Evaluating Marginalized Likelihood of Substellar-Companion Models with Radial Velocity Data by means of Diffusive Nested Ensemble Sampling

ABSTRACT

Fully marginalized likelihood, or bayesian evidence, is of great significance in bayesian data analysis, because fully marginalized likelihood directly gives us the probability of a model or a mixture of models being true, furthermore it also provides us with a properly normalized posterior distribution, which enables population study in the framework of hierarchical models. This paper contains an exposition of diffusive nested sampling as an approach to computing the marginalized likelihood. We also refine diffusive nested sampling by applying an affine invariant ensemble sampler. We show that the modified diffusive nested sampling performs efficiently and returns the correct results on trial problems. We apply the algorithm to the problem of multi-companion model fitting of the radial velocity data of HIP 88048. HIP 88048 has two confirmed companions announced by Quirrenbach in 2011. We are able to evaluate the fully marginalized likelihood of 1-, 2-, 3-, and 4-companion models given the radial velocity data of HIP 88048 under our choice of the prior distribution. We find that the 2-companion model indeed has the largest marginalized likelihood, among from 1- up to 4-companion models.

Subject headings: methods: nested sampling — methods: markov chain monte carlo — methods: data analysis — bayesian decision theory — stars: individual (HIP 88048)

1. Introduction

As of today, hundreds of extrasolar planets have been discovered through various methods. Among them, about 60% have been found through radial velocity (RV) surveys (Mitchell *et al.* 2013). Bayesian inference is usually employed to make parameter estimations given the RV data. One of the advantages of bayesian data analysis is that it quantifies uncertainty naturally through the use of probability (Gelman *et al.* 2004). However, one piece of the puzzle that usually eludes bayesian statistician is how to evaluate the fully marginalized likelihood.

Fully marginalized likelihood, or simply marginalized likelihood, is defined as follow: Given data $\mathcal{D}$, the posterior probability of model m_n is

$$P(\mathrm{m}_n \mid \mathcal{D}) \equiv \frac{P(\mathcal{D} \mid \mathrm{m}_n)\, P(\mathrm{m}_n)}{\sum_n P(\mathcal{D} \mid \mathrm{m}_n)\, P(\mathrm{m}_n)} \,, \tag{1}$$

where $P(\mathrm{m}_n)$ is the prior probability of model m_n and $P(\mathcal{D} \mid \mathrm{m}_n)$ is the likelihood of data $\mathcal{D}$ given model m_n, which is *marginalized likelihood* of model m_n over the model's whole parameter space.

More explicitly, suppose model m_n has parameters $\boldsymbol{\theta}_n = (\theta_1, \ldots, \theta_{d_n})$, where d_n is the dimension. Let $\pi_n(\boldsymbol{\theta}_n)$ be the prior of $\boldsymbol{\theta}_n$, and $L_n(\mathcal{D} \mid \boldsymbol{\theta}_n)$ be the likelihood of data $\mathcal{D}$ given $\boldsymbol{\theta}_n$ within model m_n. The posterior probability of $\boldsymbol{\theta}_n$ given data $\mathcal{D}$ is

$$P(\boldsymbol{\theta}_n \mid \mathcal{D}\,,\, \mathrm{m}_n) \equiv \frac{P(\mathcal{D} \mid \boldsymbol{\theta}_n\,,\, \mathrm{m}_n)\, P(\boldsymbol{\theta}_n \mid \mathrm{m}_n)}{P(\mathcal{D} \mid \mathrm{m}_n)} = \frac{L(\mathcal{D} \mid \boldsymbol{\theta}_n)\pi(\boldsymbol{\theta}_n)}{\int L(\mathcal{D} \mid \boldsymbol{\theta}_n)\pi(\boldsymbol{\theta}_n)\mathrm{d}\boldsymbol{\theta}_n}\ . \tag{2}$$

The denominator $P(\mathcal{D} \mid \mathrm{m}_n)$ is exactly the marginalized likelihood in Eqn. (1),

$$Z_{\mathrm{m}\,n}(\mathcal{D}) \equiv P(\mathcal{D} \mid \mathrm{m}_n) = \int L(\mathcal{D} \mid \boldsymbol{\theta}_n)\, \pi_n(\boldsymbol{\theta}_n)\, \mathrm{d}\boldsymbol{\theta}_n\ .$$

For several sections we refer to the marginalized likelihood without the context of different models, so we can drop the indexes for models. The marginalized likelihood is simply

$$Z = \int L(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\ , \tag{3}$$

where $L(\boldsymbol{\theta}) = L(\mathcal{D} \mid \boldsymbol{\theta})$.

Marginalized likelihood is of central importance in bayesian statistical analysis. First of all, marginalized likelihood makes possible the bayesian decision theory given loss functions or utilities under various scenarios. In the context of exoplanet or substellar companion study using RV data, the probability of a multi-companion model or a mixture of multi-companion models is provided directly by marginalized likelihoods of all the models under consideration. Second, marginalized likelihood gives us a properly normalized posterior distribution, which is essential in population study under the framework of bayesian hierarchical modelling. And the results from population study will have a great impact on star and planet formation theories.

To either make maximum-likelihood parameter estimation or perform bayesian statistical analysis, Markov chain Monte Carlo (MCMC) are often employed. But simple MCMC is usually not enough to undertake the task of evaluating marginalized likelihood (Ford *et al.* 2007). This is especially true in the case of multi-companion fit to the RV data. In such cases, the posterior distribution often exhibit multi-peak, 'fat tail', or other unpredictable deviation from any gaussian approximation.

Various approaches have been developed to robustly evaluate marginalized likelihood, including parallel tempering (Ford *et al.* 2007), nested sampling (Skilling 2006), and diffusive nested sampling (Brewer *et al.* 2011). We choose diffusive nested sampling to work with, because it can be viewed as a combination of both parallel tempering and nested sampling, and it is easier to incorporate diffusive nested sampling with the affine-invariant ensemble sampler (Goodman *et al.* 2010; Hou *et al.* 2012; Foreman-Mackey *et al.* 2013).

We apply our algorithm to the radial velocity data of HIP 88048 (ν Oph) from the Lick K-Giant Search (Frink 2002; Mitchell *et al.* 2003; Hekker *et al.* 2006, 2008; Quirrenbach *et al.* 2011). We choose HIP 88048 to study because it has two confirmed brown-dwarf companions of approximately

530-d period and 3210-d period (Quirrenbach *et al.* 2011). Also the noise level of HIP 88048 is low, so over-fitting should be more obvious when excessive amount of companions are added to the model. We are mainly concerned with multi-companion models, so we include all possible sub-stellar companions in our priors, and do not confine ourselves to the planet regime. Let k represent a model with k companions, and there are 5 parameters per companion. Because our data are from a single observatory, meaning one jitter and one offset, the dimension of model k is $2 + 5k$. We evaluate the marginalized likelihood of 1-, 2-, 3-, and 4-companion model based on the data. The result shows that 2-companion model has the largest marginalized likelihood among them, consistent with previous conjectures.

In Section (2), we discuss in detail the alogrithm of diffusive nested sampling and how we apply affine-invariance ensemble sampler to it. In Section (3), we present the results of the Rosenbrock trial problem. In Section (4), we present the marginalized likelihood of different models for HIP 88048. Finally, In Section (5), we discuss our findings and future projects.

