Diffusive Nested Ensemble Sampling

ABSTRACT

Diffusive Nested Sampling provides an efficient method to evaluate the marginalized likelihood or Bayesian evidence integral. We developed an affine invariant ensemble sampling version of the Diffusive Nested Sampling method. We tested our code on models in which the evidence can be calculated analytically. We also evaluate different models' evidence integrals of radial velocity models for a star with 2 known companions and confirm that 2-companion model fitting has the largest evidence.

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

1. Introduction

When presented with competing models, where by 'model' we mean a likelihood function and a prior over parameters, Bayes theorem tells us to compare models as follows:

$$P(\mathrm{Model}_j|\mathrm{Data}) = \frac{P(\mathrm{Data}|\mathrm{Model}_j)\, P(\mathrm{Model}_j)}{\sum_j P(\mathrm{Data}|\mathrm{Model}_j)\, P(\mathrm{Model}_j)}. \tag{1}$$

This is just Bayes law. The *likelihood* in Eqn. (1) is in fact the evidence $Z_{\mathrm{Model}\ j}$ for model j.

$$Z_{\mathrm{Model}\ j} \equiv P(\mathrm{Data}|\mathrm{Model}_j) \equiv \int P(\mathrm{Data}|\theta, \mathrm{Model}_j)\, P(\theta|\mathrm{Model}_j)\mathrm{d}\theta, \tag{2}$$

where $P(\mathrm{Data}|\theta, \mathrm{Model}_j)$ is the likelihood of parameter θ for model j and $P(\theta|\mathrm{Model}_j)$ 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

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

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

$$\int \pi(\theta)\, \mathrm{d}\theta = 1, \tag{4}$$

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

However, evaluating or even estimating the evidence integral has always been challenging. Diffusive Nested Sampling proves to be an efficient and accurate method to evaluate or estimate 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 nested sampling, we change the variable in the evidence integral from parameter θ to the prior mass

$$M(L^*) \equiv \int_{L(\theta)>L^*} \pi(\theta)\,\mathrm{d}\theta, \tag{5}$$

which is, in another word, cumulant prior mass covering the area whose likelihood values are 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

$$\mathrm{d}M = \int_{L^*-\mathrm{d}L^*<L(\theta)<L^*} \pi(\theta)\,\mathrm{d}\theta = \pi(\theta) \times \text{volume of } \theta, \text{ satisfying } L^* - \mathrm{d}L^* < L(\theta) < L^*, \tag{6}$$

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

$$\int_0^1 L^*\,\mathrm{d}M = \int L(\theta)\,\pi(\theta)\,\mathrm{d}\theta. \tag{7}$$

so the evidence Z can be expressed as

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

In most cases, it is impossible to know the function $L^*(M)$ analytically and evaluate the integral analytically. Note from Eqn. (8) that M can be viewed as a random variable with uniform distribution. Nested sampling takes advantage of this to build the $L^*(M)$ curve 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 just 'threshold'. Each level defines a constrained prior,

$$p_j(\theta) = \frac{\pi(\theta)}{M_j}\,\mathbb{1}_{L(\theta)>L_j^*}, \tag{9}$$

where

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

and note that p_j is properly normalized by M_j.

2.2. Setting Level Thresholds

The nested sampler we use first attempt to make the levels' prior masses to be $\widehat{M}_j = e^{-j}$ and estimates 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 the refined prior masses M_j, described in Section (2.4). The zeroth level has $M_0 = \widehat{M}_0 = 1$ and $L_0^* = 0$. In the current code, I actually keep the L_j^* unchanged while changing M_j, so I instead put the wide hat on top of the M_j's.

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 templatelibrary (STL) of C++. $\widehat{M}_1 \approx 1/e$ is the estimate of the prior mass that level 1 covers.

To estimate the next level $\left(\widehat{M}_2,\, L_2^*\right)$, we generate N samples from the prior with likelihood larger than L_1^*. There are many different ways to do this. One would be sampling the constrained density $p_1(\theta)$ defined in Eqn. (9). Another would be sampling a mixture of $p_1(\theta)$ and prior $\pi(\theta)$, defined below in Eqn. (10). We sample the mixture 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. Mixing density functions require us to put different weights to the densities and the weights should sum up to one, as in Eqn. (11). We need to balance efficiency and the need to circumvent discontinuity, so we give the latest level more weight. In current case, we use $w_1/w_0 = e$. Like before, we get a chain of N likelihoods, rank these likelihoods in descending order and find the N/e-th likelihood, which we call L_2^*. This would give us another point on the $L^*(M)$ curve, which is approximately $(1/e^2, L_2^*)$. This is level 2 and L_2^* is the likelihood threshold of level 2. $\widehat{M}_2 \approx 1/e^2$ is the prior mass that level 2 covers.

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 levels $\left(\widehat{M}_1,\, L_1^*\right), \ldots, \left(\widehat{M}_k,\, L_k^*\right)$. The evidence integral corresponding to all these levels is

$$Z_j = \sum_{k=0}^{j-1} \int_{\widehat{M}_k}^{\widehat{M}_{k+1}} L^*(M)\,\mathrm{d}M = \int_0^{\widehat{M}_j} L^*(M)\,\mathrm{d}M.$$

The remaining integral is

$$\int_{\widehat{M}_j}^{1} L^*(M)\,\mathrm{d}M.$$

But $L^*(M) < L_{max}$ always, so the remain integral cannot be larger than $L_{max}(1 - \widehat{M}_k)$. We choose a stopping point k so that $L_{max}(1 - \widehat{M}_k) \leq \epsilon Z_k$. We usually choose $\epsilon = 10^{-6}$. It is clear that errors in estimating Z from other sources are much larger than this in practice. 'clear' may be a strong word here. I don't think it is trivial to prove this.

At the end, we get a series of points on the $L^*(M)$ curve, $\{(\widehat{M}_0, L_0^*), (\widehat{M}_1, L_1^*), (\widehat{M}_2, L_2^*), \ldots\}$.

Correspondingly, we also have a series of mixture of constrained priors,

$$p(\theta) = \sum_{j=0} w_j \, p_j(\theta), \tag{10}$$

where p_j is the constrained prior defined in Eqn. (9) and w_j are the weights of each level which sum up to 1,

$$\sum_{j=0} w_j = 1. \tag{11}$$

The choice of weight may change according to different purposes. For example, when we are building a new level, we might want to put more weight on the last level. And in the final stage, when we sample all the levels together, we might want to have the same weight on all the levels.

