mathematics

MDPI

Article

Automatic Tempered Posterior Distributions for Bayesian Inversion Problems

Luca Martino [1,*], Fernando Llorente [2], Ernesto Curbelo [2], Javier López-Santiago [3] and Joaquín Míguez [3]

[1] Department of Signal Processing, Universidad rey Juan Carlos (URJC), 28942 Madrid, Spain
[2] Department of Statistics, Universidad Carlos III de Madrid (UC3M), 28911 Madrid, Spain; felloren@est-econ.uc3m.es (F.L.); ecurbelo@est-econ.uc3m.es (E.C.)
[3] Department of Signal Processing, Universidad Carlos III de Madrid (UC3M), 28911 Madrid, Spain; jalopezs@ing.uc3m.es (J.L.-S.); jmiguez@ing.uc3m.es (J.M.)
* Correspondence: luca.martino@urjc.es

Abstract: We propose a novel adaptive importance sampling scheme for Bayesian inversion problems where the inference of the variables of interest and the power of the data noise are carried out using distinct (but interacting) methods. More specifically, we consider a Bayesian analysis for the variables of interest (i.e., the parameters of the model to invert), whereas we employ a maximum likelihood approach for the estimation of the noise power. The whole technique is implemented by means of an iterative procedure with alternating sampling and optimization steps. Moreover, the noise power is also used as a tempered parameter for the posterior distribution of the the variables of interest. Therefore, a sequence of tempered posterior densities is generated, where the tempered parameter is automatically selected according to the current estimate of the noise power. A complete Bayesian study over the model parameters and the scale parameter can also be performed. Numerical experiments show the benefits of the proposed approach.

Keywords: Bayesian inference; importance sampling; MCMC; inversion problems

Citation: Martino, L.; Llorente, F.; Curbelo, E.; López-Santiago, J.; Míguez, J. Automatic Tempered Posterior Distributions for Bayesian Inversion Problems. *Mathematics* **2021**, *9*, 784. https://doi.org/10.3390/math9070784

Academic Editor: Daniel-Ioan Curiac

Received: 27 February 2021
Accepted: 1 April 2021
Published: 6 April 2021

1. Introduction

The estimation of unknown parameters from noisy observations is an essential problem in signal processing, statistics, and machine learning [1–3]. Within the Bayesian signal processing framework, these problems are addressed by constructing posterior probability distributions of the unknowns. Given the posterior, one often wants to make inference about the unknowns, e.g., if we are estimating parameters, finding the values that maximize their posterior, or the values that minimize some cost function given the uncertainty of the parameters. Unfortunately, obtaining closed-form solutions, usually expressed as integrals of the posterior, is infeasible in most practical applications. Therefore, developing approximate computational techniques, such as importance sampling and Markov chain Monte Carlo (MCMC) algorithms, is often required [4–6].

The so-called *tempering* of the posterior distributions is a well-known procedure for improving the performance of Monte Carlo (MC) algorithms [7–10]. Tempering is performed by modulating an artificial scale parameter or by sequentially including new data. There are several reasons for the improvement in performance: improving mixing, discovering modes, fostering the exploration of the inference space, etc. In the first iterations of the MC scheme, a posterior density with a bigger scale is considered. The artificial scale parameter (often called *temperature*) is reduced along the iterations, until considering the true posterior distribution. However, the user should select a *temperature schedule*, i.e., a decreasing rule for the scale parameter, which is usually chosen in an heuristic way [4,5]. In the literature, the tempering procedure has gained particular attention for the estimation of the marginal likelihood (also known as Bayesian model evidence) [9,11,12].

Furthermore, the joint inference of parameters (denoted as $\boldsymbol{\theta}$) of observation models, $\mathbf{f}(\boldsymbol{\theta})$, and scale parameters of the likelihood function (that, in the scalar case, is usually denoted as σ) can be a hard task. Indeed, "wrong choices" of σ values can easily jeopardize the sampling of $\boldsymbol{\theta}$. In this work, we introduce a procedure to tackle this problem.

To be specific, in this work, we design an adaptive importance sampling (AIS) scheme [13] for Bayesian inversion problems, where an automatic tempering procedure is implemented. We assume that the vector of observations $\mathbf{y}$ is obtained by a nonlinear transformation $\mathbf{f}(\boldsymbol{\theta})$ of the variables of interest $\boldsymbol{\theta}$, perturbed by additive Gaussian noise with unknown power σ^2. The nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$ usually represents a complex physical model, a computer code, etc. The resulting posterior densities are usually highly multi-modal and complex distributions. Furthermore, the inference task in the joint space $[\boldsymbol{\theta}, \sigma]$ is particularly challenging. We introduce a split strategy to tackle this problem, involving an optimization approach over σ and a sampling scheme for $\boldsymbol{\theta}$. More specifically, we design an iterative procedure where these two tasks are alternated. Additionally, the current maximum likelihood (ML) estimate of the noise power, $\widehat{\sigma}^2_{\texttt{ML}}$, is employed as a tempering parameter, starting from high values and then "cooling down" according to the ML estimates at each iteration. Therefore, the proposed scheme deals with a sequence of tempered posteriors according to the current estimation $\widehat{\sigma}^2_{\texttt{ML}}$. It is important to observe that, given a fixed vector $\boldsymbol{\theta}$, the ML estimator $\widehat{\sigma}^2_{\texttt{ML}}$ can be obtained analytically.

Furthermore, the complete Bayesian analysis regarding the joint posterior of $\boldsymbol{\theta}$ and σ is also possible (as discussed in Section 5). This is obtained by implementing a proper re-weighting of the samples generated by the proposed algorithm, called Automatic Tempering AIS (ATAIS), without any additional evaluations of the observation model. An approximation of the marginal posterior of σ is provided as well. The advantages of the proposed scheme are shown in two numerical experiments, one of them considering a complex astronomical model.

2. Problem Statement

Let us denote the observed measurements as $\mathbf{y} = [y_1, ..., y_K]^\top \in \mathbb{R}^K$, and the variable of interest that we wish to infer as $\boldsymbol{\theta} = [\theta_1, ..., \theta_M]^\top \in \boldsymbol{\Theta} \subseteq \mathbb{R}^M$. Furthermore, let us assume the observation model

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) + \mathbf{v}, \tag{1}$$

where we have a nonlinear mapping,

$$\mathbf{f}(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), ..., f_K(\boldsymbol{\theta})]^\top : \boldsymbol{\Theta} \to \mathbb{R}^K \quad \text{with} \quad \boldsymbol{\Theta} \subseteq \mathbb{R}^M, \tag{2}$$

and a Gaussian perturbation noise,

$$\mathbf{v} = [v_1, ..., v_K]^\top \sim \mathcal{N}(\mathbf{v}|\mathbf{0}, \sigma^2 \mathbf{I}_K), \tag{3}$$

with $\sigma > 0$, and $\mathbf{I}_K$ denotes the K-dimensional identity matrix. The model can be easily extended to a matrix of observations $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_K]^\top \in \mathbb{R}^{d_y \times K}$ instead of a vector, if the nonlinear mapping is of type $\mathbf{F}(\boldsymbol{\theta}) = [\mathbf{f}_1(\boldsymbol{\theta}), ..., \mathbf{f}_K(\boldsymbol{\theta})]^\top : \boldsymbol{\Theta} \subseteq \mathbb{R}^M \to \mathbb{R}^{d_y \times K}$. The noise variance σ^2 is unknown, in general. The mapping $\mathbf{f}(\boldsymbol{\theta})$ may be analytically unknown: the only assumption is that we are able to evaluate it pointwise. The likelihood function is

$$\begin{aligned} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) &= \frac{1}{(2\pi\sigma^2)^{K/2}} \exp\left(-\frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})||^2\right), && (4) \\ &= \frac{1}{(2\pi\sigma^2)^{K/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^{K} (y_k - f_k(\boldsymbol{\theta}))^2\right). && (5) \end{aligned}$$

Note that we have two types of variables of interest: the vector $\boldsymbol{\theta}$ contains the parameters of the nonlinear mapping $\mathbf{f}(\boldsymbol{\theta})$, whereas σ is a scale parameter of the likelihood function.

Given the vector of measurements $\mathbf{y}$, we wish to make inferences regarding the hidden parameters $\boldsymbol{\theta}$ and the noise power σ^2, obtaining at least some point estimators $\widehat{\boldsymbol{\theta}}$ and

$\widehat{\sigma}^2$. We are also interested in performing uncertainty and correlation analysis for the components of $\boldsymbol{\theta}$. Furthermore, we aim to perform model selection, i.e., to compare, select, or properly average different models.