2. Diffusive Nested Sampling

2.1. Importance Sampling

The basic idea behind diffusive nested sampling is analogous to importance sampling. To evaluate the marginalized likelihood (3), the simplest way is direct Monte Carlo integration. Direct integration would use N samples from the prior, $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta})$, and make estimator

$$\widehat{Z} \equiv \frac{1}{N} \sum_{k=1}^{N} L(\boldsymbol{\theta}_k) \; . \tag{4}$$

This approach is 'correct' in the sense that the $\widehat{Z}$ converges to Z in the hypothetical limit $N \to \infty$. But it is impractical in situations where the data severely constrain $\boldsymbol{\theta}$. In that case, it is exceedingly unlikely that $\boldsymbol{\theta}_k$ drawn 'at random' from $\pi(\boldsymbol{\theta})$ is a good fit to the data. Mathematically, this means that all but a very small part of $\boldsymbol{\theta}$ space contributes little to the integral (3).

Diffusive nested sampling, on the other hand, uses an adaptive version of importance sampling, which is a Monte Carlo variance reduction strategy for situations like this. Diffusive nested sampling tries to find a 'better' probability $p(\boldsymbol{\theta})$ using values from the likelihood function $L(\boldsymbol{\theta})$ so as to focus on the part of the $\boldsymbol{\theta}$ space that contributes the most to the integral (3). The corresponding probability ratio is

$$R(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \; . \tag{5}$$

The marginalized likelihood (3) becomes

$$Z = \int_0^1 L(\boldsymbol{\theta})\, R(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \; . \tag{6}$$

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

$$\widehat{Z} \equiv \frac{1}{N} \sum_{k=1}^{N} L(\boldsymbol{\theta}_k) R(\boldsymbol{\theta}_k) \ . \tag{7}$$

$p(\boldsymbol{\theta})$ is better in the sense that $\widehat{Z}$ has small variance.

However, the analogy between diffusive nested sampling and importance sampling stops here. In diffusive nested sampling, we usually cannot apply Eqn. (7) directly as we do in importance sampling, because we only know $p(\boldsymbol{\theta})$ approximately.

2.2. Prior Mass and Constrained Prior

If L^* is some likelihood value, we define the corresponding *prior probability mass* to be

$$M(L^*) \equiv \int_{L(\boldsymbol{\theta})>L^*} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \mathrm{Prob}_\pi(L > L^*) \ , \tag{8}$$

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

$$\mathrm{d}M = \int_{L^*-\mathrm{d}L^*<L(\boldsymbol{\theta})<L^*} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \ . \tag{9}$$

In other words, $\mathrm{d}M$ is $\pi(\boldsymbol{\theta})$ times the incremental volume in $\boldsymbol{\theta}$ space that satisfies $L^* - \mathrm{d}L^* < L(\boldsymbol{\theta}) < L^*$. Multiplying both sides by L^* and integrating, we get

$$\int_0^1 L^* \, \mathrm{d}M = \int L(\boldsymbol{\theta}) \, \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \ . \tag{10}$$

so the marginalized likelihood (3) can be expressed as

$$Z = \int_0^1 L^*(M) \, \mathrm{d}M \ . \tag{11}$$

In most cases, it is impossible to know the function $L^*(M)$ analytically. But based on the definition of M, see Eqn. (8), if we generate N samples from the prior $\pi(\boldsymbol{\theta})$, $M(L^*)$ can be estimated via the proportion of samples which have likelihood larger than L^*.

Diffusive nested sampling is a two-stage algorithm. The first stage finds several points (M_j, L_j^*) of the function $L^*(M)$, where

$$M(L_j^*) \equiv M_j \equiv e^{-j} \ . \tag{12}$$

We call (M_j, L_j^*) the *levels* and L_j^* the *level thresholds*. Note that although we set $M_j \equiv e^{-j}$ and then look for the corresponding L_j^*, in practice, once we find L_j^*, we view M_j only as an crude

approximation of the true prior mass that L_j^* covers. We will use M_j^* (with an asterisk) to stand for the true prior mass of level j. For every (M_j, L_j^*) pair , the *constrained priors* are defined as

$$p_j(\boldsymbol{\theta}) \equiv \frac{\pi(\boldsymbol{\theta})}{M_j} \mathbb{1}_{L(\boldsymbol{\theta})>L_j^*} , \tag{13}$$

where $\mathbb{1}$ is the indicator function

$$\mathbb{1}_{L(\boldsymbol{\theta})>L_j^*} \equiv \begin{cases} 1 & \text{if } L(\boldsymbol{\theta}) > L_j^* , \\ 0 & \text{otherwise .} \end{cases} \tag{14}$$

We can also define a weighted sum of all the available constrained priors as a *mixture of constrained priors*

$$p(\boldsymbol{\theta}) \equiv \sum_j w_j \, p_j(\boldsymbol{\theta}) , \tag{15}$$

where

$$\sum_j w_j = 1 . \tag{16}$$

The details about finding levels and the choice of weights will be discussed in Section (2.3). The problem here is that, because $M_j \equiv e^{-j}$ is only an approximation of the true prior mass M_j^*, $p_j(\boldsymbol{\theta})$ is not properly normalized as a probability density function. But this does not affect the final result, because when we sample the parameter $\boldsymbol{\theta}$ space, we only stay within one single level, that is to say keeping the value j fixed while sampling $\boldsymbol{\theta}$. Details about this sampling will be discussed in Section (2.4).

The second stage of diffusive nested sampling finds a better estimator $\widehat{M_j^*}$ of the true M_j^* covered by L_j^* by counting the visits to each level and whether the walker has a likelihood larger than the next level's threshold. This procedure is called refining the levels, and will be discussed in Section (2.5). We will also discuss the variance and covariance of $\widehat{M^*}$'s in this section. These will be important for understanding the error bar of our estimation of marginalized likelihood Z.

We present computations using these ideas that construct of order 100 levels. This 'localizes' the marginalized likelihood (3) to a region of parameter space that contains total prior probability $\sim e^{-100}$.

2.3. Setting Level Thresholds

The zeroth level has $M_0 = M_0^* = 1$ and $L_0^* = 0$. No likelihood bound means M_0 covers the whole prior space. To find L_1^*, we generate N samples $\boldsymbol{\theta}_k$ from the prior density $\pi(\boldsymbol{\theta})$. We choose L_1^* so that the number of $\boldsymbol{\theta}_k$ with $L(\boldsymbol{\theta}_k) > L_1^*$ is N/e. This may be done with the *quick find* algorithm that is part of the standard template library (STL) of C++. $M_1 \equiv e^{-1}$ is the approximate prior

mass that level 1 covers. [1]

To find the next level $\left(M_2 \equiv e^{-2},\, L_2^*\right)$, 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 $p_1(\boldsymbol{\theta})$ defined in Eqn. (13). Another would be sampling a mixture of $p_1(\boldsymbol{\theta})$ and prior $\pi(\boldsymbol{\theta})$. 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 $p_1(\boldsymbol{\theta})$ 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. $M_2 \equiv e^{-2}$ is the approximate prior mass that level 2 covers.

There is a simple stopping criterion to tell how many levels are enough, assuming we have a good estimate of L_{max}. Suppose we already have $n_{\rm levels}$ new levels besides level 0. The marginalized likelihood 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 $L_{max}\, M_J \leq \epsilon Z_J$. 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, because $L^*(M)$ is 'flat' enough.

