On the safe use of prior densities for Bayesian model selection

F. Llorente[†], L. Martino[⋆], E. Curbelo[†], J. Lopez-Santiago[†], D. Delgado[†]

[†] Universidad Carlos III de Madrid, Leganés (Spain).

[⋆] Universidad Rey Juan Carlos, Fuenlabrada (Spain).

Abstract

The application of Bayesian inference for the purpose of model selection is very popular nowadays. In this framework, models are compared through their marginal likelihoods, or their quotients, called Bayes factors. However, marginal likelihoods show strong dependence on the prior choice, even when the data are very informative, unlike the posterior distribution. Furthermore, when the prior is improper, the marginal likelihood of the corresponding model is undetermined. In this work, we aim to raise awareness about the issue of prior sensitivity of the marginal likelihood and its role in model selection. We also comment on the use of uninformative priors, which are very common choices in practice. Several practical suggestions are provided and possible solutions allowing the use of improper priors are discussed. The connection between the marginal likelihood approach and the well-known information criteria is also presented. We describe all the issues and possible solutions by illustrative numerical examples (providing some related code). One of them involving a real-world application on exoplanet detection.

Keywords: Model selection, Marginal likelihood, Bayesian evidence, improper priors, information criteria, BIC, AIC, posterior predictive.

1 Intro

In the last decades, we observe a growing trend in the use of Bayesian approaches to the problem of inferring the parameters of physical models describing natural processes. Although Bayesian inference has historically been used (e.g. [1, 2]), it is only now becoming more widespread. Nowadays, we can find applications of Bayesian inference methods in fields such as remote sensing [3, 4], astronomy [5, 6, 7, 8], cosmology [9, 10], or optical spectroscopy [11, 12].

One of the most common problems we may encounter in Bayesian inference is that of model selection. For this purpose, the determination of the Bayes factor is often used. This involves the approximation of the *Bayesian evidence*, a.k.a., *marginal likelihood*, of the several models. The marginal likelihood shows a strong dependence on the choice of the prior probability density functions (pdfs). Many papers propose very uninformative (usually uniform) prior pdfs, in order

to avoid biasing the exploration of the parameter space (see, e.g., [13]). In some cases, the selected prior pdfs are diffuse or even *improper* [14].

In a first part of this work, we show some issues in Bayesian model selection (or hypothesis testing) based on the marginal likelihood computation [15, 16, 17]. First of all, the results can be *strongly* affected by the choice of the prior. Indeed, the marginal likelihood is very sensitive to the variations on the prior density, much more than the corresponding posterior distribution. Secondly, this issue becomes even more dramatic when improper priors are employed: the Bayesian inference with improper priors is allowed if the corresponding posterior proper, whereas Bayesian model selection with improper priors is *not* allowed/possible, due to the fact the marginal likelihood is actually not completely specified (it is defined up to an arbitrary constant). We describe all these issues by mathematical considerations and several and illustrative numerical examples. One of them involves a real-world application for detecting exo - objects (orbiting other stars) based on a radial velocity model.
Furthermore, in a second part of this work, we show some possible solutions presented in the literature, such the *partial and/or intrinsic* Bayes factors [15], remarking potential benefits and possible drawbacks. An alternative to the marginal likelihood approach for Bayesian model selection, called *posterior predictive* framework [18, Ch. 6][19], is also described. Finally, the relationship between the information criteria [20], such as Bayesian-Schwarz information criterion (BIC) Akaike information criterion (AIC), and the marginal likelihood approach is discussed in Appendix A. Therefore, the contribution is twofold: we provide (a) a gentle guide for interested practitioners (with several warnings and advices), and (b) and a useful work for more expert researchers looking for practical solutions and/or possible alternatives. Some related code is also provided (see Section 5).

2 Background

2.1 Problem statement

In many applications, the goal is to make inference about a variable of interest, $\boldsymbol{\theta} = \theta_{1:D_\theta} = [\theta_1, \theta_2, \ldots, \theta_{D_\theta}] \in \Theta \subseteq \mathbb{R}^{D_\theta}$, where $\theta_d \in \mathbb{R}$ for all $d = 1, \ldots, D_\theta$, given a set of observed measurements, $\mathbf{y} = [y_1, \ldots, y_{D_y}] \in \mathbb{R}^{D_y}$. In the Bayesian framework, one complete model $\mathcal{M}$ is formed by a likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$ and a prior probability density function (pdf) $g(\boldsymbol{\theta}|\mathcal{M})$. All the statistical information is summarized by the posterior pdf, i.e.,

$$\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}) g(\boldsymbol{\theta}|\mathcal{M})}{p(\mathbf{y}|\mathcal{M})}, \tag{1}$$

where

$$Z = p(\mathbf{y}|\mathcal{M}) = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}) g(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}, \tag{2}$$

is the so-called *marginal likelihood*, a.k.a., *Bayesian evidence* [1, 2]. This quantity is important for model selection purpose, as we show below. However, usually $Z = p(\mathbf{y}|\mathcal{M})$ is unknown and

difficult to approximate, so that in many cases we are only able to evaluate the unnormalized target function,

$$\pi(\boldsymbol{\theta}|\mathbf{y},\mathcal{M}) = \ell(\mathbf{y}|\boldsymbol{\theta},\mathcal{M})g(\boldsymbol{\theta}|\mathcal{M}) \propto \bar{\pi}(\boldsymbol{\theta}|\mathbf{y},\mathcal{M}). \tag{3}$$

Model Selection and testing hypotheses. Let us consider now M possible models (or hypotheses), $\mathcal{M}_1, ..., \mathcal{M}_M$, with prior probability mass $p_m = \mathbb{P}(\mathcal{M}_m)$, $m = 1, ..., M$. Note that, we can have variables of interest $\boldsymbol{\theta}^{(m)} = [\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_{D_m}^{(m)}] \in \Theta_m \in \mathbb{R}^{D_m}$, with possibly different dimensions in the different models. The posterior of the m-th model is given by

$$p(\mathcal{M}_m|\mathbf{y}) \;=\; \frac{p_m p(\mathbf{y}|\mathcal{M}_m)}{p(\mathbf{y})} \propto p_m Z_m \tag{4}$$

where $Z_m = p(\mathbf{y}|\mathcal{M}_m) = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta},\mathcal{M}_m)g(\boldsymbol{\theta}|\mathcal{M}_m)d\boldsymbol{\theta}$, and $p(\mathbf{y}) = \sum_{m=1}^{M} p(\mathcal{M}_m)p(\mathbf{y}|\mathcal{M}_m)$. Moreover, the ratio of two marginal likelihoods

$$\mathrm{BF}_{mm'} = \frac{Z_m}{Z_{m'}} = \frac{p(\mathbf{y}|\mathcal{M}_m)}{p(\mathbf{y}|\mathcal{M}_{m'})} = \frac{p(\mathcal{M}_m|\mathbf{y})/p_m}{p(\mathcal{M}_{m'}|\mathbf{y})/p_{m'}}, \tag{5}$$

also known as *Bayes factors*, represents the posterior to prior odds of models m and m'. If some quantity of interest is common to all models, the posterior of this quantity can be studied via *model averaging* [21], i.e., a complete posterior distribution as a mixture of M partial posteriors linearly combined with weights proportionally to $p(\mathcal{M}_m|\mathbf{y})$ (see, e..g, [22, 23]). Therefore, in all these scenarios, we need the computation of Z_m for all $m = 1, ..., M$.

Remark 1. *For the sake of simplicity, hereafter we skip the dependence on $\mathcal{M}_m$ in the notation. For instance, we denote the posterior density as $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and the marginal likelihood as $Z = p(\mathbf{y})$. Thus, we write*

$$Z = \int_\Theta \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{6}$$

Remark 2. *From Eq. (6), we can see clearly that Z is an average of likelihood values $\ell(\mathbf{y}|\boldsymbol{\theta}')$, weighted according to the prior pdf $g(\boldsymbol{\theta}')$.*

Clearly, the results of the Bayesian inference depends on the choice of the prior density and the actual number of data D_y.

2.2 Type of prior densities

The prior distribution in Bayesian inference should express the belief about the quantity of interest before that some data are observed. In this case, the prior is often defined as *informative*. An informative prior pdf can be determined from previous information, past experiments or by other sources of information (different from the observation model). As an alternative, when a family of conjugate priors exists, choosing a prior from that family simplifies calculation of the posterior distribution. However, in many scenarios, additional information and conjugate

priors are not available. Therefore, *uninformative* priors are employed. Uninformative priors can express "objective" information such as "the variable is positive" or "the parameter is less than some threshold value". Below, we describe some classes of uninformative priors.

Diffuse priors. The simplest idea for determining a non-informative prior is to assign equal probabilities to all possible outcomes, such as uniform densities in bounded support. In un unbounded domain, one can employ *vague* priors, i.e., densities with probability mass spread in all the state space, with a great scale parameter. A more extreme alternative is to use *improper* prior when it is possible (see the description below).

