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 depends on the prior choice. For model selection, even diffuse priors can be actually very informative, unlike for the parameter estimation problem. Furthermore, when the prior is improper, the marginal likelihood of the corresponding model is undetermined. In this work, we discuss 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 discussed and many possible solutions, proposed in the literature, to design objective priors for model selection are described. Some of them also allow the use of improper priors. The connection between the marginal likelihood approach and the well-known information criteria is also presented. We describe the main issues and possible solutions by illustrative numerical examples, providing also 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. (Robert & Casella, 2004; Liu, 2004)), it is only now becoming more widespread. Nowadays, we can find applications of Bayesian inference methods in fields such as remote sensing (Martino, Elvira, et al., 2021; Llorente et al., 2021), astronomy (Feroz et al., 2019; Anfinogentov et al., 2021), cosmology (Ashton & Talbot, 2021; Ayuso et al., 2021), or optical spectroscopy (Emmert et al., 2019; Von Toussaint, 2011).

[*] Corresponding author: felloren@est-econ.uc3m.es.

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 clear dependence on the choice of the prior probability density functions (pdfs). Many papers propose diffuse (usually uniform) prior pdfs, in order to avoid biasing the exploration of the parameter space (see, e.g., (Pascoe et al., 2020)). In some cases, the selected prior pdfs are diffuse or even *improper* (Gregory, 2011). These ideas have been borrowed from the Bayesian parameter estimation problem, where they are adequate and *objective* choices. However, in model selection, the situation is more complex and we describe below.

In a first part of this work, we describe some issues in Bayesian model selection (or hypothesis testing) based on the marginal likelihood computation (Llorente et al., 2020; Chib & Jeliazkov, 2001; Bos, 2002). First of all, we show how the results can be affected by the choice of the prior. The typical solution for parameter estimation of using a diffuse prior (which is said *uninformative* in this scenario) cannot be considered an objective choice for the marginal likelihood computation. With an objective choice, we refer to prior selection that attempts to bring impartiality in the model selection problem, and diffuse prior can be actually a very informative prior for model selection. Secondly, this issue becomes even more dramatic when improper priors are employed: the Bayesian parameter estimation with improper priors is allowed if the corresponding posterior is proper, whereas Bayesian model selection with improper priors is *not* allowed, due to the fact the marginal likelihood is 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 as hierarchical approaches, *likelihood-based priors*, and the *partial, intrinsic, fractional* Bayes factors (Llorente et al., 2020; O'Hagan, 1995), remarking potential benefits and possible drawbacks of each of them. An alternative to the marginal likelihood approach for Bayesian model selection, called *posterior predictive* framework (Vehtari et al., 2017, Ch. 6)(Piironen & Vehtari, 2017), is also described. Finally, the relationship between the information criteria (Konishi & Kitagawa, 2008), such as Bayesian-Schwarz information criterion (BIC), Akaike information criterion (AIC), and the marginal likelihood approach is discussed in Appendix B. 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.

2 Problem statement

In many applications, the goal is to make inference about a variable of interest, $\boldsymbol{\theta} = \theta_{1:D_\theta} = [\theta_1, \theta_2, \dots, \theta_{D_\theta}] \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{D_\theta}$, where $\theta_d \in \mathbb{R}$ for all $d = 1, \dots, D_\theta$, given a set of observed measurements, $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$. 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

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

is the so-called *marginal likelihood*, a.k.a., *Bayesian evidence* (Robert & Casella, 2004; Liu, 2004). This quantity is important for model selection purposes, 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 $\boldsymbol{\theta}_m = [\theta_{m,1}, \theta_{m,2}, \dots, \theta_{m,D_{\theta_m}}] \in \boldsymbol{\Theta}_m \in \mathbb{R}^{D_{\theta_m}}$, with possibly different dimensions in the different models. The posterior probability 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 $Z_m = p(\mathbf{y}|\mathcal{M}_m) = \int_{\Theta_m} \ell(\mathbf{y}|\boldsymbol{\theta}_m,\mathcal{M}_m)g(\boldsymbol{\theta}_m|\mathcal{M}_m)d\boldsymbol{\theta}_m$, 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* (Hoeting et al., 1999), 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, (Martino et al., 2017; Urteaga et al., 2016)). 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*

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

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 depend on the choice of the prior density, the model prior probabilities and the actual number of data D_y.

3 Important definitions and classifications

In this section, we describe some preliminary definitions that are necessary for a clear description of the issues in Bayesian model selection and the corresponding possible solutions (described in the rest of the work).

3.1 Levels in Bayesian inference

Generally speaking, in Bayesian inference we can distinguish between two types of problems or levels of inference (MacKay, 2003, Ch. 28), described below:

- **Level-1: estimation and prediction problems.** In the first level, given the i-th model $\mathcal{M}_i$, we are interested in making inferences regarding parameter $\boldsymbol{\theta}_i$ by focusing on its posterior pdf $\bar{\pi}(\boldsymbol{\theta}_i|\mathbf{y}, \mathcal{M}_i) \propto \ell(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i)g(\boldsymbol{\theta}_i|\mathcal{M}_i)$. This is also denoted as "Level-1 of inference" in the literature.