Bayesian inference in the complete space. We consider independent prior densities $g_\theta(\boldsymbol{\theta})$ and $g_\sigma(\sigma)$ over the unknowns. Therefore, the complete posterior density is

$$p(\boldsymbol{\theta}, \sigma|\mathbf{y}) = \frac{1}{p(\mathbf{y})} p(\boldsymbol{\theta}, \sigma, \mathbf{y}) = \frac{1}{p(\mathbf{y})} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) g_\sigma(\sigma), \tag{6}$$

The marginal likelihood is

$$Z = p(\mathbf{y}) = \int_{\mathbb{R}^+} \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) g_\sigma(\sigma) d\boldsymbol{\theta} d\sigma, \tag{7}$$

This quantity is often needed for model selection. While Z is generally unknown, we can usually evaluate pointwise the un-normalized posterior $\pi(\boldsymbol{\theta}, \sigma|\mathbf{y}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) g_\sigma(\sigma)$, i.e., $p(\boldsymbol{\theta}, \sigma|\mathbf{y}) \propto \pi(\boldsymbol{\theta}, \sigma|\mathbf{y})$. More generally, the computation of integrals of the form

$$I(h) = \int_{\mathbb{R}^+} \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}, \sigma) p(\boldsymbol{\theta}, \sigma|\mathbf{y}) d\boldsymbol{\theta} d\sigma, \tag{8}$$

where $h : \boldsymbol{\Theta} \times \mathbb{R}^+ \to \mathbb{R}$ is an integrable function, is usually required. We consider a Monte Carlo quadrature approach for approximating the integral above and, more generally, provide a particle approximation of the joint posterior $p(\boldsymbol{\theta}, \sigma|\mathbf{y})$.

Main observation. Generating random samples from a complicated posterior in Equation (6) and efficiently computing the integrals as in Equations (7) and (8) is very often a hard task. Moreover, this task becomes more difficult when we try to perform a joint inference where scale parameters are involved, i.e., σ, and parameters of the nonlinearity, i.e., $\boldsymbol{\theta}$. Indeed, "wrong choices" of σ values can easily jeopardize the sampling of $\boldsymbol{\theta}$. In the next section, we describe a strategy that we propose to tackle this problem. Before do so, however, we need to recall some additional definitions.

Conditional and marginal posteriors. In other to design efficient computational schemes, it is often useful to consider the conditional posteriors, for instance,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, \sigma) = \frac{p(\boldsymbol{\theta}, \mathbf{y}, \sigma)}{p(\mathbf{y}, \sigma)} &= \frac{\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) g_\sigma(\sigma)}{p(\mathbf{y}|\sigma) g_\sigma(\sigma)}, \\ &= \frac{\ell(\mathbf{y}|\sigma, \boldsymbol{\theta}) g_\theta(\boldsymbol{\theta})}{p(\mathbf{y}|\sigma)}. \end{aligned} \tag{9}$$

In the next section, we will see that the idea underlying the proposed scheme is to split the space $[\boldsymbol{\theta}, \sigma]$, restricting the sampling problem only to $\boldsymbol{\theta}$ and considering an optimization problem with respect to σ. The conditional marginal likelihood is obtained by integrating out one of the two variables, e.g.,

$$Z(\sigma) = p(\mathbf{y}|\sigma) = \int_{\boldsymbol{\theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{10}$$

The integral above cannot be computed analytically, in general. We can also consider marginal posteriors, for instance, the marginal posterior of σ is

$$p(\sigma|\mathbf{y}) = \frac{p(\mathbf{y}|\sigma) g_\sigma(\sigma)}{p(\mathbf{y})} = \frac{Z(\sigma) g_\sigma(\sigma)}{Z}. \tag{11}$$

Note that the joint posterior in Equation (6) can be also written as

$$p(\boldsymbol{\theta}, \sigma|\mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y}, \sigma) p(\sigma|\mathbf{y}). \tag{12}$$

Outline of the proposed approach. The underlying idea of this work is to divide the inference study in two parts. In the first part (Sections 3 and 4), we focus on the study of the conditional posterior $p(\boldsymbol{\theta}|\mathbf{y},\sigma)$ given a fixed σ. Then, in the second part (Section 5), we also estimate the marginal posterior $p(\sigma|\mathbf{y})$. Finally, using (12), we can obtain a final approximation of the complete posterior $p(\boldsymbol{\theta},\sigma|\mathbf{y})$. Estimations of $Z(\sigma)$ and Z are also obtained.

3. Key Observations and Proposed Approach

3.1. Split Inference

In the first part of work, we assume a uniform proper (or improper) prior over $\boldsymbol{\theta}$, i.e., $g_\theta(\boldsymbol{\theta}) \propto 1$ in $\boldsymbol{\Theta}$. The possible use of a general choice of $g_\theta(\boldsymbol{\theta})$ is discussed in Section 4.1. Let $\boldsymbol{\theta}_{\mathtt{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y},\sigma)$ denote the MAP estimator of $\boldsymbol{\theta}$. Generally, $\boldsymbol{\theta}_{\mathtt{MAP}}$ should be a function of σ, i.e., $\boldsymbol{\theta}_{\mathtt{MAP}} = \boldsymbol{\theta}_{\mathtt{MAP}}(\sigma)$. However, due to the choice of likelihood function (and the uniform prior) considered in this paper, we have that

$$\begin{aligned}\boldsymbol{\theta}_{\mathtt{MAP}} &= \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{y},\sigma), \\ &= \arg\min_{\boldsymbol{\theta}} ||\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})||^2, \qquad \text{with} \quad g_\theta(\boldsymbol{\theta}) \propto 1, \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta},\end{aligned}$$

which does not depend on σ, i.e., we have that $\boldsymbol{\theta}_{\mathtt{MAP}}$ maximizes the conditional posterior $p(\boldsymbol{\theta}|\mathbf{y},\sigma)$ for any σ. See Appendix A for further details.

Furthermore, the variance of the conditional posterior $p(\boldsymbol{\theta}|\mathbf{y},\sigma)$ grows when σ increases. In this sense, with larger σ, the density $p(\boldsymbol{\theta}|\mathbf{y},\sigma)$ is "broader"; thus, it is easier for Monte Carlo methods to explore the space (namely, we have a *tempering effect*). Based on these considerations, we can run Monte Carlo schemes (specifically IS algorithms) on $p(\boldsymbol{\theta}|\mathbf{y},\sigma_0)$ with a large value σ_0 for estimating $\boldsymbol{\theta}_{\mathtt{MAP}}$ more efficiently. Furthermore, apart from estimating $\boldsymbol{\theta}_{\mathtt{MAP}}$, we are also interested in studying the conditional posterior $p(\boldsymbol{\theta}|\mathbf{y},\sigma_{\mathrm{ML}})$, where

$$\sigma_{\mathrm{ML}} = \arg\max_{\sigma} \ell(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{MAP}},\sigma).$$

The value σ_{ML} can be obtained in closed-form (see Appendix A). In fact, for any $\boldsymbol{\theta}$, we have

$$\ell(\mathbf{y}|\boldsymbol{\theta},\sigma) \quad \propto \quad \left(\frac{1}{\sigma^2}\right)^{\frac{K}{2}} \exp\left(-\frac{||\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})||^2}{2\sigma^2}\right), \tag{13}$$

which has the form of an *Inverse Gamma* density for σ^2 and it has a *unique* mode at $\sqrt{\frac{1}{K}||\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})||^2}$, where $\boldsymbol{\theta}$ is a fixed. Therefore, finally we have

$$\sigma_{\mathrm{ML}} = \sqrt{\frac{1}{K}||\mathbf{y}-\mathbf{f}(\boldsymbol{\theta}_{\mathtt{MAP}})||^2}.$$

This can serve as a point estimator of the noise power in the system, and also as a threshold value to stop the tempering of the conditional posterior, as we show in the following section.