With total J levels, the weighted sum of constrained priors or the mixture of constrained priors is defined as

$$p(\boldsymbol{\theta}) \equiv \sum_{j=0}^{J} w_j\, p_j(\boldsymbol{\theta})\ , \tag{17}$$

where $p_j(\boldsymbol{\theta})$ is the constrained prior defined in Eqn. (13) and w_j are the weights of each level which sum up to 1,

$$\sum_{j=0}^{J} w_j = 1\ . \tag{18}$$

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

$$w_j \propto \exp\left(\frac{j-J}{\lambda}\right)\ , \tag{19}$$

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

---

[1] We use fraction to approximate e^{-1} which is subject to round-off error. But luckily we are able to find N's that make $[N/e]/e$ extremely close to e^{-1}, where $[\cdot]$ stands for rounding. For example, $N = 1084483$ and N/e is rounded to be 398959. $\log\left(398959/1084483\right) = -0.999999999999823$. So we treat M_j and e^{-j} as synonyms in this paper.

2.4. Nested Sampling by Stretch Move

This section describes a sampler for weighted distribution (17). Consider the pair $(j, \boldsymbol{\theta})$ to be random variable, with joint probability density

$$p(j, \boldsymbol{\theta}) = w_j \, p_j(\boldsymbol{\theta}) \; . \tag{20}$$

So $p(\boldsymbol{\theta})$ is $p(j, \boldsymbol{\theta})$ marginalized over j.

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

$$p(\boldsymbol{\theta} \mid j) = p_j(\boldsymbol{\theta}) \; . \tag{21}$$

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

$$p(j, \boldsymbol{\theta}) = \left(w_j \frac{1}{M_j} \mathbb{1}_{L(\boldsymbol{\theta}) > L_j^*} \right) \pi(\boldsymbol{\theta}) \; . \tag{22}$$

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 a given $\boldsymbol{\theta}$ is

$$j_{max}(\boldsymbol{\theta}) = \max \{ \, j \text{ with } L_j < L(\boldsymbol{\theta}) \} \; . \tag{23}$$

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

$$p(j \mid \boldsymbol{\theta}) = C(\boldsymbol{\theta}) \frac{w_j \mathbb{1}_{j < j_{max}}(\boldsymbol{\theta})}{M_j} \; . \tag{24}$$

Sampling the distribution (24) is very straightforward. But one should keep in mind that M_j is only an estimation and may deviate from the true M_j^* a lot, we may not get the expected number of visits to each level. The remedy is that, in such cases, we can use the actual number of visits to each level 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 j space, and does not affect the sampling of the parameter $\boldsymbol{\theta}$ space and the estimation of Z.

The distribution (21) can be sampled with the affine invariant stretch move sampler from the emcee package (Foreman-Mackey *et al.* 2013). This has the advantage of being able to sample highly anisotropic distributions without problem dependent tuning (Hou *et al.* 2012). 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 (Goodman *et al.* 2010) for more exposition. The acceptance rule is

$$\text{Prob(accept)} = \max \left(z^{d-1} \frac{p_j(\boldsymbol{\theta}_k')}{p_j(\boldsymbol{\theta}_k)}, \, 1 \right) \; . \tag{25}$$

Here d 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 pseudo-code 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 (25) for the acceptance rule
  if accept, $\boldsymbol{\theta}_k \to \boldsymbol{\theta}'_k$
  resample j_k from the distribution (24)
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. Suppose $X \sim p(x)$ and $Y \sim q(y)$ are two random variables in $\mathbb{R}^n$. And we propose $X' = Y + z(X - Y)$ and accept/reject according to the stretch move rule (25). Then the new (X, Y) pair also are independent samples from p and q respectively. The proof is to repeat the justification given in (Goodman *et al.* 2010) and notice that the distribution of the partner Y does not affect the acceptance rule (25) at all.

2.5. Refining Levels

Assume we have constructed J levels following previous sections. With the algorithm described in Section (2.4), we sample $p(j, \boldsymbol{\theta})$ to obtain a long enough chain of both the visited level indexes and the likelihoods of the walkers during those visits. For a sample $(j, \boldsymbol{\theta})$, we define the following function

$$X(j, \boldsymbol{\theta}) \equiv \begin{cases} 1 & \text{if } L(\boldsymbol{\theta}) > L^*_{j+1} \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{26}$$

In this case, $X(j, \boldsymbol{\theta})$'s with fixed j can be seen as samples from a Bernoulli distribution

$$X(j, \boldsymbol{\theta}_1), \ldots, X(j, \boldsymbol{\theta}_{n_j}) \sim \text{Bernoulli}(R^*_j) \,, \tag{27}$$

where n_j is the number of samples that visit level j and the success probability is

$$R^*_j = \frac{M^*_{j+1}}{M^*_j} \,. \tag{28}$$

The mean of $X(j,\boldsymbol{\theta})$'s with fixed j is an unbiased estimator of R_j^*,

$$\widehat{R_j^*} \equiv \frac{n_j^{j+1}}{n_j}\,, \tag{29}$$

$$\mathrm{E}\left[\widehat{R_j^*}\right] = R_j^*\,, \tag{30}$$

where

$$n_j^{j+1} \equiv \sum_{k=1}^{n_j} X(j,\boldsymbol{\theta}_k)\,. \tag{31}$$

Because we know $M_0^* = 1$ exactly, the estimator for M_j^* is

$$\widehat{M_j^*} \equiv \prod_{i=0}^{j-1} \widehat{R_j^*} = \prod_{i=0}^{j-1} \frac{n_j^{j+1}}{n_j}\,, \qquad 0 < j < J\,. \tag{32}$$

Since X_j's are samples from Bernoulli distribution, the variance of $\widehat{R_j^*}$ is

$$\mathrm{Var}\left[\widehat{R_j^*}\right] \approx \frac{\tau}{n_j}\,\widehat{R_j^*}\,(1-\widehat{R_j^*})\,, \tag{33}$$

where τ is the auto-correlation time of the samples. R_j^* is replaced with the estimator $\widehat{R_j^*}$, which is asymptotically correct. We will talk about how to estimate τ below. The variance of $\widehat{M_j^*}$ is

$$\begin{aligned}\mathrm{Var}\left[\widehat{M_j^*}\right] &= \mathrm{Var}\left[\widehat{M_{j-1}^*}\,\widehat{R_{j-1}^*}\right] \\ &\approx \mathrm{Var}\left[\widehat{M_{j-1}^*}\right]\,\mathrm{Var}\left[\widehat{R_{j-1}^*}\right] + \widehat{M_{j-1}^*}^2\,\mathrm{Var}\left[\widehat{R_{j-1}^*}\right] + \mathrm{Var}\left[\widehat{M_{j-1}^*}\right]\,\widehat{R_{j-1}^*}^2\,,\end{aligned} \tag{34}$$

where we have replaced the expectations with the estimator values, and assumed $\widehat{M_{j-1}^*}$ and $\widehat{R_{j-1}^*}$ are independent. We can evaluate Eqn. (34) inductively. We also need the covariance between $\widehat{M_j^*}$'s to estimate the error bar of the estimator of Z. Assuming $k > j$, the covariance between $\widehat{M_j^*}$ and $\widehat{M_k^*}$ is

$$\begin{aligned}\mathrm{Cov}\left[\widehat{M_j^*},\widehat{M_k^*}\right] &= \mathrm{E}\left[\left(\widehat{M_j^*} - M_j^*\right)\,\left(\widehat{M_k^*} - M_k^*\right)\right] \\ &= \mathrm{E}\left[\left(\widehat{M_j^*} - M_j^*\right)\,\left(\widehat{M_j^*}\,\widehat{R_j^*}\,\ldots\,\widehat{R_{k-1}^*} - M_j^*\,R_j^*\,\ldots\,R_{k-1}^*\right)\right] \\ &= \mathrm{E}\left[\widehat{M_j^*}^2\,\widehat{R_j^*}\,\ldots\,\widehat{R_{k-1}^*}\right] - \left(M_j^*\right)^2\,R_j^*\,\ldots\,R_{k-1}^* \\ &\approx \mathrm{Var}\left[\widehat{M_j^*}\right]\,\widehat{R_j^*}\,\ldots\,\widehat{R_{k-1}^*}\,.\end{aligned} \tag{35}$$

