Submitted to the Annals of Applied Statistics

ADAPTIVE ANNEALED IMPORTANCE SAMPLING FOR MULTI-MODAL POSTERIOR EXPLORATION AND MODEL SELECTION WITH APPLICATION TO EXTRASOLAR PLANET DETECTION*

BY BIN LIU[†,§] , MERLISE A. CLYDE[†,§], TOM LOREDO[‡,¶] AND JIM BERGER[†,§]

Duke University[§] *and Cornell University*[¶]

The search for extrasolar planets presents several statistical challenges including model selection and inference within models that involve multi-modal posterior distributions. To address these problems, we propose an adaptive annealed importance sampling algorithm (AAIS) which facilitates simulation from multi-modal joint posterior distribution and provides an effective and easy-to-implement method for estimating marginal likelihoods that are required for Bayesian model comparison. Using a sequential importance sampling framework, we construct mixtures of Student t distributions to approximate a sequence of "annealed" distributions that gradually approximates the target posterior density. Borrowing ideas of birth/death and split/merge steps from reversible jump Markov-chain Monte Carlo, we propose an adaptive online method to increase/decrease the number of mixture components guided by the effective sample size of the importance sample. We use simulation studies and several examples from exo-planet searches to demonstrate the greater efficiency of the method. The combination of annealing and heavier tails of the Student t components in the mixture greatly facilitate capturing the "spikey" posterior densities present in the highly nonlinear models in the exo-planet problem, while importance sampling permits straightforward estimation of marginal likelihoods and posterior model probabilities to address the uncertainty of whether planets are indeed present.

1. Introduction. Since the beginning of recorded time we have wondered whether we are alone or whether there are other planets that might support life. The scientific quest for extrasolar or exoplanets (planets beyond our own solar system) began in the mid 19th century, but it was only as recent as 1992 that astonomers have been able to confirm the existence of other planetary systems. Since then `describe resources, telescopes,`

*This work has been supported by Statistical and Applied Mathematical Sciences Institute through National Science Foundation grant DMS–042240. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

†Partially supported by National Science Foundation Grant AST–0507481

‡Partially supported by National Science Foundation Grant AST–XXXXX

Keywords and phrases: Bayes Factor, Model Selection, Nonlinear Regression

etc have been devoted to identifying stars with orbiting planets, and in particular, those that might have earth-like planets. As of June 2010 over 450 exoplanets have been detected, the majority using techniques that rely on measurements of a star's radial velocity The presence of a planet orbiting a star will lead to a periodic wobble in the the star's radial velocity, while if there are no planets present the expected velocity is constant although measurements fluctuate because of stellar jitter and other explanations for noise. Based on a short sequence of noisy measurements, the challenge is to detemine whether there are no planets or one, two or more planets, and if there are planets to characterize their orbital parameters. cite work of Gregory and Ford, but mention limitations? If the data do not provide strong evidence either in favor or against the presence of planets, additional observations may be needed to provide confirmation. Determining which stars to observe and when to schedule scarce telescope time optimally is critical given limited resources; Bayesian methods in particular are ideally suited for such sequential updating of information/evidence and decision making in such sampling problems.

The primary question of whether there are no planets or one or more may be poised as a Bayesian model selection problem, where the collection of models under consideration are the zero-planet mode $\mathcal{M}_0$, one-planet $\mathcal{M}_1$, two-planet $\mathcal{M}_2$, etc. In Bayesian statistics, the posterior probability of a model $\mathcal{M}_j$ requires accurately calculating the integrated or marginal likelihood and in the setting of exo-planet is extremely challenging; see **?** for a general overview or **?** for some of the issues that arise in astronomy. Out of available methods, importance sampling (IS) generally produces the best estimates of marginal likelihoods in terms of bias and variance. The key to building an efficient IS algorithm is to design an importance function that mimics the integrand. Building such a function can be quite challenging even in lower to modest dimensional settings, with multi-modal posterior distributions that arise in the exo-planet setting the task is all the more difficult.

In this paper, we propose a novel adaptive method for constructing an importance function using a mixture of Student-t distributions that evolves to the target posterior distribution. We present in Section 2 an overview of the scientific issues that inspired this work, namely inferring the number of planets around distant stars. In the exoplanet application, each model $\mathcal{M}_p$ has $2+5p$ parameters where p is the number of planets in the model, so even the simplest single planet model has a seven dimensional parameter space to integrate over. While low dimensional compared to many modern problems of model selection, the models are highly nonlinear making Laplace approxi-

mations untenable given the modest sample sizes and multi-modal posterior distribution; even sampling from posterior distributions using Markov chain Monte Carlo methods is in itself a challenging problem. In Section 3 we review importance sampling and how it may be used to calculate marginal likelihoods for model selection. In Section 4 we describe the Adaptive Annealed Importance Sampling method developed for the exoplanet problem. This method integrates several key features of other methods: mixture models, annealling, and adaptation via a sequence of distributions to build a flexible importance distribution that mimics the posterior distribution, but incorporates ideas successfully used in trans-dimensional methods to adapt the number of kernels in the mixture by utilizing split and merge moves. In Section 5 we present two simulation studies that demonstrate the efficiency of the method compared to other approaches. The first involves a challenging flared helix in three dimensions, while the second seven dimensional problem that was selected to reflex the target function in the exoplanet problem. In Section 6 we illustrate the method in the analysis of several real star systems with one and two planets. Finally, we conclude with discussion of possible extensions of the method in Section 7.

2. Radial Velocity Models in the Search for Exoplanets. In a two-body system such as a star and planet, the pair rotate together about a point lying somewhere on the line connecting their centers of mass. If one of the bodies (the star) radiates light, the frequency of this light measured by a distant observer will vary cyclically with a period equal to the orbital period. This Dopplar effect is understood well enough that astronomers can translate between frequency shift and the star's velocity toward or away from earth.

If a star does not host any orbiting planets, then the radial velocity (RV) measurements v_i will be roughly constant over any period of time, varying only due to the "stellar jitter" s^2, the random fluctuations in a star's luminosity. Under the zero-planet model, $\mathcal{M}_0$, the RV measurements are assumed to have Gaussian distributions

$$v_i \mid \mathcal{M}_0 \sim \mathcal{N}\left(C, \sigma_i^2 + s^2\right). \tag{1}$$

with mean C, the constant center-of-mass velocity of the star relative to the earth, and variances $\sigma_i^2 + s^2$. The parameters C and s are both in the same units as the velocity measurements, typically meters per second (m/s). The additional variance component σ_i^2 is a calculated error due to the measurement procedure `more details, ref, justification for additive form`.

The observed radial velocities (RV) v_i at time t_i for a single planet model, $\mathcal{M}_1$, are also assumed to be Gaussian distributed with

$$v_i \mid \mathcal{M}_1 \sim \mathcal{N}\left(C_1 + \triangle V(t_i|\phi_1), \sigma_i^2 + s_1^2\right), \tag{2}$$

where the velocity shift $\triangle V(t_i|\phi_1)$ due to the presence of a single planet is a family of curves parameterized by the 5 dimensional vector $\phi_1 \equiv (K_1, P_1, e_1, \omega_1, \mu_1)$

$$\triangle V(t|\phi_1) = K_1[\cos(\omega_1 + T(t)) + e\cos(\omega_1)] \tag{3}$$

where $T(t)$ is the "true anomaly at time t" given by

$$T(t) = 2\arctan\left[\tan(\frac{E(t)}{2})\sqrt{\frac{1+e_1}{1-e_1}}\right]. \tag{4}$$

and $E(t)$ is called the "eccentric anomaly at time t", which is the solution to the transcendental equation

$$E(t) - e_1\sin(E(t)) = \text{mod}\left(\frac{2\pi}{P_1}t + \mu_1, 2\pi\right). \tag{5}$$

The five orbital parameters that comprise ϕ_1 are the velocity semi-amplitude K_1, the orbital period P_1, the eccentricity e_1, $(0 \leq e_1 \leq 1)$, the argument of periastron ω_1, $(0 \leq \omega \leq 2\pi)$ and the mean anomaly at time $t = 0$, μ_1, $(0 \leq \mu_1 \leq 2\pi)$. The parameters C_1 K_1 and s_1 have units m/s; the velocity semi-amplitude K_1 is usually restricted to be non-negative to avoid identification problems, while C_1 may be positive or negative. The eccentricity parameter e_1 is unitless, with $e_1 = 0$ corresponding to a circular orbit, and larger e_1 leading to more eccentric orbits. Periastron is the point at which the planet is closest to the star and the argument of periastron ω_1, measures the angle at which we observe the elliptical orbit. The mean anomoly μ_1 is an angular distance of a planet from periastron.

If there are $p \geq 1$ planets, the expected velocity is given by $C_p + \triangle V(t_i|\phi_1, \ldots, \phi_p)$ with overall velocity shift $\triangle V$ approximated as the sum of the velocity shifts of the individual planets:

$$\triangle V(t_i|\phi_1, \ldots, \phi_p) = \sum_{j=1}^{p} K_j[\cos(\omega_j + T_i(t_i)) + e_j\cos(\omega_j)] \tag{6}$$

where the planets' mutual gravitational interactions are assumed to be negligible. With p planets, there are a total of $2+5p$ parameters, $\theta_p = \{\mathcal{C}_{\surd}, s_p^2, \phi_1, \ldots, \phi_p\}$ for each of the models $\mathcal{M}_p$. Of course, we do not know how many planets there are *a priori* - indeed, finding the number of planets p and characterizing their oribital parameters is a major aim.

2.1. *Bayesian Methods for Identifying the Number of Planets.* Determining the number of planets in a system is, from a statistical point of view, a model choice problem. Bayesian model selection requires calculation of marginal likelihoods of models or "evidence" provided by the data for each model:

$$m(\mathcal{M}_p) = \int_{\Theta_p} p(\mathbf{v} \mid \theta_p, \mathcal{M}_p) p(\theta_p \mid \mathcal{M}_p) d\theta_p \tag{7}$$

which entails integrating the sampling model of the data $\mathbf{v} = (v_1, \ldots v_n)^T$ with respect to the prior distribution of model specific parameters θ_p. Bayes Factors for comparing a p planet model to the 0 planet model may be expressed as

$$\mathsf{BF}(\mathcal{M}_p : \mathcal{M}_0) = \frac{m(\mathbf{v} \mid \mathcal{M}_p)}{m(\mathbf{v} \mid \mathcal{M}_0)} \tag{8}$$

where the Bayes factor $\mathsf{BF}(\mathcal{M}_0, \mathcal{M}_0) = 1$, while the posterior probability of the p planet model is of the form

$$p(\mathcal{M}_p \mid \mathbf{v}) = \frac{\mathsf{BF}(\mathcal{M}_p : \mathcal{M}_o) O(\mathcal{M}_p : \mathcal{M}_0)}{\sum_{j=1}^{p_{\max}} \mathsf{BF}(\mathcal{M}_j : \mathcal{M}_o) O(\mathcal{M}_j : \mathcal{M}_0)} \tag{9}$$

where $O(\mathcal{M}_p : \mathcal{M}_0)$ is the prior odds of having p planets to 0 planets and $p_{\max}$ is the maximum number of planets for the system. This requires specifying a prior distribution on θ_p for each of the models in order to obtain marginal likelihoods and Bayes factors.