3.2. An Iterative Scheme

Consider that we start with a large value σ_0, which can be viewed as a coarse approximation of σ_{ML}, so we denote it $\sigma_0 = \widehat{\sigma}_{\mathrm{ML}}^{(0)}$. Let $\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(1)}$ denote an estimate of $\boldsymbol{\theta}_{\mathtt{MAP}}$ obtained by working w.r.t. $p(\boldsymbol{\theta}|\mathbf{y},\widehat{\sigma}_{\mathrm{ML}}^{(0)})$. We use this current estimation to obtain the next value of σ, i.e., $\widehat{\sigma}_{\mathrm{ML}}^{(1)} = \sqrt{\frac{1}{K}||\mathbf{y}-\mathbf{f}(\widehat{\boldsymbol{\theta}}_{\mathtt{MAP}}^{(1)})||^2}$. In general, $\widehat{\sigma}_{\mathrm{ML}}^{(1)}$ is a better estimator of σ_{ML} than $\widehat{\sigma}_{\mathrm{ML}}^{(0)}$, as we have tried to evaluate of the smallest error between $\mathbf{f}(\boldsymbol{\theta})$ and the data, $\mathbf{y}$, i.e., $||\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})||$, which is related to the power of the noise perturbation in the system. For instance, assuming zero noise, we would have $||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_{\texttt{MAP}})|| = 0$, recalling that $g_\theta(\boldsymbol{\theta}) \propto 1$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We can iterate this procedure for $t = 1, \ldots, T$:

1 Estimate $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$ by Monte Carlo (e.g., an IS scheme) by approximately maximizing $p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(t-1)})$.
2 Compute

$$\widehat{\sigma}_{\texttt{ML}}^{(t)} = \sqrt{\frac{1}{K}||\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)})||^2}. \tag{14}$$

With this iterative scheme, we have that $\widehat{\sigma}_{\texttt{ML}}^{(T)} \to \sigma_{\texttt{ML}}$ as T grows, thus we eventually perform IS with respect to the density of interest $p(\boldsymbol{\theta}|\mathbf{y}, \sigma_{\texttt{ML}})$. Furthermore, a non-increasing sequence of values $\widehat{\sigma}_{\texttt{ML}}^{(0)} \geq \widehat{\sigma}_{\texttt{ML}}^{(1)} \geq \cdots \geq \widehat{\sigma}_{\texttt{ML}}^{(T)}$ is produced, which facilitates the estimation of $\boldsymbol{\theta}_{\texttt{MAP}}$, and ensures the IS estimation of $p(\boldsymbol{\theta}|\mathbf{y}, \sigma_{\texttt{ML}})$ is performed efficiently by using the set of intermediate, tempered (i.e., wider) distributions $p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(t)})$ for $t = 0, 1, ..., T$. Finally, a particle approximation of $p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(T)})$ is obtained, i.e.,

$$p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(T)}) = \sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}_t^{(n)} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(n)}),$$

where $\sum_{t=1}^{T} \sum_{n=1}^{N} \widetilde{w}_t^{(n)} = 1$. Note that $\widetilde{w}_t^{(n)}$ are the final corrected weights obtained at the end of the algorithm (see Algorithm 1).

4. Automatic Tempering Adaptive Importance Sampling (ATAIS)

In this section, we describe an adaptive importance sampler with an *automatic tempering* approach which follows the procedure given above. At each iteration t of the algorithm, we have an ML approximation of σ, i.e., $\widehat{\sigma}_{\texttt{ML}}^{(t-1)}$. Considering Equation (9), we define the un-normalized *tempered conditional posterior* at the t-th iteration,

$$\pi_t(\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \widehat{\sigma}_{\texttt{ML}}^{(t-1)}) g_\theta(\boldsymbol{\theta}), \tag{15}$$

where we assume $g_\theta(\boldsymbol{\theta}) \propto 1$ in $\boldsymbol{\Theta}$. For other generic choice of $g_\theta(\boldsymbol{\theta})$, see the discussion in Section 4.1. At each iteration, we consider $p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(t-1)}) \propto \pi_t(\boldsymbol{\theta})$ as the target distribution. The dependence on the iteration t occurs because $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ varies with t. The ATAIS algorithm is outlined in Algorithm 1. The resulting scheme is an adaptive IS algorithm which combines sampling schemes and stochastic optimization. It is important to remark that if $\widehat{\sigma}_{\texttt{ML}}^{(0)}$ is bigger than the true ML value, we generate a non-increasing sequence of $\widehat{\sigma}_{\texttt{ML}}^{(t)}$, i.e., $\widehat{\sigma}_{\texttt{ML}}^{(0)} \geq \widehat{\sigma}_{\texttt{ML}}^{(1)} \geq ...\widehat{\sigma}_{\texttt{ML}}^{(t)} \geq \widehat{\sigma}_{\texttt{ML}}^{(t+1)}$, etc. Note that this is true as we have assumed a uniform prior $g_\theta(\boldsymbol{\theta})$. To see this, recall that $\widehat{\sigma}_{\texttt{ML}} = \sqrt{\frac{1}{K}||\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_{\texttt{MAP}})||^2}$. Improving $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ means that the squared error $||\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_{\texttt{MAP}})||^2$ is smaller, as shown in Equation (14), which implies that $\widehat{\sigma}_{\texttt{ML}}$ always decreases (provided that we start with $\widehat{\sigma}_{\texttt{ML}} > \sigma_{\texttt{ML}}$).

IS steps. A set of N samples $\{\boldsymbol{\theta}_t^{(n)}\}_{n=1}^{N}$ are drawn from a (normalized) proposal density $q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with mean $\boldsymbol{\mu}_t$ and a covariance matrix $\boldsymbol{\Sigma}_t$. An importance weight

$$w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)},$$

is assigned to each sample.

Proposal adaptation. A particle estimation of the conditional MAP estimator of $\boldsymbol{\theta}$ is given by $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$. The value of current MAP approximation $\pi_t(\widehat{\boldsymbol{\theta}}_t)$ is then compared with the value of global MAP estimator obtained so far denoted as $\pi_{\texttt{MAP}}$.

If $\pi_t(\widehat{\boldsymbol{\theta}}_t) \geq \pi_{\texttt{MAP}}$, all the global MAP estimators are updated and the proposal pdf is moved at $\widehat{\boldsymbol{\theta}}_t$, i.e., we set

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_t, \quad \pi_{\texttt{MAP}} = \pi_t(\widehat{\boldsymbol{\theta}}_t), \quad \boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t. \tag{16}$$

Algorithm 1: ATAIS: AIS with automatic tempering.

```
1. Initializations: Choose $N$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, and obtain an initialization for $\widehat{\sigma}_{\text{ML}}^{(0)}$, and set $\pi_{\texttt{MAP}} = 0$.
2. For $t = 1, \ldots, T$:
   (a) Sampling:
       i.  Draw $\boldsymbol{\theta}_t^{(1)}, ..., \boldsymbol{\theta}_t^{(N)} \sim q(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$.
       ii. Assign to each sample the weights
           $w_t^{(n)} = \frac{\pi_t(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}, \qquad n = 1, ..., N.$   (17)
   (b) Current maximum estimations:
       i.  Obtain $\widehat{\boldsymbol{\theta}}_t = \arg\max_n \pi_t(\boldsymbol{\theta}_t^{(n)})$, and compute $\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$
       ii. Compute $\widehat{\sigma}_t = \sqrt{\frac{1}{K}||\mathbf{y} - \widehat{\mathbf{r}}_t||^2}$.
   (c) Global maximum estimations:
       i.  If $\widehat{\sigma}_t \leq \widehat{\sigma}_{\text{ML}}^{(t-1)}$, then set $\widehat{\sigma}_{\text{ML}}^{(t)} = \widehat{\sigma}_t$. Otherwise, set $\widehat{\sigma}_{\text{ML}}^{(t)} = \widehat{\sigma}_{\text{ML}}^{(t-1)}$.
       ii. If $\pi_t(\widehat{\boldsymbol{\theta}}_t) \geq \pi_{\texttt{MAP}}$, then set $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_t$ and $\pi_{\texttt{MAP}} = \pi_t(\widehat{\boldsymbol{\theta}}_t)$. Otherwise, $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t-1)}$ and keep the value of $\pi_{\texttt{MAP}}$.
   (d) Adaptation: Set
       $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)},$   (18)
       $\boldsymbol{\Sigma}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)} (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t)^\top (\boldsymbol{\theta}_t^{(n)} - \bar{\boldsymbol{\theta}}_t) + \epsilon \mathbf{I}_M,$   (19)
       where $\bar{w}_t^{(n)} \frac{w_t^{(n)}}{\sum_{i=1}^{N} w_t^{(i)}}$ are the normalized weights, $\bar{\boldsymbol{\theta}}_t = \sum_{n=1}^{N} \bar{w}_t^{(n)} \boldsymbol{\theta}_t^{(n)}$ and $\epsilon > 0$ is a small scalar value .
3. Output: Return the final estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(T)}$, $\widehat{\sigma}_{\text{ML}}^{(T)}$, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$, for all $t$ and $n$, with the corrected weights
   $\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}.$   (20)
```

Otherwise, we keep the previous values of $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)} = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t-1)}$, $\pi_{\texttt{MAP}}$, and $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1}$. The covariance matrix $\boldsymbol{\Sigma}_t$ is adapted by considering the empirical covariance of the weighted samples. Note that we set $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(t)}$ instead of using the empirical mean of the samples (as in other classical AIS schemes). This is because we have noticed that this choice provides better and more robust results, especially as the dimension of the problem grows.

