

# Diffusive Nested Ensemble Sampling

## ABSTRACT

Diffusive Nested Sampling proves to be an efficient and most importantly accurate method to evaluate bayesian evidence integral. We developed an affine invariant ensemble version of the Diffusive Nested Sampling Method to evaluate bayesian evidence integral. We tested our code on models whose evidences can be calculated analytically and got satisfactory results. As an example, we also evaluate different models' evidence integrals of exoplanet radial velocity fitting for 'star 122' 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

In bayesian decision theory, one should return the best mixture of competing models instead of simply deciding which model is the best. To achieve this, the evaluation of evidence integral  $Z$  is required. Bayesian decision theory can be summarized as

$$P(\text{Model}_j|\text{Data}) = \frac{P(\text{Data}|\text{Model}_j)P(\text{Model}_j)}{\sum_j P(\text{Data}|\text{Model}_j)P(\text{Model}_j)}, \quad (1)$$

which is just the Bayes law. The *likelihood* in Eqn. (1) is in fact the evidence  $Z_j$  for model  $j$ .

$$Z_j = P(\text{Data}|\text{Model}_j) = \int P(\text{Data}|\theta, \text{Model}_j)P(\theta|\text{Model}_j)d\theta, \quad (2)$$

where  $P(\text{Data}|\theta, \text{Model}_j)$  is the likelihood of parameter  $\theta$  for model  $j$  and  $P(\theta|\text{Model}_j)$  is the prior of parameter  $\theta$  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  $j$  in  $Z_j$  because the discussion applies to all models. So the evidence can be written as

$$Z = \int L(\theta)\pi(\theta)d\theta. \quad (3)$$

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

$$\int \pi(\theta)d\theta = 1, \quad (4)$$

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 affine invariant ensemble sampler (Goodman *et al.* 2010) 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  $\theta$  to the prior mass

$$M(L^*) = \int_{L(\theta) > L^*} \pi(\theta) d\theta, \quad (5)$$

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  $L^*$  and it ranges from 0 to 1. And the mapping between  $M$  and  $L^*$  is a bijection. An infinitesimal increment of  $M$  is

$$dM = \int_{L^* - dL^* < L(\theta) < L^*} \pi(\theta) d\theta = \pi(\theta) \times \text{volume of } \theta, \text{ satisfying } L^* - dL^* < L(\theta) < L^*. \quad (6)$$

Multiply both sides with  $L^*$  and integrate. It is easy to see that

$$\int_0^1 L^* dM = \int L(\theta) \pi(\theta) d\theta. \quad (7)$$

so evidence  $Z$  can be expressed as

$$Z = \int_0^1 L^*(M) dM. \quad (8)$$

So the integral  $Z$  is the area below the  $L^*(M)$  curve. 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 and proposes to build the  $L^*(M)$  curve statistically.

### 2.1. Level and Constrained Prior

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

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

where

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

and note that  $p_{L_j^*}$  is properly normalized by  $M_j$ .

## 2.2. Setting Level Thresholds

We already know the right-most point on the curve,  $(M_0 = 1, L_0^* = 0)$ , which is level 0, because there is no restriction on the likelihood function with  $L_0^* = 0$  and from the definition of prior mass, Eqn. (5),  $M(\theta)$  should cover the whole parameter space. And most likely another point on the left-most side of the curve,  $(M_{max} = 0, L_{max})$ .

To find the first level  $(M_1, L_1^*)$ , we generate  $N$  samples from the prior density  $\pi(\theta)$ ,  $\theta_1, \dots, \theta_N$ . We calculate the likelihoods  $L(\theta_k)$ . We choose  $L_1^*$  so that the number of  $k$  with  $L(\theta_k) > L_1^*$  is  $N/e$ . A better way to do this is with the algorithm called **quick find**. In the STL it is described here: [http://www.cplusplus.com/reference/algorithm/nth\\_element/](http://www.cplusplus.com/reference/algorithm/nth_element/).  $M_1 \approx 1/e$  is the prior mass that level 1 covers (the prior mass of the parameters whose likelihoods are larger than  $L_1^*$ ).  $M_1$  is a random variable and its expectation and variance will be given below.

One way to estimate the next level  $(M_2, L_2^*)$  would be to generate  $N$  samples from the constrained prior density  $p_{L_1^*}(\theta)$  defined in Eqn. (9). But instead, we generate samples from a mixture of  $p_{L_1^*}(\theta)$  and prior  $\pi(\theta)$  until we have  $N$  samples with likelihood larger than the previous level's threshold  $L_1^*$ . We sample the mixture because the area covered by level 1 may be disconnected in parameter space and only sampling the constrained prior  $p_{L_1^*}(\theta)$  may get us stucked in only one or few of those disconnected areas. 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. Again,  $M_2 \approx 1/e^2$  is the prior mass that level 2 covers. Like  $M_1$ ,  $M_2$  is also a random variable. The variance of  $M_1$  and  $M_2$  can be calculated by using order statistics.