2.2. *Priors Distributions.* We adopt the prior distributions recommended by Ford and Gregory (2006); Bullard (2009), which are based in part on their approximate realism, but also their mathematical tractability. For all models, intercept parameter C_p and stellar jitter parameter s_p, are taken as being *a priori* independent, where C_p is uniform over a finite set $[C_{\min}, C_{\max}]$

$$p_C(C) = \begin{cases} \dfrac{1}{C_{\max} - C_{\min}} & \text{for } C_{\min} \leq C \leq C_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{10a}$$

and $\log(s_{\min}+s_p)$ has a uniform distribution on the interval $(\log(s_{\min}), \log(s_{\max})]$,

$$p_s(s) = \begin{cases} \dfrac{1}{\log\left(1 + \frac{s_{\max}}{s_{\min}}\right)} \cdot \dfrac{1}{s_{\min} + s} & \text{for } 0 < s \leq s_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{10b}$$

The joint prior on (C_p, s_p) may be viewed as a modified independent Jeffrey's prior as $C_{\min} \to -\infty$, $C_{\max} \to \infty$, $s_{\min} \to 0$, $s_{\max} \to \infty$, which leads to well defined posterior distributions and Bayes Factors even in the limit.

For each of the p planet models, in the absence of other prior information, we take each ϕ_j to have independent identical prior distributions. Many of the parameters in ϕ_p allow informative marginal prior distributions to be specified, however, the nonlinear relationships induce strong correlations among many of the parameters, which leads to a difficult task for joint prior elicitation. For circular orbits ($e = 0$), ω and μ_0 are in fact unidentifiable. To simplify this task, we specify independent prior distribution for the components of each ϕ_p in a transformed parameter space,

$$\begin{aligned} x &= e\cos\omega \qquad y \quad = e\sin\omega \qquad z = (\omega + \mu_0) \mod 2\pi \\ \dot{P} &= \log P \qquad \dot{K} \quad = \log K \end{aligned}$$

leading to $\dot{\phi}_p \equiv (\dot{K}_p, \dot{P}_p, x_p, y_p, z_p)^T$. The Poincaré variables x_p and y_p greatly reduce the very strong correlations between μ_p and ω_p, which is particularly important for low eccentricity orbits, where the parameters are nearly unidentifiable. The use of z further reduces correlations between the parameters ω and μ_0 when $e \ll 1$, but has little effect for large e. Bullard (2009) recommended using the log transformation of P and K as posterior distributions were more Gaussian in these coordinates, which led to improved posterior simulation. In the transformed parameter space, the prior distribution for ϕ is

$$p_{\dot{\phi}}(\dot{\phi}) = c_\phi \cdot \exp\dot{K} \cdot \frac{1}{1 + \frac{\exp\dot{K}}{K_{\min}}} \cdot \frac{1}{\sqrt{x^2 + y^2}} \tag{11a}$$

for $\log(K_{\min}) < \dot{K} \leq \log(K_{\max})$, $\log(P_{\min}) \leq \dot{P} \leq \log(P_{\max})$, $x^2 + y^2 < 1$, and $0 \leq z \leq 2\pi$, where the normalizing constant is

$$c_\phi = \frac{1}{\log\left(1 + \frac{K_{max}}{K_{\min}}\right)} \cdot \frac{1}{K_{\min}} \cdot \frac{1}{\log\left(\frac{P_{\max}}{P_{\min}}\right)} \cdot \left(\frac{1}{2\pi}\right)^2. \tag{11b}$$

The constants in the prior distribution (see Table 1) are set based upon the physical realities (e.g., an orbit with too small a period will result in the planet getting consumed by the star) (Bullard, 2009) of the form given in 11a with hyperparameters from Table 1.

3. Importance Sampling for Marginal Likelihood Estimation. Beyond conjugate models, the integrals defining the marginal likelihood in Equation

$P_{\min}$	1 day	$P_{\max}$	1,000 years
$K_{\min}$	1 m/s	$K_{\max}$	2128 m/s
C_{min}	−2128 m/s	C_{max}	2128 m/s
$s_{\min}$	1 m/s	$s_{\max}$	2128 m/s

TABLE 1
Prior hyperparameters for the distribution of ϕ_p.

(7) are generally intractable and must be estimated by numerical methods. Importance sampling (IS) (**?**) is a well known Monte Carlo technique for estimating such integrals. Given an importance distribution Q that is abolutely continuous with respect to the target posterior distribution the marginal likelihood may be re-expressed as an expectation with respect to Q

$$m(\mathcal{M}) = \mathbb{E}_Q\left[\frac{p(\mathbf{v} \mid \theta, \mathcal{M})p(\theta \mid .\mathcal{M})}{q(\theta)}\right] \tag{12}$$

$q(\theta)$ is the associated probability density function for distribution Q from which samples may be obtained relatively easily. Given a random sample of size N $\theta^{(1)}, \ldots, \theta^{(N)}$ from Q, an unbiased and simulation consistent Monte Carlo estimate of the marginal likelihood is provided by

$$\hat{m}(\mathcal{M}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{v} \mid \theta^{(i)}, \mathcal{M})p(\theta^{(i)} \mid \mathcal{M})}{q(\theta^{(i)})} = \frac{1}{N}\sum_{i=1}^{N} w^{(i)} \tag{13}$$

where $w^{(i)}$ are known as the IS weights. The distribution of the error $|\hat{m}(\mathcal{M}) - m(\mathcal{M})|$ can be approximated via the central limit theorem

$$\mathrm{Var}_q\left\{\hat{\mathsf{Z}}\right\} = \mathbb{E}_q\left\{\left(\hat{\mathsf{Z}} - \mathsf{Z}\right)^2\right\} = \frac{\sigma_q^2}{N} < \infty, \tag{14}$$

when $\hat{\mathsf{Z}}$ has finite variance. Here

$$\sigma_q^2 = \mathrm{Var}_q\left\{\frac{L(\theta|y)p(\theta|\mathcal{M})}{q(\theta)}\right\} = \int_{R^d}\left(\frac{L(\theta|y)p(\theta|\mathcal{M})}{q(\theta)} - \mathsf{Z}\right)^2 q(\theta)d\theta = \mathsf{Z}^2\mathbb{E}_q\left\{\left(\frac{\pi(\theta) - q(\theta)}{q(\theta)}\right)^2\right\}. \tag{15}$$

If the importance density is such that $\sigma_q^2 < \infty$, then via the central limit theorem

$$\frac{\hat{\mathsf{Z}} - \mathsf{Z}}{\sigma_q/\sqrt{N}} \to X \tag{16}$$

as $N \to \infty$, where $X \sim \mathcal{N}(0, 1)$. The above results then implies that $\hat{Z} - Z = \mathcal{O}(\frac{1}{\sqrt{N}})$.

One can construct confidence intervals to assess the quality of a particular estimate of Z (Lefebvre *et al.*, 2009). With a large degree of certainty, the interval $\hat{\mathrm{Z}} \pm 3\sigma_q/\sqrt{N}$ contains Z when N is large and so the half length $3\sigma_q/\sqrt{N}$ indicates how accurate $\hat{\mathrm{Z}}$ is. Since σ_q is unknown, a natural estimator of σ_q^2 is the usual consistent variance estimator

$$\sigma_q^2 \simeq \frac{1}{N}\sum_{i=1}^{N}\left(w^i - \hat{\mathrm{Z}}\right)^2. \tag{17}$$

The efficiency of importance sampling depends largely on how closely the importance function mimics the shape of the target, $L(y|\theta)p(\theta)$. A number of things can go wrong. If the importance function has thinner tails than the target, the estimate will tend to be unstable because the weights can be arbitrarily large; on the other hand if the importance function has much fatter tails than the target, too many of the θ_i's will be drawn from the tails, resulting in a seemingly stable but biased estimate. The best case is an importance function with slightly heavier tails than the target. Of course asymptotically it should not matter which importance function is used; for samples of practical size, though, a poor choice can be disastrous.

These problems are only compounded in high-dimensional settings where the target is often extremely complicated. Unless one is very lucky indeed it is usually impossible to get accurate results in real-life situations with dimension greater than 3.

4. Annealed Adaptive IS Method for Constructing Importance Function. In this section, we propose the adaptive mixture modeling approach, which is used for constructing importance function that resembles the posterior and then is used in (13) for calculating the marginal likelihood.

Let $\pi(\theta)$ denote the posterior distribution, which is proportional to $L(y|\theta)p(\theta)$. Note that computing $L(y|\theta)p(\theta)$ for any θ is feasible, but that we are not able to directly sample from the distribution it defines, i.e. $\pi(\theta)$. The aim is to find a distribution, $q(\theta)$, that approximates $\pi(\theta)$, which we are also able to evaluate.

We first select an initial proposal distribution $q_0(\theta)$, then define a series of intermediate distributions between $q_0(\theta)$ and $\pi(\theta)$, by

$$\pi_t(\theta) \propto q_0(\theta)^{1-\lambda_t}\pi(\theta)^{\lambda_t}, \qquad t = 1, \ldots, T, \quad \text{and} \quad 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_T = 1. \tag{18}$$

Finally we propose an iterative method to construct $q(\theta)$, as shown in Algorithm 1.

Algorithm 1. Sequential Proposal Construction

- Use $\pi_1(\theta)$ as the target distribution, with $q_0(\theta)$ as the proposal. Tune parameters of $q_0(\theta)$ to obtain an updated mixture density function, denoted $q_1(\theta)$, which approximates $\pi_1(\theta)$.
- Similarly as in step 1, tune parameters of $q_1(\theta)$ to obtain an updated proposal,$q_2(\theta)$, which is used to approximate $\pi_2(\theta)$.
- iterates the above process, i.e., tuning parameters of $q_t(\theta)$ and then getting the updated mixture density function $q_{t+1}(\theta)$, from $t = 2$ to $t = T - 2$.
- Use $\pi_T(\theta)$ as the target distribution, tune parameters of $q_{T-1}(\theta)$ to obtain $q_T(\theta)$, which approximates $\pi_T(\theta) = \pi(\theta)$. Then $q_T(\theta)$ is actually the importance function which we want.