In the last two steps, we have used the fact that $\widehat{M_j^*}$ is independent with $\widehat{R_j^*},\ldots,\widehat{R_{k-1}^*}$, and we have replaced R^* with $\widehat{R^*}$. We can rewrite the product of $\widehat{R^*}$'s as the ratio between $\widehat{M_k^*}$ and $\widehat{M_j^*}$,

$$\mathrm{Cov}\left[\widehat{M_j^*},\widehat{M_k^*}\right] \approx \mathrm{Var}\left[\widehat{M_j^*}\right]\,\frac{\widehat{M_k^*}}{\widehat{M_j^*}}\,, \qquad j < k\,. \tag{36}$$

To estimate auto-correlation time τ, we need to establish a reasonable time series. Assume our ensemble sampler has an ensemble of size S, and at step t, the walkers in the ensemble are in state: $(j_1, \boldsymbol{\theta}_1)$, ..., $(j_S, \boldsymbol{\theta}_S)$. The ensemble mean of X defined in Eqn. (26) at step t forms a good time series,

$$\bar{X}_t \equiv \frac{1}{S} \sum_{s=1}^{S} X(j_s, \boldsymbol{\theta}_s) \,. \tag{37}$$

We take the chain of $\bar{X}_t$ and use acor[2] package to estimate auto-correlation time τ. Because of the nature of diffusive nested sampling, the auto-correlation time τ obtained this way is a good measure for correlation between samples. We will use τ for all purposes in this paper.

2.6. Marginalized Likelihood and Error Bar

We take the mean of likelihoods sandwiched between two levels,

$$\bar{L}_j \equiv \frac{1}{l_j} \sum_{L_j^* \le L(\boldsymbol{\theta}) < L_{j+1}^*} L(\boldsymbol{\theta}) \,, \tag{38}$$

where l_j is the number of samples withe the likelihood sandwiched between level j and level $j+1$. For 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 marginalized likelihood can be expressed as

$$\widehat{Z} \equiv \sum_{j=0}^{J} \bar{L}_j \left(\widehat{M_j^*} - \widehat{M_{j+1}^*} \right) \,. \tag{39}$$

The variance of $\widehat{Z}$ can be estimated via

$$\begin{aligned} \mathrm{Var}\left[\widehat{Z}\right] \approx & \sum_{j=0}^{J} \bar{L}_j^2 \, \mathrm{Var}\left[\widehat{M_j^*} - \widehat{M_{j+1}^*}\right] \\ & + 2 \sum_{0 \le i < j < J} \bar{L}_i \, \bar{L}_j \, \mathrm{Cov}\left[\widehat{M_i^*} - \widehat{M_{i+1}^*}, \widehat{M_j^*} - \widehat{M_{j+1}^*}\right] \\ & + \mathrm{Var}\left[\bar{L}_j\right] \left(\widehat{M_j} - \widehat{M_{j+1}}\right]^2 \,, \end{aligned} \tag{40}$$

where the variance and covariance of the difference of adjacent prior masses can be easily calculated with Eqn. (34) and Eqn. (36) and

$$\mathrm{Var}\left[\bar{L}_j\right] \approx \frac{1}{{l_j}^2/\tau} \sum_{L_j^* \le L(\boldsymbol{\theta}) < L_{j+1}^*} \left(L(\boldsymbol{\theta}) - \bar{L}_j\right)^2 \,, \tag{41}$$

where τ_j is the auto-correlation time for $L(\boldsymbol{\theta})$ chain that satisfies $L_j^* \le L(\boldsymbol{\theta}) < L_{j+1}^*$. In Eqn. (40), we have assumed that $\widehat{M_j^*} - \widehat{M_{j+1}^*}$ and $\bar{L}_j$ are independent, and ignored high-order terms.

[2] https://github.com/dfm/acor

3. Trial Problem with Rosenbrock Function

We apply our algorithm to an integral of Rosenbrock function, because the two variables of Rosenbrock functions are highly correlated and deviate from gaussian distribution greatly, which we hope will mimic some properties of the posterior of real problems. To be specific, we will try to evaluate the following integral

$$Z = \int \pi_R(\boldsymbol{\theta})\, L_R(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\,,$$

where

$$\pi_R(\boldsymbol{\theta}) = \frac{1}{100}\, \mathbb{1}_{[-5,\,5]}(\theta_1)\, \mathbb{1}_{[-5,\,5]}(\theta_2)\,, \tag{42}$$

and

$$L_R(\boldsymbol{\theta}) = \exp\left(-\frac{100\,(\theta_2 - \theta_1^2)^2 + (1-\theta_1)^2}{20}\right)\,. \tag{43}$$

This integral can only be solved numerically. With mesh size 10000 in both dimensions, the quadrature is

$$Z = 3.13323 \pm 0.00007 \times 10^{-2}\,. \tag{44}$$

The error is estimated by integrating the 2nd-order terms of Taylor expansion of the integrand.

For this trial problem, 10 levels are built and the ensemble size is 20. We repeat the experiment for 1000 times to get multiple independent results, Z_1, Z_2, …, Z_{1000}. The mean of these results is

$$\bar{Z} = \frac{1}{1000}\sum_{k=1}^{1000} Z_k = 3.1334\,, \tag{45}$$

consistent with Eqn. (44). The direct estimation of variance gives

$$\widehat{\sigma^2}_P = \frac{1}{1000}\sum_{k=1}^{1000}(Z_k - \bar{Z})^2 = 8.4 \times 10^{-10}\,. \tag{46}$$

Our estimation of the variance via Eqn. (40) is

$$\widehat{\sigma^2}_R = 8.7 \times 10^{-10}\,. \tag{47}$$

The two estimations of variance are consistent with each other up to the leading order.

4. Multi-Companion Model Fit for HIP 88048

The host star HIP 88048 (ν Ophiuchi) is a K0III star, and has $3.04\,M_\odot$ and $15.1\,R_\odot$ (Sato *et al.* 2012). We have totally 131 radial velocity data for HIP 88048. The parameters of its two confirmed brown-dwarf companions are shown in Tab. (1).

For keplerian fit to RV data, the parameters are $\{K_1, \omega_1, \phi_1, e_1, \varpi_1, \ldots, K_n, \omega_n, \phi_n, e_n, \varpi_n, v_0, S\}$. The meanings of these symbols will become clear in the following sections. We try to use non-informative priors, and exclude non-physical parameter combinations. When we set boundaries to the parameters, we have in mind that sub-stellar companions as large as brown dwarfs have to be include, so the boundaries are wider than if one is only concerned with planet-size companions.