There is a simple stopping criterion to tell how many levels are enough. Suppose we already have levels  $(M_1, L_1^*), \dots, (M_k, L_k^*)$ . The evidence integral corresponding to all these levels is

$$Z_k = \sum_{j=0}^{k-1} \int_{M_j}^{M_{j+1}} L^*(M) dM$$

The remaining integral is

$$\int_{M_k}^1 L^*(M) dM .$$

But  $L^*(M) < L_{max}$  always, so the remain integral cannot be larger than  $L_{max}(1 - M_k)$ . We choose a stopping point  $k$  so that  $L_{max}(1 - M_k) \leq \epsilon Z_k$ . We usually choose  $\epsilon = 10^{-6}$ .

At the end, we get a series of points on the  $L^*(M)$  curve,  $\{(M_0, L_0^*), (M_1, L_1^*), (M_2, L_2^*), \dots\}$ . Correspondingly, we also have a series of mixture of constrained priors,

$$p(\theta) = \sum_{j=0} w_j p_{L_j^*}, \quad (10)$$

where  $p_{L_j^*}$  is the constrained prior defined by level  $j$ ,

$$p_{L_j^*}(\theta) = \frac{\pi(\theta)}{M_j} \mathbb{1}_{L(\theta) > L_j^*}, \quad j = 0, 1, 2, \dots \quad (11)$$

and  $w_j$  are the weights of each level which sum up to 1,

$$\sum_{j=0} w_j = 1. \quad (12)$$

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.

### 2.3. Affine Invariant Stretch Move

To implement the affine invariant ensemble sampler to diffusive nested sampling, we assign different levels to different walkers in the ensemble. This probably would cause concern since all the walkers would be from different density and this might render stretch move invalid. But in the stretch move, the helper walker does not interfere with the walker which it helps, even if they are from different density distribution.

Our goal is to sample the mixture of the constrained priors Eqn. (10). We realize it by updating the walkers according to their own constrained priors Eqn. (9) and updating the indices of the levels of those walkers according to their weights and other restrictions. So the algorithm consists of two part: updating the ensemble of walkers followed or preceded by updating all the level indeces of those walkers. The first part can be summarized as: **JG: probably  $X_{new} = \dots$ , and  $\alpha = z$ . I don't think the metropolis part is stated correctly. FH: It should be correct now.**

- randomly choose a helping walker  $Y$
- propose a new walker with stretch move:  $X_{new} \rightarrow Y + \alpha(X_{old} - Y)$ , where  $\alpha$  is a random variable from some distribution (Goodman *et al.* 2010).
- if the proposed  $X_{new}$  has a likelihood smaller than its current threshold  $L^*$ , reject the proposal.
- else, accept the proposed  $X_{new}$  with probability:  $\max\left(z^{\dim-1} \frac{p_{L^*}(X_{new})}{p_{L^*}(X_{old})}, 1\right)$ .

The second part can be summarized as (Brewer *et al.* 2011):

- propose a new level for a walker in the ensemble, with proposal probability, Eqn. (13):  $i \rightarrow j$
- if  $j \geq i$ , accept the proposal if, in parameter space, the likelihood of the walker is larger than  $L_j^*$ ; reject if it is smaller.

- else, accept the proposal with probability  $\min\left(\frac{M_i w_j}{M_j w_i}, 1\right)$ .

The proposal probabilities for the second part,  $T_{i \rightarrow j}$ , can be written as entries of matrix  $T$ ,

$$(T_{i \rightarrow j}) = (T_{ij}) \begin{pmatrix} 0.5 & 0.5 & & & & \\ 0.5 & 0 & 0.5 & & & \\ & 0.5 & 0 & 0.5 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 0.5 & 0 & 0.5 \\ & & & & 0.5 & 0 & 0.5 \\ & & & & & 0.5 & 0.5 \end{pmatrix} \quad (13)$$

## 2.4. Refining Level Masses

Constructing levels and sampling a mixture of constrained priors  $p(\theta)$ , Eqn. (10), are equivalent to finding a probability density function  $p(\theta)$  similar in shape with  $L(\theta)\pi(\theta)$  and performing importance sampling. We can refine the prior masses  $M$ 's of the levels to have a more accurate  $p(\theta)$  and thus a more accurate importance function  $\frac{L(\theta)\pi(\theta)}{p(\theta)}$ .