We see that the annealing strategy is adopted in the above iterative procedure. In what follows, we propose an adaptive method to handle the task of updating $q_{t-1}(\theta)$ to $q_t(\theta)$ in Algorithm 1.

4.1. *Multivariate Student's T-mixture Model.* Considering possible complexities in the structure of $\pi_t(\theta)$, we employ a mixture model to define $q_t(\theta)$. Note that any continuous probability density function can be approximated by a finite mixture model, provided that the mixture has a sufficient number of components along with correctly estimated parameters (Bishop, 2005; Zeevi and Meir, 1997). In particular, we equip $q_t(\theta)$ with the Student's T distribution as the mixture component, making it easier to detect modes that are far apart thanks to its fat tail structure. Let $\mathcal{S}_d(\mu, \Sigma, v)$ denote a d-variate Student's T distribution, where μ is the center, Σ the positive definite inner product matrix, and v is the degrees of freedom. We just fix v to be a constant, e.g., 5 in the following. So the mixture model we use is:

$$q(\theta|\psi) = \sum_{m=1}^{M} \alpha_m \mathcal{S}_d(\theta|\mu_m, \Sigma_m) \tag{19}$$

where $\psi = \{M, \alpha_1, \ldots, \alpha_M, \mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M\}$ is the mixture parameter. Here α_m is the probability mass of the mth component. It satisfies that $\alpha_m \geq 0, \sum_{m=1}^{M} \alpha_m = 1$. So the task of fitting a T-mixture model to a probability density translates into seeking suitable parameter values for ψ.

4.2. *Proposal Adaptation via IS plus EM.* Let's focus on iteration t in Algorithm 1. We first simulate an random sample, θ^n, from $q_{t-1}(\theta)$, with the importance weight calculated by

$$w^n = \frac{\frac{\pi_t(\theta^n)}{q_{t-1}(\theta^n)}}{\sum_{n=1}^{N} \frac{\pi_t(\theta^n)}{q_{t-1}(\theta^n)}}, \tag{20}$$

where N denotes the sample size. The resulting weighted particle set then can be seen as a particle approximation to the current target distribution $\pi_t(\theta)$, i.e.

$$\pi_t(\theta) \simeq \sum_{n=1}^{N} w^i \delta(\theta^i), \tag{21}$$

where $\delta(\cdot)$ is the delta mass function.

Here we adopt the Kullback-Leibler (KL) divergence to measure the distance between any two distributions. Then the aim is to obtain a proposal, $q_t(\theta)$, that minimizes the KL distance with respect to $\pi_t(\theta)$, which is shown to be

$$\mathcal{D}[\pi_t(\theta)||q_t(\theta|\psi)] = \mathbb{E}_{\pi_t(\theta)}\left[\log \frac{\pi_t(\theta)}{\sum_{m=1}^{M} \alpha_m \mathcal{S}_d(\theta|\mu_m, \Sigma_m)}\right]. \tag{22}$$

where $\mathbb{E}_f$ denotes expectation with respect to a function f. It's clear that minimizing (22) in (α, μ, Σ) is equivalent to maximizing

$$\int \pi_t(\theta) \log\left(\sum_{m=1}^{M} \alpha_m \mathcal{S}_d(\theta|\mu_m, \Sigma_m)\right). \tag{23}$$

Substituting Equation (21) in the above, we have

$$\sum_{i=1}^{N} w^i \log\left(\sum_{m=1}^{M} \alpha_m \mathcal{S}_d(\theta^i|\mu_m, \Sigma_m)\right). \tag{24}$$

At this stage, the task translates into a mixture-density parameter estimation problem, that is, how to optimize the parameter values of ψ to make the resulting mixture density function fit the data $\{\theta^n, w^n\}_{n=1}^N$ as well as possible.

Given data component θ^n, the posterior probability of each mixture component can be obtained according to the Bayes rule:

$$\rho_m(\theta^n) = \alpha_m \mathcal{S}_d(\theta^n; \mu_m, \Sigma_m)/q(\theta^n), \tag{25}$$

where α_m is regarded as the prior, and $\mathcal{S}_d(\theta^n; \mu_m, \Sigma_m)$ is the associated likelihood. The proportion mass of this component can then be updated by

$$\alpha_m = \sum_{n=1}^{N} w^n \rho_m(\theta^n). \tag{26}$$

And each pair of $\{\mu_m, \Sigma_m\}$ can be updated by the following Equations:

$$\mu_m = \frac{\sum_{n=1}^{N} w^n \rho_m(\theta^n) u_m(\theta^n) \theta^n}{\sum_{n=1}^{N} w^n \rho_m(\theta^n) u_m(\theta^n)}, \tag{27}$$

$$\Sigma_m = \frac{\sum_{n=1}^{N} w^n \rho_m(\theta^n) u_m(\theta^n) C_n}{\sum_{n=1}^{N} w^n \rho_m(\theta^n)}, \tag{28}$$

where

$$u_m(\theta) = \frac{v + d}{v + (\theta - \mu_m)^T (\Sigma_m)^{-1} (\theta - \mu_m)} \tag{29}$$

and $C_n = (\theta^n - \mu_m)(\theta^n - \mu_m)'$, where $'$ denotes matrix transposition.

The operations from (25) to (28) constitute one iteration of the EM algorithm, which is used to do mixture density parameter estimation (Peel and McLachlan, 2000). Let $q_t(\theta)$ be a T-mixture model, with the updated parameter values $\{\alpha_m, \mu_m, \Sigma_m\}$, and it's expected to be an approximation to $\pi_t(\theta)$. The efficiency of $q_t(\theta)$, as an approximation of $\pi_t(\theta)$, can be measured by the KL divergence using (24), or the effective sample size, which is defined to be

$$\text{ESS} = \frac{1}{\sum_{n=1}^{N} w_n^2}. \tag{30}$$

Note that these importance weights are associated with the random sample drawn from $q_t(\theta)$. The ESS was originally proposed by Kong, Liu and Wong (1994) to measure the overall efficiency of an IS algorithm. Heuristically it measures how many i.i.d samples are equivalent to the N weighted samples. The ESS is related to the coefficient of variation (CV) of the importance weights (see Kong, Liu and Wong (1994)), which finds applications in Ardia, Hoogerheide and van Dijk (2008); Cappé *et al.* (2008); Cornebise, Moulines and Olsson (2008). It's shown that ESS is determined by the associated proposal and the particle size, N, while, once N is big enough, the ratio of ESS to N should be stable. So in this paper, we use ESS $/N$ as a quantity to monitor the efficiency of $q_t(\theta)$. When ESS $/N$ is big enough, let Algorithm 1 go to next iteration. If ESS $/N$ is smaller than a given threshold, we continue to update $q_t(\theta)$ until it satisfies the requirement, i.e. the resulting ESS $/N$ is

big enough. In this way, we can make sure that the finally obtained $q_t(\theta)$ is really a good approximation to $\pi_t(\theta)$.

Suppose that currently ESS $/N$ is not big enough, we consider two approaches that are used to update $q_t(\theta)$ again. First we check whether the particle with the biggest importance weight, denoted $\theta^W, W \in \{1, \ldots, N\}$, is located in the tail of the proposal. If it is, which means $q_t(\theta^W)$ is smaller than the proposal densities of most other particles, we then select to perform an operation called component-splitting (see Section 4.3.2), otherwise, we perform the IS plus EM procedure again. The reason for doing this is that the IS plus EM procedure is exactly a local search approach, because it fixes the number of mixture components to be a constant and the EM itself is a local optimization method, while the component-splitting procedure will break the local model structure by splitting one mixture component into two, thus facilitates to find the global or, at least a better local optimum (see Section 4.3.2).

4.3. *Online Approach to Tune the Number of the Mixture Components.* In the above IS plus EM procedure, the number of components M is imposed to be constant, while this assumption is of course not reasonable. To deal with this issue, we propose a series of adaptive mechanisms to adapt M online, which include operations called components-merging, components-splitting and components-deletion, respectively. We'll provide respective conditions, on which each of these procedures needs to be performed.

4.3.1. *Components Merging.* This step targets for merging mixture components that have much mutual information among each other. To begin with, we first introduce the concept, mutual information. Suppose that a set of equally weighted data components $\{\theta^n\}_{n=1}^N$ are available, whose distribution is modeled by a mixture density function $q = \sum_{m=1}^M \alpha_m f_m$. Given any data component θ^n, the posterior probability of each mixture component f_m can be calculated by

$$\rho_m(\theta^n) = \alpha_m f_m(\theta^n)/q(\theta^n). \tag{31}$$

If there are two mixture components f_i, f_j, for which it satisfies that $\rho_i(\theta^n) = \rho_j(\theta^n)$ for any $n \in \{1, \ldots, N\}$, then f_i and f_j are regarded to contain completely overlapped information. Inspired by the components merging criterion proposed in Ueda *et al.* (2000); Wang *et al.* (2004), we define the mutual information between f_i and f_j to be:

$$\mathrm{MI}(i,j) = \frac{(\Upsilon_i - \bar{\Upsilon}_i)^T(\Upsilon_j - \bar{\Upsilon}_j)}{\|\Upsilon_i - \bar{\Upsilon}_i\| \cdot \|\Upsilon_j - \bar{\Upsilon}_j\|}, \tag{32}$$

where $\Upsilon_m = [\rho_m(\theta^1), \ldots, \rho_m(\theta^N)]'$, $'$ denotes transposition of a matrix, $\|\cdot\|$ denotes the Euclidean norm, and $\bar{\Upsilon}_m = \frac{1}{N}\sum_{n=1}^N \rho_m(\theta^n)\mathbf{1}_N$. Here $\mathbf{1}_n$ denotes the $n-$dimensional column vector with all elements being 1.

Note that $\mathsf{MI}(i,j) \in [-1,1]$, and $\mathsf{MI}(i,j) = 1$ iff f_i and f_j contain completely overlapped information.