- **Level-2: model selection problems.** In the second type of problem, we focus on the model posterior distribution $p(\mathcal{M}_i|\mathbf{y}) \propto p(\mathcal{M}_i)Z_i = p(\mathcal{M}_i)\int_{\Theta_i} \ell(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i)g(\boldsymbol{\theta}_i|\mathcal{M}_i)$ for all i. This is also known as "Level-2 of inference".

More *levels* of inference can be recognized in the so-called hierarchical Bayesian approaches. However, conceptually there are the two *main* levels of inference since they are associated to the two main inference scenarios: parameter estimations and model selection. We will see that the prior choice has a different impact in each of the different levels.

3.2 Type of model comparison

In the literature, we can distinguish different types of model selection, as we summarize below. The type of model selection problem can affect the user's choice of a suitable prior density.

- **Basic model selection:** In this scenario, we compare different likelihood functions (i.e., observation models). The likelihood functions can represent completely different models, living even in different parameter space. In this scenario, the parameters $\boldsymbol{\theta}_m$ of each model can have a completely different physical or statistical interpretation.

- **Selection in nested models:** Nested models are models that belong to the same parametric family, but the *size* of the model $|\boldsymbol{\Theta}_m| = D_{\theta_m}$ is also unknown and must be inferred as well, jointly with the parameter $\boldsymbol{\theta}_m$. Namely, we have a sequence of likelihoods defined in an increasing dimensional space, such as $\ell(\mathbf{y}|\theta_1, \mathcal{M}_1)$, $\ell(\mathbf{y}|\theta_1, \theta_2, \mathcal{M}_2)$, $\ell(\mathbf{y}|\theta_1, \theta_2, \theta_3, \mathcal{M}_3)$, etc.
  Famous applications which belong to this scenario are the following: *variable selection* (e.g., selecting a subset of relevant features/variables in regression or classification), *order selection* (e.g., in polynomial regression or ARMA models etc.), *clustering* (when the number of clusters are unknown) and *dimension reduction* problems (Bishop, 2006).

3.3 Type of prior densities

The literature is plenty of works devoted to the specification and classification of different priors. The interested readers can find gentle reviews in, e.g.,(Kass & Wasserman, 1996; Consonni et al., 2018; Mikkola et al., 2021). Here, we provide with a brief summary of concepts related to the choice of the priors $g(\boldsymbol{\theta}|\mathcal{M})$ over the parameters, and how this choice can affect the analysis in the two levels of inference, discussed above.

3.3.1 Subjective priors

If the user or practitioner has some belief or any a-priori knowledge about the quantity of interest (before the data are observed), then this information should be included in the analysis by the addition of a suitable prior density. This prior is called as *informative* (or more precisely, in our opinion, *subjective-informative*). We can distinguish three main classes of subjective priors:

- **Priors including beliefs.** An informative prior pdf can be determined from previous information, past experiments or by other sources of information (different from the observation model). Prior elicitation ideas can be used to transform such knowledge into a prior density. See (Mikkola et al., 2021) for a review on different approaches for prior elicitation.

- **Priors as regularizers.** In this case, the practitioner/researcher desires to force that the final solution satisfies some properties established in advance, such *smoothness* (designing specific structure in covariance matrices in Gaussian priors, e.g., see (Martino & Read, 2021)), *sparsity* (this is the case of LASSO regularized, i.e., Laplacian priors (Bishop, 2006)) etc. Moreover, the regularization effect produced by the prior usually yields more computational stability and, hence, reducing the numerical issues.

- **Conjugate priors.** A prior can also be set with goal of reducing the computation required by the posterior analysis. Indeed, when a family of conjugate priors exists, choosing a prior from that family simplifies the calculation of the posterior distribution, avoiding the use of costly computational techniques.

3.3.2 Objective priors

In many scenarios, additional information and conjugate priors are not available, and “objective” choice of priors could be desired (Consonni et al., 2018). A first (and perhaps primitive) approach for obtaining an objective prior are related to the concept of *uninformative* priors, representing the absence of a-priori knowledge (Kass & Wasserman, 1996; Consonni et al., 2018). A second approach is related to the idea of constructing priors by the use of formal rules and automatized procedures based on desirable criteria and properties. The term *objective* prior aims at encompassing both group of priors above (Consonni et al., 2018). Below, we give more specific definitions.

Uninformative priors. Generally, a prior is defined as *uninformative*, if it has been chosen in order to have a minimal impact on the posterior density (Mikkola et al., 2021). In this sense, for the inference problem of parameter estimation (**Level-1**), a uniform prior over all the support $\boldsymbol{\Theta}$ is the is maximal expression of uninformative prior. In fact, the inference would be completely data-driven. On the contrary, we will show that in model selection (**Level-2**), this prior is highly informative. Below, we describe some classes of uninformative priors (or attempts of uninformative priors) for the **Level-1** of inference, i.e., parameter estimation.

- **Uniform prior over $\boldsymbol{\Theta}$ when $|\boldsymbol{\Theta}| < \infty$.** If $\boldsymbol{\Theta}$ is bounded, the simplest idea for determining a non-informative prior (for **Level-1**, parameter estimation) is to assign equal probabilities to all possible outcomes, such as uniform densities in the bounded support, i.e., $g(\boldsymbol{\theta}) \propto 1 \ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$.