Improper priors. The use of improper priors, i.e., such that $\int_\Theta g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, is allowed for inference when $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, since the corresponding posteriors are proper. However, this is an issue for the model selection using the marginal likelihood Z. Indeed, the prior $g(\boldsymbol{\theta}) = ch(\boldsymbol{\theta})$ is not completely specified, since $c > 0$ is arbitrary. Some possible solutions are given in Section 4.3. A uniform density over an unbounded domain is a clear example of improper prior.

Reference and Jeffreys priors. Priors densities can also be designed according to some other principle such as invariance after transformations, symmetry or maximizing entropy given some constraints. Examples of this family are the *reference* priors or *Jeffreys* priors [1, 24]. Often, they are also *improper* priors. An example is $g(\sigma) \propto 1/\sigma$ for $\sigma > 0$ which is a Jeffreys improper prior, which is usually applied for a variable that represents a standard deviation.

Below we discuss how the choice of the prior affects **(a)** the inference of $\boldsymbol{\theta}$, and **(b)** the estimation of the Bayesian evidence Z for the model selection problem.

3 Dependence on the choice of the prior density

In this section, we show that the marginal likelihood Z is highly sensitivity to the choice of prior density (even with strong data) [25]. It is also more sensitive than the posterior $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ to variations on the prior density. Here, we first show all the values that the evidence Z can take changing the prior pdf and then we discuss its sensitivity. Finally, we describe further issues in the use of improper priors.

Bounds of the evidence Z. Let us denote the maximum and minimum value of the likelihood function as $\ell_{\min} = \ell(\mathbf{y}|\boldsymbol{\theta}_{\min}) = \min_{\boldsymbol{\theta}\in\theta} \ell(\mathbf{y}|\boldsymbol{\theta})$, and $\ell_{\max} = \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}) = \max_{\boldsymbol{\theta}\in\theta} \ell(\mathbf{y}|\boldsymbol{\theta})$, respectively. Note that

$$Z = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}) \int_\Theta g(\boldsymbol{\theta})d\boldsymbol{\theta} = \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}).$$

Similarly, we can obtain $Z \geq \ell(\mathbf{y}|\boldsymbol{\theta}_{\min})$. The maximum and minimum value of Z are reached, for instance, with two degenerate choices of the prior, $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\max})$ and $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\min})$.

Hence, for every other choice of $g(\boldsymbol{\theta})$, we have

$$\ell(\mathbf{y}|\boldsymbol{\theta}_{\min}) \leq Z \leq \ell(\mathbf{y}|\boldsymbol{\theta}_{\max}). \tag{7}$$

Namely, depending on the choice of the prior $g(\boldsymbol{\theta})$, we can have any value of Bayesian evidence contained in the interval $[\ell(\mathbf{y}|\boldsymbol{\theta}_{\min}), \ell(\mathbf{y}|\boldsymbol{\theta}_{\max})]$.
The two possible extreme values correspond to the worst and the best model fit, respectively. We can obtain $Z = \ell(\mathbf{y}|\boldsymbol{\theta}_{\min})$ with the choice $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\min})$ (which applies the greatest possible penalty to the model), and we obtain $Z = \ell(\mathbf{y}|\boldsymbol{\theta}_{\max})$, with the choice $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\max})$ (which does not apply any penalization to the model complexity, i.e., we have the maximum overfitting). Indeed, $Z = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is by definition an average of the likelihood values weighted according to the prior.

Remark 3. *Depending on the choice of the prior, the evidence Z can take any possible value in the interval $[\ell(\mathbf{y}|\boldsymbol{\theta}_{min}), \ell(\mathbf{y}|\boldsymbol{\theta}_{max})]$. Hence, in this sense, the choice of the prior is equivalent to choice a penalization term for the model complexity.*

Note that Remark 3 above it is strictly connected to Remark 2. For the relationship with the well-known Bayesian-Schwarz information criterion (BIC) and the Akaike information criterion (AIC), see Appendix A.

Consistency of Bayesian inference. Here, we consider a fixed observation model and a fixed prior but we vary the number of data D_y. More specifically, we focus on the behavior of the inference and model selection as $D_y \to \infty$. Just as an example, let us consider two Bayesian point estimators, such as the minimum mean square error (MMSE) estimator

$$\widehat{\boldsymbol{\theta}}_{\texttt{MMSE}} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} \bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \tag{8}$$

and the maximum-a-posteriori (MAP) estimator

$$\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \bar{\pi}(\boldsymbol{\theta}|\mathbf{y}). \tag{9}$$

Both estimators depends on the posterior density, and as a consequence on the choice of the prior pdf. However, as the number of data grows, $D_y \to \infty$, under mild conditions, both estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MMSE}}$ and $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ and are consistent, i.e., they convergence to the true value of the parameters $\boldsymbol{\theta}_{\texttt{true}}$ (recovering frequentist arguments). This is due to, under mild conditions (see Bernstein-von Mises theorem [1, 2, 25]), the probability mass represented by the posterior $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ becomes more concentrate around the true value of the parameter, as the number of data grows $D_y \to \infty$. This means that for large amounts of data, one can use the posterior distribution to make, from a frequentist point of view, valid statements about estimation and uncertainty.
Similarly, under the assumption that one of $\mathcal{M}_m$ is the true generating model, the Bayes factor (i.e., the marginal likelihood approach) will choose the correct model as the number of data grows, $D_y \to \infty$ [26].

Remark 4. *Therefore, as number of data grows, $D_y \to \infty$, we can reduce the dependence on the prior and obtain the correct results. However, the marginal likelihood Z is particularly sensitive to changes of the prior pdf, as we show in the next section.*

Robustness of the Bayesian inference. Here, we consider a fixed and finite number of data D_y, but we assume variations in the choice of the prior density. The goal is to analyze the robustness of the parameter and evidence estimations. Below, by means of illustrative examples, we show that the posterior $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is more robust under prior changes and/or variations, with respect the marginal likelihood Z, as we state in the following remark.

Remark 5. *Both the posterior density $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and the marginal likelihood $Z = p(\mathbf{y})$ depend on the prior choice. However, unlike the posterior density, the marginal likelihood even with strong data is highly sensitivity to the choice of prior density. Namely, under the assumption of strong data and if we vary the prior density, the estimators $\widehat{\boldsymbol{\theta}}_{MMSE}$, $\widehat{\boldsymbol{\theta}}_{MAP}$ does not change drastically, where the marginal likelihood Z can suffer significant variations [27, 25].*

Remark 6. *Diffuse priors tend to produce smaller values of the marginal likelihood Z. This is due to the integration would consider many values of $\boldsymbol{\theta}$ that do not explain well the data, i.e., the likelihood at $\boldsymbol{\theta}$ is small. Hence, a good model can display a low value of Z only because we choose a prior that is very spread out. Conversely, a worse model can display a bigger value of Z due to choosing a concentrated prior [28, 15].*

The next illustrative example shows the robustness of the posterior (and the corresponding estimators) and the sensitivity of the evidence Z, under prior changes.

Illustrative example 1. Let us consider the following Gaussian conjugate model for θ,

$$\ell(\mathbf{y}|\theta) = \mathcal{N}(\mathbf{y}|\theta, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(y_i|\theta, \sigma^2)$$
$$g(\theta) = \mathcal{N}(\theta|\mu_0, \sigma_0^2).$$

Hence, the posterior is also Gaussian, $\bar{\pi}(\theta|\mathbf{y}) = \mathcal{N}(\theta|\mu_{\text{post}}, \sigma^2_{\text{post}})$, where

$$\mu_{\text{post}} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2} \right)$$
$$\sigma^2_{\text{post}} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right),$$

where $\bar{y}$ denotes the sample mean of $\mathbf{y}$. The marginal likelihood is given by

$$Z = (2\pi n \sigma_n^2)^{-\frac{n}{2}} \left(\frac{\sigma_0^2}{\sigma_n^2} + 1 \right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\frac{v_y + \bar{y}^2}{\sigma_n^2} + \frac{\mu_0^2}{\sigma_0^2} - \frac{1}{\frac{1}{\sigma_n^2} + \frac{1}{\sigma_0^2}} \left(\frac{\bar{y}}{\sigma_n^2} + \frac{\mu_0}{\sigma_0^2} \right)^2 \right) \right),$$

where $\sigma_n = \frac{\sigma}{\sqrt{n}}$ and v_y denotes the sample variance of $\mathbf{y}$. We consider a single data point ($n = 1$), $\mathbf{y} = y = 2.078$. We fix μ_0 and vary σ_0. In Figure 1, we show the corresponding posterior for