In our algorithm framework, each data point θ^n is weighted by w^n. Accordingly, Υ_m should be $\sum_{n=1}^N w^n \rho_m(\theta^n)$, and the mutual information between i, j turns out to be:

$$\mathsf{MI}(i,j) = \frac{(\Upsilon_i - \bar{\Upsilon}_i)' D(w) (\Upsilon_j - \bar{\Upsilon}_j)}{\sqrt{\sum_{n=1}^N w^n (\Upsilon_i(\theta^n) - \bar{\Upsilon}_i)^2}\sqrt{\sum_{n=1}^N w^n (\Upsilon_j(\theta^n) - \bar{\Upsilon}_j)^2}} \tag{33}$$

where

$$D(w) = diag([w^1, \ldots, w^N]) \tag{34}$$

We judge if two components, i and j, need to be merged into one by comparing $\mathsf{MI}(i,j)$ with a threshold $0 < T_{merge} < 1$. If $\mathsf{MI}(j,k) > T_{merge}$, we then merge the ith component into j, and set $M = M - 1$; otherwise, we maintain the original mixture. The merging operation is specified to be:

$$\alpha_j = \alpha_i + \alpha_j \tag{35}$$

$$\mu_j = \frac{\alpha_i \mu_i + \alpha_j \mu_j}{\alpha_j} \tag{36}$$

$$\Sigma_j = \frac{\Sigma_i \mu_i + \Sigma_j \mu_j}{\alpha_j} \tag{37}$$

which is a linear combination of the two old components, and has been used in Wang *et al.* (2004). We traverse each pair of the original components to do the merging judgements and perform the above merging operation if the merging condition satisfies.

We execute the above merging operation when superfluous components may exist, e.g., when $q_0(\theta)$ is initialized with too many overlapped components, or too many new components appear at one iteration of Algorithm 1 via component-splitting (see Section 4.3.2).

4.3.2. *Component-splitting.* As mentioned in Section 4.2, it may happen that, after performing the IS plus EM procedure, ESS $/N$ was not big enough, and the maximum weight particle was located in the tail of $q_t(\theta)$. In that case, we perform the mechanism proposed here, called component-splitting. This procedure is used to break the local structure of $q_t(\theta)$ by splitting one component into two, whereby the component number, M, will increase to $M + 1$.

This Component-splitting approach is performed as follows. First, when we simulate draws from $q_t(\theta)$ for evaluating the efficiency of $q_t(\theta)$, we record the component origination of each draw. We then see the parent component of the maximum weight sample, θ^W, denoted $f_p = \mathcal{S}_d(\theta|\mu_p, \Sigma_p)$, and know that which samples were originated from f_p. We encapsulate the samples that come from f_p to a data array, $\theta_{local} = \{\theta^1, \ldots, \theta^{pN}\}$, then split f_p into two components– $f_{c1} = \mathcal{S}_d(\theta|\mu_{c1}, \Sigma_{c1})$ and $f_{c2} = \mathcal{S}_d(\theta|\mu_{c2}, \Sigma_{c2})$, and finally replace f_p in $q_t(\theta)$ by f_{c1} and f_{c2}.

Next we give operations to get $\mu_{c1}, \Sigma_{c1}, \mu_{c2}, \Sigma_{c2}$. We first initialize $\mu_{c1} = \theta^W$, $\mu_{c2} = \mu_p$, $\Sigma_{c1} = \Sigma_{c2} = \Sigma_p$, and then specify a local mixture $q_{local} = \alpha_{c1} f_{c1} + \alpha_{c2} f_{c2}$, where $\alpha_{c1} = \alpha_{c1} = 0.5$, finally we perform a local IS plus EM procedure to update $\alpha_{c1}, \alpha_{c2}, \mu_{c1}, \Sigma_{c1}, \mu_{c2}, \Sigma_{c2}$. Such local IS plus EM procedure starts by evaluating the local weight for each θ^i in θ_{local} via

$$w^i = \frac{\frac{\pi_t(\theta^i)}{q_{local}(\theta^i)}}{\sum_{n=1}^{pN} \frac{\pi_t(\theta^n)}{q_{local}(\theta^n)}}. \tag{38}$$

Substituting $\{\theta^i, w^i\}_{i=1}^{pN}$ and the local mixture parameters, $\alpha_{c1}, \alpha_{c2}, \mu_{c1}, \Sigma_{c1}, \mu_{c2}, \Sigma_{c2}$ in the Equations from (26) to (28), we then can obtain the updated parameter values for the local mixture q_{local}. Suppose that f_p is assigned a proportion mass α_p in $q_t(\theta)$, we assign f_{c1} and f_{c2} proportion masses $\alpha_{c1} = \alpha_p \alpha_{c1}$ and $\alpha_p \alpha_{c2}$, respectively, for the updated $q_t(\theta)$, in which f_p is replaced by f_{c1} and f_{c2}.

This component-splitting method makes use of the local behavior of the target distribution around the maximum weight sample. It adds mixture proposal probability mass to those areas of the parameter space which is near the maximum weight sample, where there is relatively little mixture proposal probability mass. As the local IS plus EM procedure is actually a local mixture learning process, by which the local structure around the maximum weight sample will be modeled well by the updated local mixture, which then will elicit a better global mixture representation of the whole target distribution, which may lead to a satisfactory ESS $/N$ value. Note that the above component-splitting procedure may be repeated provided that the condition of executing it satisfies.

4.3.3. *Component-Deletion.* As mentioned above, when simulating draws from $q_t(\theta)$, we can record the component origination of each sample. For any mixture component that generates no sample, we deem it has taken no effect in modeling the current target distribution, thus we just delete it in current mixture proposal. Its probability masses is then redistributed strengthening the remaining components.

4.4. *Practical Issues to be Considered.* EM based mixture modeling approach provides a convenient and flexible utility for estimating or approximating distributions, while the mixture log-likelihood presents a well known problem, i.e., degeneracy toward infinity (Biernacki and Chrétien, 2003) (actually EM also suffers from the problem of slow convergence, so it may converge to a local optimum (Biernacki and Chrétien, 2003), however, we don't need to worry about that, because on one side the proposed component-splitting procedure facilitates to search the global optimum, and, one the other side, the objective here is just to find a good mixture importance function, which is not necessary to be the global optimum). To prevent the mixture log-likelihood from arising to infinity, we assign an inverse-wishart prior on the covariance matrix of each mixture component, then perform EM to get the *maximum a posterior* (instead of maximum likelihood) estimate of the covariance matrix.

In the component-splitting step, the parameter updating of f_{c1} and f_{c2} is based upon pN particles originated from f_p. In case of pN being too small, e.g. smaller than a threshold specified beforehand, denoted N_s, we need to generate another set of $N_s - pN$ samples from f_p, then do the parameter updating for f_{local} based upon the N_s particles. By doing this, we can guarantee that we are using a sufficient number of random draws to capture the local structure in the posterior, that will then be represented by f_{local}.

In the above mentioned component-splitting step, when we replace f_p in $q_t(\theta)$ by f_{c1} and f_{c2}, we assign the proportion mass of f_p to f_{c1} and f_{c2} proportionally to their original weights, i.e. α_{c1} and α_{c2}, respectively. Note that if α_p is a very small value, e.g., 0.0001, the new added mixture components, f_{c1} and f_{c2}, will only take an insignificant effect in the updated $q_t(\theta)$, while f_{c1} and f_{c2} may contain new and important information about the posterior. To guarantee that f_{c1} and f_{c2} can play a role in $q_t(\theta)$, at lease to some extent, we propose the following operations. We first specify a threshold $\alpha_{threshold}$, which may be, e.g., 0.1 or 0.2. Then, if $\alpha_p > \alpha_{threshold}$, we assign weights to f_{c1} and f_{c2} as usual as shown in Section 4.3.2. Otherwise, we assign $\alpha_{threshold}\alpha_{c1}$ and $\alpha_{threshold}\alpha_{c2}$ to f_{c1} and f_{c2}, respectively. So the sum of weights of f_{c1} and f_{c2} in $q_t(\theta)$ will be $\alpha_{threshold}$. To make sure the sum of the weights of the mixture components equals to 1, we reduce the weights of the components, other than f_{c1} and f_{c2}, proportionally to their original weights.

4.5. *Relationships to Existing Methods.* As the proposed AAIS method utilizes the idea of 'annealing' as a way of searching isolated modes, it has close relationships to a series of MCMC methods, which are based on the

"thermodynamic integration" technique for computing marginal likelihoods, and are discussed under the name of Bridge and Path sampling in Gelman and Meng (1998), parallel tempering MCMC in Gregory and Fischer (2010); Gregory (2005) and Annealed Importance Sampling (AIS) in Neal (2001). The central idea of these methods is to divide the ratio of two normalization constants, which we can think of for our purposes as the ratio of marginal likelihoods, Z_t and Z_0, which are associated to $\pi(\theta)$ and $q_0(\theta)$, respectively, into a series of easier ones, parameterised by inverse temperature, λ. In detail:

$$\frac{\mathsf{Z}_t}{\mathsf{Z}_0} = \frac{\mathsf{Z}_1}{\mathsf{Z}_0}\frac{\mathsf{Z}_2}{\mathsf{Z}_1}\cdots\frac{\mathsf{Z}_t}{\mathsf{Z}_{t-1}}. \tag{39}$$

Each of the ratios is estimated using samples from a MCMC method. For example, to compute $\frac{\mathsf{Z}_t}{\mathsf{Z}_{t-1}}$, we should have samples that are drawn from equilibrium from the distribution defined by $\pi_{t-1}(\theta) \propto \pi(\theta)^{\lambda_{t-1}} q_0(\theta)^{1-\lambda_{t-1}}$. Therefore, these methods require us to devise a single Metropolis proposal at each temperature, which should be a troublesome issue in practice, since it is not prior known the target distribution's structure at each temperature. In addition, the involvement of MCMC results in annoying questions, e.g. how long should the 'burn-in' period last, and how many samples are needed to get a reliable estimate, which should be considered at each temperature.

Striking differently, the proposed AAIS is developed from the perspective of mixture modeling, that is, how to adapt a mixture model to resemble the posterior, while the similar temperature idea, i.e., 'annealing', is also involved target for searching isolated peaky modes. We only need to select an initial mixture proposal q_0, the annealing temperatures λ_t, and up to three thresholds, after which, the mixture adaptation process is completely adaptive, and finally the marginal likelihood is simply obtained by substituting the resulting mixture as the importance function in (12). Such property of easy-to-implementation results from the adaptive techniques involved in the algorithm, including the EM based parameter estimation for each mixture component, and the proposed online approach used for adjusting the cardinality of the mixture, i.e. the number of mixture components.

