis, 4, 833

This preprint was prepared with the AAS LaTeX macros v5.2.