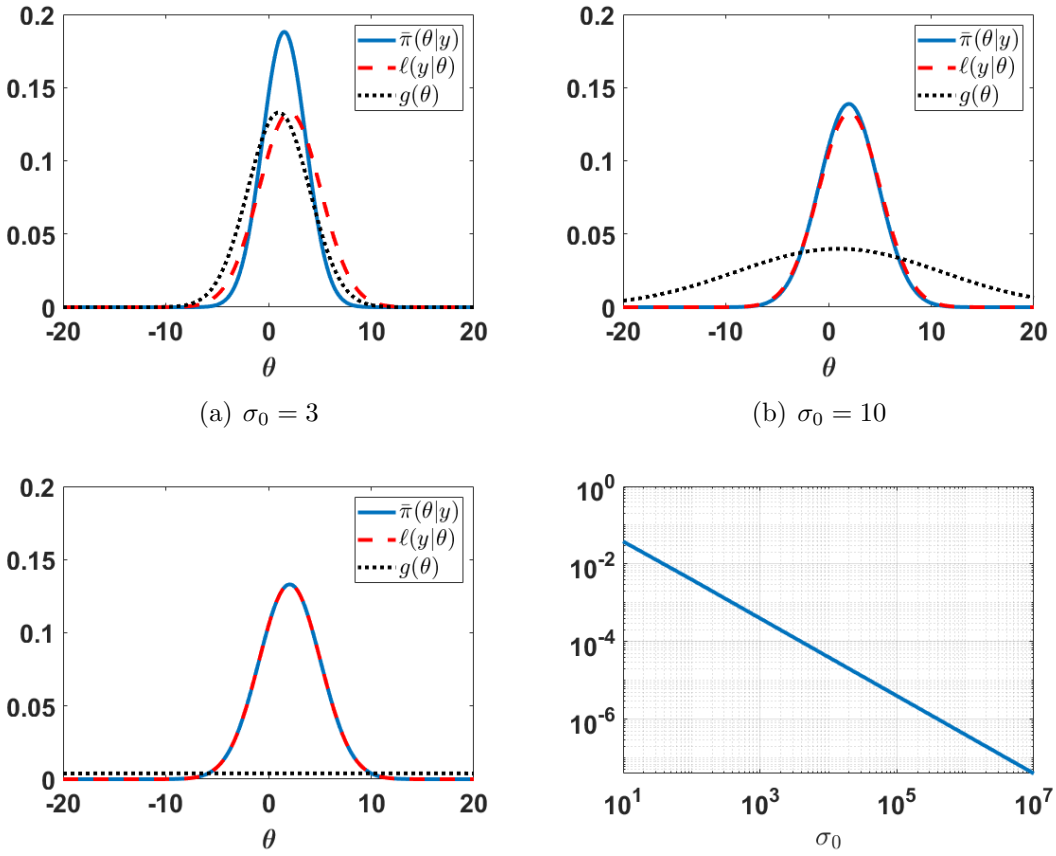
$\sigma_0 = 3, 10, 100$ in solid line, whereas the likelihood is depicted with dashed line and the prior is shown with dotted line. The evolution of the corresponding marginal likelihood Z versus σ_0 is given in Figure 1(d).
As σ_0 grows, the posterior pdf approaches the likelihood as depicted in Figures 1(a)-(b)-(c). Then, for big values of σ_0, the posterior is insensitive to the increasing of the prior dispersion. If we consider $\sigma_0 \to \infty$ (corresponding to an *improper* prior), the posterior pdf coincides with the likelihood function, and the inference (e.g., the estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MMSE}}$ and $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$) is completely driven by the observed data. In this example both estimators $\widehat{\boldsymbol{\theta}}_{\texttt{MMSE}}$ and $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ converges to the maximum of the likelihood function as $\sigma_0 \to \infty$. Note also that from Figure 1(a) to Figure 1(c) that variation of the posterior is also negligible.
On the contrary, as σ_0 grows, the marginal likelihood decreases approaching zero as shown in Figure 1(d) (instead of converging to the normalizing constant of the likelihood, as someone could expect). Based on this fact, many authors claim that the use of Bayes factor is not reliable. Next, we provide another illustrative example showing that using an increasingly diffuse prior makes us eventually select the wrong model over the true one.

The illustrative example 2 below shows how the sensitivity (with a fixed D_y we change the prior) in the estimation of Z can affect the model selection results. Moreover, we show also the consistency of the Bayesian factors fixing the prior and increasing the number of data D_y, i.e., $D_y \to \infty$.

Illustrative example 2. Consider two models, $\mathcal{M}_1 = \{\ell_1(y|\theta) = \theta^y e^{-\theta}/y!,\ g_1(\theta)\}$ and $\mathcal{M}_2 = \{\ell_2(y|\phi) = \phi(1-\phi)^y,\ g_2(\phi)\}$ with different priors $g_1(\theta)$ and $g_2(\phi)$. Suppose we have generated D_y independent data $\mathbf{y} = (y_1, \ldots, y_{D_y})$ from $\mathcal{M}_1$ with $\theta_{\text{true}} = 2$. We use a uniform prior $g_2(\phi) = 1$ for $\phi \in (0,1)$, and also a uniform prior $g_1(\theta) = \frac{1}{L}$ for $\theta \in (0, L)$. The goal of this example is to show the how the Bayes factor below

$$BF_{12} = \frac{Z_1}{Z_2} = \frac{\int_0^L \ell_1(\mathbf{y}|\theta)\frac{1}{L}d\theta}{\int_0^1 \ell_2(\mathbf{y}|\phi)d\phi} \tag{10}$$

is affected when L increases, i.e., $g_1(\theta)$ becomes more diffuse. We fix L, generate D_y data from model $\mathcal{M}_1$ and compute BF_{12}. Hence, the model $\mathcal{M}_1$ is the *correct* one, since we generate the data according to $\mathcal{M}_1$.
For each $L = 10^\alpha$ ($\alpha = 1, \ldots, 6$), we repeat this process in 100 different runs, and count *the number of errors* in model selection, i.e., whenever $BF_{12} < 1$ (note it should be greater than 1 since $\mathcal{M}_1$ is the true model). Table 1 shows the results when $D_y = 30$. Specifically, we show the maximum and minimum values of BF_{12}, obtained in the 100 simulations, along with the number of error. We observe that, as L increases, i.e., we use a more diffuse prior, the model $\mathcal{M}_2$ is (wrongly) selected more often. In fact, with $L = 10^6$, the Bayes factor always selects $\mathcal{M}_2$ over $\mathcal{M}_1$. Table 2 shows results when $D_y = 100$, where we observe that the number of errors is very low even for large L, namely, in this example having more data compensates the potential drawbacks of the use of diffuse priors. In addition, in Figure 2, we have computed the number of errors (over the 100 different runs) for fixed $L = 10^5$ versus the number of data D_y. We see that, for a given prior width, increasing D_y reduces the number of times we choose the wrong model. This example



(a) $\sigma_0 = 3$ (b) $\sigma_0 = 10$

(c) $\sigma_0 = 100$ (d) Marginal likelihood Z versus σ_0

Figure 1: In **(a)-(c)**, we show the posterior for a Gaussian prior $\mathcal{N}(\mu_0, \sigma_0^2)$ with three different choices of σ_0. In **(d)**, we show the corresponding marginal likelihood versus σ_0 in log-scale. Note that increasing σ_0 (i.e. prior is more diffuse) does not change the shape of the posterior, but the marginal likelihood is indeed decreasing.

shows that even if the number data grows D_y, diffuse priors can affect the results.
Clearly, keeping fixed the (proper) priors, and including the *enough* number of data D_y in our study, we can obtain the correct results (see Figure 2). However, the number of *enough* data is unknown and depends on the specific problem. Furthermore, the joint use of a huge amount of data often jeopardized the performance of the computational methods employed for estimating the evidence Z [15, 17].

Remark 7. *We have already discussed the sensitivity of Z to variations of the prior density. Even more caution is needed in the case of employing improper priors. We have seen that the use of improper priors, $\int_\Theta g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, is allowed for inference when $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, since the corresponding posteriors are proper. However, this is an issue for the model selection with Z. Indeed, the prior $g(\boldsymbol{\theta}) = ch(\boldsymbol{\theta})$ is not completely specified, since $c > 0$ is arbitrary. We discuss it and some possible solutions in the next sections.*

Table 1: Model comparison for $D_y = 30$. Minimum and maximum BF_{12} under true model $\mathcal{M}_1$ (Poisson) for 100 simulations

True model = $\mathcal{M}_1$ (with $\theta_{\text{true}} = 2$)			
L	min	max	Errors in model choice, over 100 simulations
10	0.094	4.77×10^{5}	3
10^2	0.059	2.49×10^{4}	15
10^3	0.0012	1.46×10^{3}	31
10^4	1.06×10^{-4}	339.86	67
10^5	1.02×10^{-4}	41.05	84
10^6	1.59×10^{-6}	0.7080	100

Table 2: Model comparison for $D_y = 100$. Minimum and maximum BF_{12} under true model $\mathcal{M}_1$ (Poisson) for 100 simulations

True model = $\mathcal{M}_1$ ($\theta_{\text{true}} = 2$)			
L	min	max	Errors in model choice, over 100 simulations
10	41.27	9.05×10^{13}	0
10^2	6.93	1.55×10^{13}	0
10^3	14.45	2.21×10^{11}	0
10^4	7.94×10^{-4}	3.75×10^{11}	3
10^5	0.5214	1.36×10^{12}	2
10^6	7.98×10^{-4}	2.07×10^{8}	7

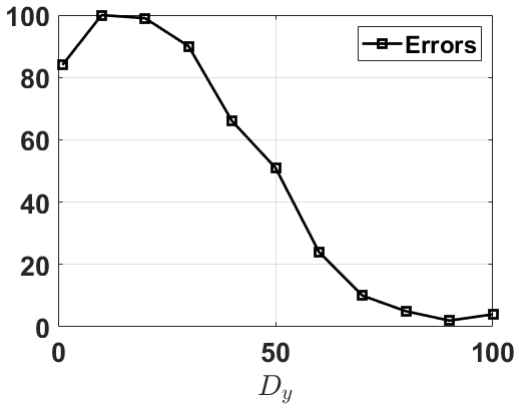


Figure 2: Number of errors in model selection, i.e., selecting the wrong model ($BF_{12} < 1$), out of 100 independent runs, when using $g_1(\theta) = \frac{1}{L},\ \theta \in (0, L)$ with $L = 10^5$ (i.e., fixing the prior), for different number of data D_y. We can see that keeping, fix the priors, as D_y grows we decide for the true model. However, fixing D_y, changing L we can always adulterate the result of the study penalizing more and the model 1, as shown in Tables 1-2.

