Diffusive Nested Ensemble Sampling

ABSTRACT

This paper contains an exposition and some refinements of nested sampling as an approach to computing Bayesian evidence for Bayesian model selection. We show that an affine invariant ensemble sampler is effective in some cases. We use the modified algorithm to study multi-planet fits to radial velocity data for star *****. Computations show that there is an order of magnitude more Bayesian evidence for a 5 planet model than for models with fewer or more planets.

1. A

More precisely, suppose model j has parameters $\boldsymbol{\theta}_j = (\theta_1, \ldots, \theta_{n_j})$. Let $\pi_j(\boldsymbol{\theta}_j$ be the prior, and let $L_j(\mathcal{D} \mid \boldsymbol{\theta}_j)$ be the probability of data $\mathcal{D}$ in model j with parameters $\boldsymbol{\theta}_j$. If model j is correct, then the probability of data $\mathcal{D}$ is

$$Z_j(\mathcal{D}) = \int L(\mathcal{D} \mid \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j \ . \tag{1}$$ Zj

A Bayesian approach to model selection is to use prior probabilities P_j for model j. The posterior probability of model j is

$$P(j \mid \mathcal{D}) = \frac{Z_j(\mathcal{D})P_j}{\sum_k Z_k P_k} \ .$$

This gives $P(j \mid \mathcal{D})$ (after a normalization) as the product of the prior and the Bayesian evidence Z_j. Our goal is to estimate Z_j using MCMC and nested sampling.

We have in mind the application to estimating the number of planets about a star. Let j represent a model with j planets. The parameters for model j consist of two common parameters (velocity offset and jitter) and five orbital parameters per planet. This gives

$$n_j = 2 + 5j$$

in this case.

2. B

For several sections we refer to the evidence integral without the model selection context. The evidence integral is simply

$$Z = \int L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \ , \tag{2}$$ Z

where $\pi(\boldsymbol{\theta})$ is some probability density function and $L(\boldsymbol{\theta}) = L(\mathcal{D} \mid \boldsymbol{\theta})$ is some likelihood function. Nested sampling is an alternative to direct Monte Carlo integration of the evidence integral (Z). Direct estimation would use N samples, $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta})$, and make the estimate

$$Z \approx \frac{1}{N} \sum_{k=1}^{N} L(\boldsymbol{\theta}_k) \; .$$

This approach is "correct" in the sense that the right side converges to Z in the hypothetical limit $N \to \infty$. But is is impractical in situations where the data severely constrain $\boldsymbol{\theta}$. In that case, it is exceeding unlikely that $\boldsymbol{\theta}$ drawn "at random" from π is a good fit to the data. Mathematically, this means that all but a very small the part of $\boldsymbol{\theta}$ space contributes contributes little to the integral (Z).

Importance sampling IS is a Monte Carlo variance reduction strategy for situations like this. Nested sampling constructs an importance function using level surfaces of the likelihood function. If L^* is some likelihood "level", the corresponding probability mass is

$$M(L^*) = \int_{L(\boldsymbol{\theta})>L^*} \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = \mathrm{Prob}_\pi(L > L^*) \; . \tag{3}$$ M

blabla

Nested sampling is a two stage algorithm. The first stage constructs an importance function by estimating the levels L_j with

$$M(L_j) = M_j = e^{-j} \; . \tag{4}$$ Lj

Once the L_j are known (approximately), the *constrained priors*

$$p_j(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{M_j} \mathbb{1}_{L(\boldsymbol{\theta})>L_j} \; , \tag{5}$$ pj

$\mathbb{1}_{***}$ is the indicator function

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

These are the probability densities of $\boldsymbol{\theta}$ conditioned on $L(\boldsymbol{\theta}) > L_j$.

The second phase of nested sampling estimates the integral (Z) using a weighted sum of constrained priors

$$p(\boldsymbol{\theta}) = \sum_j w_j p_j(\boldsymbol{\theta}) \; . \tag{6}$$ p

The corresponding probability ratio is

$$R(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} = \frac{1}{\sum\limits_{L_j<L(\boldsymbol{\theta})} w_j} \; . \tag{7}$$ R

Simple algebra shows that evidence integral (2) is

$$Z = \int L(\boldsymbol{\theta}) R(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \; . \tag{8}$$ ZR

An idealized nested sampling algorithm would draw N samples $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta})$, and form the approximation

$$Z \approx \frac{1}{N} \sum_{k=1}^{n} L(\boldsymbol{\theta}_k) R(\boldsymbol{\theta}_k) \; . \tag{9}$$ Zns

We present computations using these ideas that construct on the order of 100 levels. This "localizes" the evidence integral (2) to a region of parameter space whose prior probability is something like e^{-100}.

The following sections contain details of various aspects of the algorithm we use. Section *** explains how we estimate the levels L_j. Section *** explains sampling from the constrained prior using the emcee algorithm. Section *** gives the weights we use in practice and how we sample the weighted distribution $p(\boldsymbol{\theta})$.