Automatic tempering. As we showed in the previous section, the current ML estimator of σ can be obtained analytically as

$$\widehat{\sigma}_t = \sqrt{\frac{1}{K}||\mathbf{y} - \widehat{\mathbf{r}}_t||^2}, \tag{21}$$

where $\widehat{\mathbf{r}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$. If the current ML estimator $\widehat{\sigma}_t$ is smaller than the current global one $\widehat{\sigma}_{\texttt{ML}}^{(t-1)}$, i.e., $\widehat{\sigma}_t < \widehat{\sigma}_{\texttt{ML}}^{(t-1)}$, then we update $\widehat{\sigma}_{\texttt{ML}}^{(t)} = \widehat{\sigma}_t$, Otherwise, we keep the value of $\widehat{\sigma}_{\texttt{ML}}^{(t)} = \widehat{\sigma}_{\texttt{ML}}^{(t-1)}$. Actually, with a uniform prior $g_\theta(\boldsymbol{\theta})$, every time that we update $\widehat{\boldsymbol{\theta}}_{\text{MAP}}^{(t)}$, we also update $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ (see footnote in the previous page).

ATAIS outputs. After T iterations, a final correction of the weights is needed, i.e.,

$$\widetilde{w}_t^{(n)} = w_t^{(n)} \frac{\pi_{T+1}(\boldsymbol{\theta}_t^{(n)})}{\pi_t(\boldsymbol{\theta}_t^{(n)})}, \tag{22}$$

in order to obtain a particle approximation of the measure of the final conditional posterior $p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(T)}) \propto \pi_{T+1}(\boldsymbol{\theta})$. Thus, the algorithm returns the final estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}^{(T)}, \widehat{\sigma}_{\texttt{ML}}^{(T)}$, and all the weighted samples $\{\boldsymbol{\theta}_t^{(n)}, \widetilde{w}_t^{(n)}\}$, for all $n = 1, ..., N$ and $t = 1, ..., T$. Other outputs can be obtained with a postprocessing of the weighted samples, as shown below. Note that Equation (22) does not require any additional evaluation of the model, and the error is $e_t^{(n)} = ||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_t^{(n)})||^2$. Moreover, we can also use $e_t^{(n)}$ and $\{\boldsymbol{\theta}_t^{(n)}\}$ for building a particle approximation of any other conditional posterior $p(\boldsymbol{\theta}|\mathbf{y}, \sigma)$. This allows the study of the marginal posterior $p(\sigma|\mathbf{y})$ and provides the complete Bayesian inference, as we show in the next section.

4.1. With a Generic Prior $g_\theta(\boldsymbol{\theta})$

The ATAIS algorithm is based on the fact that $\boldsymbol{\theta}_{\texttt{MAP}}$ does not depend on σ. This allows us to progressively estimate it by targeting the sequence of tempered posteriors $\pi_t(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{y}, \widehat{\sigma}_{\texttt{ML}}^{(t)})$ that share all the same MAP. However, in the case $g_\theta(\boldsymbol{\theta})$ is not uniform, we generally have one $\boldsymbol{\theta}_{\texttt{MAP}}(\sigma)$ for each $p(\boldsymbol{\theta}|\mathbf{y}, \sigma)$, and we could have that the sequence of $\boldsymbol{\theta}_{\texttt{MAP}}(\widehat{\sigma}_{\texttt{ML}}^{(t)})$ will not approach $\boldsymbol{\theta}_{\texttt{MAP}}(\sigma_{\texttt{ML}})$.

If the data are informative and the prior $g_\theta(\boldsymbol{\theta})$ is chosen such it is vague with respect to the likelihood, the position of $\boldsymbol{\theta}_{\texttt{MAP}}(\sigma)$ is not very sensitive to the value of σ. Namely, we have $\boldsymbol{\theta}_{\texttt{MAP}}(\widehat{\sigma}_{\texttt{ML}}^{(1)}) \approx \boldsymbol{\theta}_{\texttt{MAP}}(\widehat{\sigma}_{\texttt{ML}}^{(2)}) \approx \cdots \approx \boldsymbol{\theta}_{\texttt{MAP}}(\sigma_{\texttt{ML}})$, and thus our algorithm can be applied in this context. When the data are not informative, we should use an even more vague prior (i.e., wider than the likelihood function) in order to maintain the usefulness of the algorithm.

5. Complete Bayesian Inference with ATAIS

Let us assume we have a proper prior $g_\theta(\boldsymbol{\theta})$ and we introduce another proper prior $g_\sigma(\sigma)$ for σ. The outputs of the ATAIS algorithm can serve to approximate the normalizing constant of the joint posterior $p(\boldsymbol{\theta}, \sigma|\mathbf{y}) \propto \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) g_\sigma(\sigma)$, i.e., the so-called marginal likelihood or Bayesian model evidence, given by

$$Z = \int_{\mathbb{R}^+} \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta}) g_\sigma(\sigma) d\boldsymbol{\theta} d\sigma = \int_{\mathbb{R}^+} Z(\sigma) g_\sigma(\sigma) d\sigma, \tag{23}$$

where we have denoted $Z(\sigma) = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma) g_\theta(\boldsymbol{\theta} d\boldsymbol{\theta}$, usually called conditional marginal likelihood. The quantity Z is useful for model selection purposes. Furthermore, a complete Bayesian study of the joint posterior $p(\boldsymbol{\theta}, \sigma|\mathbf{y})$ can be provided as well.

Approximation of $Z(\sigma) = p(\mathbf{y}|\sigma)$. After the T iterations of ATAIS, we can also approximate the conditional marginal likelihood $Z(\sigma) = p(\mathbf{y}|\sigma)$ without additional evaluations of the target function. Indeed, saving the error values at each particle obtained for the computation of the likelihood function during ATAIS,

$$e_t^{(n)} = ||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_t^{(n)})||^2,$$

We can compute the IS weights,

$$\rho_t^{(n)}(\sigma) = \frac{\frac{1}{(2\pi\sigma^2)^{\frac{K}{2}}} \exp\left(-\frac{e_t^{(n)}}{2\sigma^2}\right) g_\theta(\boldsymbol{\theta}_t^{(n)})}{q(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}, \tag{24}$$

for a generic value of σ Thus, the IS estimator of the conditional marginal likelihood $Z(\sigma)$ is given by the arithmetic mean of the weights $\rho_t^{(n)}(\sigma)$,

$$\widehat{Z}(\sigma) = \widehat{p}(\mathbf{y}|\sigma) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{n=1}^{N} \rho_t^{(n)}(\sigma). \tag{25}$$

Approximation of Z**.** Drawing $\sigma^{(r)} \sim g_\sigma(\sigma)$, for $r = 1, ..., R$, (or considering a deterministic grid, e.g., as a Riemannian integration), we can approximate the global marginal likelihood Z by applying simple Monte Carlo to the integral in Equation (23),

$$\widehat{Z} = \frac{1}{R} \sum_{r=1}^{R} \widehat{Z}(\sigma^{(r)}). \tag{26}$$

Approximation of $p(\sigma|\mathbf{y})$**.** An approximation of the marginal posterior $p(\sigma|\mathbf{y}) = \frac{p(\mathbf{y}|\sigma)g_\sigma(\sigma)}{p(y)}$ can be also obtained as

$$p(\sigma|\mathbf{y}) \approx \widehat{p}(\sigma|\mathbf{y}) = \frac{\widehat{Z}(\sigma)g_\sigma(\sigma)}{\widehat{Z}}, \tag{27}$$

which can be used to approximate, e.g., the MAP of $p(\sigma|\mathbf{y})$ by $\sigma_{\texttt{MAP-marg}} \approx \arg\max_\sigma \widehat{Z}(\sigma)g_\sigma(\sigma)$. Other different moments of $p(\sigma|\mathbf{y})$ can be computed by a deterministic quadrature (as the problem is now one-dimensional) or applying noisy Monte Carlo approaches.