Our goal is to evaluate the evidence Z, defined in Eqn. (7). The incentive to build levels and eventually have a mixture of constrained priors $p(\theta)$, defined in Eqn. (10), can be understood in the context of importance sampling. Note that the integrand in Eqn. (7) is $L(\theta)\,\pi(\theta)$, if we can find a probability density function which is similar in shape with $L(\theta)\,\pi(\theta)$ and can be sampled, we can evaluate the integral in Eqn. (7) by importance sampling. And $p(\theta)$ is that probability density function. Effectively, we generate samples from $p(\theta)$ and the estimator for Z is

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^{N} \frac{L(\theta_i)\,\pi(\theta_i)}{p(\theta_i)}, \quad \text{where } \theta_i\text{'s are i.i.d. from } p(\theta). \tag{12}$$

I moved up the part making analogy with importance sampling. It was a little abrupt for it to appear in such a late section.

2.3. Nested Sampling by Stretch Move

In this section, the algorithm to sample Eqn. (10) is described. It is impractical to sample the sum in Eqn. (10) directly, because some of the p_j's are considerably smaller than the others and directly taking the sum will render those p_j's useless and the corresponding levels never visited. To overcome this, we sample 1. the parameter space and 2. the levels separately.

To sample the parameter space, we use stretch move which has the feature of affine invariance (Goodman *et al.* 2010). Stretch move is an efficient ensemble sampler with low auto-correlation time and few tuning parameters. To make use of stretch move in nested sampling, we assign a level index to every walker in the ensemble. (Note that the level assigned to a walker should have a threshold smaller than the likelihood of the walker.) So a walker with level index j can be seen as a sample from distribution p_j. When we update the ensemble using stretch move, we keep the level indices unchanged. The stretch move can be described in pseudo-code as: I didn't elaborate why it is ok to have walkers in different levels (different distribution). The proof takes too much space.

- to propose a new location for walker X whose level index is j, randomly choose a helping walker Y in the ensemble different from X. Note that Y may have a very different level index from j and one may impose a restriction that the level of Y is close to X.
- propose a new location with stretch move: $X_{new} = Y + \alpha\,(X_{old} - Y)$, where α is a random variable from some distribution (Goodman *et al.* 2010).
- accept the proposed X_{new} with probability: $\max\left(z^{\dim-1}\,\frac{p_j(X_{new})}{p_j(X_{old})}, 1\right)$.

To sample the levels is much more straightforward, because there are only finite number of states to be sampled. Simple Metropolis-Hastings would suffice. When we update the indices of the ensemble, we keep the ensemble unchanged in parameter space. This step can be described in pseudo-code as (Brewer *et al.* 2011):

- ramdomly propose a new level for walker X with original level j in the ensemble: $j \to k$.
- if $k \geq j$
  - if the likelihood of the walker X is larger than L_k^*, accept the proposal with probability $\min\left(\frac{w_k}{w_j}, 1\right)$.
  - else, reject the proposal.
- else, accept the proposal with probability $\min\left(\frac{\widehat{M}_j}{\widehat{M}_k}\frac{w_k}{w_j}, 1\right)$.

We need to distinguish whether the proposed level index k for walker X is larger or smaller than the original level j. This is because X is seen as a sample from distribution p_j and thus p_k, and if $k > j$ and $L(X) < L_k^*$, X is not in the support of p_k and consequently ill-defined. So we have to reject such cases that $k > j$ and $L(X) < L_k^*$. But we do not need to worry about this if $k < j$, because the proposed level has a smaller threshold than the original one and X is always in the support. This is also why there is an extra term $\frac{\widehat{M}_j}{\widehat{M}_k}$ in the acceptance ratio when $k < j$ in order to maintain detailed balance.

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

At the end of Section (2.2), in the comparison with importance sampling, we notice that the density function $p(\theta)$ is the denominator of the importance function, which is

$$\frac{L(\theta)\,\pi(\theta)}{p(\theta)},$$

in Eqn. (12). So we need an accurate $p(\theta)$ to have an accurate estimation of Z. This is why we would like to *refine* the prior masses $\widehat{M}$'s. Refining the prior masses is not the only way to have a more accurate $p(\theta)$, we can also refine the likelihood thresholds instead. Maybe in discussion, we can list this as a future project.

Assume we have constructed J levels, $\{(\widehat{M_1}, L_1^*), (\widehat{M_2}, L_2^*), \ldots, (\widehat{M_J}, L_J^*)\}$. We sample the mixture of constrained priors $p(\theta)$, defined in Eqn. (10), using these levels to obtain a supposedly long chain of both the visited level indices and the likelihoods of the walkers during those visits, we can then get the refined prior masses $\{M_1, M_2, \ldots, M_J\}$ using this chain. Assuming that the sampler has visited level $j-1$ for n_{j-1} times and during those n_{j-1} visits there are n_{j-1}^j times that the likelihoods exceed level j's threshold L_j^*, we can refine $\widehat{M_j}$ to be M_j as following (Brewer *et al.* 2011)

$$M_j = M_{j-1} \frac{n_{j-1}^j + C\, \widehat{M_j}/\widehat{M_{j-1}}}{n_{j-1} + C}, \tag{13}$$

where M_j is the refined prior masses, $\widehat{M_j}$ is the un-refined prior masses and C is a constant that reflects one's confidence in the accuracy of the $\widehat{M}$'s before this refinement. We use $C = 10^4$. As mentioned in Section (2.2), $M_0 = \widehat{M_0} = 1$ and we start from refining $\widehat{M_1}$, so M_{j-1} will be known when we refine $\widehat{M_j}$.