- **Locally-uniform priors.** If $\boldsymbol{\Theta}$ is unbounded, one can employ a *vague* priors, i.e., densities with probability mass spread in all the state space, with a great scale parameter (this is the reason of the name "locally uniform"). The priors built using this philosophy have been given different names such as *diffuse*, *vague*, *flat*, *weakly-informative*, etc. (Consonni et al., 2018) A more extreme alternative is to use *improper* priors when it is possible (see the description below).

- **Improper priors.** Let us consider again that $\boldsymbol{\Theta}$ is unbounded. The use of improper priors, i.e., such that $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, is allowed for **Level-1** inference when $\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, since the corresponding posteriors are proper. However, this is an issue for model selection (**Level-2** inference), where we use the marginal likelihood Z. Indeed, the prior $g(\boldsymbol{\theta}) = c \cdot h(\boldsymbol{\theta})$ is not completely specified, since $c > 0$ is arbitrary.

Other authors design priors using formal rules which are theoretical and practically appealing. In this sense, these type of priors are *informative but not subjective*. Some example is given below.

Reference and Jeffreys priors. Prior 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 (Robert & Casella, 2004; Jeffreys, 1998). 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. More generally, the Jeffrey's prior is constructed by taking $g(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{\frac{1}{2}}$ where $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix.

Below we discuss how the choice of the prior affects **(a)** the inference of $\boldsymbol{\theta}$ (**Level-1**), and **(b)** the estimation of the Bayesian evidence Z for the model selection problem (**Level-2**).

4 Dependence on the choice of the prior density

In this section, we show how the marginal likelihood Z is sensitivity to the choice of prior density (Bernardo & Smith, 1994a). Here, first we show all the possible values that the evidence Z can take changing the prior pdf. Then, we present some reassuring asymptotic results. Finally, we describe further issues in the use of improper priors.

4.1 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

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

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 $g(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\min})$, where $\delta(\boldsymbol{\theta})$ denotes the Dirac point mass at 0. 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, $Z = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$ 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 prior $g(\boldsymbol{\theta})$ induces a penalization term for the model complexity. See also Appendix A for further details.*

Remark 4. *Choosing a prior $g(\boldsymbol{\theta})$, we fix our bias-variance trade-off (a point between the maximum under-fitting and maximum over-fitting). In this sense, the sensitivity of Z could be also considered as a benefit, i.e., an additional degree of freedom for improving the bias-variance trade-off.*

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 B.

4.2 Asymptotic considerations in Bayesian inference

Throughout this section, we consider the priors have been selected and fixed, whereas the number of data D_y diverges to infinity, i.e., $D_y \to \infty$. We summarize the basic consistency properties of Bayesian inference in both inference problems (i.e., Level-1, estimation, and Level-2, model selection) and discuss the asymptotic behavior of Bayes factors and posterior model probabilities. For sake of simplicity, we assume that weak regularity conditions are satisfied for these results to hold (Dawid, 2011; Kass & Raftery, 1995; Rossell & Rubio, 2021; Bernardo & Smith, 1994b).

It is important to distinguish and describe two scenarios: **(a)** one where the true (unknown) distribution of the data $\ell_{\text{true}}(y|\boldsymbol{\theta}_{\text{true}})$ is included in the M possible models ($\mathcal{M}$-closed scenario), **(b)** and the other one where $\ell_{\text{true}}(y|\boldsymbol{\theta}_{\text{true}})$ is *not* included in the possible set of models ($\mathcal{M}$-open scenario).

- **$\mathcal{M}$-closed scenario.** When one of the models under consideration, say $\mathcal{M}_{i_{\text{true}}}$, contains $\ell_{\text{true}}(y|\boldsymbol{\theta}_{\text{true}})$, i.e., there is $\boldsymbol{\theta}_{\text{true}}$ such that $\ell_{\text{true}}(y|\boldsymbol{\theta}_{\text{true}}) = \ell(y|\boldsymbol{\theta}_{\text{true}}, \mathcal{M}_{i_{\text{true}}})$.

- **$\mathcal{M}$-open scenario.** When none of the models contains $\ell_{\text{true}}(y|\boldsymbol{\theta}_{\text{true}})$ (*miss-pecification*), then we can define $\boldsymbol{\theta}_i^* = \arg\min_{\boldsymbol{\theta}_i} KL(\ell_{\text{true}}, \ell(\cdot|\boldsymbol{\theta}_i, \mathcal{M}_i))$, which is the parameter that minimizes the Kullback-Leibler (KL) divergence between $\ell_{\text{true}}(y|\boldsymbol{\theta}_{\text{true}})$ and $\ell(y|\boldsymbol{\theta}_i, \mathcal{M}_i)$. Furthermore, we can define

$$\mathcal{M}_{i^*} = \arg\min_i KL(\ell_{\text{true}}, \ell(\cdot|\boldsymbol{\theta}_i^*, \mathcal{M}_i)), \tag{8}$$

  as the model that is closest in KL divergence to the true distribution of the data.