4.1. Prior Distribution

For amplitude K, we have

$$\pi(K) = \frac{1}{\log \frac{K_b+K_0}{K_a+K_0}} \frac{1}{K+K_0}\,, \qquad K_a < K < K_b\,, \tag{48}$$

where we choose $K_a = 0\,\mathrm{m\,s^{-1}}$, $K_b = 10000\,\mathrm{m\,s^{-1}}$, and $K_0 = 10\,\mathrm{m\,s^{-1}}$. We choose a large K_b as boundary, because we have seen some companion imposing maximum speed on its host stars with such magnitude. We use angular speed ω as a parameter instead of period, because it is much more straightforward. For angular speed ω, we have

$$\pi(\omega) = \frac{1}{\log \frac{\omega_b+\omega_0}{\omega_a+\omega_0}} \frac{1}{\omega+\omega_0}\,, \qquad \omega_a < \omega < \omega_b\,, \tag{49}$$

where we choose $\omega_a = 0\,\mathrm{rad\,d^{-1}}$, $\omega_b = \pi\,\mathrm{rad\,d^{-1}}$, and $\omega_0 = 0.01\,\mathrm{rad\,d^{-1}}$. For eccentricity e, we use a beta distribution with one of the hyper-parameters 1 and the other one 5, so that the distribution has more weight at around 0, and also goes all the way to 1. The prior for e is

$$\pi(e) = 5\,(1-e)^4\,, \qquad 0 < e < 1\,. \tag{50}$$

For both longitude of ascending node ϖ and longitude of periastron ϕ, we simply use uniform distribution between 0 and 2π as their priors. For the velocity offset v_0, we use uniform distribution

	HIP 88048 b	HIP 88048 c
$K\,(\mathrm{m\,s^{-1}})$	288.1 ± 1.3	175.8 ± 1.6
$P\,(\mathrm{day})$	529.9 ± 0.2	3211 ± 35
$\phi\,(\mathrm{rad})$	4.130 ± 0.032	3.859 ± 0.046
e	0.1298 ± 0.0045	0.195 ± 0.012
$\varpi\,(\mathrm{rad})$	1.732 ± 0.032	1.768 ± 0.039
$m \sin i\,(M_J)$	24	26
$a\,(\mathrm{AU})$	1.9	6.2

Table 1:: Parameters for the two companions of HIP 88048 in 2-companion model fit.

between $-5000\,\mathrm{m\,s^{-1}}$ and $5000\,\mathrm{m\,s^{-1}}$ as its prior. For jitter square S, we have

$$\pi(S) = \frac{1}{\log \frac{S_b+S_0}{S_a+S_0}} \frac{1}{S+S_0}\,, \qquad S_a < S < S_b\,, \tag{51}$$

where we choose $S_a = 0\,\mathrm{m^2\,s^{-2}}$, $S_b = 100000\,\mathrm{m^2\,s^{-2}}$, and $S_0 = 100\,\mathrm{m^2\,s^{-2}}$. For K, ω, and S, the priors we use are what most people call Jeffery's priors.

We exclude the cases when the orbits of companions cross from the prior distribution. Actually, our requirement is stronger than this. We require that the radius at apoastron of an inner orbit is smaller than the radius at periastron of an outer orbit. So for n-companion model, the overall prior is

$$\pi(\boldsymbol{\theta}) \propto \mathbb{1}(\text{orbits not crossed})\pi(v_0)\,\pi(S) \prod_{i=1}^{n} \left[\pi(K_i)\,\pi(\omega_i)\,\pi(\phi_i)\,\pi(e_i)\,\pi(\varpi_i)\right]\,, \tag{52}$$

where $\mathbb{1}$ is an indicator function, if orbits not crossed is true, $\mathbb{1}$ is 1; otherwise, 0. For the sake of sampling, we also require the periods of companions to be in ascending order. This would introduce another normalization factor, $1/n!$. But in fact, we do not need the normalization of $\pi(\boldsymbol{\theta})$, because we only need the prior mass, which is a proportion.

4.2. Likelihood

The likelihood function is

$$L(\boldsymbol{\theta}) = (2\pi)^{N/2} \prod_{i=1}^{N} \left[(\sigma_i^2+S)^{-1/2}\,\exp\left(-\frac{(v_i - v_{\mathrm{rad}}(t_i,\boldsymbol{\theta}))^2}{2(\sigma_i^2+S)}\right)\right]\,, \tag{53}$$

where $\{t_i,\, v_i,\, \sigma_i\}$ are the data, and N is the size of data. The formula for $v_{\mathrm{rad}}(t,\boldsymbol{\theta})$ is given by (Ohta *et al.* 2005),

$$v_{\mathrm{rad}}(t,\boldsymbol{\theta}) = v_0 + \sum_{i=1}^{k} \left[K_i\,(\sin\,(f_i+\varpi_i) + e_i\,\sin\varpi_i)\right]\,, \tag{54}$$

where the true anomaly f is a function of t, and it satisfies

$$\cos f = \frac{\cos E - e}{1 - e\,\cos E}, \tag{55}$$

and

$$\omega\,t + \phi = E - e\,\cos E\,. \tag{56}$$

I have omitted the companion indexes in Eqn. (55) and Eqn. (56). The likelihood needs to be properly normalized to get the correct marginalized likelihood.

4.3. Marginalized Likelihood

We fit the RV data with 1-, 2-, 3-, and 4-companion model. The fits and residuals are shown in Fig. (1). Note that the 2-companion model fit is a much better fit than the 1-companion model, but 2-companion model fit almost has no visible difference from 3- and 4-companion model. The histograms of some of the physical parameters are shown in Fig. (2), Fig. (3), Fig. (4), and Fig. (5). For 3- and 4-companion model, only the extra companions' histograms are shown. Note that the periods of the extra companions in 3- and 4-companion model are all badly bound. But they do show lots of peaks, and some of the peaks are significant. The histograms for eccentricities of the extra companions in 3- and 4-companion model are basically the prior for eccentricity given in Eqn. (50).

Z_n means the marginalized likelihood for n-companion fit. The results are summarized below.

$$\begin{aligned} Z_1 &= 3.72 \pm 0.01 \times 10^{-356}\,; \\ Z_2 &= 1.95 \pm 0.03 \times 10^{-237}\,; \\ Z_3 &= 9.2 \pm 0.3 \times 10^{-238}\,; \\ Z_4 &= 2.1 \pm 0.2 \times 10^{-238}\,. \end{aligned} \tag{57}$$

These marginalized likelihoods all have the same unit $\left(\mathrm{m\,s^{-1}}\right)^N$, where N is the number of observations. The units are omitted to make the presentation clean. The 2-companion model has the largest marginalized likelihood. Z_2 is much larger than Z_1, because the 2-companion model is much better fit than the 1-companion model. Z_3 and Z_4 are only slightly smaller than Z_2, because 3- and 4-companion models actually improve the fit but are penalized for including more parameters.

Probability theory only provides us with the probabilities of different states of nature, in our case, different companion models, given the data and any prior information. It does not tell us what decision one should make based on these probabilities (Jaynes 2003). In order to make decisions, we need to have a loss function and take any other additional 'evidence' into account. In such a general situation, it is hard to specify a loss function, because loss functions depend on the purposes of our decisions and what strategies we use to make our decisions. Nevertheless, we are able to talk about additional evidence at hand, which are the histograms of the physical parameters of the companions in the models. In Fig. (4), we can see that the histograms of the posterior of K of the 3rd companion in 3-companion model clearly favors small amplitude ($< 10\,\mathrm{m\,s^{-1}}$), which is barely larger than the error bar in the data. Furthermore, the posterior distribution has almost no constraint on the 3rd companion's eccentricity, the marginal posterior of which is very close to the prior given in Eqn. (50). However, the histogram of the period of the 3rd companion does show a peak at about 54 days, even though the period is badly constrained overall. Periodogram performed on the data residual of the 2-companion model also confirms this periodic signal. At this stage, we do not know exactly what it is. But based on the marginal likelihoods given in Eqn. (57) and the histograms of the posterior of various as additional evidence, we are confident that we should decide that 2-companion model is true, even without explicitly specifying any loss function

and decision strategy.