Complete Bayesian analysis. We can approximate the integral of interest as

$$I = \int_{\mathbb{R}^+} \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}, \sigma) p(\boldsymbol{\theta}, \sigma|\mathbf{y}) d\boldsymbol{\theta} d\sigma, \tag{28}$$

$$= \int_{\mathbb{R}^+} \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta}, \sigma) p(\boldsymbol{\theta}|\mathbf{y}, \sigma) p(\sigma|\mathbf{y}) d\boldsymbol{\theta} d\sigma \tag{29}$$

$$\approx \frac{1}{J} \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{n=1}^{N} \bar{\rho}_t^{(n)}(\sigma^{(j)}) h(\boldsymbol{\theta}_t^{(n)}, \sigma^{(j)}), \tag{30}$$

where

$$\bar{\rho}_t^{(n)}(\sigma^{(j)}) = \frac{\rho_t^{(n)}(\sigma^{(j)})}{\sum_{\tau=1}^{T} \sum_{i=1}^{N} \rho_\tau^{(i)}(\sigma^{(j)})}, \tag{31}$$

and $\sigma^{(j)}$ are generated by applying a noisy MCMC with invariant density $\widehat{p}(\sigma|\mathbf{y}) \propto \widehat{Z}(\sigma)g_\sigma(\sigma)$. Note that the samples $\boldsymbol{\theta}_t^{(n)}$ do not depend on the index j (they do not change) as we are *recycling* the particles generated by ATAIS and reusing evaluations $e_t^{(n)} = ||\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_t^{(n)})||^2$.

6. Simulations

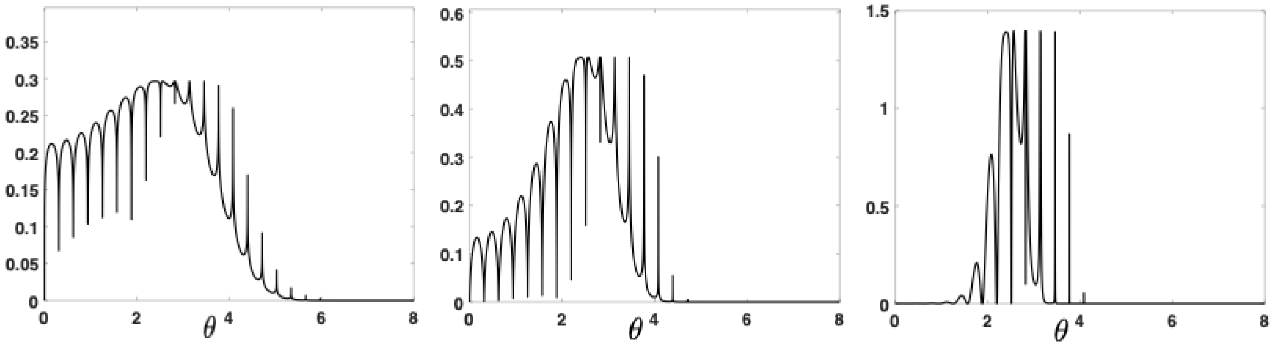
We test the proposed scheme in two numerical examples: The first numerical experiment is a simple bidimensional example (which is easy to be reproduced). The second experiment considers a real-world application, i.e., a radial velocity model of exoplanet systems which is often employed in astronomy applications (with a dimension of the inference problem of 6 and 11).

6.1. First Numerical Analysis

For the sake of simplicity, let us consider $\theta \in \mathbb{R}$ and an observation model given by the equation

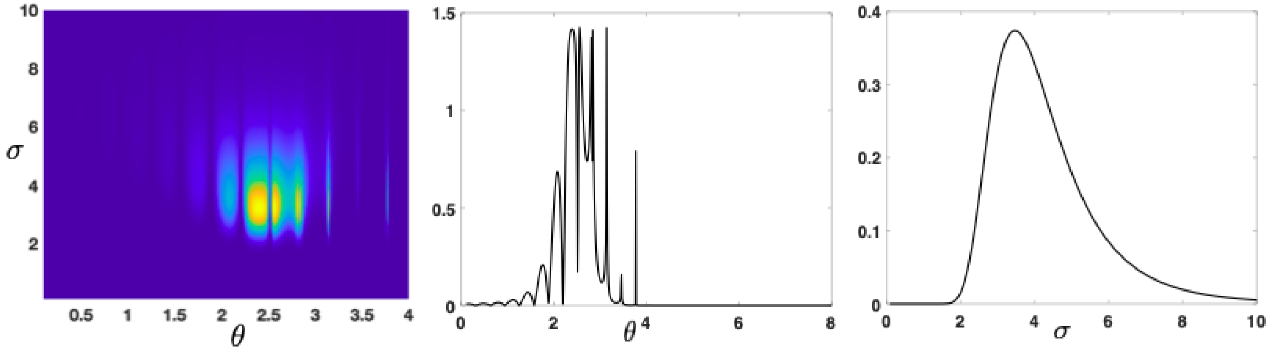
$$y_k = \theta^2 + \log(|\sin(10\theta)|) + v_k,$$

so that $f(\theta) = \theta^2 + \log(|\sin(10\theta)|)$, and $v_k \sim \mathcal{N}(0, \sigma^2)$. We consider $\theta_{\texttt{true}} = 2.5$, and $\sigma_{\texttt{true}} = 4$. We generate $K = 8$ observations from the model above. We also consider a uniform prior for θ in $(0, 20]$. The conditional posterior $p(\theta|\mathbf{y}, \sigma_{\texttt{true}})$ is shown in Figure 1c. We can observe that $p(\theta|\mathbf{y}, \sigma_{\texttt{true}})$ is highly multimodal. Figure 1 also depicts the conditional posteriors $p(\theta|\mathbf{y}, \sigma)$ with $\sigma \in \{10, 20\}$. Considering also a uniform prior over σ in $(0, 20]$, we have also a bidimensional joint posterior over $[\theta, \sigma]$, which is depicted in Figure 2a.



(**a**) $\sigma = 20$ (**b**) $\sigma = 10$ (**c**) $\sigma = \sigma_{\texttt{true}} = 4$

Figure 1. Conditional posteriors corresponding to different values of σ: (**a**) $\sigma = 20$, (**b**) $\sigma = 10$, and (**c**) $\sigma = \sigma_{\texttt{true}} = 4$.



(**a**) Joint posterior (**b**) Marginal posterior $p(\theta|\mathbf{y})$ (**c**) Marginal posterior $p(\sigma|\mathbf{y})$

Figure 2. The bidimensional joint posterior $p(\theta, \sigma|\mathbf{y})$ and the two marginal posteriors $p(\theta|\mathbf{y})$, $p(\sigma|\mathbf{y})$ in Equation (11), computed by using a thin grid approximation.

In this bidimensional example, it is possible to obtain the ground-truths using an expensive thin grid. We show the ground-truths of the different pdfs in Table 1. Moreover, the true value of the complete evidence $Z = p(\ \mathbf{y}) = 1.5983 \times 10^{-9}$. As the prior over σ is uniform, the maximum likelihood of σ is $\sigma_{\texttt{ML}} = \sigma_{\texttt{MAP-joint}} = 3.23$. The two marginal posteriors are shown in Figure 2b,c.

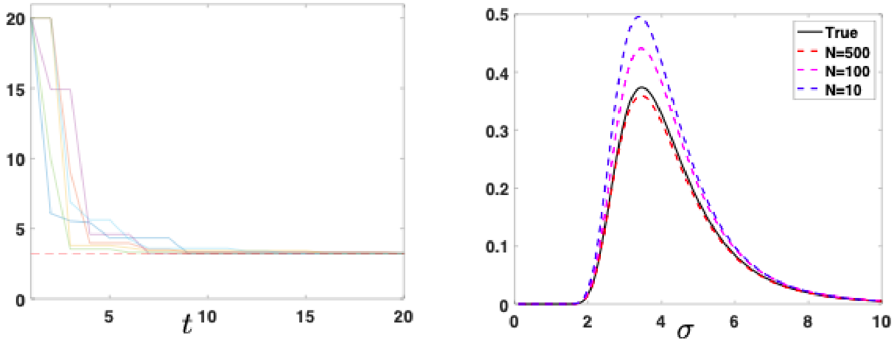
Table 1. Summary of pdfs and ground-truths for the first numerical experiment.