3.1 Issues with improper priors for model selection

So far we have considered proper priors, i.e., $\int_\Theta g(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$. The use of improper priors is common in Bayesian inference to represent weak prior information. Consider $g(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta})$ where $h(\boldsymbol{\theta})$ is a non-negative function whose integral over the state space does not converge, $\int_\Theta g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_\Theta h(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$. In that case, $g(\boldsymbol{\theta})$ is not completely specified. Indeed, we can have different definitions $g(\boldsymbol{\theta}) = ch(\boldsymbol{\theta})$ where $c > 0$ is (the inverse of) the "normalizing" constant, not uniquely determinate since c formally does not exist. Regarding the parameter inference and posterior definition, the use of improper priors poses no problems as long as $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, indeed

$$\begin{aligned}\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{Z}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\ell(\mathbf{y}|\boldsymbol{\theta})ch(\boldsymbol{\theta})}{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})ch(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}},\\ &= \frac{1}{Z_h}\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta}) \end{aligned} \tag{11}$$

where $Z = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$, $Z_h = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $Z = cZ_h$. Note that the unspecified constant $c > 0$ is canceled out, so that the posterior $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is well-defined even with an improper prior if $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. However, the issue is not solved when we compare different models, since $Z = cZ_h$ depends on c. For instance, the Bayes factors depend on the undetermined constants $c_1, c_2 > 0$ [29],

$$\mathrm{BF}(\mathbf{y}) = \frac{c_1}{c_2}\frac{\int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta})h_1(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta})h_2(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{Z_1}{Z_2} = \frac{c_1 Z_{h_1}}{c_2 Z_{h_2}}, \tag{12}$$

so that different choices of c_1, c_2 provide different preferable models. There exists various approaches for dealing with this issue. Below we describe some relevant ones.

4 Safe approaches for Bayesian model selection

In a Bayesian inference, the best scenario is surely when the user has strong beliefs that can be translated into informative priors. When this additional information is not available, a careful strategy should be employed due to the dependence and sensitivity of the evidence Z with the prior choice $g(\boldsymbol{\theta})$.

We define as a safe scenario, an approach where the choice of the priors is virtually not favoring any of the models, and the results are not depending on some unspecified constant $c > 0$ (as in the case of using improper priors). Below, we describe some scenarios and some possible solutions for reducing, in some way, the dependence of the model comparison on the choice of the priors. In Section 4.4, we also discuss an alternative approach for model selection in Bayesian statistics [18, Ch. 6][19].

Same priors. Generally, we are interested in comparing two or more models. The use of the same (even improper) priors is suitable when the models have the same parameters (and hence also share the same support space). With this choice, the resulting comparison seems fair and

reasonable. However, this scenario is very restricted in practice. An example is when we have nested models, which share some common parameters. As noted in [26, Sect. 5.3], in the context of testing hypothesis, some authors have considered improper priors on nuisance parameters that appear on both null and alternative hypothesis. Since the nuisance parameters appear on both models, the multiplicative constants cancel out in the Bayes factor.

4.1 Hierarchical modeling

Hierarchical models are formed by multiple levels with the purpose of estimating also the *hyper*-parameters of the assumed prior densities. More specifically, additional prior pdfs (called often *hyper*-priors) over the the *hyper*-parameters of the priors are considered [30, 25]. Below, we provide just a summary of the new terms:

- Hyper-parameters: parameters of the prior distributions,
- Hyper-priors: prior distributions of hyper-parameters.

The underlying idea is to vary the hyper-parameters of the prior pdfs and performs different inference problems. Namely, fixing the hyper-parameters and studying the posterior, we have one inference problem. Then, we change the hyper-parameters and studying the corresponding posterior, we have another inference problem.
Let us consider now that our prior pdfs can be expressed as a parametric (or non-parametric) family of functions. We can vary the parameters in this family and even make inference on those variables. In this sense, we reduce the dependence on the choice of the prior, since we are not actually considering a unique prior *but* a family of them. Moreover, several authors claim hat the resulting model seems to be robust, with even the posterior distribution less sensitive to the more flexible hierarchical priors [25].
Mathematically speaking, let us denote $g(\boldsymbol{\theta}|\boldsymbol{\nu})$ our family of priors over $\boldsymbol{\theta}$ with hyper-parameters $\boldsymbol{\nu} \in \mathbb{R}^{\xi}$. Assuming and hyper-prior $h(\boldsymbol{\nu})$, the complete posterior is given by the following expression,

$$\bar{\pi}(\boldsymbol{\theta}, \boldsymbol{\nu}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\nu})h(\boldsymbol{\nu})}{Z_{\texttt{new}}}, \tag{13}$$

where

$$Z_{\texttt{new}} = p(\mathbf{y}) = \int_{\Theta}\int_{\mathbb{R}^{\xi}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\nu})h(\boldsymbol{\nu})d\boldsymbol{\theta}d\boldsymbol{\nu}, \tag{14}$$

$$= \int_{\mathbb{R}^{\xi}} Z(\boldsymbol{\nu})h(\boldsymbol{\nu})d\boldsymbol{\nu}, \tag{15}$$

is a Bayesian evidence that takes into account all the members of the prior family. Note that we have set $Z(\boldsymbol{\nu}) = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\nu})d\boldsymbol{\theta}$ following Eq. (6). Clearly, the model selection scheme based on $Z_{\texttt{new}}$ is more robust than a model selection approach based on a single marginal likelihood $Z = Z(\boldsymbol{\nu})$, considering only one possible value of $\boldsymbol{\nu}$ (i.e., only a unique prior). However, the

computation of $Z_{\texttt{new}}$ is more complex than the computation of a single $Z(\boldsymbol{\nu})$, since we have to approximate an higher dimensional integral [15]. Hence, this approach can be much more computational demanding. However, in some cases, it could a suitable solution: see, as an example, the numerical experiment in Section 5.3.

4.2 Likelihood-based priors

When $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, we can build a proper prior based on the data and the observation model. For instance, we can choose $g_{\text{like}}(\boldsymbol{\theta}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}$, then the marginal likelihood is

$$Z = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) g_{\text{like}}(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{\int_\Theta \ell^2(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{16}$$

We can consider $g_{\text{like}}(\boldsymbol{\theta})$ a non-invasive prior in the sense that does not incorporate any additional information, but is only based on the data. This idea is connected to *posterior predictive approach*, that is described in Section 4.4.

Less informative likelihood-based priors can be constructed using a tempering effect with a parameter $0 < \beta \leq 1$ or considering only a subset of data, denoted as $\mathbf{y}_{\text{sub}}$. For instance, when $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})^\beta d\boldsymbol{\theta} < \infty$ or $\int_\Theta \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, then we can choose $g_{\text{like}}(\boldsymbol{\theta}) \propto \ell(\mathbf{y}|\boldsymbol{\theta})^\beta$ or $g_{\text{like}}(\boldsymbol{\theta}) \propto \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})$, the marginal likelihood is

$$Z = \frac{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta+1}d\boldsymbol{\theta}}{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})^\beta d\boldsymbol{\theta}}, \quad \text{or} \quad Z = \frac{\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})\ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{17}$$

This is also the key idea underlying the partial and intrinsic Bayes factors described in the next section.

4.3 How to compute Bayes factors with improper priors

If one desire to use improper priors, a way to remove the unspecified constants c_m (where m denotes the subindex of the model) consists of building a likelihood-based prior (similarly we have described above) and compute/approximate its normalizing constant. Instead of using the complete likelihood, a partial likelihood approach (based only in a subset of data) has been suggested in the so called "Partial Bayes Factors" [31, Sect. 2].