We can justify the refinement Eqn. (13) and estimate the uncertainty of the M_j and the uncertainty of interval $M_{j-1} - M_j$, if n_{j-1} and n_{j-1}^j are large enough so we can ignore C, the expectation and variance of M_j are

$$\begin{aligned} \mathrm{E}(M_j) &= \mathrm{E}(M_{j-1}\, R), \\ \mathrm{Var}(M_j) &= \mathrm{Var}(M_{j-1}\, R), \end{aligned}$$

where $R = M_j/M_{j-1}$. By the central limit theorem, the expectation and variance of R can be estimated with

$$\mathrm{E}(R) = \frac{n_{j-1}^j}{n_{j-1}}, \tag{14}$$

$$\mathrm{Var}(R) = \frac{\mathrm{E}(R)\,(1 - \mathrm{E}(R))}{n_{j-1}/\tau}, \tag{15}$$

where τ is the auto-correlation time of the chain There should be a citation here for auto-correlation time. But I couldn't find a proper one. In the last paper we wrote, we had a whole subsection for auto-correlation time. But I don't think we should do that here.. And

$$\mathrm{E}(M_j) = \mathrm{E}(M_{j-1})\, \frac{n_{j-1}^j}{n_{j-1}}, \tag{16}$$

$$\mathrm{Var}(M_j) = \mathrm{Var}(M_{j-1})\, \mathrm{Var}(R) + \mathrm{Var}(M_{j-1})\, \mathrm{E}(R)^2 + \mathrm{E}(M_{j-1})^2\, \mathrm{Var}(R). \tag{17}$$

We can see that Eqn. (16) is consistent with Eqn. (13) if C is small compared with n_j and n_j^{j-1}. Similarly we can get the expectation and variance of $M_j - M_{j+1}$,

$$\mathrm{E}(M_{j-1} - M_j) = \mathrm{E}(M_{j-1}) \left(1 - \frac{n_{j-1}^j}{n_{j-1}}\right), \tag{18}$$

$$\mathrm{Var}(M_{j-1} - M_j) = \mathrm{Var}(M_{j-1})\,\mathrm{Var}(R) + \mathrm{Var}(M_{j-1})\,(1 - \mathrm{E}(R))^2 + \mathrm{E}(M_{j-1})^2\,\mathrm{Var}(R). \tag{19}$$

Note that in Eqn. (17) and Eqn. (19), $\mathrm{E}(R)$ is not replaced by Eqn. (14) to keep the notation unclustered. Lots of things are defined here. If I don't define these, the notations become too long.

While sampling the mixture of constrained priors Eqn. (10), it is possible that the weights w_j are not sampled as desired. And the acceptance probability $\min\left(\frac{M_i}{M_j}\frac{w_j}{w_i}, 1\right)$ in the algorithm described in Section (2.3) deviates from what it should be. As a result, some of the n_j's and n_j^{j+1}'s are too small to make a meaningful refinement to the prior masses. 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. In my case, this kind of treatment was only needed when there are kinks in the $L^*(M)$ curve. And if there is a kink between Level i and Level i+1, M_{i+1} is usually overestimated. This is perhaps because a kink means a protuberant peak inside current prior mass M_i and it is hard for the sampler to *get inside* that peak so it wanders outside of the peak a lot. I NEVER needed this kind of treatment when there ISN'T a kink in the $L^*(M)$ curve, like the test cases where I only deal with gaussian and uniform. But Brendon seems to encounter this a lot more frequently than I do. And he has to be more experienced in this than me. So I am reluctant to say that 'kink' is what causes all the fuss.

2.5. Computing Evidence

We take the mean of likelihoods sandwiched between two levels,

$$\bar{L}_j = \frac{1}{n_j} \sum_{L_j^* \le L(\theta) < L_{j+1}^*} L(\theta), \tag{20}$$

where θ represents samples. With one extra level whose likelihood threshold L_{J+1}^* is the optimum likelihood and whose prior mass M_{J+1} is 0, the estimation of the evidence is

$$Z \approx \sum_{j=0}^{J} \bar{L}_j\,(M_j - M_{j+1}). \tag{21}$$

The variance of the evidence Z can be estimated via

$$\mathrm{Var}(Z) \approx \sum_{j=0}^{J} \bar{L}_j \mathrm{Var}(M_j - M_{j+1}), \tag{22}$$

where the value of $\mathrm{Var}(M_j - M_{j+1})$ is given in Eqn. (19). There are a million issues with this estimation. Because clearly, variance of Z should also include $Var(\bar{L})\,(M_j - M_{j+1})$ and the covariance of $\mathrm{Cov}\left(M_j - M_{j+1},\, \bar{L}\right)$. But even this is under the assumption that my estimation of $\mathrm{Var}(M_j - M_{j+1})$ is correct. Maybe we can figure of some scheme to estimate $\mathrm{Var}(\bar{L})$ but I don't think we can get the covariance in any way.

3. 2-d Gaussian Testing Case

The algorithm was tested on a 2-d gaussian likelihood and a 2-d uniform prior. The likelihood is

$$L(\theta_1, \theta_2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\theta_1^2 + \theta_2^2}{2\sigma^2}\right),\ \sigma = 1, \tag{23}$$

where θ_1 and θ_2 are the parameters. The prior is

$$\pi(\theta_1, \theta_2) = \frac{1}{400},\ \theta_1 \in [-10,\, 10],\ \theta_2 \in [-10,\, 10], \tag{24}$$

and 0 otherwise. This prior basically is a square whose sides' length is 20 and whose area is 400. The evidence is approximately inverse of that area,

$$\text{evidence} \approx \frac{1}{400}, \tag{25}$$

where the approximation is equivalent to equality up to machine error because gaussian distribution has extremely thin tail. The likelihood threshold of any level can be analytically calculated in this model. As a matter of fact, the whole $L^*(M)$ curve can be built analytically,

$$\log L^*(M) = -\log 2\pi - \frac{200M}{\pi}, \tag{26}$$

where the number 200 comes from half the area that the prior covers. Note that $\log L^*$ is a linear function of M.

3.1. Testing Level Thresholds Setting

The test is to see if the algorithm can build levels matching the analytically calculated ones. Recall that to find a new level, one needs to generate a chain of N_1 likelihoods larger than previous level's threshold. (N_1 is used so not to be confused with N_2 used later.) Each new level requires N_1 likelihoods. Two N_1's are tested, $N_{1a} = 10,000$ and $N_{1b} = 100,000$. The larger N_{1b} should give a smaller variance than N_{1a}. For both N_{1a} and N_{1b}, 6 levels are built for $10,000$ times in order to check the statistical features of these levels.

For $N_{1a} = 10,000$, after ranking the chain of N_{1a} likelihoods in descending order, the $J_{1a} = 3,678$-th likelihood is picked as the new level's threshold. For $N_{1b} = 100,000$, after ranking the chain of N_{1b} likelihoods. the $J_{1b} = 36,787$-th likelihood is picked as the next level's threshold.