Suppose we have  $J$  levels,  $\{(M_1, L_1^*), (M_2, L_2^*), \dots, (M_J, L_J^*)\}$ . Sampling the mixture of constrained priors Eqn. (10), we can refine the prior masses  $\{M_1, \dots, M_J\}$  by keeping track of all the levels the walkers have visited as well as the likelihoods of those visits. Assuming 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  $M_j$  as following (Brewer *et al.* 2011)

$$M_j^* = M_{j-1}^* \frac{n_{j-1}^j + C M_j / M_{j-1}}{n_{j-1} + C}, \quad (14)$$

where  $M^*$ 's are the refined prior masses and  $C$  is a constant that reflects one's confidence in the accuracy of the levels before this refinement. We use  $C = 10^4$ . Of course  $M_0^* = 1$  and we start from refining  $M_1$ , so  $M_{j-1}^*$  will be known when we refine  $M_j$ . If  $n_{j-1}$  and  $n_{j-1}^j$  dwarf  $C$ , the variance of  $M_j^*$  is approximately

$$\text{var} \left( \frac{M_j^*}{M_{j-1}^*} \right) \approx \frac{1/e(1 - 1/e)}{n_{j-1}}. \quad (15)$$

While sampling the mixture of constrained priors Eqn. (10), it is possible that the weights  $w_j$  are not sampled as desired. This happens when the prior masses of levels are not estimated accurately. As a result, the acceptance probability  $\min\left(\frac{M_i w_j}{M_j w_i}, 1\right)$  deviates from desired. Paradoxically, We need the weights to be sampled as desired in order to refine the prior masses to be more accurate. In such cases, we can implement certain reinforcing mechanism. (Brewer *et al.* 2011)

## 2.5. Computing Evidence

We take the mean of likelihoods sandwiched between two levels,

$$\bar{L}_j = \frac{1}{n_j} \sum_{L_j^* \leq L(\theta) < L_{j+1}^*} L(\theta), \quad (16)$$

where  $\theta$  represents samples. Again with the extra level whose likelihood threshold  $L_{J+1}^*$  is the maximum likelihood and prior mass  $M_{J+1}$  is 0, the evidence is

$$Z = \sum_{j=0}^J \bar{L}_j (M_j^* - M_{j+1}^*). \quad (17)$$

## 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

$$L(\theta_1, \theta_2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\theta_1^2 + \theta_2^2}{2\sigma^2}\right), \quad \sigma = 1, \quad (18)$$

where  $\theta_1$  and  $\theta_2$  are the parameters. The prior is

$$\pi(\theta_1, \theta_2) = \frac{1}{400}, \quad \theta_1 \in [-10, 10], \quad \theta_2 \in [-10, 10], \quad (19)$$

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,

$$\text{evidence} \approx \frac{1}{400}, \quad (20)$$

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,

$$\log L^*(M) = -\log 2\pi - \frac{200M}{\pi}, \quad (21)$$

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  $\left\{ \frac{J_{1a}}{N_{1a}+1}, \left( \frac{J_{1a}}{N_{1a}+1} \right)^2, \dots \right\}$ . And for  $N_{1b}$ , the expectation of the prior masses that each level covers are  $\left\{ \frac{J_{1b}}{N_{1b}+1}, \left( \frac{J_{1b}}{N_{1b}+1} \right)^2, \dots \right\}$ . 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 expectation and variance of the corresponding (logarithm of) likelihood thresholds can then be calculated straightforwardly, because  $\log L^*$  is a linear function of  $M$  (from Eqn. (21)). 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

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{10/2}} \exp\left(-\frac{\boldsymbol{\theta}^2}{2\sigma^2}\right), \quad \sigma = 1, \quad (22)$$

where  $\theta_1$  and  $\theta_2$  are the parameters. The prior is

$$\pi(\boldsymbol{\theta}) = \frac{1}{20^{10}}, \quad \theta_j \in [-10, 10], \quad j = 1, 2, \dots, 10, \quad (23)$$

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  $9.77 \times 10^{-14}$ . 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

$$\pi(A) = \frac{1}{10000} (\text{m s}^{-1})^{-1}, \quad A \in [0, 10000] \text{ m s}^{-1}, \quad (24)$$

$$\pi(\omega) = 1 (\text{rad d}^{-1})^{-1}, \quad \omega \in [0, 1] \text{ rad d}^{-1}, \quad (25)$$

$$\pi(\phi) = \frac{1}{2\pi} \text{rad}^{-1}, \quad \phi \in [0, 2\pi] \text{ rad}, \quad (26)$$

$$\pi(e) = 1, \quad e \in [0, 1], \quad (27)$$

$$\pi(\varpi) = \frac{1}{2\pi} \text{rad}^{-1}, \quad \varpi \in [0, 2\pi] \text{ rad}. \quad (28)$$