5. Discussion

One of the biggest advantages of diffusive nested sampling is its ability to 'jump' between peaks in the probability density one is studying. For example, the various peaks in the posterior probability densities shown in Fig. (4) and Fig. (5) pose no threat to diffusive nested sampling. The affine invariant ensemble sampler is very efficient in sampling multivariate probability distribution hat has highly correlated parameters, which is often the case in multi-companion model fitting for the RV data. The combination of these two, in principle, should be able to overcome any challenging problems. However, the algorithm is still subject to curse of dimension and hard likelihood functions. In such cases, one needs considerable amount of computing power to evaluate marginal likelihoods. In the case of 3- and 4-companion models, it takes the sampler much longer time to reach equilibrium and start to take samples.

Even though diffusive nested sampling deviates from traditional importance sampling, the estimator $\widehat{Z}$, defined in Eqn. (39), is an unbiased estimator of marginalized likelihood Z. However, it does present a challenge to put an error bar on $\widehat{Z}$. We have to make lots of assumptions to arrive at our estimation of $\mathrm{Var}\widehat{Z}$, given in Eqn. (40), and we confirm the reliability of this estimation on trial problem discussed in Section (3). Nonetheless, these assumptions may not be true in difficult problems, and we may underestimate the error bar in such problems.

For any bayesian approach, the statistical outcome will be affected more or less by the choice of prior distributions. But the sensitivity to the choice of priors is probably reflected no more in any other situations than in that of the marginalized likelihood. For example, if I double the boundary K_b in Eqn. (48), I will actually double the volume but the new volume has almost zero likelihood in it. So the only thing that changes is the normalization in Eqn. (48). Consequently, the marginalized likelihood of n-companion model will then decease by a factor of 2^{-n}, further discrediting large number of companions. But one should keep in mind that the choice of prior distributions should reflect any prior information. For instance, since we know majority of substellar companions have small eccentricity, it would be insane to have a prior that favors large eccentricity. Our choice of priors, though pedagogical, does reflect, to certain degree, the general distributions of substellar companions' physical parameters. In order to obtain a more appropriate prior distribution, one would hierarchically model the whole population to infer any hyper-parameters in the prior probability density function, which comes back to the importance of the capability of evaluating marginalized likelihood.

Because directly evaluating the marginalized likelihood was rarely done before, it is difficult to independently confirm our results using other numerical tools. We are developing an algorithm based on bayesian information criterion (BIC) to achieve this goal.

REFERENCES

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

Ford, E. B., Gregory, P. C., 2007, Statistical Challenges in Modern Astronomy IV, G.J. Babu and E.D. Feigelson (eds.), San Francisco: Astron. Soc. Pacific

Foreman-Mackey, D., Hogg, D. W., Lang, D., Goodman, J., 2013, http://arxiv.org/abs/1202.3665

Frink, S., *et al.*, 2002, ApJ, 576, 478

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D.B., 2004, Bayesian Data Analysis (2nd ed.; Chapman & Hall/CRC)

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

Hekker, S., *et al.*, 2006, A&A, 454, 943

Hekker, S., *et al.*, 2008, A&A, 480, 215

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

Jaynes, E. T., 2003, Probability Theory The Logic of Science (Cambridge University Press)

Mitchell, D. S., *et al.*, 2003, BAAS, 35, 1234

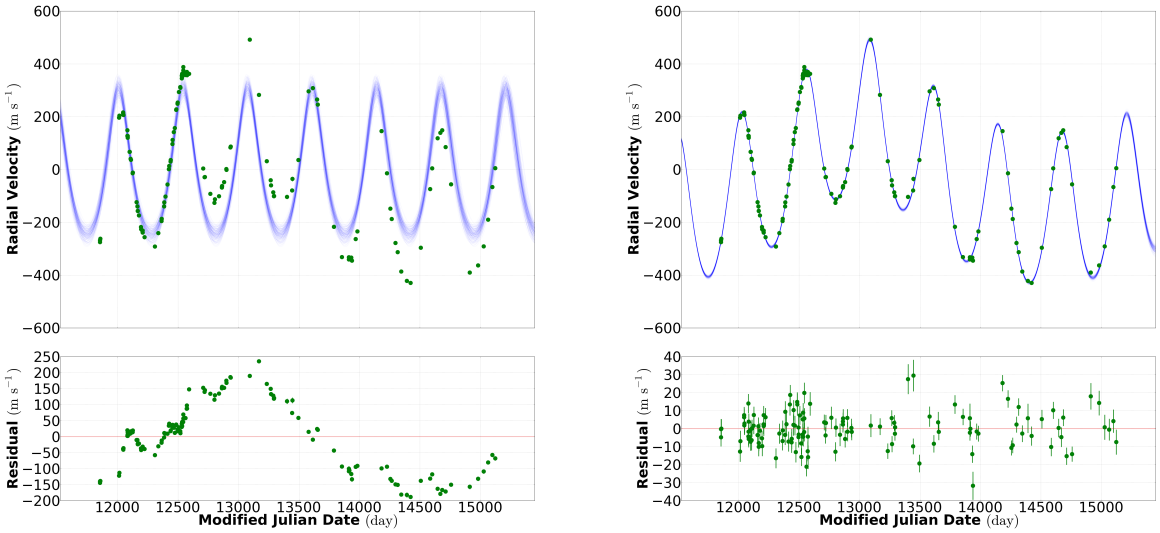
Mitchell, D. S., *et al.*, 2013, http://arxiv.org/abs/1305.5107

Ohta, Y., Taruya, A. & Suto, Y., 2005, ApJ, 622, 1118

Quirrenbach, A., Reffert, S. & Bergmann, C., 2011, AIP Conference Proceedings, 1331, 102

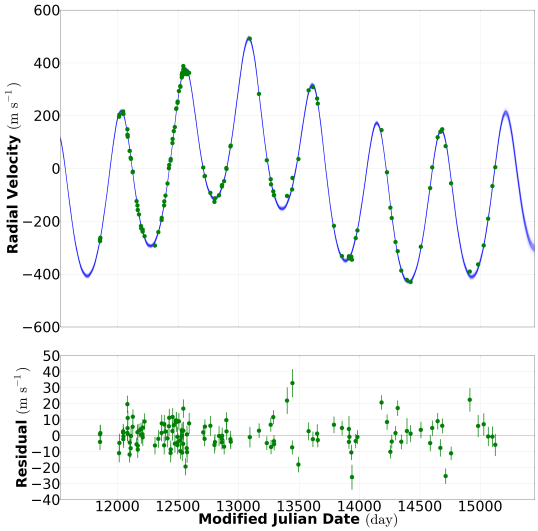
Sato, B., *et al.*, 2012, PASJ, 64(6), 135