Consistency in Level-1. Consider the posterior distribution $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ for a fixed model $\mathcal{M}$ (that is, a particular observation model and a fixed prior). In the $\mathcal{M}$-closed scenario, the posterior $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ concentrates around $\boldsymbol{\theta}_{\text{true}}$ as $D_y \to \infty$ (see Bernstein-von Mises theorem (Robert & Casella, 2004; Liu, 2004; Bernardo & Smith, 1994a)). Then the two Bayesian point estimators, the posterior mean $\widehat{\boldsymbol{\theta}}_{\texttt{mean}} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, and the maximum-a-posteriori (MAP) estimator $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$. converge to $\boldsymbol{\theta}_{\texttt{true}}$ (recovering frequentist arguments). 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. In the $\mathcal{M}$-open scenario (i.e. when the model is misspecified), then the asymptotic limits of the estimators $\widehat{\boldsymbol{\theta}}_{\texttt{mean}}$ and $\widehat{\boldsymbol{\theta}}_{\texttt{MAP}}$ approach the best-fitting parameters $\boldsymbol{\theta}_i^*$ (Bernardo & Smith, 1994b; Rossell & Rubio, 2021).

Consistency in Level-2. In the $\mathcal{M}$-closed scenario, as the sample size diverges, $D_y \to \infty$, the posterior model distribution concentrates around the true model, that is, $p(\mathcal{M}_{i_{\text{true}}}|\mathbf{y}) \to 1$ (Kass & Raftery, 1995; Dawid, 2011). In the $\mathcal{M}$–open scenario, the posterior model distribution concentrates on the model closest in KL divergence, that is, $p(\mathcal{M}_{i^*}|\mathbf{y}) \to 1$, as $D_y \to \infty$ (Dawid, 2011; Rossell & Rubio, 2021).

Remark 5. *Under regularity conditions, Bayesian parameter estimation and model selection are consistent. Specifically, as $D_y \to \infty$, in the $\mathcal{M}$-closed scenario, Bayesian inference gives the correct answer by selecting the true model $\mathcal{M}_{i_{true}}$, and also converging to $\boldsymbol{\theta}_{true}$. In the $\mathcal{M}$-open scenario, Bayesian inference gives the best approximate answer, converging to the KL minimizers under each model $\boldsymbol{\theta}_i^*$ and selecting the model with overall minimal KL divergence $\mathcal{M}_{i^*}$.*

Furthermore, in specific application frameworks and under fairly general conditions, asymptotic expressions of quotients of posterior model probabilities and Bayes factors have been derived; see, e.g., (Dawid, 2011; Rossell & Rubio, 2021). An important observation is that the leading terms in those expressions do not depend on the prior densities. Namely, in the asymptotic regime, Bayesian model selection is more sensitive to the sample size D_y than to the prior specifications (Dawid, 2011; Rossell & Rubio, 2021). Namely, Bayes factors depend more strongly on the sample size D_y than on the prior dispersion, as we show in the example in Section 6.3. As we can see in Figure 4(b), there exists a reasonable"default range" of the prior dispersion parameter that provides good results. Such default ranges could be obtained, for instance, by using a measure of predictive accuracy (Rossell & Rubio, 2021).

These results for the asymptotic regime are reassuring and comforting. However, in the finite sample size regime (i.e., D_y fixed) the results of Bayesian model selection are indeed affected by the prior choice: as we already discussed in Sect. 4.1, the marginal likelihood can take any value in the interval $[\ell_{\text{min}}, \ell_{\text{max}}]$. Below, we discuss this issue in the context of increasingly diffuse priors, and compare it with Bayesian parameter estimation.

4.3 Robustness of the Bayesian inference to prior dispersion

In this section, we keep the (finite) number of data D_y fixed, and we vary the spread of the prior density (changing some parameter of the prior). Below, we consider an illustrative example to show the perceived differences in robustness of Bayesian parameter estimation (Level-1) and Bayesian model selection (Level-2).

4.3.1 Illustrative example

Here we provide an alternative formulation of the Lindley-Bartlett paradox (Lindley, 1957; Villa & Walker, 2017; Robert, 2014)) which shows the well-known robustness of the posterior distribution (Level-1) when increasingly diffuse priors are employed. These priors are common for parameter estimation where they are seen as uninformative. However, in model selection (Level-2), actually such prior are highly informative: an increasingly diffuse prior penalizes more and more the considered model.
Let us assume a likelihood function that is integrable in every subset of an unbounded Θ, that is, for all $A \subseteq \boldsymbol{\Theta}$, $\int_{A\in\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. In particular, when $A = \boldsymbol{\Theta}$, the integral corresponds to the "area below" the likelihood function

$$S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty. \tag{9}$$

Hence, in this scenario, the normalized likelihood is a proper pdf on $\boldsymbol{\Theta}$. Then, we consider a uniform and proper prior defined on the hyper-volume B, i.e.,

$$g(\boldsymbol{\theta}) = \frac{1}{|B|}\mathbf{1}_B(\boldsymbol{\theta}),$$

where $|B|$ represents the volume of B. Hence, the posterior pdf is

$$\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})\mathbf{1}_B(\boldsymbol{\theta})}{\int_B \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{10}$$

which is the normalized likelihood restricted to the set B.