For N_{1a}, the expectation of the prior masses that each level covers are $\left\{\frac{J_{1a}}{N_{1a}+1}, \left(\frac{J_{1a}}{N_{1a}+1}\right)^2, \ldots\right\}$. And for N_{1b}, the expectation of the prior masses that each level covers are $\left\{\frac{J_{1b}}{N_{1b}+1}, \left(\frac{J_{1b}}{N_{1b}+1}\right)^2, \ldots\right\}$. The variance of the prior masses can also be easily calculated. For example, the variance of the 1st level's covered mass is $\frac{(J_{1a})(N_{1a}-J_{1a})}{(N_{1a}+1)^2(N_{1a}+2)}$ for N_{1a} and $\frac{(J_{1b})(N_{1b}-J_{1b})}{(N_{1b}+1)^2(N_{1b}+2)}$ for N_{1b}. The expection and variance of the corresponding (logarithm of) likelihood thresholds can then be calculated straightfowardly, because $\log L^*$ is a linear fuction of M (from Eqn. (26)). The mean values of the 6 levels' thresholds are listed in Tab. (2) for both N_{1a} and N_{1b} together with the corresponding analytical values of the thresholds. The standard deviations with their analytical values are listed in Tab. (3). The histograms of level1 and level 6 are visualized in Fig. (1)

3.2. Testing Constrained Prior Mixture

We draw samples from a mixture of all the constrained priors defined by true levels listed in Tab. (4). Every adjacent two levels define a bin. For each sample, we find two adjacent levels whose thresholds sandwich the likelihood of the sample and that sample can be put into the bin defined by those two levels. The prior masses of samples inside each bin should follow uniform distribution, which is consistent with our testing result, illustrated in Fig. (2). The variances of the likelihoods of samples can vary dramatically among different bins, Fig. (3).

4. High-Dimension Gaussian Testing Case

In 10-d case, the likelihood is

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{10/2}} \exp\left(-\frac{\boldsymbol{\theta}^2}{2\sigma^2}\right),\ \sigma = 1, \tag{27}$$

where θ_1 and θ_2 are the parameters. The prior is

$$\pi(\boldsymbol{\theta}) = \frac{1}{20^{10}},\ \theta_j \in [-10,\ 10],\ j = 1, 2, \ldots, 10, \tag{28}$$

and 0 otherwise. We make 30 levels in 10-d case so that the last level will cover approximately $1/20^{10}$ of the total prior hyper-volume. To build each level, a likelihood chain of length N_1 is generated. After all the levels are built, we sample mixture of constrained priors for N_2 times, using these samples to refine the prior masses and evaluate the evidence.

4.1. Testing Prior Mass Refinement

We repeat the prior mass refinement and evidence evaluation for $1,000$ times for different N_1's and N_2's. The mean evidences for all N_1's and N_2's are very close to the true evidence 9.77×10^{-14}.

The standard deviations are summarized in Tab. (5).

5. Exoplanet for Star 122

Star 122 has two confirmed companions (citation). We fit the radial velocity data of Star 122 with 1-companion, 2-companion and 3-companion model and evaluate the evidence integrals for these models to see if the 2-companion model has a larger evidence than the other two.

For simplicity, uniform priors are used. For the 5 orbital parameters, the priors are

$$\pi(A) = \frac{1}{10000}\,(\mathrm{m\,s^{-1}})^{-1}, \qquad A \in [0, 10000]\,\mathrm{m\,s^{-1}}, \tag{29}$$

$$\pi(\omega) = 1\,(\mathrm{rad\,d^{-1}})^{-1}, \qquad \omega \in [0, 1]\,\mathrm{rad\,d^{-1}}, \tag{30}$$

$$\pi(\phi) = \frac{1}{2\pi}\,\mathrm{rad}^{-1}, \qquad \phi \in [0, 2\pi]\,\mathrm{rad}, \tag{31}$$

$$\pi(e) = 1, \qquad e \in [0, 1], \tag{32}$$

$$\pi(\varpi) = \frac{1}{2\pi}\,\mathrm{rad}^{-1}, \qquad \varpi \in [0, 2\pi]\,\mathrm{rad}. \tag{33}$$

The offset and jitter have priors

$$\pi(v_0) = \frac{1}{10000}\,(\mathrm{m\,s^{-1}})^{-1}, \qquad v_0 \in [-5000, 5000]\,\mathrm{m\,s^{-1}}, \tag{34}$$

$$\pi(s^2) = \frac{1}{100000}\,(\mathrm{m^2\,s^{-2}})^{-1}, \qquad s^2 \in [0, 100000]\,\mathrm{m^2\,s^{-2}}. \tag{35}$$

We do not include linear trend as a parameter because there is clearly a massive long-period companion whose period can be fitted very accurately.

The optimal-fit parameters for 1-companion, 2-companion and 3-companion models are summarized in Tab. (1). The fits are shown in Fig. (5) and Fig. (6). The evidence of 1-companion model is $\exp(-486.4139 \pm 0.0434)$. The evidence of 2-companion model is $\exp(-399.9947 \pm 0.0848)$. The evidence of 3-companion model is $\exp(-403.7936 \pm 0.1604)$. The levels of these models are shown in Fig. (4). So the 1-companion model is extremely unlikely. And 2-companion model is about 44 times more likely than the 3-companion model.

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

	1-companion model	2-companion model	3-companion model
log likelihood	-446.054	-322.223	-314.403
$A_1\,(\mathrm{m\,s^{-1}})$	4036.14	4016.07	4015.09
$\omega_1\,(\mathrm{rad\,d^{-1}})$	4.87235×10^{-4}	5.24207×10^{-4}	5.16891×10^{-4}
$\phi_1\,(\mathrm{rad})$	4.68676	4.56655	4.59078
e_1	0.745968	0.734338	0.736879
$\varpi_1\,(\mathrm{rad})$	4.16403	4.17599	4.17536
$A_2\,(\mathrm{m\,s^{-1}})$		134.792	132.058
$\omega_2\,(\mathrm{rad\,d^{-1}})$		0.0210239	0.0210218
$\phi_2\,(\mathrm{rad})$		5.39148	5.37908
e_2		0.0503521	0.0370734
$\varpi_2\,(\mathrm{rad})$		3.50770	3.53115
$A_3\,(\mathrm{m\,s^{-1}})$			12.9190
$\omega_3\,(\mathrm{rad\,d^{-1}})$			9.64646×10^{-3}
$\phi_3\,(\mathrm{rad})$			3.09382
e_3			0.366577
$\varpi_3\,(\mathrm{rad})$			2.38103
$v_0\,(\mathrm{m\,s^{-1}})$	-234.629	-247.814	-239.267
$s^2\,(\mathrm{m^2\,s^{-2}})$	8562.04	288.578	251.668

Table 1: The optimal-fit parameters for all 3 models are summarized here. The 1st row lists the log likelihoods of the optimal-fit parameters of these 3 models.