Partial Bayes Factors [31, Sect. 2]. The idea behind the partial Bayes factors consists of using a subset of data to build proper priors and, jointly with the remaining data, they are used to calculate the Bayes factors. This is related to the likelihood-based prior approach, described above.
Let us consider a different improper prior $g_m(\boldsymbol{\theta}) = c_m h_m(\boldsymbol{\theta})$, for each model. The method starts by dividing the data in two subsets, $\mathbf{y} = (\mathbf{y}_{\text{train}}, \mathbf{y}_{\text{test}})$. The first subset $\mathbf{y}_{\text{train}}$ is used to obtain partial posterior distributions,

$$\bar{g}_m(\boldsymbol{\theta}|\mathbf{y}_{\text{train}}) = \frac{c_m}{Z^{(m)}_{\text{train}}} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}), \tag{18}$$

using the improper prior. The partial posterior $\bar{g}_m(\boldsymbol{\theta}|\mathbf{y}_{\text{train}})$ is then employed as prior. Note that

$$Z_{\text{train}}^{(m)} = c_m \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Recall that the complete posterior of m-th model is

$$\bar{\pi}_m(\boldsymbol{\theta}|\mathbf{y}) = \bar{\pi}_m(\boldsymbol{\theta}|\mathbf{y}_{\text{test}}, \mathbf{y}_{\text{train}}) = \frac{c_m}{Z_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}), \tag{19}$$

where Z_m is the standard marginal likelihood, i.e.,

$$Z_m = c_m \int_{\Theta_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Note that both $Z_{\text{train}}^{(m)}$ and Z_m both depend on the unspecified constant c_m. Considering the conditional likelihood $\ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}, \mathbf{y}_{\text{train}})$ of the remaining data $\mathbf{y}_{\text{test}}$, we can study another posterior density conditioned only to $\mathbf{y}_{\text{test}}$,

$$\bar{\pi}_{\text{test}}^{(m)}(\boldsymbol{\theta}|\mathbf{y}_{\text{test}}) = \frac{1}{Z_{\text{test|train}}^{(m)}} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}, \mathbf{y}_{\text{train}}) \bar{g}_m(\boldsymbol{\theta}|\mathbf{y}_{\text{train}}), \tag{20}$$

where $\bar{g}_m(\boldsymbol{\theta}|\mathbf{y}_{\text{train}})$ in (18) plays the role of a prior pdf.

Remark 8. *In case of conditional independence of the data given $\boldsymbol{\theta}$, we have $\ell_m(\mathbf{y}_{test}|\boldsymbol{\theta}, \mathbf{y}_{train}) = \ell_m(\mathbf{y}_{test}|\boldsymbol{\theta})$.*

Furthermore, we can write $Z_{\text{test|train}}^{(m)}$ as function of the standard evidence Z_m,

$$\begin{aligned}
Z_{\text{test|train}}^{(m)} &= \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}, \mathbf{y}_{\text{train}}) \bar{g}_m(\boldsymbol{\theta}|\mathbf{y}_{\text{train}}) d\boldsymbol{\theta}, \\
&= \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}, \mathbf{y}_{\text{train}}) \frac{c_m}{Z_{\text{train}}^{(m)}} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}, \\
&= \frac{c_m}{Z_{\text{train}}^{(m)}} \int_{\Theta_m} \ell_m(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}, \mathbf{y}_{\text{train}}) \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}, \\
&= \frac{c_m}{Z_{\text{train}}^{(m)}} \int_{\Theta_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}, \\
&= \frac{Z_m}{Z_{\text{train}}^{(m)}}.
\end{aligned} \tag{21}$$

Remark 9. *Note that $Z_{test|train}^{(m)}$ does not depend on c_m. In fact, we have*

$$\begin{aligned}
Z_{test|train}^{(m)} = \frac{Z_m}{Z_{train}^{(m)}} &= \frac{\int_{\Theta_m} \ell_m(\mathbf{y}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta_m} \ell_m(\mathbf{y}_{train}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \\
&= \frac{\int_{\Theta_m} \ell_m(\mathbf{y}_{test}, \mathbf{y}_{train}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta_m} \ell_m(\mathbf{y}_{train}|\boldsymbol{\theta}) h_m(\boldsymbol{\theta}) d\boldsymbol{\theta}},
\end{aligned} \tag{22}$$

i.e., in the numerator, we consider all the data whereas, in the denominator, only $\mathbf{y}_{train}$.

Let us consider now two models, for simplicity. Therefore, considering $\bar{g}_1(\boldsymbol{\theta}|\mathbf{y}_{\text{train}})$, $\bar{g}_2(\boldsymbol{\theta}|\mathbf{y}_{\text{train}})$, as proper priors, we can define the following *partial* Bayes factor

$$\begin{aligned}\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}) &= \frac{Z^{(1)}_{\text{test}|\text{train}}}{Z^{(2)}_{\text{test}|\text{train}}} = \frac{\frac{Z_1}{Z^{(1)}_{\text{train}}}}{\frac{Z_2}{Z^{(2)}_{\text{train}}}}, \\ &= \frac{\frac{Z_1}{Z_2}}{\frac{Z^{(1)}_{\text{train}}}{Z^{(2)}_{\text{train}}}} = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}_{\text{train}})}. \quad \textit{(“Bayes law for Bayes Factors”).} \end{aligned} \tag{23}$$

Therefore, one can approximate firstly $\text{BF}(\mathbf{y}_{\text{train}})$, secondly $\text{BF}(\mathbf{y})$ and then compare the model using the partial Bayes factor $\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}})$.

Remark 10. *The trick here consists in computing two normalizing constants for each model, instead of only one. The first normalizing constant is used for building an auxiliary proper prior, depending on* $\mathbf{y}_{train}$*. The difference with the likelihood-based prior approach in previous section is that* $\mathbf{y}_{train}$ *is used only once (in the auxiliary proper prior).*

The main drawback of the partial Bayes factor approach is the dependence on the choice of $\mathbf{y}_{\text{train}}$ (which could affect the selection of the model). The authors suggest finding the *minimal* suitable training set $\mathbf{y}_{\text{train}}$, but this task is not straightforward. A training dataset $\mathbf{y}_{\text{train}}$ is called *proper*, if $\int_{\Theta_m} \ell_m(\mathbf{y}_{\text{train}}|\boldsymbol{\theta})h_m(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$ for all models, and it is called *minimal* if is proper and no subset of $\mathbf{y}_{\text{train}}$ is proper. Two alternatives in the literature have been proposed, the fractional Bayes factors and the intrinsic Bayes factors.

Fractional Bayes Factors [31]. Instead of using a training data, it is possible to use power posteriors, i.e.,

$$\text{FBF}(\mathbf{y}) = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}|\beta)}, \tag{24}$$

where the denominator is

$$\text{BF}(\mathbf{y}|\beta) = \frac{\int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta})^\beta g_1(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta})^\beta g_2(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{c_1 \int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta})^\beta h_1(\boldsymbol{\theta})d\boldsymbol{\theta}}{c_2 \int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta})^\beta h_2(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{25}$$

with $0 < \beta < 1$, and $\text{BF}(\mathbf{y}|1) = \text{BF}(\mathbf{y})$. Note that the value $\beta = 0$ is not admissible since $\int_{\Theta_m} h_m(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$ for $m = 1, 2$. Again, since both $\text{BF}(\mathbf{y})$ and $\text{BF}(\mathbf{y}|\beta)$ depend on the ratio $\frac{c_1}{c_2}$, the fractional Bayes factor $\text{FBF}(\mathbf{y})$ is independent on c_1 and c_2 by definition.

Intrinsic Bayes factors [32]. The partial Bayes factor (23) will depend on the choice of (minimal) training set $\mathbf{y}_{\text{train}}$. These authors solve the problem of choosing the training sample by averaging the partial Bayes factor over all possible minimal training sets. They suggest using the arithmetic mean, leading to the *arithmetic* intrinsic Bayes factor, or the geometric mean, leading to the *geometric* intrinsic Bayes factor. Note that the Intrinsic Bayes factor is in some sense related to the well-known cross-validation approach (see numerical example in Section 5.2).

Remark 11. *The main drawback of these approaches is their dependence on the likelihood function and, for this reason, usually they tend to select more complex models, overweighting the likelihood values. See the numerical experiment in Section 5.2.*

4.4 Posterior predictive approach

The marginal likelihood approach is not the unique approach for model selection in Bayesian statistics. Here, we discuss an alternative strategy, that is based on the concept of prediction [18, Ch. 6][19]. This approach is more robust with respect to the choice of the prior density, so that it can be considered as a possible solution to the issues described above.
After fitting a Bayesian model, a popular approach for model checking (i.e. assessing the adequacy of the model fit to the data) consists in measuring its predictive accuracy [18, 19]. Hence, a key quantity in these approaches is the posterior predictive distribution of generic different data $\widetilde{\mathbf{y}}$ given $\mathbf{y}$,

$$\begin{aligned} p(\widetilde{\mathbf{y}}|\mathbf{y}) = E_{\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})}[\ell(\widetilde{\mathbf{y}}|\boldsymbol{\theta})] &= \int_{\Theta} \ell(\widetilde{\mathbf{y}}|\boldsymbol{\theta})\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \\ &= \frac{1}{Z}\int_{\Theta} \ell(\widetilde{\mathbf{y}}|\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned} \tag{26}$$

Considering $\widetilde{\mathbf{y}} = \mathbf{y}$, we can observe that exists a clear connection with likelihood-based priors described in Section 4.2. Indeed, if we assume $g(\boldsymbol{\theta}) \propto 1$ and $\widetilde{\mathbf{y}} = \mathbf{y}$, Eq. (26) becomes Eq. (16).
Note that the posterior predictive distribution in Eq. (26) is an expectation w.r.t. the posterior, which is robust to the prior selection with informative data, unlike the marginal likelihood as we shown in Section 3. Therefore, this approach is less affected by the prior choice.
Note that we can consider posterior predictive distributions $p(\tilde{\mathbf{y}}|\mathbf{y})$ for vectors $\tilde{\mathbf{y}}$ smaller than y (i.e., with less components). The posterior predictive checking is based on the main idea of considering some simulated data $\widetilde{\mathbf{y}}_i \sim p(\widetilde{\mathbf{y}}|\mathbf{y})$, with $i = 1, \ldots, L$, and comparing with the observed data $\mathbf{y}$. After obtaining a set of fake data $\{\tilde{\mathbf{y}}_i\}_{i=1}^{L}$, we have to measure the discrepancy between the true observed data $\mathbf{y}$ and the set $\{\widetilde{\mathbf{y}}_i\}_{i=1}^{L}$. This comparison can be made with test quantities and graphical checks (e.g., posterior predictive p-values) [18].