Level-1 of inference. As we increase the volume of B, more and more mass of the likelihood is considered. Roughly speaking, for a $|B|$ great enough, the posterior is insensitive to further increase the size of B. Indeed, as $|B| \to \infty$, we have that $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ becomes closer and closer to

$$\bar{\pi}^*(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{S}. \tag{11}$$

Namely, in the limit where $B = \boldsymbol{\Theta}$, the prior $g(\boldsymbol{\theta})$ becomes equivalent to an improper uniform prior on $\boldsymbol{\theta}$, for which the Bayesian estimators coincide with their frequentist counterparts. The posterior $\bar{\pi}^*(\boldsymbol{\theta}|\mathbf{y})$ contains only the information contained in the likelihood function, which is not affected or distorted by the prior. In this sense, when can be used (i.e., S is finite), a uniform improper prior is the maximal expression of a non-informative prior for the Level-1 of inference.

Level-2 of inference. We focus now on the marginal likelihood Z which, in this case, is given by

$$Z = \frac{\int_B \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}{|B|}. \tag{12}$$

Now, consider increasing B until we cover all parameter space. In this situation,

$$|B| \to \infty, \quad \text{but} \quad \int_B \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \to S,$$

Hence,

$$\lim_{|B| \to \infty} Z = 0. \tag{13}$$

We see that the marginal likelihood of a model with a increasingly-diffuse uniform proper prior becomes null. This is because increasing the spread of the prior penalizes more and more the considered model. Note that, in Level-2 of inference, a diffuse uniform prior is actually highly informative.

Now, we can already deduce some conclusions, highlighted below.

Remark 6. *In Level-1 inference, if $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$ is finite, we can use a diffuse proper or improper prior as non-informative choice. Moreover, under the assumption of strong data*[1]*, and if we vary the prior density, the estimators $\widehat{\boldsymbol{\theta}}_{mean}$, $\widehat{\boldsymbol{\theta}}_{MAP}$ do not change drastically. In the case, with the use of an improper uniform prior we can recover the frequentist results (Consonni et al., 2018).*

Remark 7. *In Level-2 inference, the concept of non-informative prior cannot be applied. Any choice of prior (also a diffuse, flat one) is actually very informative. If $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$ is finite,*

[1] With "strong data", we refer to a dataset under which the likelihood function is very concentrated (many data or data is very informative).

diffuse priors tend to produce smaller values of the marginal likelihood Z (Cameron & Pettitt, 2014; Bernardo & Smith, 1994a). 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 (Bernardo & Smith, 1994b; MacKay, 2003; R Oaks et al., 2019; Llorente et al., 2020).

Remark 8. *The evidence Z contains an implicit penalization of the model complexity. See Appendices A-B and (MacKay, 2003, Ch. 28)(Knuth et al., 2015).*

4.4 Issues with improper priors for model selection

In the previous section, we have already discussed the sensitivity of Z to variations of the spread of the prior density, and the fact a diffuse prior is highly informative in the Level-2 of inference. Even more caution is needed in the case of employing improper priors.

We have seen that the use of improper priors, $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, is allowed for Level-1 inference when $\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, since the corresponding posteriors are proper. However, improper priors are not allowed for the Level-2 (model selection). We describe this fact below and some possible solutions in the rest of the work.

The use of improper priors is common in Level-1 of inference to represent weak a-priori 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, $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$. 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 $\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, indeed

$$\begin{aligned}\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{Z}\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})ch(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})ch(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}}, \\ = \frac{1}{Z_h}\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta}) \qquad (14)\end{aligned}$$

where $Z = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$, $Z_h = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$ 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 $\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. However, the issue is not solved when we compare different models, since $Z = cZ_h$ depends on undetermined value c. For instance, the Bayes factors depend on the undetermined constants $c_1, c_2 > 0$ (D. J. Spiegelhalter & Smith, 1982),

$$\text{BF}(\mathbf{y}) = \frac{c_1}{c_2}\frac{\int_{\Theta_1} \ell_1(\mathbf{y}|\boldsymbol{\theta})h_1(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Theta_2} \ell_2(\mathbf{y}|\boldsymbol{\theta})h_2(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{Z_1}{Z_2} = \frac{c_1 Z_{h_1}}{c_2 Z_{h_2}}, \qquad (15)$$

so that different choices of c_1, c_2 provide different preferable models. There exists various approaches for dealing with this issue, as we show in the next section. More generally, we describe different solutions for a safe choice of the priors in the Level-2 of inference.

5 Objective 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 of the evidence Z with the prior choice $g(\boldsymbol{\theta})$. Moreover, we have seen that in model selection (Level-2), the concept of non-informative prior cannot be directly applied, since any kind of prior is actually informative in Level-2. For instance, diffuse priors can be very informative in the Level-2 of inference.