The offset and jitter have priors

$$\pi(v_0) = \frac{1}{10000} (\text{m s}^{-1})^{-1}, \quad v_0 \in [-5000, 5000] \text{ m s}^{-1}, \quad (29)$$

$$\pi(s^2) = \frac{1}{100000} (\text{m}^2 \text{s}^{-2})^{-1}, \quad s^2 \in [0, 100000] \text{ m}^2 \text{s}^{-2}. \quad (30)$$

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.



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

## REFERENCES

- Goodman, J., Weare, J., 2010, Comm. App. Math. and Comp. Sci., 5, 65
- Brewer, B. J., Pártay, L. B. & Csányi, G, 2011, Statistics and Computing, 21, 649
- 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  $9.77 \times 10^{-14}$ , 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.

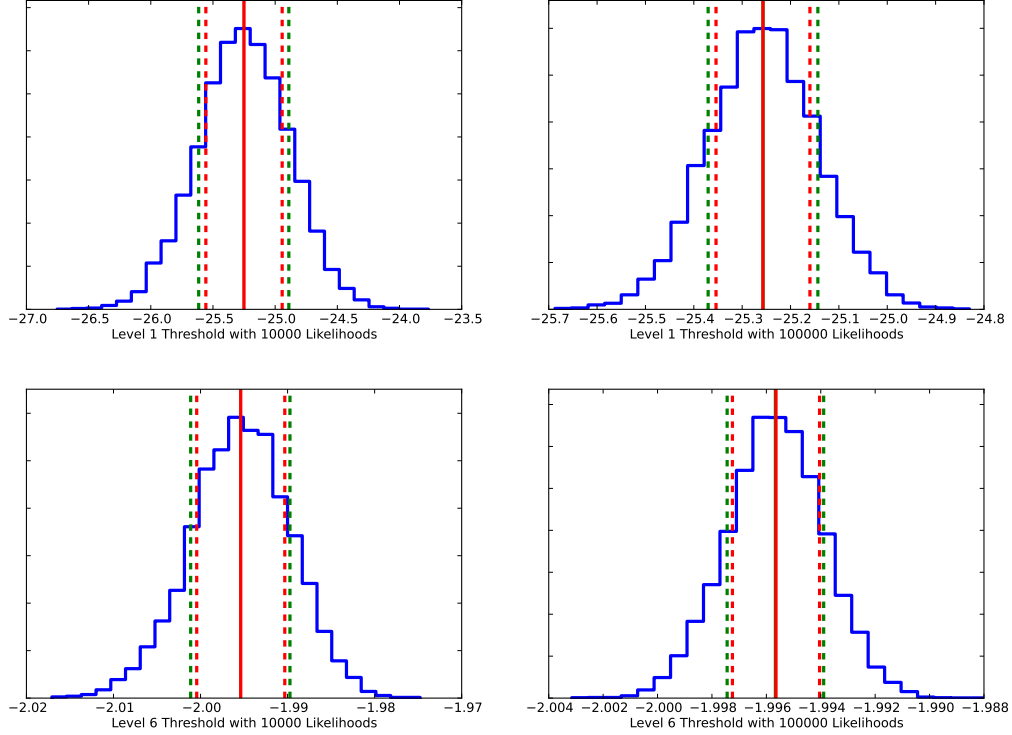


Fig. 1.— Level 1 (upper) and Level 6 (lower): Both histograms are plotted with 25 bins and 10,000 samples. (Here samples mean repetitions of the experiment, not the samples of likelihood to build every single level.) The left-hand side is the histogram of levels built with  $N_{1a}$  likelihoods and 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.