5 Numerical experiments

In this section, we provide different numerical simulations testing different models, prior pdfs and possible solutions. One of them is a well-known model based on the radial velocity technique for detecting exo-objects orbiting other stars. Some related code is also provided.[1]

[1]Related Matlab code is available at `http://www.lucamartino.altervista.org/Code_Llorente_Priors.m`

5.1 First numerical analysis: harmonic detection

In this section, we study a comparison between two different models already considered in other works [33, 34]. In the first model, the underlying signal is just a constant value,

$$\mathcal{M}_1 : \qquad y_i = B + \varepsilon_i, \tag{27}$$

whereas in the second model the underlying signal contains also a periodic piece,

$$\mathcal{M}_2 : \qquad y_i = B + A_1 \sin\left(2\pi\left(\frac{t}{P_1} + t_1\right)\right) + \varepsilon_i, \tag{28}$$

where in both $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i = 1, ..., D_y$. We assume $B = 1$ $A_1 = 0.9$, $P_1 = 3$, $t_1 = 0$, $\sigma = 1$, and generate 50 values from the model $\mathcal{M}_2$. Thus, $\mathcal{M}_2$ is the true model. We also consider two scenarios where different priors over A and P are employed.

First experiment. We consider uniform proper priors for all the parameters: $B \sim \mathcal{U}([-10, 10])$, $A_1 \sim \mathcal{U}([0.1, 100])$, $P_1 \in\sim \mathcal{U}([0.3, 30])$ and $t_1 \sim \mathcal{U}([0, 1])$. Then, considered the generated data from $\mathcal{M}_2$, the corresponding likelihood functions and priors, we can approximate the Bayesian evidence Z_1 of the model $\mathcal{M}_1$, and Z_2 of the model $\mathcal{M}_2$. By applying numerical integration and/o Monte Carlo methods [33], the marginal likelihood of both models has been computed obtaining $\log Z_1 = -35.74$ and $\log Z_2 = -36.33$, i.e., the wrong model $\mathcal{M}_1$ would be chosen ($Z_1 > Z_2$).

Second experiment. We change the prior over A and P considering two proper uniform densities *but* in a logarithmic scale, i.e., which are two modified Jeffreys priors [14]. The prior of A is

$$g(A) \propto \frac{1}{A}, \quad \text{with} \quad A \in [1, 100],$$

and $g(A) = 0$ otherwise. Moreover, we have

$$g(P) \propto \frac{1}{P}, \quad \text{with} \quad P \in [0.3, 30],$$

and $g(P) = 0$ otherwise. The rest of the experiment is the same as above. In this case, $\log Z_2 = -31.14$ (approximated with the computational method described in [33, 35]). Hence, in this case $Z_2 > Z_1$, and we select the true model $\mathcal{M}_2$. We have seen as the change of the forms of the prior pdfs (keeping fixed the intervals of the analysis) affects the results.

5.2 Second numerical analysis: partial and intrinsic BFs

Consider again the models in the illustrative example 2 in Section 3:

$$\mathcal{M}_1 = \{\ell_1(y|\theta) = \theta^y e^{-\theta}/y!,\ g_1(\theta)\},$$
$$\mathcal{M}_2 = \{\ell_2(y|\phi) = \phi(1-\phi)^y,\ g_2(\phi)\}$$

with different priors $g_1(\theta)$ and $g_2(\phi)$. Previously in Section 3, we considered two uniform and proper priors $g_2(\phi) = 1,\ \phi \in (0,1)$, and $g_1(\theta) = \frac{1}{L},\ \theta \in (0,L)$. Hence, the Bayes factor BF_{12} is well defined. Here, we replace $g_1(\theta)$ with an *improper* uniform prior $\widetilde{g}_1(\theta) \propto 1$, $\theta \in (0,\infty)$ for model $\mathcal{M}_1$. Our goal is to replicate Tables 7 and 8 using this improper prior for $\mathcal{M}_1$.
In this situation, the Bayes factor is not well-defined due to the arbitrary constant in $\widetilde{g}_1(\theta)$. Hence, we need to resort to *partial* Bayes factors [31, Sect. 2], where we compute the posterior of a single observation y_i, denoted by a sub-index i, (training set) under prior $\widetilde{g}_1(\theta)$, i.e., $\bar{\pi}_1(\theta|y_i) \propto \ell_1(y_i|\theta)\widetilde{g}_1(\theta)$, and use $\bar{\pi}_1(\theta|y_1)$ now as a proper prior in the computation of BF_{12}. In order to avoid the dependence on the training sample, we use the approach in [32], called the *intrinsic* Bayes factor (IBF). Let denote $\mathbf{y}_{-i}$ the vector of all D_y data, $\mathbf{y}$, without the i-th component y_i, i.e., $\mathbf{y}_{-i}$ is a vector of $D_y - 1$ components. The IBF consists in averaging over all possible training samples, resulting in the following intrinsic Bayes factor,

$$\mathrm{IBF}_{12} = \frac{1}{D_y}\sum_{i=1}^{D_y} \frac{\int_0^\infty \ell_1(\mathbf{y}_{-i}|\theta)\bar{\pi}_1(\theta|y_i)d\theta}{\int_0^1 \ell_2(\mathbf{y}|\phi)d\phi} = \frac{1}{D_y}\sum_{i=1}^{D_y} \frac{\int_0^\infty \ell_1(\mathbf{y}|\theta)d\theta / \int_0^\infty \ell_1(y_i|\theta)d\theta}{\int_0^1 \ell_2(\mathbf{y}|\phi)d\phi}. \tag{29}$$

Note that the cost of computing IBF_{12} increases with D_y. For this experiment, we generate data from both models with different values of θ and ϕ, that is, we alternatively consider $\mathcal{M}_1$ and $\mathcal{M}_2$ as the true model. Specifically, we used $\theta_{\text{true}} \in \{5, 2\}$ and $\phi_{\text{true}} \in \{0.5, 0.2, 0.8\}$. We compute IBF_{12} in 100 different runs for the chosen values of θ and ϕ, and we show the results in Table 3 and Table 4 for $D_y = 30$ and $D_y = 100$, respectively[2]. We show the maximum and minimum values of IBF_{12}, obtained in the 100 simulations, along with the number of errors. When $\mathcal{M}_1$ is the true model, $\mathrm{IBF}_{12} < 1$ corresponds to an error, and conversely, when $\mathcal{M}_2$ is the true model, $\mathrm{IBF}_{12} > 1$ corresponds to an error.

The results clearly show that the use of intrinsic Bayes factors allows for correctly selecting $\mathcal{M}_1$ when it is indeed the true model, with very few errors in model selection for the considered values of θ_{true} and both $D_y = 30$ and $D_y = 100$. On the contrary, when $\mathcal{M}_2$ is the true model, the use of intrinsic Bayes factors makes more probable selecting $\mathcal{M}_1$ (as the most likely model) for some values of ϕ_{true}. Note, for instance, that the number of errors when $\phi_{\text{true}} = 0.8$ is 66, that is, more than half of the times we would wrongly select $\mathcal{M}_1$ over $\mathcal{M}_2$. This is consistent with the idea underlying partial/intrinic Bayes factors, where the proper prior is built using part of the data. Indeed, it tends to artificially increase the marginal likelihood of the model where the likelihood-based prior is applied (since the resulting prior has larger overlap with the likelihood). Increasing the number of data improves the results, as proves the 43 errors in model selection obtained when $\phi = 0.8$ and $D_y = 100$. Another way to reduce the problem is to apply the likelihood-based priors (using the same number of data in the construction of the prior) to both models.

5.3 Exoplanet detection

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 [14, 36, 37, 38]. The problem consists in fitting a dynamical model

[2]Related Matlab code is available at `http://www.lucamartino.altervista.org/Code_Llorente_Priors.m`

Table 3: Model comparison for $D_y = 30$. Minimum and maximum IBF_{12} under true model $\mathcal{M}_1$ (Poisson model) and $\mathcal{M}_2$ (geometric model), over 100 independent runs.