We define as a safe scenario, an approach where the choice of the priors is virtually not favoring any of the models (i.e., in some sense, the choice of the priors seeks to obtain *impartiality* in the model selection problem (Gelman & Hennig, 2017)), 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 a *subjective* choice of the priors. Many solutions proposed in the literature are data-driven approaches (see Section 5.3). In Section 5.3.3, we also discuss an alternative approach for model selection in Bayesian statistics (Vehtari et al., 2017, Ch. 6)(Piironen & Vehtari, 2017).

5.1 Same priors in nested models

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 parameter 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 (Kass & Raftery, 1995, Sect. 5.3), in the context of testing hypothesis, some authors consider 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.

5.2 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 *hyper*-parameters of the priors are considered (Gelman et al., 2013; Bernardo & Smith, 1994a). 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 perform 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 (Bernardo & Smith, 1994a).

Mathematically speaking, let us denote $g(\boldsymbol{\theta}|\boldsymbol{\nu})$ our family of priors over $\boldsymbol{\theta}$ with hyper-parameters $\boldsymbol{\nu} \in \mathbb{R}^{\xi}$. Below, we discuss two possible solutions.

Empirical Bayes approach. In this case we can compute the evidence in Eq. (6)) as function of $\boldsymbol{\nu}$, i.e., $Z(\boldsymbol{\nu}) = p(\mathbf{y}|\boldsymbol{\nu}) = \int_{\mathbb{R}^{\xi}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\nu})d\boldsymbol{\theta}$, and then set

$$\boldsymbol{\nu}^* = \arg\max_{\boldsymbol{\nu}} Z(\boldsymbol{\nu}). \tag{16}$$

Thus, we can use $g(\boldsymbol{\theta}|\boldsymbol{\nu}^*)$ as prior over the parameter $\boldsymbol{\theta}$, in our inference (Liang et al., 2008; Petrone et al., 2014). Note that, in this approach, the choice of the prior is in some sense *data-driven*, since $\boldsymbol{\nu}^*$ is obtained by the maximization of $p(\mathbf{y}|\boldsymbol{\nu})$ (see also Section 5.3).

Full Bayesian approach. Assuming an hyper-prior $g_h(\boldsymbol{\nu})$, the complete posterior is given by the following expression,

$$\bar{\pi}(\boldsymbol{\theta}, \boldsymbol{\nu}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\nu})g_h(\boldsymbol{\nu})}{Z_{\texttt{new}}}, \tag{17}$$

where

$$\begin{aligned} Z_{\texttt{new}} = p(\mathbf{y}) &= \int_{\boldsymbol{\Theta}} \int_{\mathbb{R}^{\xi}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\nu})g_h(\boldsymbol{\nu})d\boldsymbol{\theta}d\boldsymbol{\nu}, \qquad (18) \\ &= \int_{\mathbb{R}^{\xi}} Z(\boldsymbol{\nu})g_h(\boldsymbol{\nu})d\boldsymbol{\nu}, \qquad (19) \end{aligned}$$

is a Bayesian evidence that takes into account all the members of the prior family. Clearly, the model selection scheme based on $Z_{\texttt{new}}$ could be consider more robust than a model selection approach based on a single marginal likelihood $Z = Z(\boldsymbol{\nu})$, only computing 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 (Llorente et al., 2020). Also in the empirical Bayes scheme, we need to compute several values $Z(\boldsymbol{\nu})$'s for different $\boldsymbol{\nu}$'s, in order to perform the optimization in (16).[2] Hence, this approach can be much more computational demanding. Moreover, the hierarchical framework moves (in some sense) the problem "to another level", where we have to choose the hyper-prior $g_h(\boldsymbol{\nu})$ or, in the simplest case, we have at least to decide one possible value $\boldsymbol{\nu}^*$ for setting $g(\boldsymbol{\theta}|\boldsymbol{\nu}^*)$.

Even in this last scenario (and when S is finite), we could choose $\boldsymbol{\nu}^*$ such that the prior $g(\boldsymbol{\theta}|\boldsymbol{\nu}^*)$ is diffuse, reducing arbitrarily the value of the evidence Z (potentially approaching zero). It is also important to notice that this problem is shared with all the modern statistics, machine learning, and signal processing fields. Indeed, we have always some parameters to tune that can

[2] Note that analytical solutions are generally not available.

dramatically change the results (e.g., regularization parameters in Ridge Regression, LASSO etc. (Bishop, 2006; Martino & Read, 2021)). Hence, the real question is whether one can set these tunable parameters to reasonable values.

5.3 Data-driven and model-based approaches

Here, we describe different strategies for constructing data-driven or model-based objective priors. Some ideas for using improper priors in the Level-2 of inference, and other possible approaches for Bayesian model selection are also discussed.

5.3.1 Likelihood-based priors

In this section, we describe possible simple data-driven ideas for setting the priors, presented in an increasing order of complexity, i.e., starting from the simplest idea and describing progressively more sophisticated approaches (proposed in the literature).

Idea-1. When $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, we can build a proper prior based on the data and the observation model. For instance, we can choose $g_{\text{like}}(\boldsymbol{\theta}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}$, then the marginal likelihood is

$$Z = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) g_{\text{like}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{\int_{\boldsymbol{\Theta}} \ell^2(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{20}$$

We can consider $g_{\text{like}}(\boldsymbol{\theta})$ a non-subjective prior in the sense that does not incorporate any additional information, since is only based on the data. This idea is also connected to *posterior predictive approach*, that is described in Section 5.3.3. However, this prior can be very informative and we use the data twice, so that other approaches can be designed for dealing with these issues.