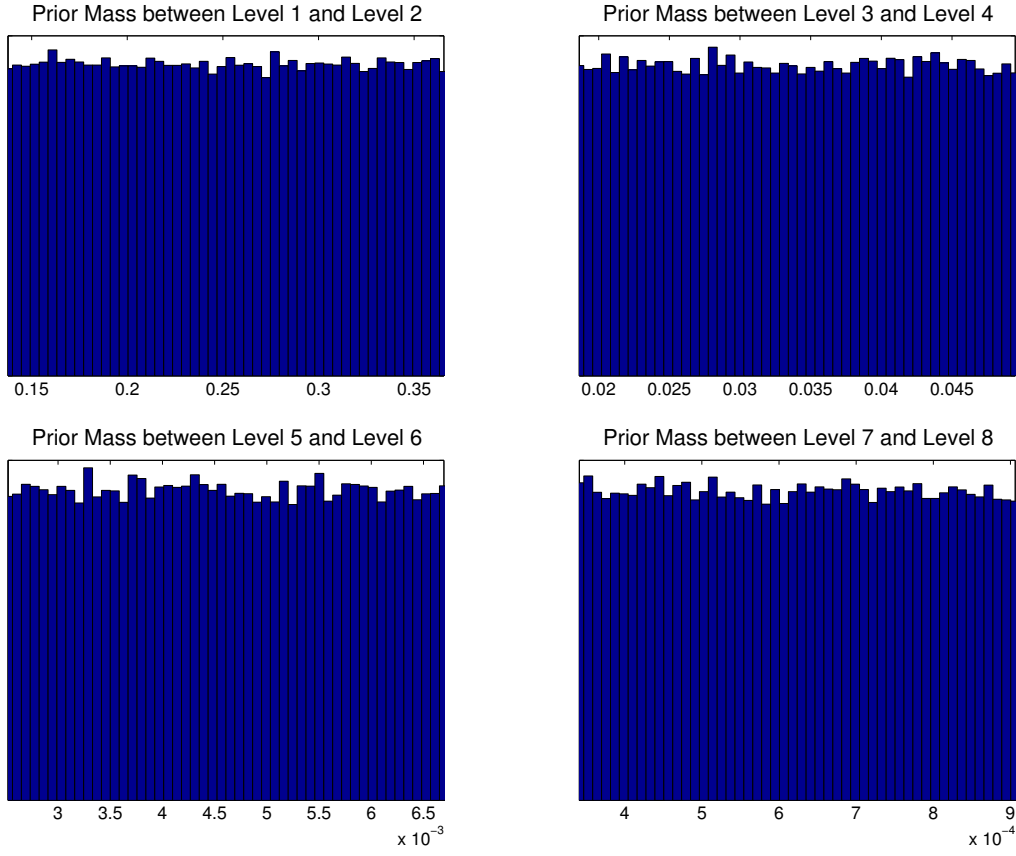


Fig. 2.— Histograms of prior masses of samples inside a bin sandwiched by two adjacent levels. 4 examples are given. All are approximately uniform distribution.