True model = $\mathcal{M}_1$				True model = $\mathcal{M}_2$			
θ	min IBF_{12}	max IBF_{12}	Errors ($\text{IBF}_{12} < 1$)	ϕ	min IBF_{12}	max IBF_{12}	Errors ($\text{IBF}_{12} > 1$)
5	6.28×10^{3}	3.95×10^{11}	0	0.2	1.61×10^{-26}	9.76	2
2	0.55	7.40×10^{6}	1	0.5	5.45×10^{-9}	884.25	30
				0.8	0.004	10.51	66

Table 4: Model comparison for $D_y = 100$. Minimum and maximum IBF_{12} under true model $\mathcal{M}_1$ (Poisson model) and $\mathcal{M}_2$ (geometric model), over 100 independent runs.

True model = $\mathcal{M}_1$				True model = $\mathcal{M}_2$			
θ	min IBF_{12}	max IBF_{12}	Errors ($\text{IBF}_{12} < 1$)	ϕ	min IBF_{12}	max IBF_{12}	Errors ($\text{IBF}_{12} > 1$)
5	2.38×10^{11}	4.52×10^{29}	0	0.5	1.98×10^{-13}	500.52	4
2	2.22×10^{3}	2.60×10^{14}	0	0.2	2.02×10^{-72}	3.34×10^{-18}	0
				0.8	0.003	6.69	43

to data acquired at different moments spanning during long time periods (up to years). The model is highly non-linear and, for certain sets of parameters, its evaluation is quite costly in terms of computation time. This is due to the fact that its evaluation involves numerically integrating a differential equation, or using an iterative procedure for solving a non-linear equation (until a certain condition is satisfied). This loop can be very long for some sets of parameters.

5.3.1 Likelihood function

When analyzing radial velocity data of an exoplanetary system, it is commonly accepted that the *wobbling* of the star around the centre 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

$$y_t = V_0 + \sum_{i=1}^{S} K_i \left[\cos\left(u_{i,t} + \omega_i\right) + e_i \cos\left(\omega_i\right)\right] + \xi_t, \tag{30}$$

with $t = 1, \ldots, T$.[3] The number of objects in the system is S. Both y_t, $u_{i,t}$ depend on time t, and ξ_t is a Gaussian noise perturbation with variance σ_e^2. We consider the noise variance σ_e^2 an unknown parameter as well. The meaning of each parameter in Eq. (30) is given in Table 5. The likelihood function is jointly defined by (30) 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

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

[3]More generally, we can have y_{t_j} with $j = 1, ..., T$.

Table 5: Description of parameters in Eq. (30).

Parameter	Description	Units
For each planet		
K_i	amplitude of the curve	$\mathrm{m\,s^{-1}}$
$u_{i,t}$	true anomaly	rad
ω_i	longitude of periastron	rad
e_i	orbit's eccentricity	...
P_i	orbital period	s
τ_i	time of periastron passage	s
Below: not depending on the number of objects/satellite		
V_0	mean radial velocity	$\mathrm{m\,s^{-1}}$

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

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

where τ_i is the time of periastron passage of the planet i and P_i is the period of its orbit (see Table 5). Then, through the Kepler's equation,

$$M_{i,t} = E_{i,t} - e_i \sin E_{i,t}, \tag{31}$$

where $E_{i,t}$ is the eccentric anomaly. Equation (31) 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 **[39].** For certain sets of parameters, this iterative procedure can be particularly slow and the computation of the likelihood becomes quite costly. We also have

$$\tan\frac{u_{i,t}}{2} = \sqrt{\frac{1+e_i}{1-e_i}}\,\tan\frac{E_{i,t}}{2}, \tag{32}$$

Therefore, the variable of interest $\boldsymbol{\theta}$ is the vector of dimension $d_\theta = 1+5S$ (where S is the number of planets),

$$\boldsymbol{\theta} = [V_0, K_1, \omega_1, e_1, P_1, \tau_1, \ldots, K_S, \omega_S, e_S, P_S, \tau_S],$$

For a single object (e.g., a planet or a natural satellite), the dimension of $\boldsymbol{\theta}$ is $d_{\boldsymbol{\theta}} = 5+1 = 6$, with two objects the dimension of $\boldsymbol{\theta}$ is $d_\theta = 11$, etc. All the Eqs. from (30) to (32) induce a likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma_e) = \prod_{t=1}^{T} \ell(y_t|\boldsymbol{\theta}, \sigma_e)$, where $\mathbf{y} = \{y_1, \ldots, y_T\}$. Given a prior density $g(\boldsymbol{\theta})$, the complete posterior is

$$p(\boldsymbol{\theta}|\mathbf{y}, \sigma_e) = \frac{1}{p(\mathbf{y}|\sigma_e)}\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma_e)g(\boldsymbol{\theta}),$$

where

$$Z = p(\mathbf{y}|\sigma_e) = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma_e)g(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Our goal is to infer the number S of planets in the system. For this purpose, we have to approximate the model evidence $Z = p(\mathbf{y}|\sigma_e)$ of each model. For simplicity, we consider the noise variance σ_e^2 is given.

5.3.2 Experiments

Let us denote $\mathcal{M}_0$ and $\mathcal{M}_1$ the models corresponding to zero and one planets. We generate a set of data $\mathbf{y}$ according to the model with one planet and parameter values $V = 5$, $K_1 = 25$, $\omega_1 = 0.61$, $e_1 = 0.1$, $P_1 = 15$, and $\tau_1 = 3$. We consider 25 total number of observations. All the data are generated with $\sigma_e^2 = 15$. The rest of trajectories are generated according to the transition model (and the corresponding measurements y_t according to the observation model). Our goal is to compute the ratio $\text{BF}_{10} = \frac{Z_1}{Z_0}$, where Z_1 and Z_0 denote respectively the marginal likelihood of the model with zero planet and the model with one planet. As we commented above, the model with zero planet has only one parameter, namely, $\boldsymbol{\theta}_0 = V_0$ with prior $\mathcal{U}([-20, 20])$. For simplicity, in the model with one planet we consider only two degrees of freedom, i.e., $\boldsymbol{\theta}_1 = [V_1, P_1]$. The rest of parameters are set to their true values. In this model, we use the same prior for V_1, while for P_1, we use $\mathcal{U}([0, P_{\text{max}}])$ with $P_{\text{max}} > 0$. We know that BF_{10} should be greater than 1 since the data were generated according to model 1. However, we aim to show that increasing P_{max} (which corresponds to use a prior that is more uninformative) makes that BF_{10} eventually becomes smaller than 1. For the computation of Z_0 and Z_1 we use a very thin grid within the prior bounds. In Figure 3, we show the Bayes factor as a function of P_{max}. For P_{max} greater than 200, we have $\text{BF}_{10} < 1$, that is, we wrongly choose the model with zero planets. This illustrates the problematic with the use of vague priors.

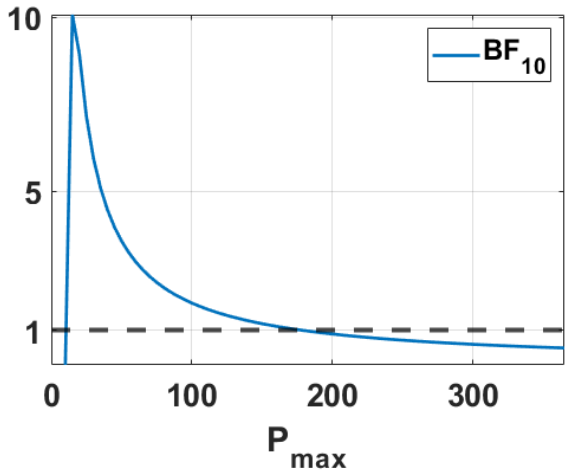


Figure 3: The Bayes factor BF_{10} as a function of prior width P_{max}. Increasing P_{max} (i.e. making the prior for P_1 more uninformative) eventually produces $\text{BF}_{10} < 1$. Note also that, when P_{max} is small (lower than $P_{\text{true}} = 15$), we have $\text{BF}_{10} < 1$, preferring the model with zero planet $\mathcal{M}_0$.

Hierarchical solution. Let us denote as $Z_1(P_{\text{max}})$, the marginal likelihood for each given value of P_{max}. Now, we consider the extended posterior where we use a prior for P_{max}, as

$h(P_{\max}) = \mathcal{U}([10, 365])$, hence the new marginal likelihood is

$$Z_{\texttt{new},1} = \int Z_1(P_{\max}) h(P_{\max}) dP_{\max}.$$

The value of $Z_{\texttt{new},1}$ is 9.1095×10^{-44}, which is greater than $Z_0 = 5.4601 \times 10^{-44}$. Hence, with this hierarchical modeling, we select the true model.

6 Conclusions

In this work, we have highlighted that the marginal likelihoods, which are fundamental quantities for Bayesian model selection, display strong dependence on the choice of the prior density. Moreover, we have explained why the use of improper priors is not suitable for model selection. More generally, we have also discussed the use of uninformative priors, whether proper (vague priors) or improper, and its effect on the model selection procedure. We have shown by means of illustrative examples the potential pitfalls of using vague priors, and we have provided and discussed possible solutions for all these scenario. We have also described an alternative for Bayesian model selection to the marginal likelihood approach, called posterior predictive. Furthermore, the connection with the information criteria has been also presented. One of the considered numerical experiment is a real-world astronomical application, consisted on detectiing the number of objects orbiting a star.