Pdf	Expectation	Variance	MAP
$p(\boldsymbol{\theta}|\mathbf{y}, \sigma_{\texttt{ML}})$	2.48	0.11	2.56
$p(\sigma|\mathbf{y})$	4.32	2.43	3.46
$p(\boldsymbol{\theta}|\mathbf{y})$	2.46	0.18	2.56

We apply ATAIS with the goal of estimating the expected value and the variance of the posterior density with respect to θ. We consider a Gaussian proposal $q(\theta|\mu_t, \lambda_t)$ with $\mu_0 = 10$ and a starting variance of $\lambda_0 = 4$. Note that μ_0 is located in a region that does not contain modes. We also start with $\widehat{\sigma}_{\texttt{ML}}^{(0)} = 20$ and $\pi_{\texttt{MAP}} = 0$ (initial conditions). The Mean Square Error (MSE) of ATAIS, averaged over 500 runs, in estimation of different moments and modes as function of N (and with $T = 10$), is given in Table 2. The ML estimation $\widehat{\sigma}_{\texttt{ML}}^{(t)}$, as function of the iteration t (with $N = 5$) for different runs, is given in Figure 3a. The approximation of the marginal posterior $p(\sigma|\mathbf{y})$, denoted $\widehat{p}(\sigma|\mathbf{y})$, is obtained as in Equation (27) in one specific run, with different $N \in \{10, 100, 500\}$ and $T = 10$. The approximations of the joint posterior $p(\boldsymbol{\theta}, \sigma|\mathbf{y})$ and the marginal posterior $p(\boldsymbol{\theta}|\mathbf{y})$, obtained by resampling the particles according to the normalized weights in Equations (31) and (24), are shown in Figure 4, i.e., using a sampling importance resampling procedure. For more details, see in [14] and Chapter 24 in [15].

Table 2. Mean Square Error (MSE) of ATAIS (averaged over 500 runs), in the estimation of the evidence, different moments and modes as function of N and $T = 10$.

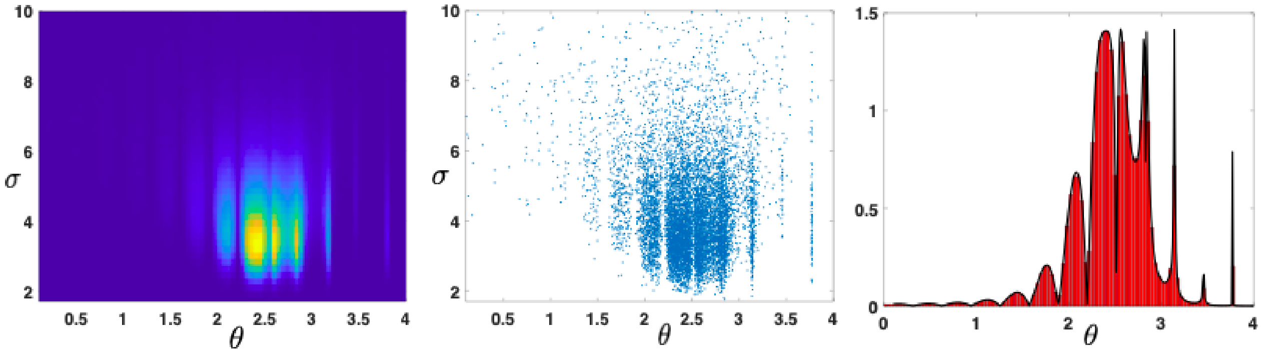
Value	$N = 10$	$N = 100$	$N = 1000$	$N = 5000$	Ground-Truths
$E[\theta\|\mathbf{y}, \sigma_{\texttt{ML}}]$	0.0311	0.0098	0.0034	0.0024	2.48
$\text{var}[\theta\|\mathbf{y}, \sigma_{\texttt{ML}}]$	0.0474	0.0370	0.0298	0.0201	0.11
$\theta_{\texttt{MAP}}$	0.0410	0.0337	0.0285	0.0127	2.56
$E[\sigma\|\mathbf{y}]$	0.9233	0.0785	0.0097	0.0023	4.32
$\text{var}[\sigma\|\mathbf{y}]$	6.1869	0.2640	0.0035	0.0010	2.43
$\sigma_{\texttt{MAP-marg}}$	0.0056	0.0004	0.0001	3×10^{-5}	3.46
$\sigma_{\texttt{ML}}$	8×10^{-5}	2×10^{-5}	5×10^{-7}	6×10^{-9}	3.23
$Z = p(\ \mathbf{y})$	2×10^{-18}	1.8×10^{-20}	1.4×10^{-20}	3.6×10^{-22}	1.6×10^{-9}



(**a**) $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ vs t

(**b**) $p(\sigma|\mathbf{y})$ for different N ($T = 10$)

Figure 3. (**a**) The maximum likelihood (ML) estimation $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ (different runs) versus the number of iterations t, with $N = 5$. (**b**) The true marginal posterior $p(\sigma|\mathbf{y})$ and different approximations, in one specific run, $\widehat{p}(\sigma|\mathbf{y})$ obtained as in Equation (27) with different $N \in \{10, 100, 500\}$ and $T = 10$ (thus, the total number of samples are NT).



(**a**) Histogram with 2×10^6 samples (**b**) 10^4 samples by ATAIS (**c**) Histogram with 2×10^6 samples

Figure 4. Approximations obtained with ATAIS. (**a**,**b**) Joint posterior $p(\boldsymbol{\theta}, \sigma|\mathbf{y})$: (a) by an histogram with 2×10^6 samples; (b) 10^4 samples from joint posterior obtained by ATAIS. (**c**) Approximation by an histogram with 2×10^6 samples, of the marginal posterior $p(\boldsymbol{\theta}|\mathbf{y})$.

6.2. Radial Velocity Curves of Exoplanets and Binary Systems

In this example, we consider an application in an astronomical model. In recent years, the problem of revealing objects orbiting other stars has acquired large attention. Different techniques have been proposed to discover exo-objects but, nowadays, the radial velocity technique is still the most used [16–19]. The problem consists in fitting a model (the so-called radial velocity curve) to data acquired at different moments spanning during long time periods (up to years). The model is highly nonlinear, and it is costly in terms of computation time (especially for certain sets of parameters). Obtaining a value to compare to a single observation involves numerically integrating a differential equation in time or an iterative procedure for solving to a nonlinear equation. Typically, the iteration is performed until a threshold is reached or 10^6 iterations are performed. The problem of radial velocity curve fitting is applied in several related applications.

Observation model—likelihood. When analyzing the radial velocity data of an exoplanetary system, it is commonly accepted that the *wobbling* of the star around the center of mass is caused by the sum of the gravitational force of each planet independently and that they do not interact with each other. Each planet follows a Keplerian orbit, and the radial velocity of the host star is given by

$$y_{r,t} = V_0 + \sum_{i=1}^{S} A_i[\cos(u_{i,t} + \omega_i) + e_i \cos(\omega_i)] + \xi_t, \tag{32}$$

with $t = 1, \dots, T$ and $r = 1, \dots, R$. In this equation, A_i is the amplitude of the curve, w_i is the argument of perigee, and e_i is the eccentricity of the orbit of the i-th planet. The parameter V_0 represents the mean velocity, and it is common for all the planets. The number of objects in the system is S, which is considered to be known in this experiment (for the sake of simplicity). Both $y_{r,t}$ and $u_{i,t}$ depend on time t, and then ξ_t is a Gaussian noise perturbation with variance σ^2. The likelihood function is defined by (32) and some indicator variables described below. The angle $u_{i,t}$ is the true anomaly of the planet i, and it can be determined from

$$\frac{du_{i,t}}{dt} = \frac{2\pi}{P_i} \frac{(1 + e_i \cos u_{i,t})^2}{(1 - e_i)^{\frac{3}{2}}} \tag{33}$$

This equation has analytical solution. As a result, the true anomaly u_t can be determined from the mean anomaly M. However, the analytical solution contains a nonlinear term that needs to be determined by iterating. First, we define the mean anomaly $M_{i,t}$ as

$$M_{i,t} = \frac{2\pi}{P_i}(t - \tau_i), \tag{34}$$

where τ_i is the time of periastron passage of the planet i and P_i is the period of its orbit. Then, through the Kepler's equation,

$$M_{i,t} = E_{i,t} - e_i \sin E_{i,t}, \tag{35}$$

where $E_{i,t}$ is the eccentric anomaly. Equation (35) has no analytic solution and it must be solved by an iterative procedure. A Newton–Raphson method is typically used to find the roots of this equation [20]. For certain sets of parameters, this iterative procedure can be particularly slow. We also have

$$\tan \frac{u_{i,t}}{2} = \sqrt{\frac{1+e_i}{1-e_i}} \tan \frac{E_{i,t}}{2}, \tag{36}$$

The variable of interest $\boldsymbol{\theta}$ is then the vector

$$\boldsymbol{\theta} = [V_0, A_1, \omega_1, e_1, P_1, \tau_1, \dots, A_S, \omega_S, e_S, P_S, \tau_S], \tag{37}$$

Then, for a single object (e.g., a planet or a natural satellite), the dimension of $\boldsymbol{\theta}$ is $M = 5 + 1 = 6$, with two objects the dimension of $\boldsymbol{\theta}$ is $M = 11$ etc.