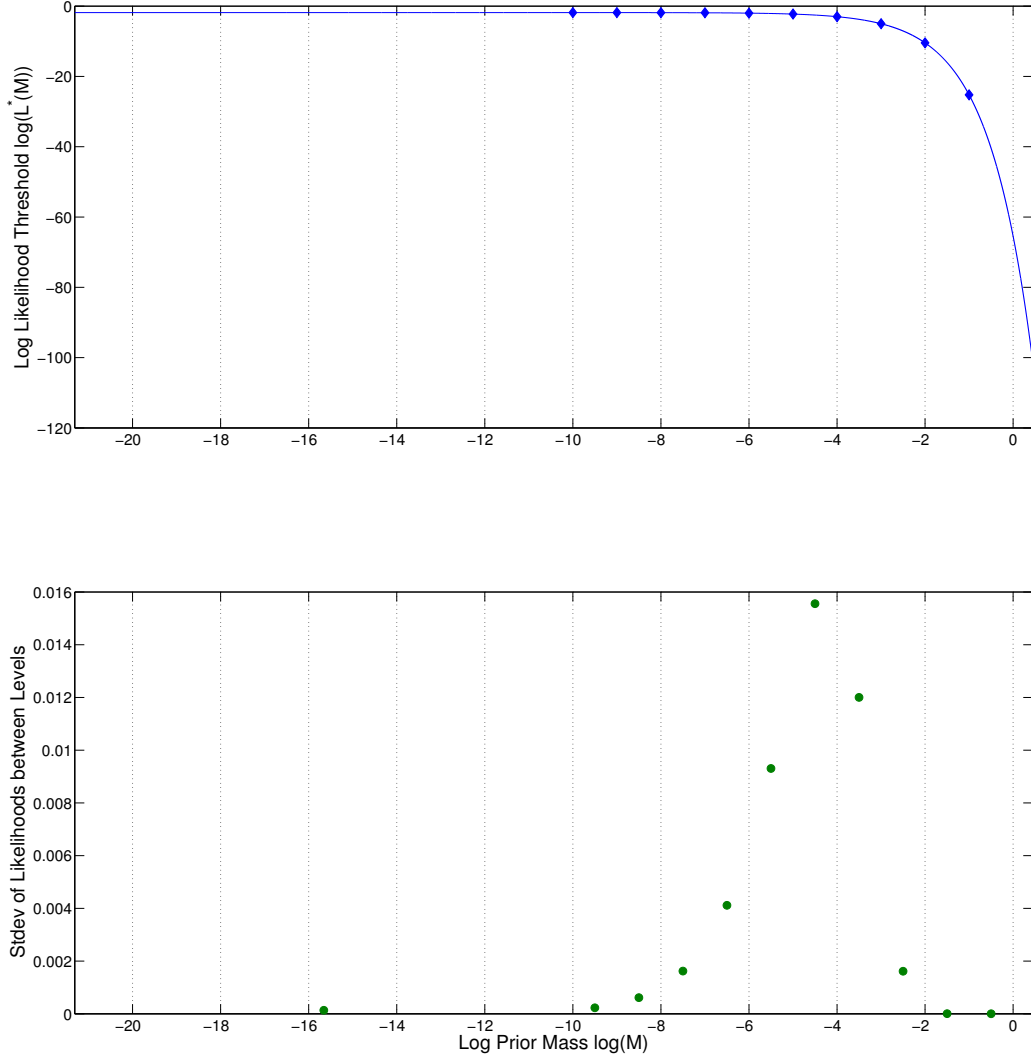


Fig. 3.— The 10 levels in the figure are the true levels summarized in Tab. (4). Notice that the standard deviation of likelihood samples between level 3 ( $\log M = -3$ ) and level 6 ( $\log M = -6$ ) will pretty much determine the standard deviation of the final result.