3. about 2.3

This section describes a sampler for the weighted distribution (6). We do not require that $M_j = e^{-j}$ but we do require that $M_j = M(L_j)$, so that $p_j(\boldsymbol{\theta})$, see (5), is properly normalized as a probability density. Suppose the levels run from $j = 0$ to $j = J$. Consider the pair $(j, \boldsymbol{\theta})$ to be random, with probability distribution

$$p(j, \boldsymbol{\theta}) = w_j p_j(\boldsymbol{\theta}) \; . \tag{10}$$ pjth

It is clear that if the pair $(j, \boldsymbol{\theta})$ has this distribution, then $\boldsymbol{\theta}$ has the distribution p. We sample $p(\boldsymbol{\theta})$ by sampling $p(j, \boldsymbol{\theta})$ and saving only the $\boldsymbol{\theta}$ values to use in (9).

We use a heat bath type MCMC strategy (also called the Gibbs sampler) to sample $p(j, \boldsymbol{\theta})$. This alternates between resampling $\boldsymbol{\theta}$ for fixed j, and resampling j for fixed $\boldsymbol{\theta}$. This requires expressions for the conditional distribution of $\boldsymbol{\theta}$ given j and vice versa. One of these is clearly

$$p(\boldsymbol{\theta} \mid j) = p_j(\boldsymbol{\theta}) \; . \tag{11}$$ bthcj

For the other one we rewrite $p(j, \boldsymbol{\theta})$ as

$$p(j, \boldsymbol{\theta}) = \left(w_j \frac{1}{M(L_j)} \mathbb{1}_{L(\boldsymbol{\theta}) > L_j} \right) \pi(\boldsymbol{\theta}) \; .$$

Only the part in parentheses depends on j. For fixed $\boldsymbol{\theta}$, the allowed j values are those with $L_j < L(\boldsymbol{\theta})$. The largest allowed j for given $\boldsymbol{\theta}$ is

$$j_{max}(\boldsymbol{\theta}) = \max \{\, j \text{ with } L_j < L(\boldsymbol{\theta}) \} \; .$$

Therefore, for a fixed $\boldsymbol{\theta}$, the j distribution is

$$p(j \mid \boldsymbol{\theta}) = C(\boldsymbol{\theta}) \, \frac{w_j \mathbb{1}_{j \le j_{max}}(\boldsymbol{\theta})}{M(L_j)} \; . \tag{12}$$ pjcth

We sample the $\boldsymbol{\theta}$ distribution (11) using the affine invariant stretch move sampler from the emcee package [?]. This has the advantage of being able to sample highly anisotropic distributions without problem dependent tuning [?]. The sampler uses an *ensemble* of L *walkers*, each of which is a pair $(j_k, \boldsymbol{\theta}_k)$ distributed by $p(j, \boldsymbol{\theta})$. The ensemble is the list $[(j_1, \boldsymbol{\theta}_1), \ldots, (j_L, \boldsymbol{\theta}_L)]$. The target ensemble distribution is that the $(j_k, \boldsymbol{\theta}_k)$ are independent samples of $p(j, \boldsymbol{\theta})$, see [?] for more exposition. The acceptance rule below is

$$\text{Prob(accept)} = \max\left(z^{n-1} \frac{p_j(\boldsymbol{\theta}'_k)}{p_j(\boldsymbol{\theta}_k)}, \, 1 \right) \; . \tag{13}$$ ar

Here n is the dimension of the parameter space. The emcee references explain why this rule works. They also give a formula and a sampling algorithm for the stretch parameter distribution $p_a(z)$ below. The following pseudocode describes the algorithm for one sweep through the ensemble.

for $k = 1, \ldots, L$ **do**
  choose $m \in \{1, \ldots, L\}$, $m \neq k$, at random // find a stretch move partner
  choose stretch $z \sim p_a(z)$
  set $\boldsymbol{\theta}'_k = \boldsymbol{\theta}_m + z(\boldsymbol{\theta}_k - \boldsymbol{\theta}_m)$ // propose new $\boldsymbol{\theta}_k$
  evaluate $p_k(\boldsymbol{\theta}'_k)$, accept or reject // see (13) for the acceptance rule
  if accept, $\boldsymbol{\theta}_k \to \boldsymbol{\theta}'_k$
  resample j_k from the distribution (12)
end for

The algorithm above is a generalization of the emcee algorithm as previously described. The difference here is that the helper walker $\boldsymbol{\theta}_m$ is drawn from a distribution that is probably distinct from the distribution of $\boldsymbol{\theta}_k$. The distributions are distinct if $j_k \neq j_m$. The justification is given by the following fact, which we would state as a lemma if we were writing for mathematicians. Suppose $X \sim p(x)$ and $Y \sim q(y)$ are two random variables in $\mathbb{R}^n$. Suppose we propose $X' = Y + z(X - Y)$ and accept/reject according to the stretch move rule. Then the new (X, Y) pair also are independent samples from p and q respectively. The proof is to repeat the justification given in [?] and notice that it works just as well if the Y distribution is $q \neq p$.