This example consists in a synthetic radial velocity curve of a planetary system with one planet or two planets (i.e., $S = 1$ or $S = 2$). More specifically, we generate simulated data with a model with two planets. The orbital parameters of the planets are listed in Table 3, where P is the period of the orbit, A is the amplitude of the curve, e is the eccentricity of the orbit, ω is the argument of perigee, and τ is the last periastron passage. A mean velocity $V_0 = 5\ \mathrm{m\,s^{-1}}$ is assumed. A Gaussian noise perturbation is added with a standard deviation $\sigma = 3\ \mathrm{m\,s^{-1}}$. To simulate observations, a total of $K = 120$ data points are selected from three random time periods (and two planets in the system). Note that the amplitude of the radial velocity curve of the second planet is close to the noise level. We run ATAIS and a standard AIS scheme with the model with one planet and with the model with two planets. The purpose of this simulation is to check the ability of the method to detect the two planets (by approximating the model evidence).

Table 3. Main orbital parameters of the two exoplanets in the simulation.

Parameter	Planet 1	Planet 2
P	15 d	115 d
A	$25\ \mathrm{m\,s^{-1}}$	$5\ \mathrm{m\,s^{-1}}$
e	0.1	0.0
ω	0.61 rad	0.17 rad
τ	3 d	24 d

We apply ATAIS and a standard AIS scheme [13] over the space $[\boldsymbol{\theta}, \sigma]$ for approximating the model evidence $Z = p(\mathbf{y})$ (marginal likelihood) of both models (one planet or two planets) with the given data (generated considering two planets). Uniform priors are considered for each parameter: $P \in [0, 365]$, $A \in [-20, 20]$, $e \in [0, 1]$, $\omega \in [0, 2\pi]$, and $\tau \in [0, 50]$ (moreover, $\sigma \in [0, 30]$ for the standard AIS scheme). The ATAIS algorithm and the standard AIS scheme have been run with $N = 10^6$ and $T = 50$ iterations for both the model with one planet and the model with two planets. In both cases, we consider the same Gaussian proposal with a starting standard deviation of 5 for each component (note that the standard AIS scheme works in higher dimensional space due the inference over σ). To decide which model is more probable, the model evidence Z of each model is estimated. More specifically, we approximate the one-planet model $\widehat{Z}_1 = \widehat{p}_1(\mathbf{y})$ and the two-planet model $\widehat{Z}_2 = \widehat{p}_2(\mathbf{y})$ with the ATAIS algorithm and the standard AIS scheme. When $\widehat{Z}_1 > \widehat{Z}_2$, we select the first model; otherwise, if $\widehat{Z}_1 < \widehat{Z}_2$, we select the second one. The true model is the two-planet model, as the simulated data were generated from that model. After

500 independent runs, the percentage of correct detection of the true model for ATAIS is $\approx 98\%$, whereas with the standard AIS scheme is only $\approx 56\%$. This is due to the difficulty of making inference jointly over $[\boldsymbol{\theta}, \sigma]$. Let us denote the Bayesian factor as $B = Z_2/Z_1$. In ATAIS, the expected value of the ratio between the model evidences (averaged over the 500 runs) is $E[B] \approx 5 \cdot 10^3$ with a relative variance of $\frac{E[(B-E[B])^2]}{E[B]^2} \approx 0.04$. In the case of the standard AIS, we have $E[B] \approx 16.32$ and $\frac{E[(B-E[B])^2]}{E[B]^2} \approx 0.15$. Therefore, for ATAIS, the model with two planets is clearly more probable than the model with one planet.

The fitted curves, corresponding to the vector of parameters $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ obtained with ATAIS, are shown in Figure 5. From the figure, it is not clear which model better fits the simulated observations (blue points), although the model with two planets seems to better fit the observations in the time period from 200 to 300 days. The values of $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$, obtained in one specific run by ATAIS, are given in Table 4. We notice that ω and τ are highly correlated and more iterations may be needed to obtain the actual global maximum, but the remaining parameters obtained from $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ are similar to the simulated values. In addition, the amplitude of the curve of the second planet is close to the intensity of the noise, making it difficult to derive the best fit for that planet. Summarizing, our results show the method is able to discriminate between a model with one planet (with six dimensions of the inference problem) and a model with two planets (with 11 dimensions of the inference problem), for this particular simulation. Finally, the evolution of the automatic tempering parameter $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ is shown in Figure 6. The dashed line is the evolution of $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ for the single-planet model, whereas the continuous line is the evolution of $\widehat{\sigma}_{\texttt{ML}}^{(t)}$ for the model with two planets. In this second model, the tempering parameter reaches a smaller value, as expected.

Table 4. The value of $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ and the variances of the marginal posteriors for the 2-planets model (with $K = 120$ data points).

Parameter	**Planet 1**		**Planet 2**	
	$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$	**Var($\boldsymbol{\theta}$| y)**	$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ **—Planet 2**	**Var($\boldsymbol{\theta}$| y)**
P	14.99 d	0.18	110.39 d	11.28
K	23.78 m s^{-1}	0.52	3.50 m s^{-1}	0.44
e	0.05	0.047	0.00	0.003
ω	7.69 rad	0.61	0.68 rad	0.82
τ	6.8 d	0.76	7.96 d	20.31

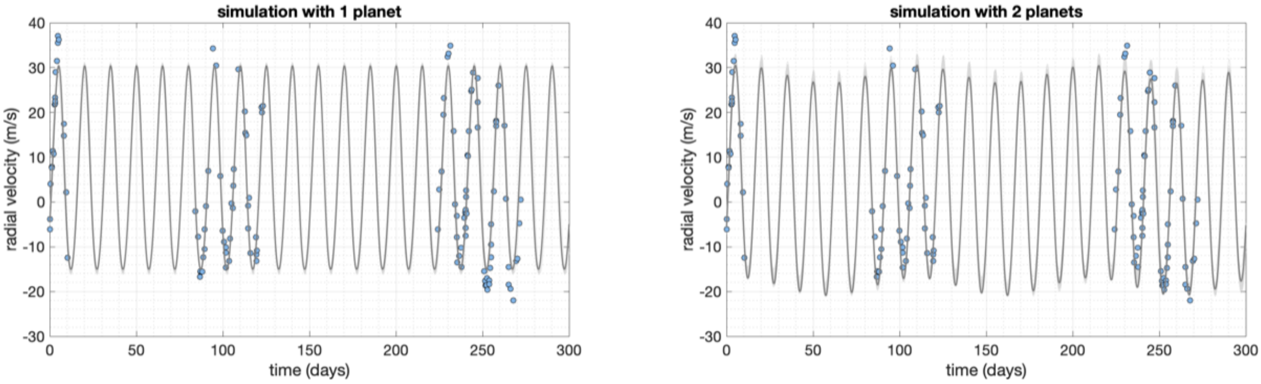


Figure 5. Comparison of the results of the ATAIS algorithm with the simulations (blue dots). Left panel shows, in gray, the radial velocity curve for $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ using a model with one planet. Right panel is like left panel but considering a model with two planets.

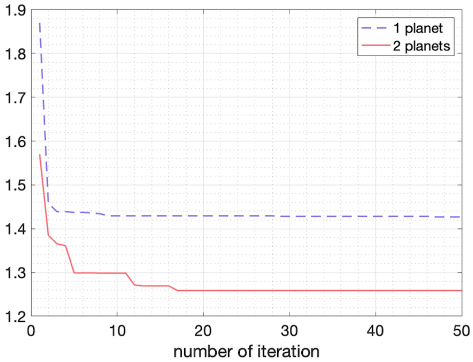


Figure 6. Evolution of the tempering parameter $\widehat{\sigma}_{\mathtt{ML}}^{(t)}$. We decide $\widehat{\sigma}_{\mathtt{ML}}^{(0)} = 50$ as starting value (the figure shows from $t = 1$), which is an arbitrary high value to help the exploration in the first iteration. However, after the first iteration, the algorithm is able to obtain reasonable values of $\widehat{\sigma}_{\mathtt{ML}}^{(1)}$. The dashed line is the evolution for the model with one planet. The continuous line is the evolution of the two-planet model.