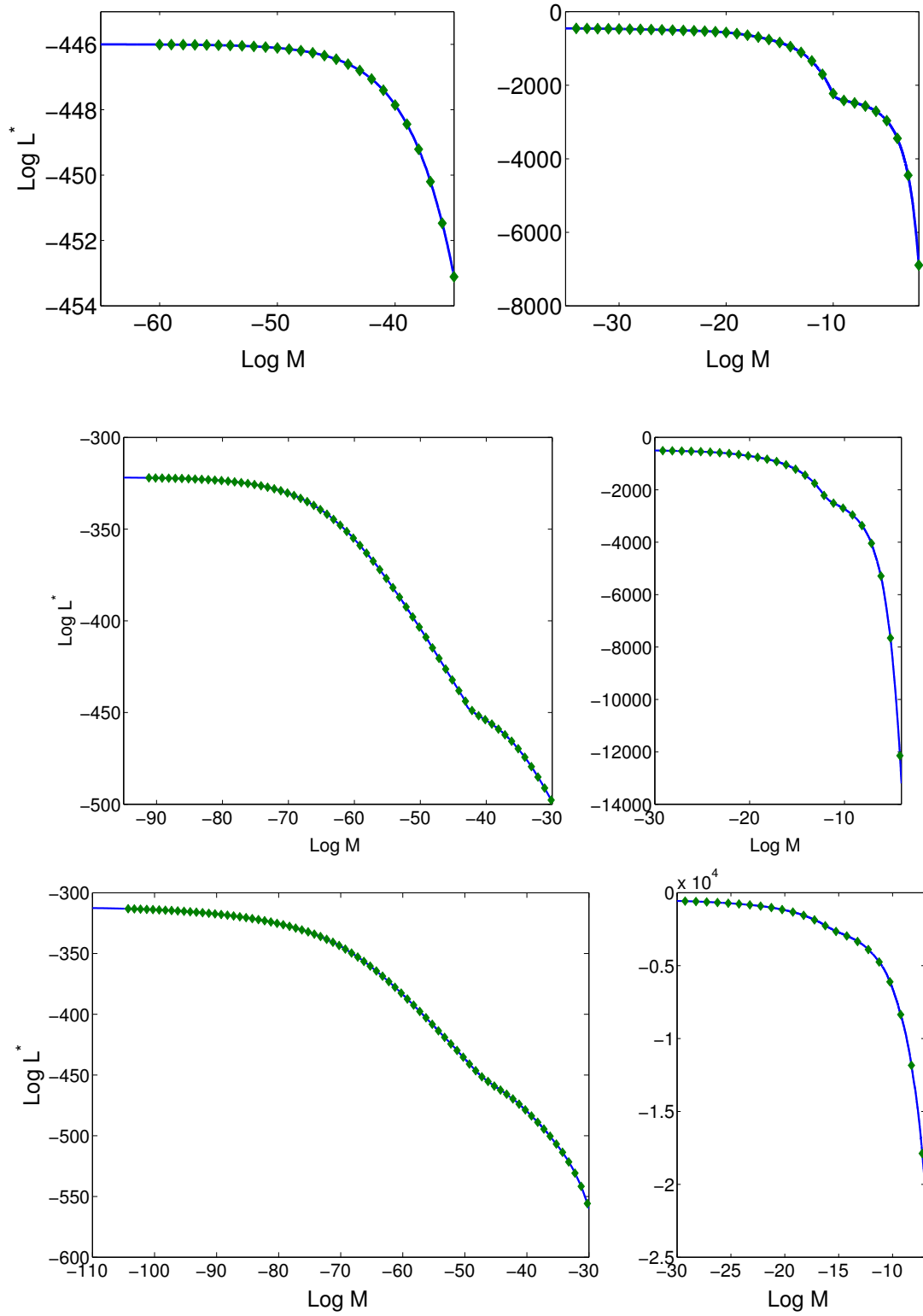


Fig. 4.— The 1st row shows the levels of 1-companion model. The 2nd row shows the levels of 2-companion model. The 3rd row shows the levels of 3-companion model.

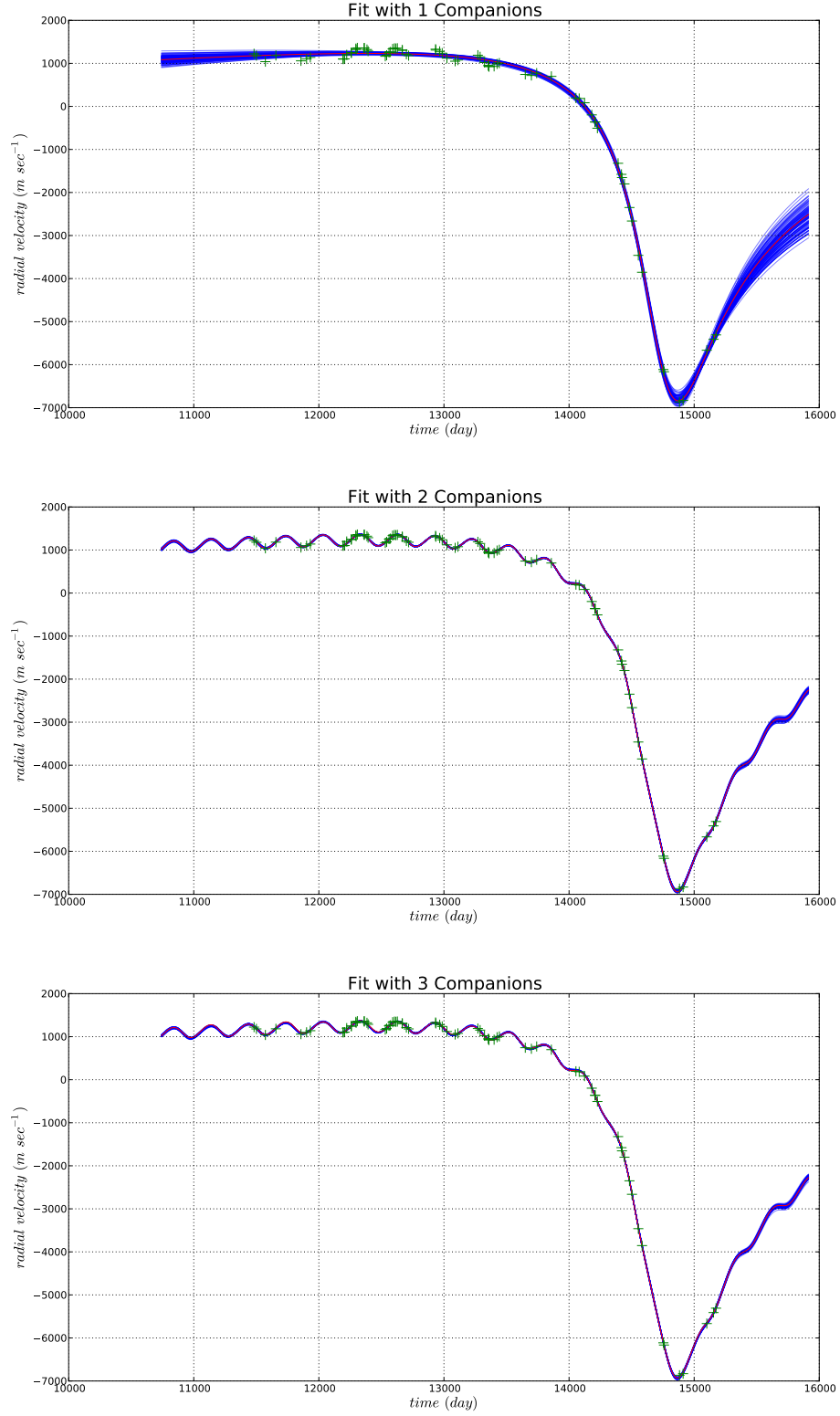


Fig. 5.— The 1st row shows the fit of 1-companion model. The 2nd row shows the fit of 2-companion model. The 3rd row shows the fit of 3-companion model. All the fits are drawn from the posterior of each model. The companion model is clearly not as good as the other two. The 2-companion and 3-companion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