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

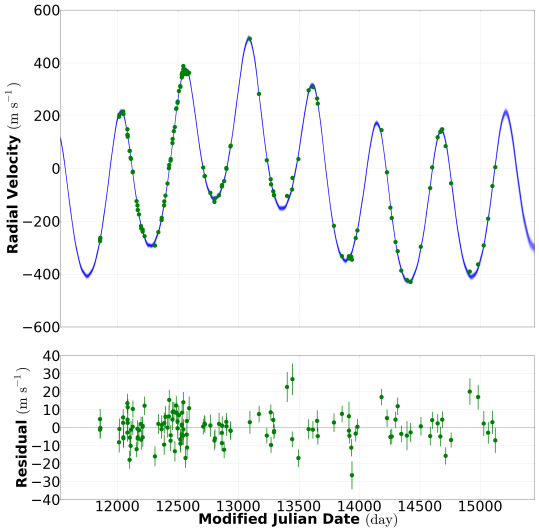


(a) 1-Companion Model

(b) 2-Companion Model



(c) 3-Companion Model



(d) 4-Companion Model

Fig. 1.—: (a) In the upper panel, the blue lines show 100 sample fits from the posterior probability distribution of the 1-companion model, and the green dots show the data. In the bottom panel, the residual data for the best fit of the 1-companion model are shown. (b) In the upper panel, data and 100 sample fits from the posterior probability distribution of the 2-companion model are shown. In the bottom panel. The residual data are shown. (c) In the upper panel, data and 100 sample fits from the posterior probability distribution of the 3-companion model are shown. In the bottom panel. The residual data are shown. (d) In the upper panel, data and 100 sample fits from the posterior probability distribution of the 4-companion model are shown. In the bottom panel. The residual data are shown.

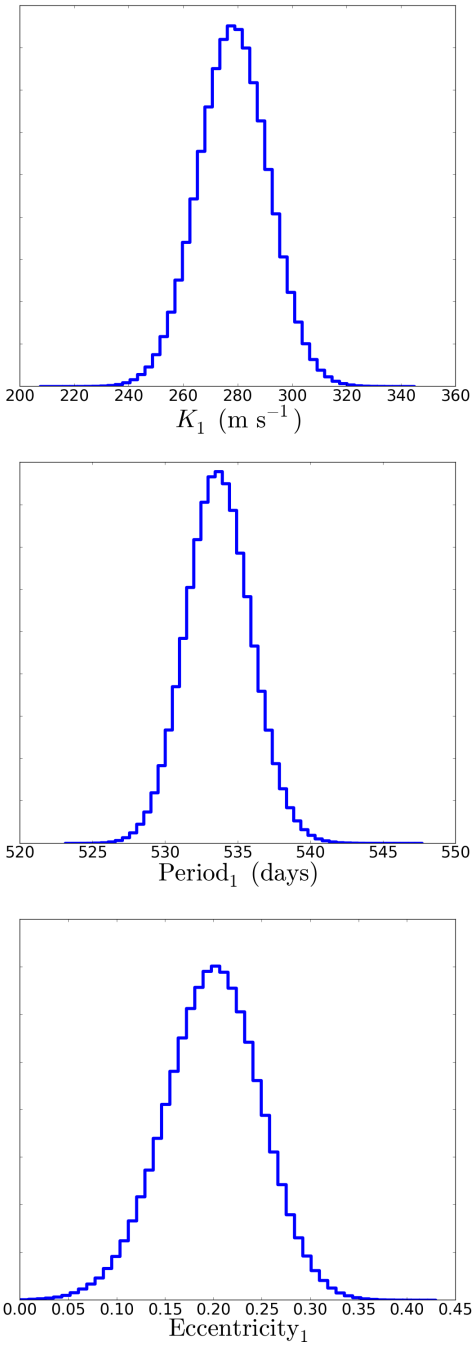


Fig. 2.—: The posterior histograms of amplitude, period and eccentricity in the 1-companion model. All parameters are well constrained by the posterior distribution.

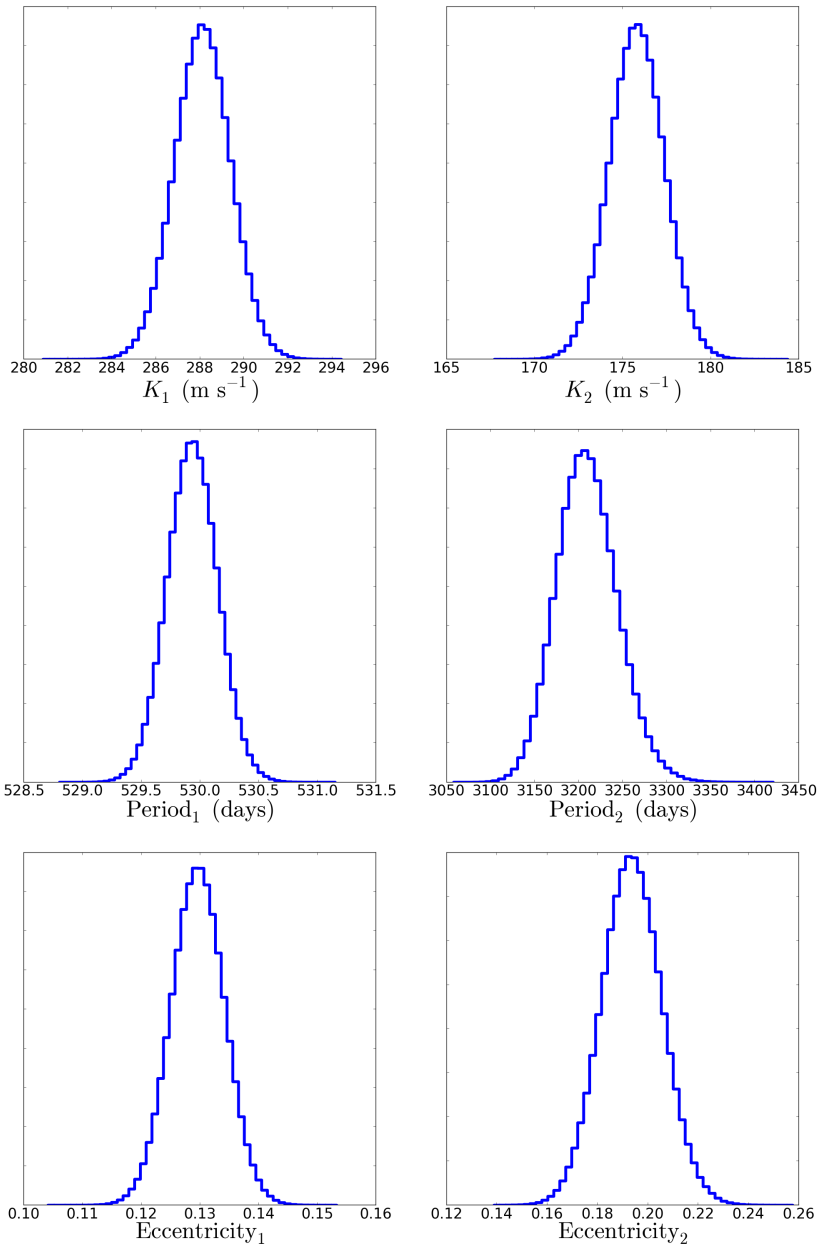


Fig. 3.—: The posterior histograms of amplitudes, periods and eccentricities of both the companions in the 2-companion model. All the parameters are well constrained by the posterior distribution.