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

	$N_{1a} = 10,000$		$N_{1b} = 100,000$	
Level	experimental mean	analytical value	experimental mean	analytical value
1	-25.2525	-25.2504	-25.2569	-25.2570
2	-10.4490	-10.4481	-10.4530	-10.4530
3	-5.00537	-5.00442	-5.00707	-5.00708
4	-3.00304	-3.00241	-3.00382	-3.00372
5	-2.26627	-2.26615	-2.26680	-2.26675
6	-1.99543	-1.99538	-1.99567	-1.99565

Table 2: The experiment was repeated $10,000$ times. The experimental mean values of the logarithm of the 6 levels' thresholds in the table are the mean of the $10,000$ repetitions. Notice that the experimental means for N_{1b} tend to be closer to analytical values than those for N_{1a} as expected.

	$N_{1a} = 10,000$		$N_{1b} = 100,000$	
Level	experimental std	analytical std	experimental std	analytical std
1	0.36	0.31	0.11	0.097
2	0.18	0.16	0.057	0.051
3	0.081	0.072	0.026	0.023
4	0.034	0.031	0.011	0.0097
5	0.014	0.013	0.0044	0.0040
6	0.0057	0.0051	0.0018	0.0016

Table 3: The experiment was repeated $10,000$ times. The experimental std of the logarithm of the 6 levels' thresholds in the table are the variance of the $10,000$ repetitions. The algorithm almost achieves the expected precision. Note that the std's for N_{1a} are approximately $\sqrt{10}$ times those for N_{1b}.

level index	log likelihood threshold	log prior mass
1	-29.8629798621251	-1
2	-15.0587589731369	-2
3	-9.61259046551731	-3
4	-7.60905703840871	-4
5	-6.87199828087570	-5
6	-6.60084951704394	-6
7	-6.50109946133118	-7
8	-6.46440346657875	-8
9	-6.45090376453600	-9
10	-6.44593750169253	-10

Table 4: True levels' thresholds and prior masses are listed in the table. 10 Levels are given. We keep many digits because these are true levels.

std for 10-d case ($\times 10^{-15}$)

	$N_1 = 10^4$	$N_1 = 3 \times 10^4$	$N_1 = 10^5$
$N_2 = 10^6$	3.1760	3.2465	3.1653
$N_2 = 3 \times 10^6$	1.8297	1.8816	1.8390
$N_2 = 10^7$	0.9887	0.9704	1.0238

Table 5: The standard deviations for different N_1's and N_2's. The experiments are repeated for $1,000$ times. Compared with evidence value 9.77×10^{-14}, all the standard deviations are reasonably small. But N_2 is clearly more important in reducing variance. Note that there are 30 levels in this case, $30 \times N_1$ and N_2 are actually comparable.

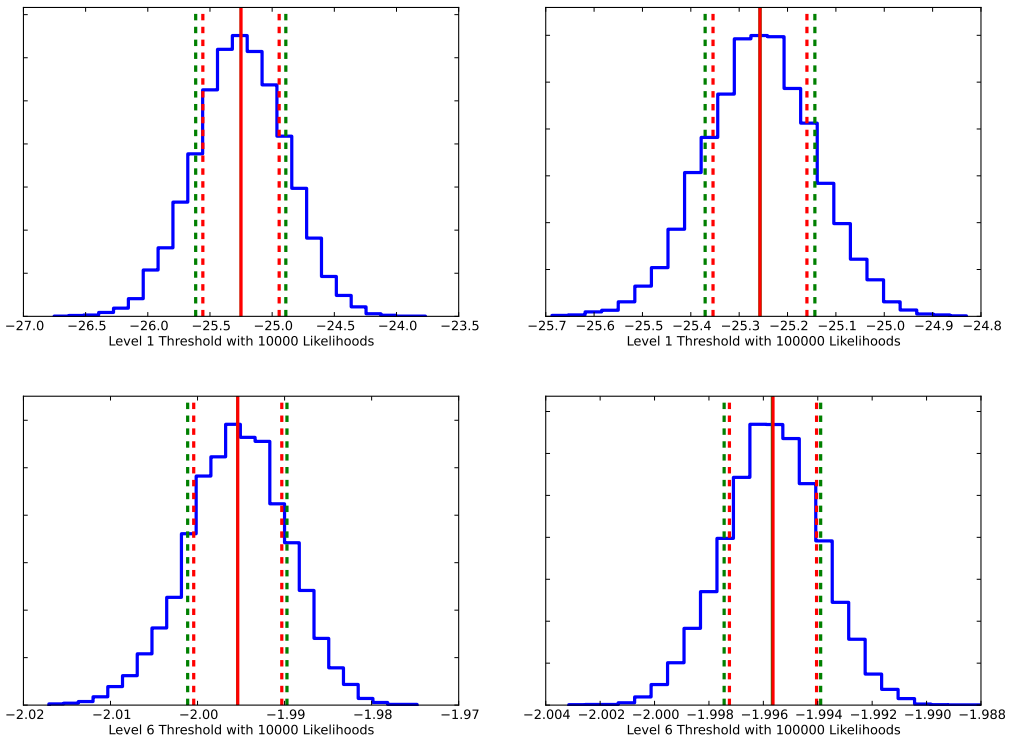

Fig. 1.— Level 1 (upper) and Level 6 (lower): Both histograms are plotted with 25 bins and 10,000 samples. (Here samples mean repetitions of the experiment, not the samples of likelihood to build every single level.) The left-hand side is the histogram of levels built with N_{1a} likelihoods and the right-hand side is the histogram of levels built with N_{1b} likelihoods. The red solid lines indicate the true values of the logarithm of the likelihood thresholds of levels and dark green solid lines indicate the experimental mean of the logarithm of the likelihood thresholds, which cannot be distinguished from the true values in these pictures. The red dashed lines indicate the theoretical standard deviation and the dark green dashed lines indicate the experimental standard deviation.