7. Conclusions

We have proposed a novel AIS scheme for Bayesian inversion problems where an automatic tempering procedure is implemented (called ATAIS). The inference of the variables of interest $\boldsymbol{\theta}$ and the noise power σ^2 is divided. A sampling strategy is considered for $\boldsymbol{\theta}$ and an optimization approach is employed for σ^2. Thus, ATAIS performs an iterative procedure, alternating sampling and optimization steps. Therefore, the proposed scheme deals with a sequence of tempered posteriors according to the current estimation of the noise power. We have also discussed the possibility of approximating the marginal posterior of σ without additional evaluations of the complex model. Furthermore, the complete Bayesian analysis regarding the complete joint posterior is possible as discussed in Section 5, again without any additional evaluations of the likelihood function.

Several simulations are provided and the application to a sophisticated astronomical model has been considered, where the number of planets in the system is detected by the analysis of the marginal likelihood. The results show the benefits of the proposed scheme. For instance, in the astronomical example, the percentage of correct detection of the true model obtained by ATAIS is ≈98%, whereas with the standard AIS scheme is only ≈56%. As future research, we plan to extend the ATAIS scheme in order to deal with an observation model with correlated noise perturbations (for instance, using a Gaussian Process). Moreover, the use of parallel AIS schemes (or MCMC algorithms) will be also considered. A combination of parallel MCMC chains and AIS schemes can be found in the so-called layered AIS method and other similar approaches [21,22]. This idea seems particularly interesting for improving the inference with radial velocity models.

Author Contributions: All the authors contribute to all the sections in the same way. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Office of Naval Research (award no. N00014-19-1-2226), the Spanish Ministry of Science and Innovation (RTI2018-099655-B- I00 CLARA and PID2019-105032GB-I00 SPGRAPH), the Foundation by the Community of Madrid in the framework of the Multiannual Agreement with the Rey Juan Carlos University in line of action 1, Encouragement of Young Ph.D. students investigation Project under Grant F661, and the regional Government of Madrid (Comunidad de Madrid, reference Y2018/TCS-4705 PRACTICO).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. On the Optimization of the Likelihood Function

Let us set $\delta = \sigma^2$ and consider to optimize of the likelihood function

$$\ell(\boldsymbol{\theta}, \delta) = \frac{1}{(2\pi\delta)^{K/2}} \exp\left(-\frac{V(\boldsymbol{\theta})}{\delta}\right).$$

Recall that, in our model, we have $V(\boldsymbol{\theta}) = ||\, \mathbf{y} - \ \mathbf{f}(\boldsymbol{\theta})||^2$. We desire to obtain

$$[\boldsymbol{\theta}_{\texttt{ML}}, \delta_{\texttt{ML}}] = \arg\max \ell(\boldsymbol{\theta}, \delta).$$

We can write the gradient and equal to zero,

$$\begin{cases} \nabla_\theta \ell(\boldsymbol{\theta}, \delta) = -\dfrac{1}{\delta} \nabla_\theta V(\boldsymbol{\theta}) \left[\dfrac{1}{(2\pi\delta)^{K/2}} \exp\left(-\dfrac{V(\boldsymbol{\theta})}{\delta}\right) \right] = \mathbf{0} \Longrightarrow \nabla_\theta V(\boldsymbol{\theta}) = \mathbf{0}, \\ \dfrac{\partial \ell(\boldsymbol{\theta}, \delta)}{\partial \delta} = \dfrac{\mathrm{e}^{-\frac{V(\boldsymbol{\theta})}{\delta}} (2V(\boldsymbol{\theta}) - \delta K)}{2^{\frac{K}{2}+1} \delta^{\frac{K}{2}+2} \pi^{K/2}} = 0 \Longrightarrow \delta = \dfrac{2}{K} V(\boldsymbol{\theta}). \end{cases} \tag{A1}$$

We have obtained that the ML solution is defined by the system of equations,

$$\begin{cases} \nabla_\theta V(\boldsymbol{\theta}_{\texttt{ML}}) = \mathbf{0} \\ \delta_{\texttt{ML}} = \dfrac{2}{K} V(\boldsymbol{\theta}_{\texttt{ML}}). \end{cases} \tag{A2}$$

References

1. Fitzgerald, W.J. Markov chain Monte Carlo methods with applications to signal processing. *Signal Process.* **2001**, *81*, 3–18. [CrossRef]
2. Andrieu, C.; de Freitas, N.; Doucet, A.; Jordan, M. An Introduction to MCMC for Machine Learning. *Mach. Learn.* **2003**, *50*, 5–43. [CrossRef]
3. Martino, L.; Míguez, J. Generalized Rejection Sampling Schemes and Applications in Signal Processing. *Signal Process.* **2010**, *90*, 2981–2995. [CrossRef]
4. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*; Springer: New York, NY, USA, 2004.
5. Liu, J.S. *Monte Carlo Strategies in Scientific Computing*; Springer: New York, NY, USA, 2004.
6. Martino, L.; Luengo, D.; Miguez, J. *Independent Random Sampling Methods*; Springer: New York, NY, USA, 2018.
7. Kirkpatrick, S., Jr.; Gelatt, C.D.; Vecchi, M.P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680. [CrossRef] [PubMed]
8. Marinari, E.; Parisi, G. Simulated Tempering: A New Monte Carlo Scheme. *Europhys. Lett.* **1992**, *19*, 451–458. [CrossRef]
9. Friel, N.; Pettitt, A.N. Marginal Likelihood Estimation via Power Posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 589–607. [CrossRef]
10. Moral, P.D.; Doucet, A.; Jasra, A. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2006**, *68*, 411–436. [CrossRef]
11. Neal, R.M. Annealed importance sampling. *Stat. Comput.* **2001**, *11*, 125–139. [CrossRef]
12. Llorente, F.; Martino, L.; Delgado, D.; Lopez-Santiago, J. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *arXiv* **2020**, arXiv:2005.08334.
13. Bugallo, M.F.; Martino, L.; Corander, J. Adaptive importance sampling in signal processing. *Digit. Signal Process.* **2015**, *47*, 36–49. [CrossRef]
14. Rubin, D.B. Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3, Ads Bernardo, Degroot, Lindley, and Smith*; Oxford University Press: Oxford, UK, 1988.
15. Gelman, A.; Meng, X.L. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*; John Wiley & Sons: New York, NY, USA, 2018.
16. Gregory, P.C. Bayesian re-analysis of the Gliese 581 exoplanet system. *Mon. Not. R. Astron. Soc.* **2011**, *415*, 2523–2545. [CrossRef]
17. Barros, S.C.C.; Brown, D.J.A.; Hébrard, G.; Gómez Maqueo Chew, Y.; Anderson, D.R.; Boumis, P.; Delrez, L.; Hay, K.L.; Lam, K.W.F.; Llama, J.; et al. WASP-113b and WASP-114b, two inflated hot Jupiters with contrasting densities. *Astron. Astrophys.* **2016**, *593*, A113. [CrossRef]
18. Affer, L.; Damasso, M.; Micela, G.; Poretti, E.; Scand ariato, G.; Maldonado, J.; Lanza, A.F.; Covino, E.; Garrido Rubio, A.; González Hernández, J.I.; et al. HADES RV program with HARPS-N at the TNG. IX. A super-Earth around the M dwarf Gl 686. *Astron. Astrophys.* **2019**, *622*, A193. [CrossRef]

19. Trifonov, T.; Stock, S.; Henning, T.; Reffert, S.; Kürster, M.; Lee, M.H.; Bitsch, B.; Butler, R.P.; Vogt, S.S. Two Jovian Planets around the Giant Star HD 202696: A Growing Population of Packed Massive Planetary Pairs around Massive Stars? *Astron. J.* **2019**, *157*, 93. [CrossRef]
20. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes in C++: The Art of Scientific Computing*; Springer: New York, NY, USA, 2002.

21. Martino, L.; Elvira, V.; Luengo, D.; Corander, J. Layered Adaptive Importance Sampling. *Stat. Comput.* **2017**, *27*, 599–623. [CrossRef]
22. Botev, Z.I.; Ecuyer, P.L.; Tuffin, B. Markov chain importance sampling with applications to rare event probability estimation. *Stat. Comput.* **2013**, *23*, 271–285. [CrossRef]