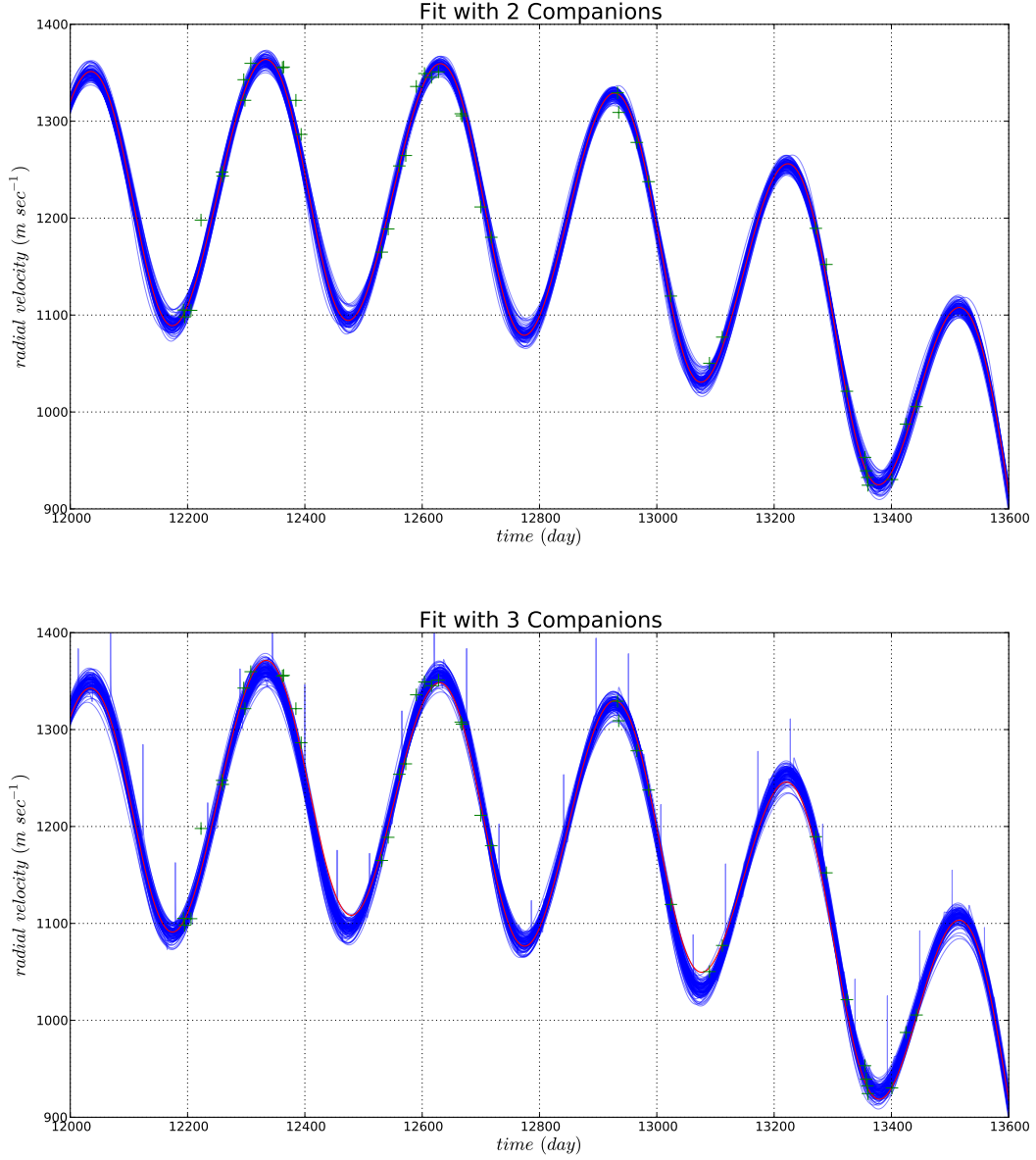


Fig. 6.— The 1st row shows the fit of 2-companion model zoomed in. The 2nd row shows the fit of 3-companion model zoomed in. The red curves in the center indicate the optimal fits. The 3-companion fit is only slightly better than the 2-companion fit. Because the optimally-fit 'well' in the 3-companion 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.)