This AAIS algorithm falls within a general algorithmic framework, called Sequential Monte Carlo Sampler (SMCS), which is proposed in Del Moral, Doucet and Jasra (2007, 2006). However, SMCS utilizes the same "thermodynamic integration" technique for calculating marginal likelihoods, so it inherits all drawbacks mentioned above. Exactly, the proposed AAIS method can be seen as a special case of SMCS, which adopts adaptive independent mixture proposals as the transition kernels (see details in Del Moral, Doucet

and Jasra (2007, 2006)), and the IS, instead of MCMC, for particle generation.

Finally, AAIS has connections to a bunch of existing adaptive IS methods in the literature, for example, Evans (1991); Oh and Berger (1993); Cappé *et al.* (2008); Cappe *et al.* (2004); West (1993); Ardia, Hoogerheide and van Dijk (2008), to name just a few. The most distinguishing point lies in that our method adopts the annealing strategy, which is well recognized as a powerful approach to search separated, especially peaky, modes in the posterior. From the mixture modeling point of view, the proposed mixture adaptation technique can be seen as an intermediate between the completely non-parametric method in West (1993) and the EM based semi-parametric methods, which is used in the adaptive mixture importance sampling (AMIS) approach (Cappé *et al.*, 2008). The AMIS approach employs a fixed number of mixture components, while our method and the nonparametric method (West, 1993) can let the data speak for themselves how many components should be contained in the mixture.

5. Simulation Studies. In this Section, we present two simulations, one in three dimensions and the other in seven, in which the target functions were built to be difficult to integrate, but their shapes can be inferred through their mathematical expressions. The objective of the simulation study is to evaluate the efficiency of the importance function yielded by AAIS, via comparing its shape with that of the real target, and comparing the calculated marginal likelihood with the real answer.

5.1. *Three Dimensional Flared Helix.* We used a cylindrical function with standard Gaussian cross-section (in $(x,\ y)$) that has been deformed (in z) into a helix with varying diameter. The helix performs three complete turns. The target function is defined to be

$$p(x,y,z) = \delta(-30 < z \leq 30)\mathcal{N}([x,y]|\mu,\Sigma) \tag{40}$$

where $\mu(1) = (z+35)\times\cos(\beta)$ and $\mu(2) = (z+35)\times\sin(\beta)$, where $\beta = (z+30)\times\pi/10$. Here $\Sigma = \text{diag}([1,1])$. Fig.1 shows the shape of this function. And the integration of this function is shown to be

$$\int\int\int p(x,y,z)dxdydz = \int\int\int p(x,y|z)dxdyp(z)dz = \int_{-30}^{30} 1dz = 60 \tag{41}$$

Treating (40) to be $L(y|\theta)p(\theta)$, the marginal likelihood in this simulation case is then 60.

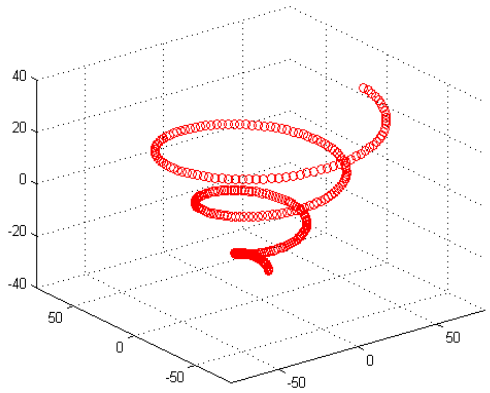

FIG 1. *The simulated 3-D flared helix target distribution*

Before running the proposed Adaptive Annealed IS (AAIS) algorithm, we select the initial proposal to be a T-mixture that is composed of 10 equally weighted components. The degree of freedom for each T component is fixed to be 5. The centers of these components are selected uniformly from a region restricted by $x \in [-100, 100], y \in [-100, 100]$ and $z \in [-30, 30]$. Then the covariance matrix for every component is identical to $\text{diag}[\sigma_x^2, \sigma_y^2, \sigma_z^2]$, where σ_a is the standard error of the argument a in the components' centers that has been just specified. The particle size $N = 2000$, $\lambda_t = 0.1t$, and $t = 1, \ldots, 10$.

The adaptive mixture importance sampling (AMIS) approach Cappé *et al.* (2008) is involved for performance comparison. We initialize it using the same setting as for the AAIS, and let it run 10 iterations to give the final mixture importance function.

Fig. 2 shows the resulting posterior samples (which are equally weighted via resampling) corresponding to AAIS and AMIS, respectively. It's shown that the samples resulted from AAIS are distributed in wider posterior area, which indicates that the importance function given by AAIS captures more structures in the posterior.

A quantitative comparison between the importance functions obtained by AAIS and AMIS is shown in Table 2. Three quantities are used for comparison, that are the ESS , the KL distance and the estimated marginal likelihood. Note that, we can easily calculate the KL distance via a simple Monte Carlo for this simulation case. As is shown in Table 2, AAIS gives correct answer to the marginal likelihood, closer KL distance with respect to the real

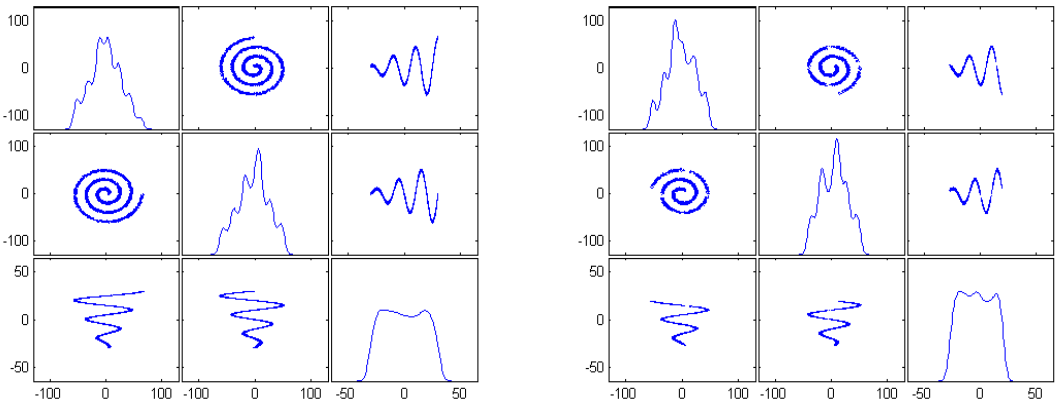

FIG 2. *Three dimensional flared helix example. Left:posterior samples produced by the proposed AAIS algorithm. Right:posterior samples produced by the AMIS algorithm (Cappé et al., 2008). In the diagonal, the curves are kernel density estimates of the posterior samples.*

	ESS /N	KL	Marginal Likelihood (real answer:60)
AMIS (Cappé *et al.*, 2008)	0.1525	6.7830	47.2 ± 3.3
AAIS proposed here	0.4459	0.1586	59.7 ± 2.0

TABLE 2

Three dimensional flared helix example. Quantitative Performance comparison.

target, and bigger ESS . So it further demonstrates that the AAIS is much better than the AMIS in dealing with this simulation case.

5.2. *Outer Product of Seven Univariate Densities.* To test the concern whether good results from a 3-D example imply good results in higher dimensions, we used as a target function the outer product of seven univariate distributions normalized to integrate to 1. These seven distributions are:

1. $\frac{3}{5}\mathcal{G}(10+x|2,3)+\frac{2}{5}\mathcal{G}(10-x|2,5)$
2. $\frac{3}{4}sk\mathcal{N}(x|3,1,5)+\frac{1}{4}sk\mathcal{N}(x|-3,3,-6)$
3. $\mathcal{S}(x|0,9,4)$
4. $\frac{1}{2}\mathcal{B}(x+3|3,3)+\frac{1}{2}\mathcal{N}(x|0,1)$
5. $\frac{1}{2}\varepsilon(x|1)+\frac{1}{2}\varepsilon(-x|1)$
6. $sk\mathcal{N}(x|0,8,-3)$
7. $\frac{1}{8}\mathcal{N}(x|-10,0.1)+\frac{1}{4}\mathcal{N}(x|0,0.15)+\frac{5}{8}\mathcal{N}(x|7,0.2)$

Here $\mathcal{G}(\cdot|\alpha,\beta)$ denotes the gamma distribution, $\mathcal{N}(\cdot|\mu,\sigma)$ denotes the normal distribution, $sk\mathcal{N}(\cdot|\mu,\sigma,\alpha)$ denotes the skew-normal distribution, $\mathcal{S}(\cdot|\mu,s,df)$ denotes the student-T distribution, $\mathcal{B}(\cdot|\alpha,\beta)$ denotes the beta distribution,

	ESS $/N$	KL	Marginal likelihood: (true value:1)
AMIS (Cappé *et al.*, 2008)	0.2454	10.8804	0.4675 ± 0.0246
AAIS proposed here	0.4948	0.4075	1.0011 ± 0.0303

TABLE 3

Seven dimensional outer product example. Quantitative Performance comparison

and $\varepsilon(\cdot|\lambda)$ denotes the exponential distribution. Dimension 2 has two modes bracketing a deep ravine, dimension 4 has one low, broad mode that overlaps a second sharper mode, and dimension 7 has three distinct, well-separated modes. Only dimension 5 is symmetric. There is a range of tail behavior as well, from Gaussian to heavy-tailed.

In this case, we initialize the proposed AAIS algorithm with a T-mixture that is composed of 50 equally weighted components. The degree of freedom for each T component is still fixed to be 5. The centers of these components for each dimension are selected uniformly $[-10, 10]$. The procedure to initialize the covariance matrix for every mixture component is identical to that used for the example shown in Section 5.1. We specify a particle size $N = 8000$, and annealing schedule $\lambda_t = 0.1t$, with $t = 1, \ldots, 10$.

The AMIS method (Cappé *et al.*, 2008) is also involved for performance comparison. The AMIS is initialized identically as for AAIS, and will run 10 iterations to give the final result.

Fig.3 depicts the scatterplot for the resulting posterior samples (which has been equally weighted by resampling). It's shown that, the AMIS missed modes in the 2nd and 7th dimension, while the proposed AAIS manages to capture all the modes in the posterior, which should be benefited from the annealing idea being involved.

Similarly as for example 1, we give a quantitative comparison for the involved algorithms in Table 3. We see that again the proposed AAIS algorithm leads to much better results for every comparison criterion.