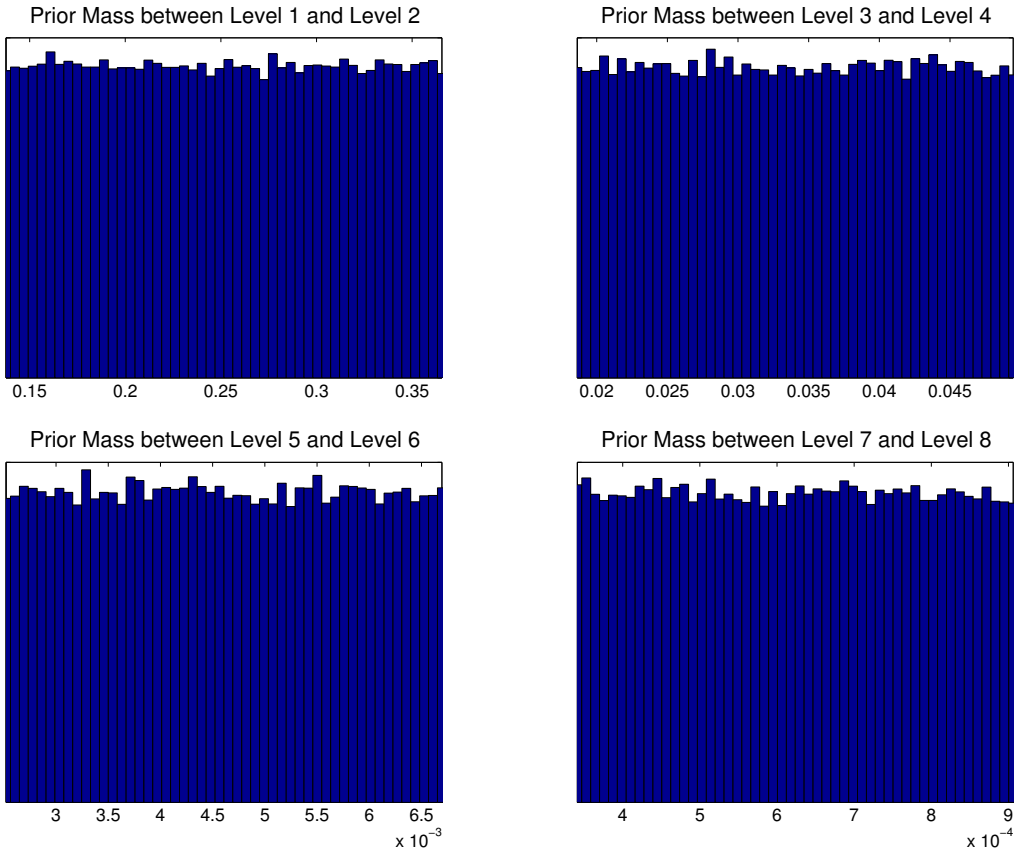

Fig. 2.— Histograms of prior masses of samples inside a bin sandwiched by two adjacent levels. 4 examples are given. All are approximately uniform distribution.

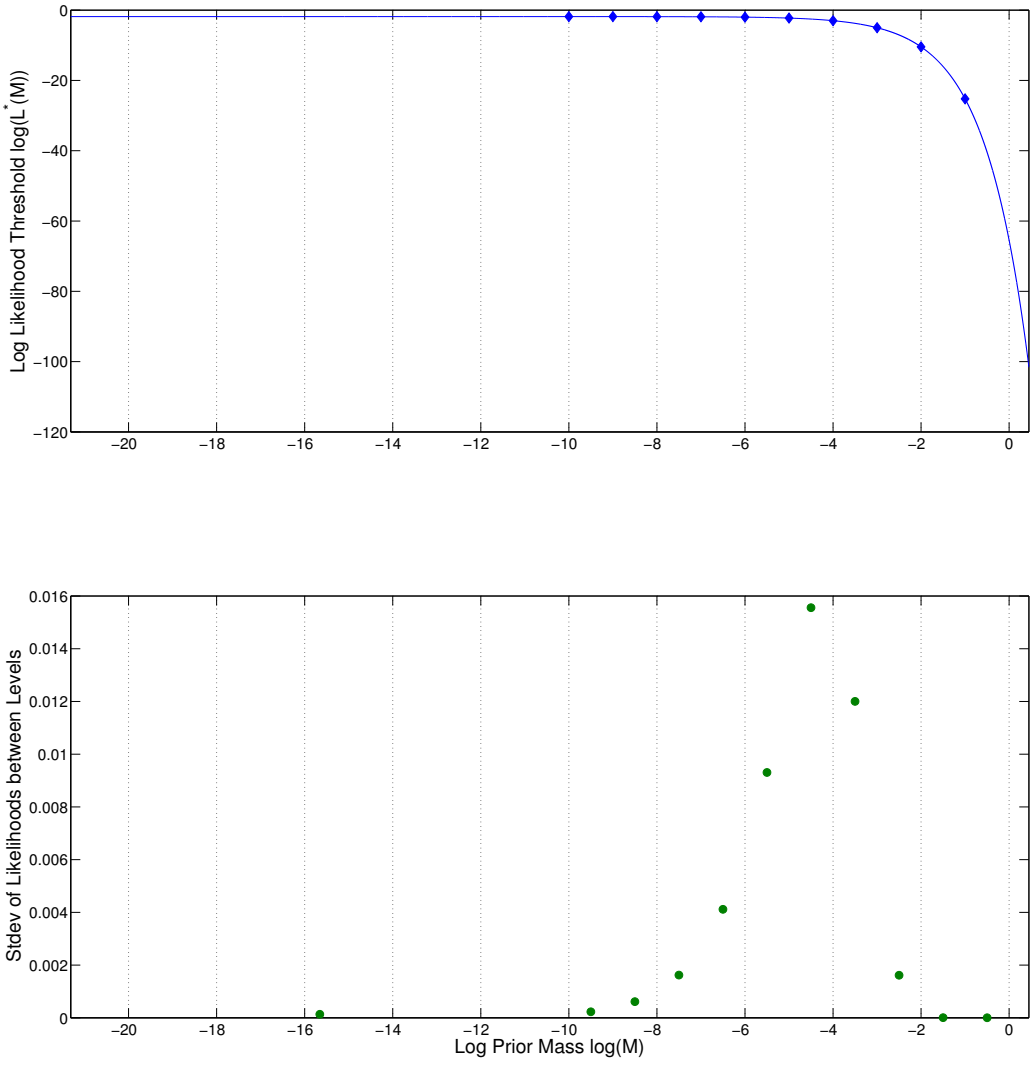

Fig. 3.— The 10 levels in the figure are the true levels summarized in Tab. (4). Notice that the standard deviation of likelihood samples between level 3 ($\log M = -3$) and level 6 ($\log M = -6$) will pretty much determine the standard deviation of the final result.