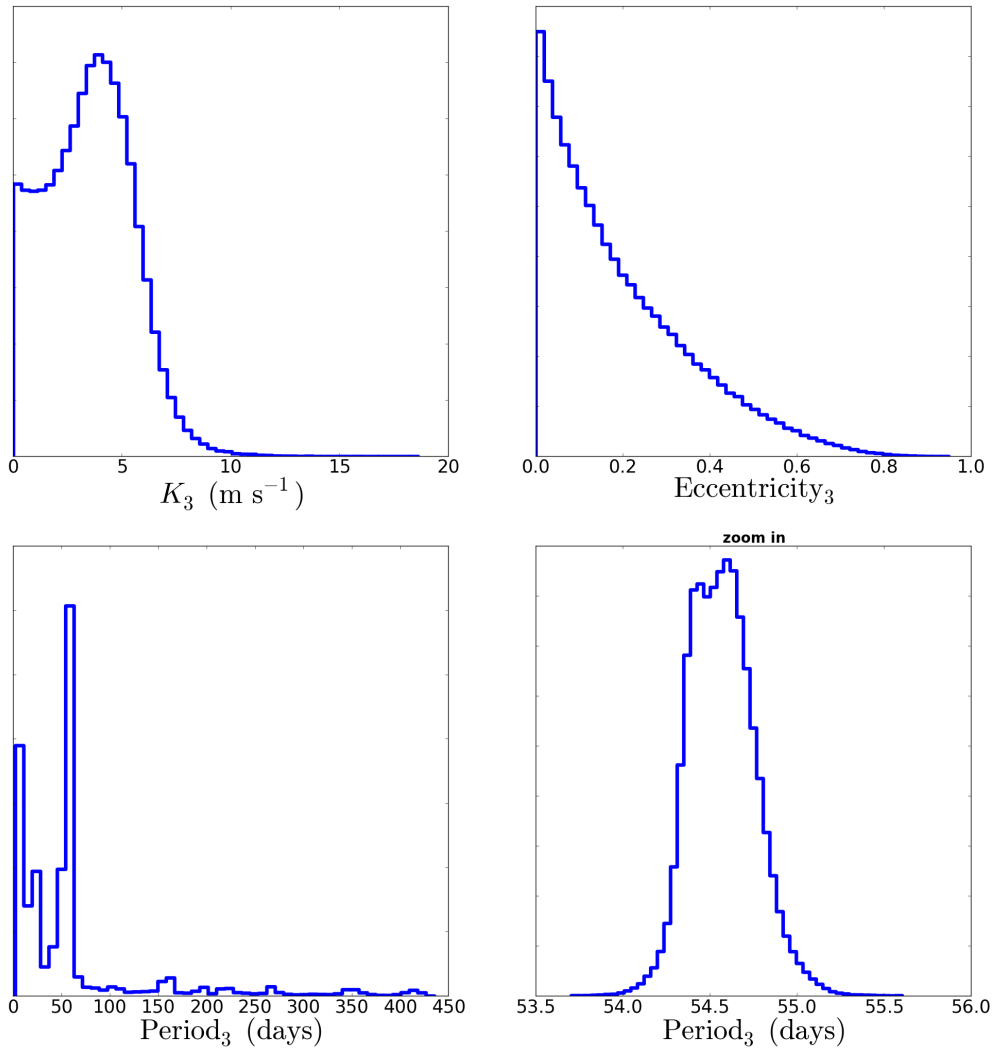


Fig. 4.—: The posterior histograms of amplitude, period and eccentricity of the 3rd companion in the 3-companion model. The histograms of the other 2 companions are almost identical to their histograms in 2-companion model, shown in Fig. 3. The histogram of the amplitude indicates that small object is favored. The histogram of period indicates that the period of the 3rd companion is poorly constrained, but there are many peaks in the histogram. The histogram on the lower right side is for the period of the 3rd companion, but zoomed in around the peak of about 54 days. The histogram of the eccentricity shows that the data almost provides no information for eccentricity of the 3rd companion, and the posterior is very close to the prior for eccentricity given in Eqn. (50).

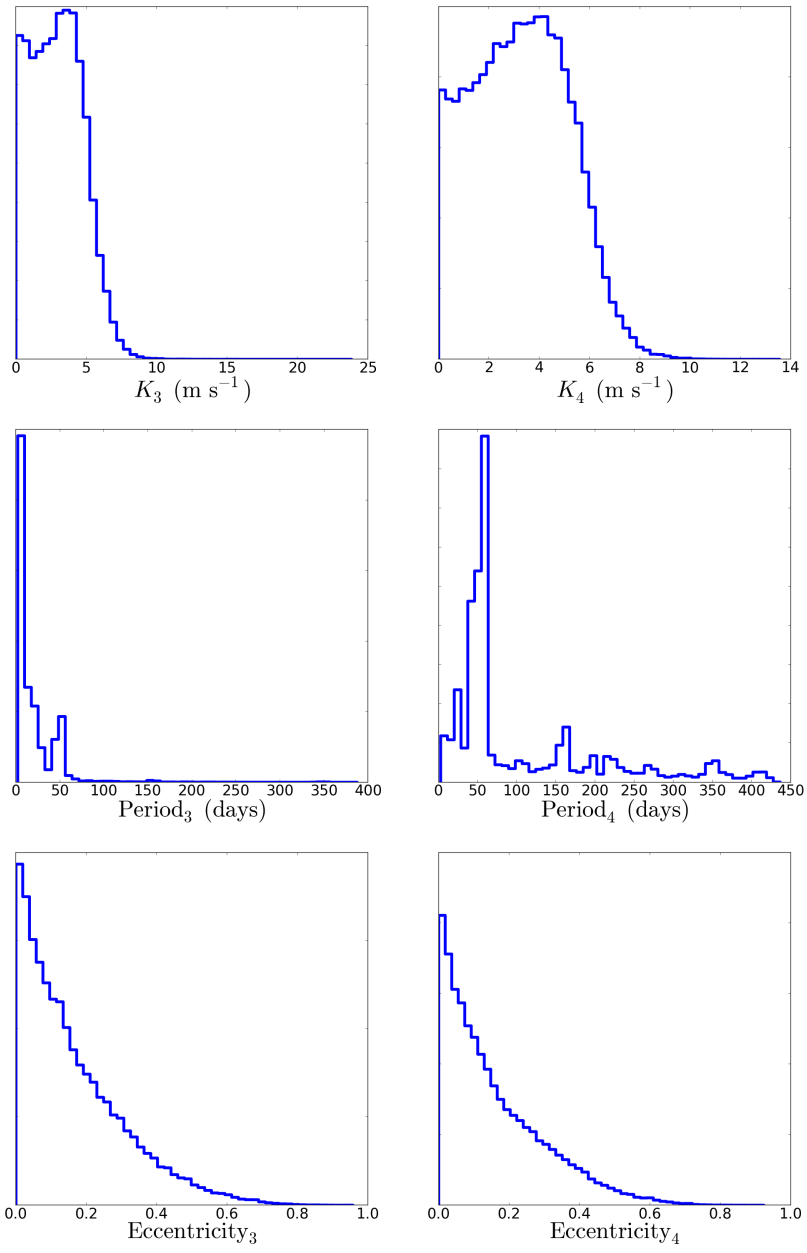


Fig. 5.—: The posterior histograms of amplitudes, periods and eccentricities of the 3rd and 4th companion in the 4-companion model. The histograms of the other 2 companions are again almost identical to their histograms in 2-companion model, shown in Fig. 3. The histograms of amplitudes indicate that both objects are small if they exist. The histograms of periods indicate that the periods are poorly constrained but they do show various peaks, including a peak around 54 days. The histograms of eccentricities are both very similar to the prior for eccentricity given in Eqn. (50).