6. Exo-Planet Examples. In this section, we perform the proposed AAIS method to deal with two real RV data set.

6.1. *Star HD73526.* This data set contains 18 data components, and were claimed to support an orbit of 190.5 ± 3.0 days (Tinney *et al.*, 2003). Gregory (2005) did a Bayesian re-analysis on this data set using a parallel tempering MCMC algorithm, and reported three possible orbits, with periods $127.88^{+0.37}_{-0.09}$, $190.4^{+1.8}_{-2.1}$ and $376.2^{+1.4}_{-4.3}$ days, respectively. Gregory (2005) also discussed the possibility of having an additional planet, and reported that the Bayes factor is less than 1 when comparing $\mathcal{M}_2$ with the $\mathcal{M}_1$.

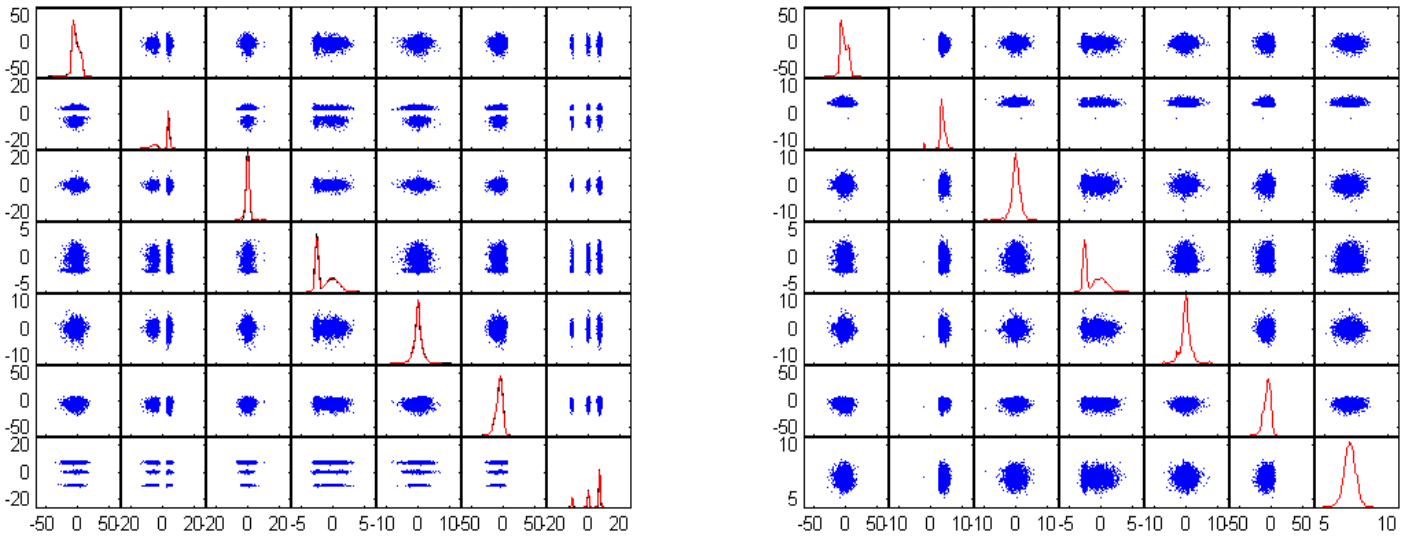

FIG 3. *Seven dimensional outer product example. Left:posterior samples produced by the proposed AAIS algorithm. Right:posterior samples produced by the AMIS algorithm (Cappé et al., 2008). In the diagonal, the curves are kernel density estimates of the posterior samples.*

	Marginal Likelihood	ESS $/N$
$\mathcal{M}_0$	$5.9013 \times 10^{-50} \pm 5.1325 \times 10^{-52}$	0.9320
$\mathcal{M}_1$	$4.4886 \times 10^{-41} \pm 3.2093 \times 10^{-42}$	0.5698
$\mathcal{M}_2$	$1.5511 \times 10^{-42} \pm 3.2878 \times 10^{-43}$	0.3458

TABLE 4

HD73526 Tinney et al. (2003) Data Case. The calculated marginal likelihoods

We use the proposed AAIS algorithm to deal with $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The marginal likelihood estimation result is summarized in Table 4. These estimates are reliable indicated by their corresponding ESS $/N$. The resulting Bayes factors are shown in Table 5. As indicated by the result, this data set support the one-planet hypothesis, which coincides the conclusions made by Tinney *et al.* (2003); Gregory (2005).

Specifically for $\mathcal{M}_1$, we show the posterior samples (equally weighted by resampling) in Fig. 4. We obtain two modes in P, that are $P_1 = 190.1^{+2.2}_{-1.5}$ and $P_2 = 375.5^{+2.0}_{-2.5}$, respectively.

6.2. *Star HD73526.* This data set contains 30 RV data components and was claimed to have two planets (Tinney *et al.*, 2006).

$\mathsf{BF}\{\mathcal{M}_1 : \mathcal{M}_0\}$	$\mathsf{BF}\{\mathcal{M}_2 : \mathcal{M}_1\}$
7.606×10^8	0.03456

TABLE 5

HD73526 (Tinney et al., 2003) Data Case. The calculated Bayes Factors

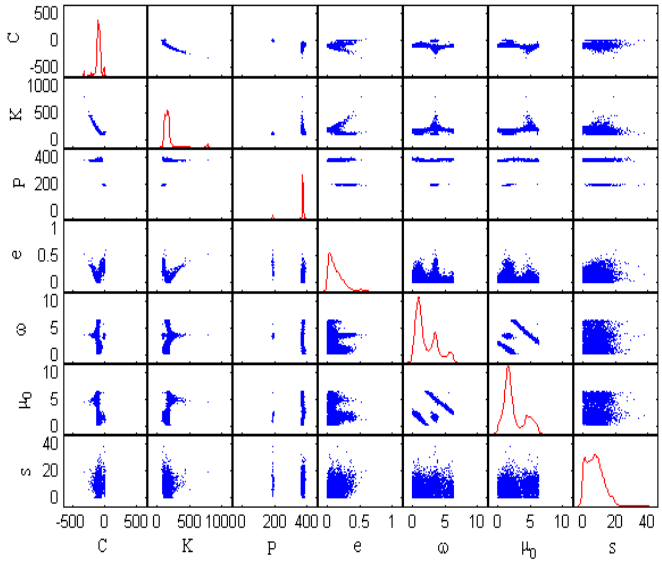

FIG 4. *HD73526 Tinney et al. (2003) Data Case. Scatter plot of the posterior samples of* $\mathcal{M}_1$

We calculate the marginal likelihoods using the proposed AAIS algorithm, and the result is shown in Table 6. Indicated by the criterion ESS $/N$, we deem that the result is reliable. The resulting Bayes factors can be seen in Table 7. So our result supports the argument in Tinney *et al.* (2006)—there are two planets underlying this data set.

The posterior samples (equally weighted by resampling) corresponding to $\mathcal{M}_1$ is shown in Fig.5, and the algorithm again captures two modes in P, that are $P_1 = 193.1162^{+2.1}_{-3.7}$ and $P_2 = 374.8732^{+6.9}_{-5.8}$.

When dealing with $\mathcal{M}_2$, we still restrict the period of the first planet to be

	Marginal Likelihood	ESS $/N$
$\mathcal{M}_0$	$8.9566 \times 10^{-77} \pm 1.0852 \times 10^{-78}$	0.9510
$\mathcal{M}_1$	$5.8519 \times 10^{-70} \pm 1.7077 \times 10^{-71}$	0.6545
$\mathcal{M}_2$	$4.8122 \times 10^{-65} \pm 1.4284 \times 10^{-66}$	0.2034

TABLE 6

HD73526 (Tinney et al., 2006) Data Case. The calculated marginal likelihoods

$\mathrm{BF}\{\mathcal{M}_1 : \mathcal{M}_0\}$	$\mathrm{BF}\{\mathcal{M}_2 : \mathcal{M}_1\}$
6.534×10^6	8.233×10^4

TABLE 7
HD73526 (Tinney et al., 2006) Data Case. The calculated Bayes Factors

smaller than that of the second planet. Fig.6 shows the posterior samples of $\mathcal{M}_2$. Given these posterior samples, we get the periods of the two planets, that are $P_1 = 187.9379^{+2.1}_{-0.8}$ and $P_2 = 377.3030^{+5.2}_{-4.5}$, respectively.

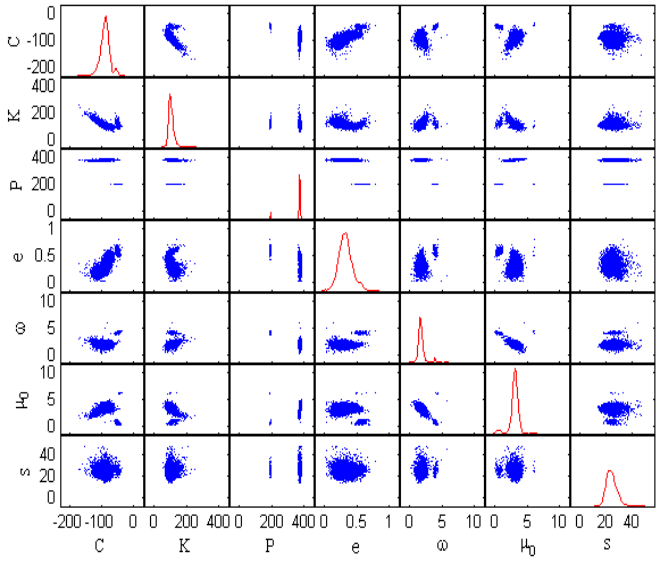

FIG 5. *HD73526 (Tinney et al., 2006) Data Case. Scatter plot of the posterior samples of* $\mathcal{M}_1$

Given samples from the posterior, we can easily get the minimum mean squared estimate (MMSE) of the RV at any given time, and then plot the RV curve. In Fig.7, we plot the estimated RV curves on basis of $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, meanwhile the real RV observations are depicted there for reference. As in shown, the RV estimation yielded by $\mathcal{M}_2$ fits the real data best.

6.3. *The HD217107 Data Case.* Star HD217107 was claimed to have two planets (Vogt *et al.*, 2005).