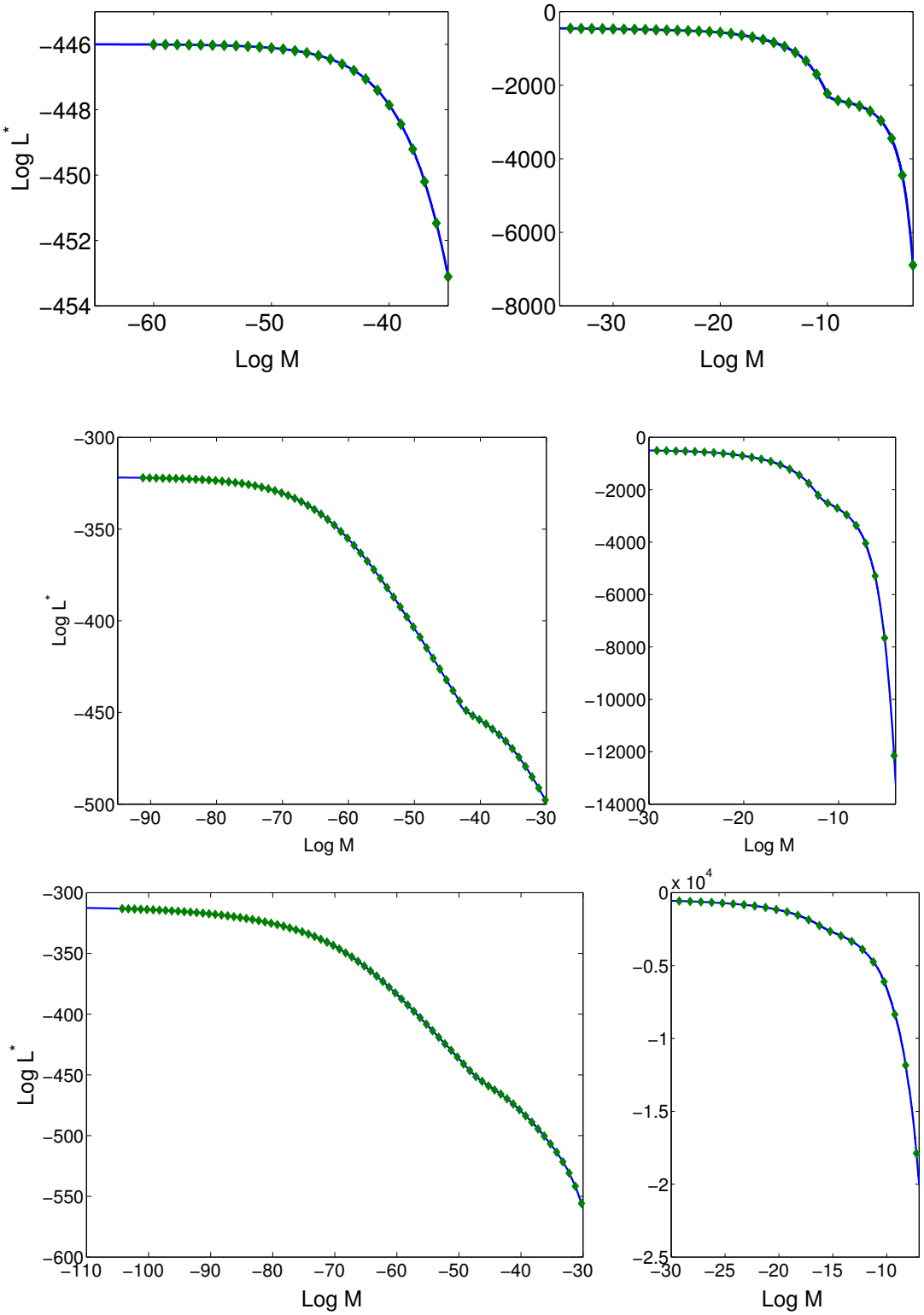

Fig. 4.— The 1st row shows the levels of 1-companion model. The 2nd row shows the levels of 2-companion model. The 3rd row shows the levels of 3-companion model.