Idea-2. 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 $\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta} d\boldsymbol{\theta} < \infty$ or $\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, 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

$$Z = \frac{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta+1} d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta} d\boldsymbol{\theta}}, \qquad \text{or} \qquad Z = \frac{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}_{\text{sub}}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{21}$$

However, we still use a subset of the data twice.

Idea-3: Data partition. In order to avoid to use part of the data twice, we can divide the data in two subsets, $\mathbf{y} = (\mathbf{y}_{\text{train}}, \mathbf{y}_{\text{test}})$. Then, if $S_{\text{train}} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}_{\text{train}}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, we use $g_{\text{like}}(\boldsymbol{\theta}) = \frac{1}{S_{\text{train}}} \ell(\mathbf{y}_{\text{train}}|\boldsymbol{\theta})$, obtaining

$$Z = \frac{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}_{\text{test}}|\boldsymbol{\theta}) \ell(\mathbf{y}_{\text{train}}|\boldsymbol{\theta})d\boldsymbol{\theta}}{S_{\text{train}}}. \tag{22}$$

If the data are conditionally independent given $\boldsymbol{\theta}$, we have that $\ell(\mathbf{y}_{\text{test}}|\boldsymbol{\theta})\ell(\mathbf{y}_{\text{train}}|\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta})$ and

$$Z = \frac{S}{S_{\text{train}}}. \tag{23}$$

In order to build the less possible informative $g_{\text{like}}(\boldsymbol{\theta})$, we can look for the *minimal* training sets $\mathbf{y}_{\text{train}} = \mathbf{y}_{\text{min}}$, i.e., the sets with a minimum number of data, such that $S_{\text{min}} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}_{\text{min}}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$ (Berger & Pericchi, 1996). The dependence on the specific partition can be alleviated by averaging over different partitions. Assume that R is the number of considered partitions. Let us also assume that for each possible training set $\mathbf{y}_{\text{train}}^{(r)}$, we have $S_{\text{train}}^{(r)} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}_{\text{train}}^{(r)}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$, for $r = 1, ..., R$. Thus, we can build R different priors $g_{\text{train}}^{(r)}(\boldsymbol{\theta}) = \frac{1}{S_{\text{train}}^{(r)}}\ell(\mathbf{y}_{\text{train}}^{(r)}|\boldsymbol{\theta})$ and then consider a mixture of posterior densities, each one with a different prior $g_{\text{train}}^{(r)}(\boldsymbol{\theta})$. In this case, we obtain $Z = \frac{1}{R}\sum_{r=1}^{R} \frac{S}{S_{\text{train}}^{(r)}}$, where recall that $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$. This approach is related to the *Partial and Intrinsic Bayes factors* (O'Hagan, 1995; Berger & Pericchi, 1996).

Connection with partial and intrinsic Bayes factors. Let $g_{\text{base}}(\boldsymbol{\theta})$ denote an improper baseline prior. We already discussed that using improper priors produce marginal likelihoods that are specified up to an arbitrary constant (see Sect. 4.4). Partial Bayes factors are solutions proposed for dealing with this issue, and are based on the same idea of training the prior using some partial likelihood (O'Hagan, 1995, Sect. 2). As a result, each model is assigned a marginal likelihood in the form of Eq. (23), but also considering the improper baseline $g_{\text{base}}(\boldsymbol{\theta})$, i.e.,

$$Z = \frac{\widetilde{Z}}{\widetilde{Z}_{\text{train}}} = \frac{\int \ell(\mathbf{y}|\boldsymbol{\theta})g_{\text{base}}(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \ell(\mathbf{y}_{\text{train}}|\boldsymbol{\theta})g_{\text{base}}(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{24}$$

Note that any arbitrary constant contained in $g_{\text{base}}(\boldsymbol{\theta})$ is canceled out in the computation of Z. Hence, the final Bayes factor (called partial Bayes factor) between any two models is

$$\text{BF}_{12}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}) = \frac{Z_1}{Z_2} = \frac{\widetilde{Z}_1/\widetilde{Z}_{\text{train},1}}{\widetilde{Z}_2/\widetilde{Z}_{\text{train},2}} = \frac{\widetilde{Z}_1/\widetilde{Z}_2}{\widetilde{Z}_{\text{train},1}/\widetilde{Z}_{\text{train},2}} = \frac{\text{BF}_{12}(\mathbf{y})}{\text{BF}_{12}(\mathbf{y}_{\text{train}})}, \tag{25}$$

where we have denoted $\text{BF}_{12}(\mathbf{y}) = \frac{\widetilde{Z}_1}{\widetilde{Z}_2}$ and $\text{BF}_{12}(\mathbf{y}_{\text{train}}) = \frac{\widetilde{Z}_{\text{train},1}}{\widetilde{Z}_{\text{train},2}}$. Clearly, we should take $\mathbf{y}_{\text{train}}$ of minimal size. As above, in order to reduce the sensitivity of the results, we can average $\text{BF}_{12}(\mathbf{y}_{\text{test}}^{(r)}|\mathbf{y}_{\text{train}}^{(r)})$ over the possible R partitions, leading to the *intrinsic* Bayes factors (Berger & Pericchi, 1996).