We analyze the observed RV data for HD217107 using our AAIS method, the calculated marginal likelihoods are shown in Table 8, each corresponding to a sufficiently big value of ESS $/N$. The resulting Bayes factors can be

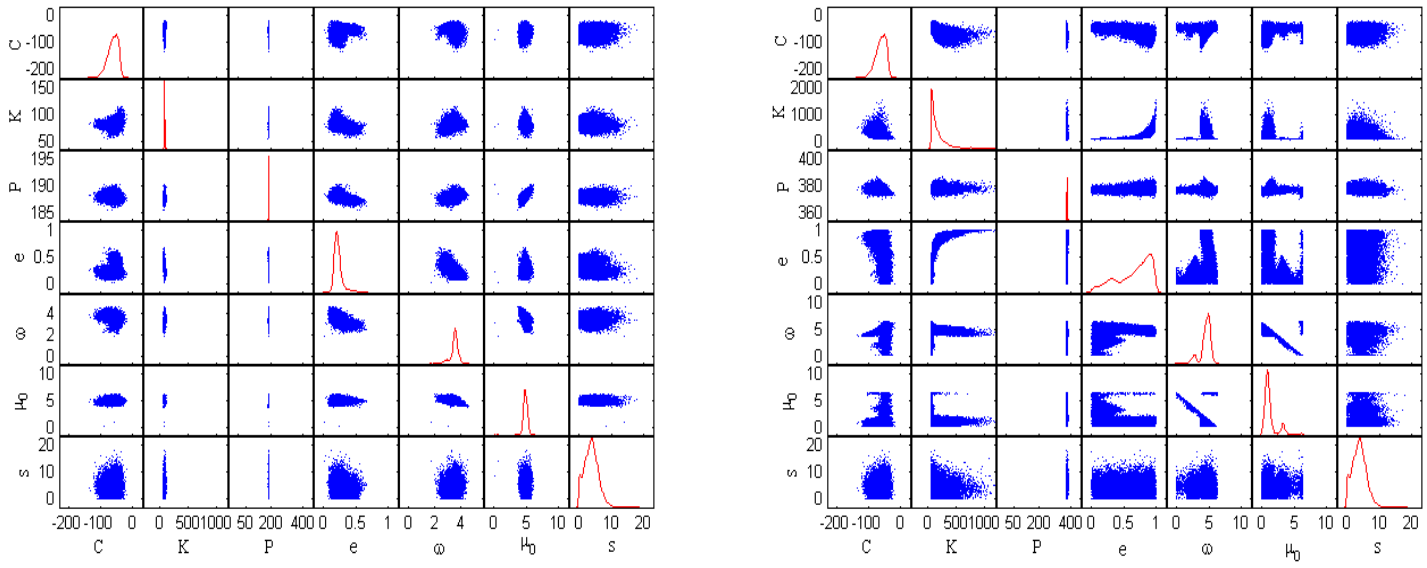

FIG 6. *HD73526 Tinney et al. (2006) Data Case. Scatter plot of the posterior samples of $\mathcal{M}_2$. Left: Scatter plot of the posterior samples of the first planet parameter; Right: Scatter plot of the posterior samples of the second planet parameter*

	Marginal Likelihood	ESS $/N$
$\mathcal{M}_0$	$3.6492 \times 10^{-168} \pm 7.0150 \times 10^{-169}$	0.9644
$\mathcal{M}_1$	$3.7389 \times 10^{-141} \pm 0.6077 \times 10^{-142}$	0.6448
$\mathcal{M}_2$	$3.0897 \times 10^{-108} \pm 1.0011 \times 10^{-109}$	0.4119

TABLE 8

HD217107 (Vogt et al., 2005) Data Case. The calculated marginal likelihoods

seen in Table 9. It's shown that $\mathcal{M}_2$ beats both $\mathcal{M}_0$ and $\mathcal{M}_1$.

6.4. *The HD37124 Data Case.* Star HD37124 was claimed to have three planets (Vogt *et al.*, 2005).

We analyze the associated RV data set using our AAIS method, the marginal likelihoods calculated by our AAIS algorithm are shown in Table 10. Indicated by the criterion ESS $/N$, the results are reliable. The resulting Bayes factors can be seen in Table 11, which support the conclusion of Tinney *et al.* (2006), i.e. this data set supports the three-planet hypothesis.

In Fig.8, we plot the estimated RV curves on basis of $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, respectively, where the real RV observations is also depicted for reference. As is shown, the RV estimation yielded by $\mathcal{M}_3$ fits the real data best.

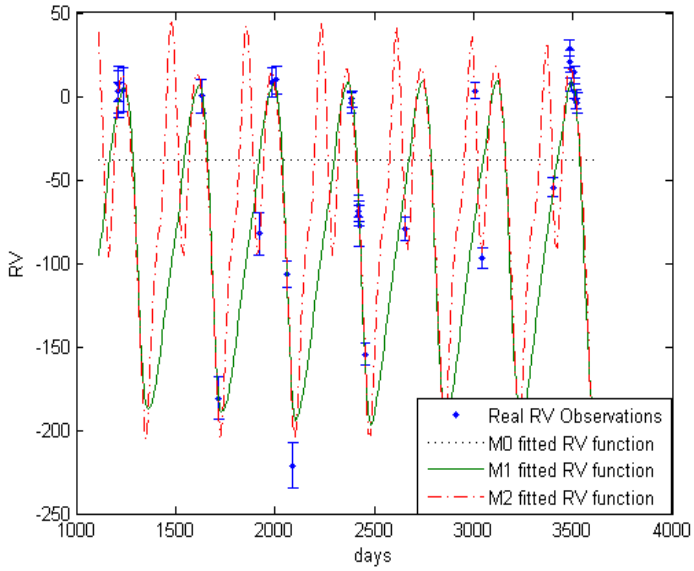

FIG 7. *Measured velocities (filled circles) vs. time for HD73526 Tinney et al. (2006). Error bars only includes internal uncertainties. The curves are MMSE estimates of the RV yielded by $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively*

$\mathrm{BF}\{\mathcal{M}_1 : \mathcal{M}_0\}$	$\mathrm{BF}\{\mathcal{M}_2 : \mathcal{M}_1\}$
1.025×10^{27}	8.264×10^{32}

TABLE 9
HD217107 (Vogt et al., 2005) Data Case. The calculated Bayes Factors

6.5. *The 47 Ursae Majoris (47 UMa) Data Case.* This 47 UMa data recently was analyzed in Gregory and Fischer (2010).

We analyze the associated RV data set using our AAIS method, the marginal likelihoods calculated by our AAIS algorithm are shown in Table 12. We obtain reliable estimates for $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, while the estimate for $\mathcal{M}_3$ is unreliable, the ESS $/N$ is 0.0089, which is not big enough. Indicated by the criterion ESS $/N$, the results are reliable. So we argue that this data set supports $\mathcal{M}_2$ more than $\mathcal{M}_0$ and $\mathcal{M}_1$, but we are not sure about the relative strength between $\mathcal{M}_2$ and $\mathcal{M}_3$. Gregory and Fischer (2010) concluded that this data set has three planets, however, gave an much ambiguous estimate for the third planet's period.

In Fig.9, we plot the estimated RV curves on basis of $\mathcal{M}_1$, $\mathcal{M}_2$ respectively, and again the real RV observations are depicted there for reference. As is shown, the RV estimation yielded by $\mathcal{M}_2$ fits better.

	Marginal Likelihood	ESS $/N$
$\mathcal{M}_0$	$2.0889 \times 10^{-110} \pm 2.1857 \times 10^{-112}$	0.9427
$\mathcal{M}_1$	$1.0717 \times 10^{-106} \pm 1.3077 \times 10^{-107}$	0.1737
$\mathcal{M}_2$	$1.9830 \times 10^{-97} \pm 1.1451 \times 10^{-98}$	0.1798
$\mathcal{M}_3$	$2.9228 \times 10^{-84} \pm 1.2563 \times 10^{-85}$	0.1305

TABLE 10

HD37124 (Vogt et al., 2005) Data Case. The calculated marginal likelihoods

BF$\{\mathcal{M}_1 : \mathcal{M}_0\}$	BF$\{\mathcal{M}_2 : \mathcal{M}_1\}$	BF$\{\mathcal{M}_3 : \mathcal{M}_2\}$
5.13×10^3	1.850×10^9	1.474×10^{13}

TABLE 11

HD37124 (Vogt et al., 2005) Data Case. The calculated Bayes Factors

	Marginal Likelihood	ESS $/N$
$\mathcal{M}_0$	$2.0198 \times 10^{-1004} \pm 9.2572 \times 10^{-1006}$	0.1002
$\mathcal{M}_1$	$3.4400 \times 10^{-896} \pm 3.10 \times 10^{-897}$	0.5643
$\mathcal{M}_2$	$1.3500 \times 10^{-816} \pm 1.77 \times 10^{-817}$	0.3324
$\mathcal{M}_3$	$2.8970 \times 10^{-825} \pm 9.1623 \times 10^{-825}$	0.0089

TABLE 12

uma47 (Gregory and Fischer, 2010) Data Case. The calculated marginal likelihoods

BF$\{\mathcal{M}_1 : \mathcal{M}_0\}$	BF$\{\mathcal{M}_2 : \mathcal{M}_1\}$	BF$\{\mathcal{M}_3 : \mathcal{M}_2\}$
1.703×10^{108}	3.924×10^{79}	?