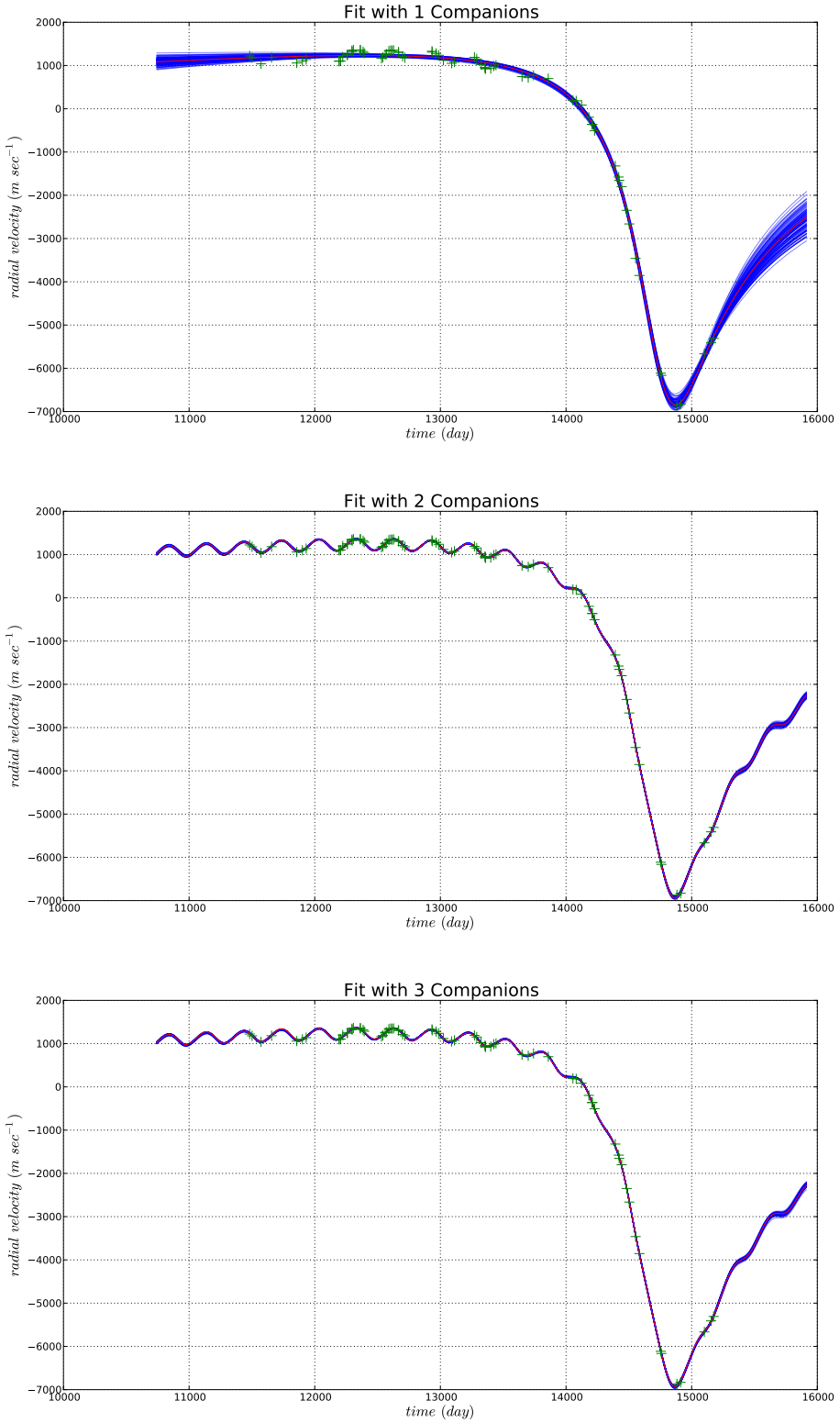

Fig. 5.— The 1st row shows the fit of 1-companion model. The 2nd row shows the fit of 2-companion model. The 3rd row shows the fit of 3-companion model. All the fits are drawn from the posterior of each model. The companion model is clearly not as good as the other two. The 2-companion and 3-companion models are not distinguishable from this view.

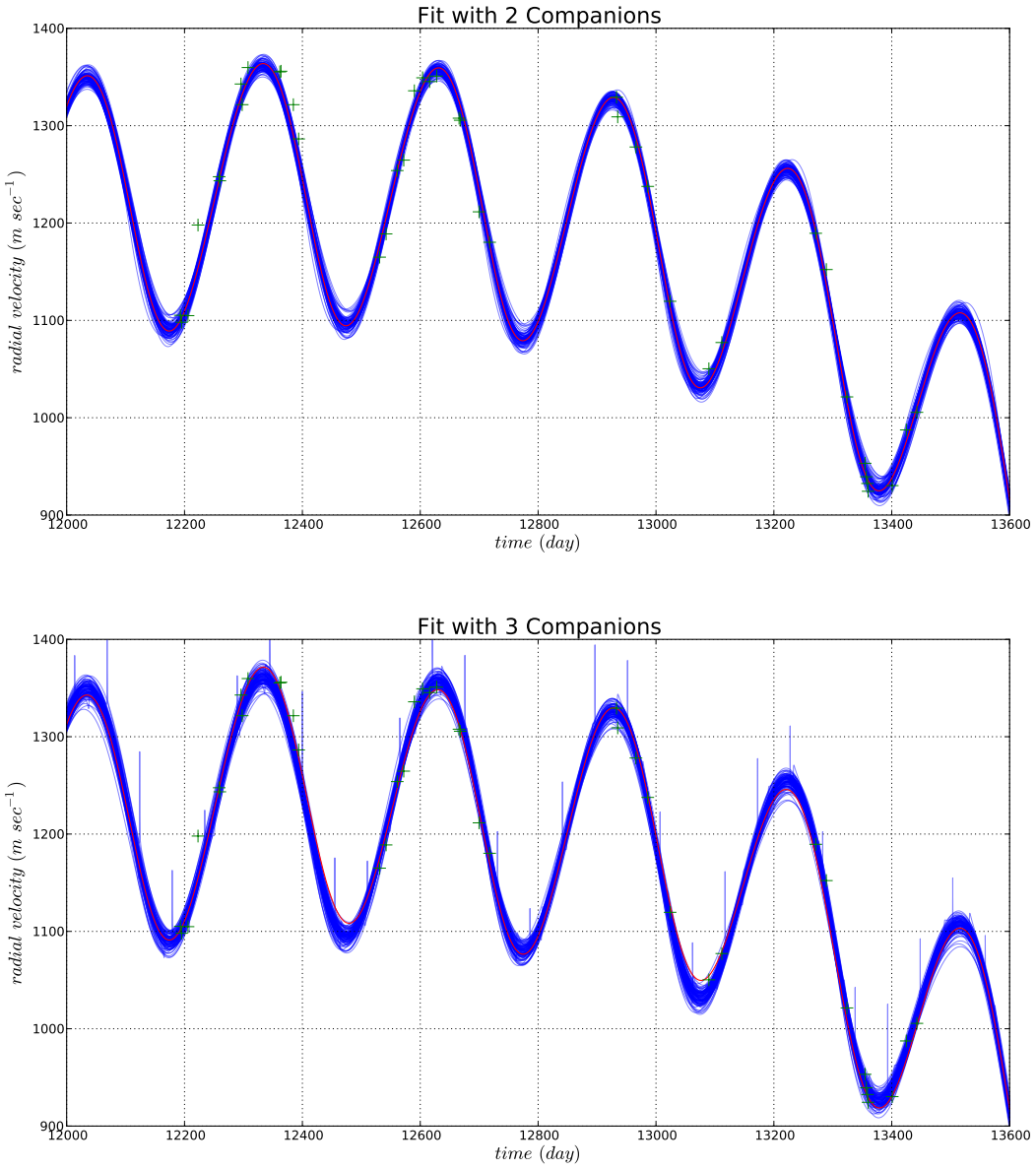

Fig. 6.— The 1st row shows the fit of 2-companion model zoomed in. The 2nd row shows the fit of 3-companion model zoomed in. The red curves in the center indicate the optimal fits. The 3-companion fit is only slightly better than the 2-companion fit. Because the optimally-fit 'well' in the 3-companion fit is too shallow. Some of the 3-companion fits actually come from local minima. (In the figure, some fits are spiky which indicates over-fitting.)