References

[1] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.

[2] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.

[3] L. Martino, V. Elvira, J. Lopez-Santiago, and G. Camps-Valls, "Compressed particle methods for expensive models with application in astronomy and remote sensing," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–15, 2021.

[4] F. Llorente, L. Martino, D. Delgado-Gomez, and G. Camps-Valls, "Deep importance sampling based on regression for model inversion and emulation," *Digital Signal Processing*, vol. 116, p. 103104, 2021.

[5] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, "Importance Nested Sampling and the MultiNest Algorithm," *The Open Journal of Astrophysics*, vol. 2, no. 1, p. 10, Nov. 2019.

[6] S. A. Anfinogentov, V. M. Nakariakov, D. J. Pascoe, and C. R. Goddard, "Solar Bayesian Analysis Toolkit—A New Markov Chain Monte Carlo IDL Code for Bayesian Parameter Inference," *Astrophysical Journal Supplement Series*, vol. 252, no. 1, p. 11, Jan. 2021.

[7] J. López-Santiago, L. Martino, M. A. Vázquez, and J. Miguez, "A Bayesian inference and model selection algorithm with an optimization scheme to infer the model noise power," *Monthly Notices of the Royal Astronomical Society*, vol. 507, no. 3, pp. 3351–3361, Nov. 2021.

[8] J. Lopez-Santiago, L. Martino, M. A. Vazquez, and J. Miguez, "A Bayesian inference and model selection algorithm with an optimization scheme to infer the model noise power," *Monthly Notices of the Royal Astronomical Society*, vol. 507, no. 3, pp. 3351–3361, 2021.

[9] G. Ashton and C. Talbot, "BILBY-MCMC: an MCMC sampler for gravitational-wave inference," *Monthly Notices of the Royal Astronomical Society*, vol. 507, no. 2, pp. 2037–2051, Oct. 2021.

[10] I. Ayuso, R. Lazkoz, and V. Salzano, "Observational constraints on cosmological solutions of f (Q) theories," *Physical review d*, vol. 103, no. 6, p. 063505, Mar. 2021.

[11] J. Emmert, S. J. Grauer, S. Wagner, and K. J. Daun, "Efficient Bayesian inference of absorbance spectra from transmitted intensity spectra," *Opt. Express*, vol. 27, no. 19, pp. 26 893–26 909, 2019.

[12] U. Von Toussaint, "Bayesian inference in physics," *Rev. Mod. Phys.*, vol. 83, pp. 943–999, 2011.

[13] D. J. Pascoe, A. Smyrli, T. Van Doorsselaere, and A. M. Broomhall, "Bayesian Analysis of Quasi-periodic Pulsations in Stellar Flares," *Astrophysical Journal*, vol. 905, no. 1, p. 70, Dec. 2020.

[14] P. C. Gregory, "Bayesian re-analysis of the Gliese 581 exoplanet system," *Monthly Notices of the Royal Astronomical Society*, vol. 415, no. 3, pp. 2523–2545, Aug. 2011.

[15] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *arXiv:2005.08334 (extended version with supplementary material)*, 2020.

[16] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis–Hastings output," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 270–281, 2001.

[17] C. S. Bos, "A comparison of marginal likelihood computation methods," in *Compstat*. Springer, 2002, pp. 111–116.

[18] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[19] J. Piironen and A. Vehtari, "Comparison of Bayesian predictive methods for model selection," *Statistics and Computing*, vol. 27, no. 3, pp. 711–735, 2017.

[20] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[21] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, 1999.

[22] L. Martino, J. Read, V. Elvira, and F. Louzada, "Cooperative parallel particle filters for online model selection and applications to urban mobility," *Digital Signal Processing*, vol. 60, pp. 172–185, 2017.

[23] I. Urteaga, M. F. Bugallo, and P. M. Djurić, "Sequential Monte Carlo methods under model uncertainty," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, 2016, pp. 1–5.

[24] H. Jeffreys, *The theory of probability*. OUP Oxford, 1998.

[25] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Wiley, 1994.

[26] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[27] E. Cameron and A. Pettitt, "Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis," *Statistical Science*, vol. 29, no. 3, pp. 397–419, 2014.

[28] J. R Oaks, K. A. Cobb, V. N Minin, and A. D. Leaché, "Marginal likelihoods in phylogenetics: a review of methods and applications," *Systematic biology*, vol. 68, no. 5, pp. 681–697, 2019.

[29] D. J. Spiegelhalter and A. F. Smith, "Bayes factors for linear and log-linear models with vague prior information," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 3, pp. 377–387, 1982.

[30] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[31] A. O'Hagan, "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 99–118, 1995.

[32] J. O. Berger and L. R. Pericchi, "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 109–122, 1996.

[33] J. López-Santiago, L. Martino, M. Vázquez, and J. Miguez, "A bayesian inference and model selection algorithm with an optimization scheme to infer the model noise power," *Monthly Notices of the Royal Astronomical Society*, vol. 507, no. 3, pp. 3351–3361, 2021.

[34] J. Buchner, "UltraNest–a robust, general purpose Bayesian inference engine," *arXiv preprint arXiv:2101.09604*, 2021.

[35] L. Martino, F. Llorente, E. Cuberlo, J. López-Santiago, and J. Míguez, "Automatic tempered posterior distributions for Bayesian inversion problems," *Mathematics*, vol. 9, no. 7, p. 784, 2021.

[36] S. C. C. Barros *et al.*, "WASP-113b and WASP-114b, two inflated hot Jupiters with contrasting densities," *Astronomy and Aastrophysics*, vol. 593, p. A113, 2016.

[37] L. Affer *et al.*, "HADES RV program with HARPS-N at the TNG. IX. A super-Earth around the M dwarf Gl 686," *arXiv:1901.05338*, vol. 622, p. A193, Feb. 2019.

[38] T. Trifonov, S. Stock, T. Henning, S. Reffert, M. Kürster, M. H. Lee, B. Bitsch, R. P. Butler, and S. S. Vogt, "Two Jovian Planets around the Giant Star HD 202696: A Growing Population of Packed Massive Planetary Pairs around Massive Stars?" *The Astronomical Journal*, vol. 157, no. 3, p. 93, Mar. 2019.

[39] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C++ : the art of scientific computing*. Springer, 2002.

[40] K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, and B. Placek, "Bayesian evidence and model selection," *Digital Signal Processing*, vol. 47, pp. 50–67, 2015.

[41] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[42] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.

[43] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde, "Bayesian measures of model complexity and fit," *J. R. Stat. Soc. B*, vol. 64, pp. 583–616, 2002.

Appendices

A Marginal likelihood and information criteria

The marginal likelihood can be expressed as

$$Z = \ell_{\max} W, \tag{33}$$

where $W \in [0, 1]$ is the *Occam factor* [40, Sect. 3]. More specifically, the Occam factor is defined as

$$W = \frac{1}{\ell_{\max}} \int_{\Theta} g(\boldsymbol{\theta}) \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{34}$$

and it is $\frac{\ell_{\min}}{\ell_{\max}} \leq W \leq 1$. The factor W measures the penalty of the model complexity *intrinsically* contained in the marginal likelihood Z: this penalization depends on the chosen prior and the number of data involved.

Considering the expression (33) and taking the logarithm, we obtain

$$\log Z = \log \ell_{\max} + \log W \tag{35}$$

Note that $\log \ell_{\max}$ is a fitting term whereas $\log W$ is a penalty for the model complexity. Instead of maximizing Z (or $\log Z$) for model selection purposes, several authors consider the *minimization* of some cost functions C derived by different information criteria [41, 42, 43]. Most of the criteria, suggested in the literature, can be expressed as

$$C = \underbrace{-2 \log \ell_{\max}}_{\text{fitting}} \underbrace{+2\eta D_\theta}_{\text{penalization}} , \tag{36}$$

where η is a real value that is often chosen as function of the number of data D_y, and D_θ is the dimension of $\boldsymbol{\theta}$, i.e., the number of parameters. The first term is a fitting term (which fosters the choice of more complex models), whereas the second one is a model penalization term (which promotes the choice of simpler models).

Remark 12. *Note that the expression of C is similar to*

$$-2 \log Z = -2 \log \ell_{\max} - 2 \log W,$$

considering Eq. (35)*, where $-2 \log W$ plays the role of the second factor $2\eta D_\theta$ in Eq.* (36)*.*

The expression (36) encompasses several well-known information criteria proposed in the literature and shown in Table 6, which differ for the choice of η.

Remark 13. *The penalty term $2\eta D_\theta$ in the information criteria is the same for every parameter. The Bayesian approach allows the choice of different penalties, assuming different priors, one for each parameter, i.e., for each component of $\boldsymbol{\theta}$.*

Table 6: Different information criterion for model selection.

Criterion	**Choice of η**
Bayesian-Schwarz information criterion (BIC) [41]	$\frac{1}{2} \log D_y$
Akaike information criterion (AIC) [43]	1
Hannan-Quinn information criterion (HQIC) [42]	$\log(\log(D_y))$