TABLE 13

uma47 (Gregory and Fischer, 2010) Data Case. The calculated Bayes Factors

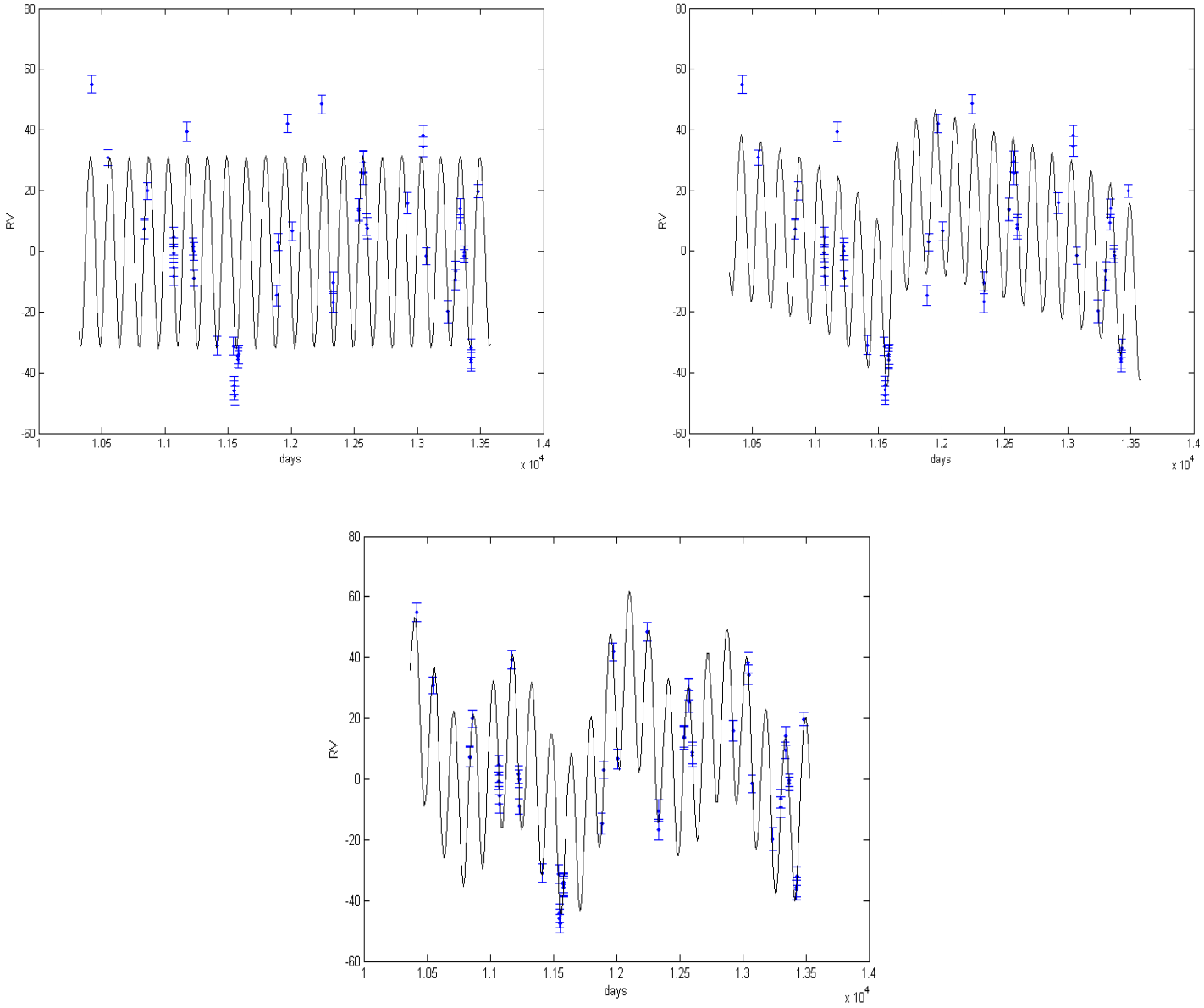

FIG 8. *Measured velocities (filled circles) vs. time for HD37124 Tinney et al. (2006). Error bars only includes internal uncertainties. The MMSE estimates of the RV curves yielded by using $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are shown in the top left, top right and the bottom panels, respectively.*

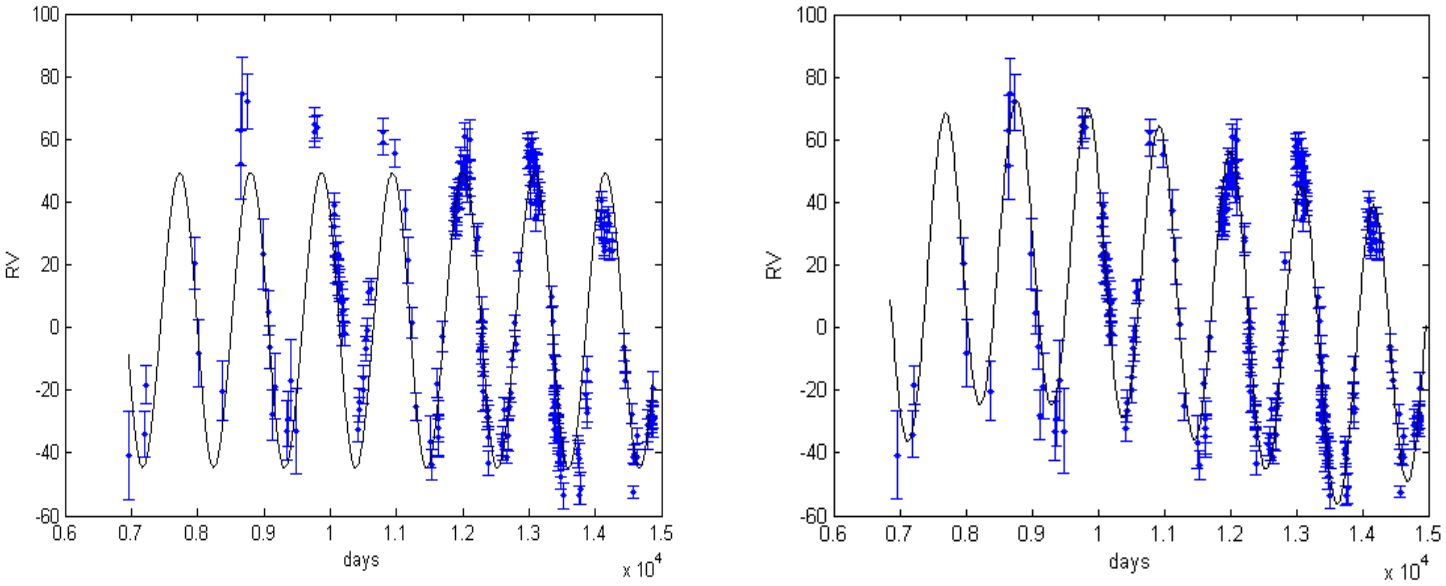

FIG 9. *Measured velocities (filled circles) vs. time for the 47 Ursae Majoris (47 UMa) (Gregory and Fischer, 2010). Error bars only includes internal uncertainties. The MMSE estimates of the RV curves yielded by using $\mathcal{M}_1$ and $\mathcal{M}_2$ are shown in the left and the right panels, respectively.*

7. Discussions and Conclusions. In this paper, we propose an adaptive mixture modeling approach to construct importance function for IS. The resulting mixture model is demonstrated to be capable of capturing separated peaky structures in the posterior, even when it is high-dimensional (an at most 17-dimensional posterior is considered in this paper). Straightforwardly, this approach facilitates simulating draws from multi-modal joint posterior distribution by IS, and provides an effective and easy-to-implement way for estimating marginal likelihood that is required for Bayesian model comparison.

References.

ARDIA, D., HOOGERHEIDE, L. F. and VAN DIJK, H. K. (2008). Adaptive Mixture of Student-t Distributions as a Flexible Candidate Distribution for Efficient Simulation. *Tinbergen Institute Discussion Papers*.

BIERNACKI, C. and CHRÉTIEN, S. (2003). Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM. *Statistics & Probability Letters* **61** 373–382.

BISHOP, C. M. (2005). *Neural networks for pattern recognition*. Oxford Univ Pr.

BULLARD, F. (2009). Exoplanet Detection: a Comparison of Three Statistics or How Long Should It Take to Find a Small Planet? PhD thesis, Duke University.

CAPPE, O., GUILLIN, A., MARIN, J. M. and ROBERT, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics* **13** 907–929.

CAPPÉ, O., DOUC, R., GUILLIN, A., MARIN, J. M. and ROBERT, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing* **18** 447–459.

CORNEBISE, J., MOULINES, E. and OLSSON, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing* **18** 461–480.

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)* **68** 411–436.

DEL MORAL, P., DOUCET, A. and JASRA, A. (2007). Sequential monte carlo for bayesian computation. *Bayesian Statistics* **8** 115–148.

EVANS, M. (1991). Chaining via annealing. *The Annals of Statistics* **19** 382–393.

FORD, E. B. and GREGORY, P. C. (2006). Bayesian Model Selection and Extrasolar Planet Detection. *Arxiv preprint astro-ph/0608328*.

GELMAN, A. and MENG, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* 163–185.

GREGORY, P. C. (2005). A Bayesian Analysis of Extrasolar Planet Data for HD 73526. *The Astrophysical Journal* **631** 1198–1214.

GREGORY, P. C. and FISCHER, D. A. (2010). A Bayesian periodogram finds evidence for three planets in 47 Ursae Majoris. *Monthly Notices of the Royal Astronomical Society* **9999**.

KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **89** 278–288.

LEFEBVRE, G., STEELE, R., VANDAL, A. C., NARAYANAN, S. and ARNOLD, D. L. (2009). Path Sampling to Compute Integrated Likelihoods: An Adaptive Approach. *Journal of Computational and Graphical Statistics* **18** 415–437.

NEAL, R. M. (2001). Annealing importance sampling. *Statistics and Computing* **11** 125–139.

OH, M. S. and BERGER, J. O. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association* 450–456.

PEEL, D. and MCLACHLAN, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* **10** 339–348.

TINNEY, C., BUTLER, R. P., MARCY, G. W., JONES, H. R. A., PENNY, A. J., MCCARTHY, C., CARTER, B. D. and BOND, J. (2003). Four New Planets Orbiting Metal-enriched Stars. *The Astrophysical Journal* **587** 423–428.

TINNEY, C. G., BUTLER, R. P., MARCY, G. W., JONES, H. R. A., LAUGHLIN, G., CARTER, B. D., BAILEY, J. A. and OTOOLE, S. (2006). 2: 1 Resonant Exoplanetary System HD 73526. *The Astrophysical Journal* **647** 594–599.

UEDA, N., NAKANO, R., GHAHRAMANI, Z. and HINTON, G. E. (2000). SMEM algorithm for mixture models. *Neural computation* **12** 2109–2128.

VOGT, S. S., BUTLER, R. P., MARCY, G. W., FISCHER, D. A., HENRY, G. W., LAUGHLIN, G., WRIGHT, J. T. and JOHNSON, J. A. (2005). Five new multicomponent planetary systems. *The Astrophysical Journal* **632** 638–658.

WANG, H. X., LUO, B., ZHANG, Q. B. and WEI, S. (2004). Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm. *Pattern Recognition Letters* **25** 1799–1809.

WEST, M. (1993). Mixture Models, Monte Carlo, Bayesian Updating, and Dynamic Models. *Computing Science and Statistics* 325–325.

ZEEVI, A. J. and MEIR, R. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks* **10** 99–110.

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
DUKE UNIVERSITY
DURHAM, NC 27705 USA
E-MAIL: bl64@duke.edu

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NC 27705 USA
E-MAIL: clyde@stat.duke.edu
berger@stat.duke.edu

DEPARTMENT OF ASTRONOMY
CORNELL UNIVERSITY
ITHACA, NY USA
E-MAIL: loredo@cornell.edu