Idea-4: Powered likelihood. Another alternative given in the literature is the following. We can use a powered likelihood $\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta}$ with $0 < \beta < 1$ to obtain the prior, and employ as likelihood also a tempered version, i.e., $\ell(\mathbf{y}|\boldsymbol{\theta})^{1-\beta}$, so that we have

$$g_{\text{like}}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}|\beta) \propto \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta}, \quad \text{and} \quad \bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \ell(\mathbf{y}|\boldsymbol{\theta})^{1-\beta}\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta}, \tag{26}$$

Note that, in this case, we do not need the conditionally independent assumption to express the marginal likelihood as ratio of normalizing constants, i.e.,

$$Z = \int \ell(\mathbf{y}|\boldsymbol{\theta})^{1-\beta}g(\boldsymbol{\theta}|\beta)d\boldsymbol{\theta} = \frac{\int \ell(\mathbf{y}|\boldsymbol{\theta})^{1-\beta}\ell(\mathbf{y}|\boldsymbol{\theta})^{\beta}d\boldsymbol{\theta}}{\int \ell(\mathbf{y}|\boldsymbol{\theta})^{\beta}d\boldsymbol{\theta}} = \frac{S}{S_{\beta}}. \tag{27}$$

Furthermore, we get rid of the indeterminacy of choosing the partition. However, a tempering value $\beta \in (0,1)$ must be selected. This idea is also employed in the so-called *Fractional Bayes Factors* (O'Hagan, 1995).

Connection with fractional Bayes factors. Fractional Bayes factors are another strategy proposed for dealing with an improper baseline $g_{\text{base}}(\boldsymbol{\theta})$. This time each model is assigned a marginal likelihood analogous to that of Eq. (27) but consider the baseline $g_{\text{base}}(\boldsymbol{\theta})$, i.e.

$$Z = \frac{\widetilde{Z}}{\widetilde{Z}_\beta} = \frac{\int \ell(\mathbf{y}|\boldsymbol{\theta}) g_{\text{base}}(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \ell(\mathbf{y}|\boldsymbol{\theta})^\beta g_{\text{base}}(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{28}$$

This marginal likelihood is free of arbitrary constants. The final Bayes factor (called fractional Bayes factor) between any two models is given as

$$\text{FBF}_{12} = \frac{Z_1}{Z_2} = \frac{\widetilde{Z}_1/\widetilde{Z}_{\beta,1}}{\widetilde{Z}_2/\widetilde{Z}_{\beta,2}} = \frac{\widetilde{Z}_1/\widetilde{Z}_2}{\widetilde{Z}_{\beta,1}/\widetilde{Z}_{\beta,2}} = \frac{\text{BF}_{12}(\mathbf{y})}{\text{BF}_{12}(\mathbf{y}|\beta)}, \tag{29}$$

where we denoted $\text{BF}_{12}(\mathbf{y}|\beta) = \frac{\widetilde{Z}_{\beta,1}}{\widetilde{Z}_{\beta,2}}$.

Remark 9. *Note that the idea above can solve the problem of using improper priors for computing marginal likelihoods. This solution is consequence of transforming the improper base-prior $g_{\texttt{base}}(\boldsymbol{\theta})$ into a proper density using some partial likelihood function.*

Idea-5: Power-prior. In the literature, other approaches with simulated data have been proposed (Consonni et al., 2018). Let $\mathbf{y}^*$ denote some imaginary data (i.e., artificial/simulated data) and consider the following *power-prior* (Ibrahim et al., 2015)

$$g_{\text{like}}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}|\mathbf{y}^*, \beta) \propto \ell(\mathbf{y}^*|\boldsymbol{\theta})^\beta g_{\texttt{base}}(\boldsymbol{\theta}), \text{ where } \quad 0 < \beta < 1. \tag{30}$$

An important special case of power priors are the well-known *g-priors*, which is an standard prior choice in linear models (Zellner, 1986; Liang et al., 2008). A mixture of g-priors is an objective choice designed for the linear regression setting, that fulfills desirable model selection criteria (Bayarri et al., 2012).

Two further generalizations have been proposed in the literature. if we consider $\mathbf{y}^*$ are not fixed, but random, we can take an additional step consisting in averaging the prior in Eq. (30) with respect to the distribution of the simulated data $\mathbf{y}^*$. The resulting prior is thus

$$g_{\text{like}}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}|\beta) = \int g(\boldsymbol{\theta}|\mathbf{y}^*, \beta) q(\mathbf{y}^*) d\mathbf{y}^*, \tag{31}$$

where $q(\mathbf{y}^*)$ is the distribution of the artificial data. With $\beta = 1$, the above expression is called *expected posterior prior* (EPPs) (Pérez & Berger, 2002). Moreover, in the case where all likelihoods (including that of the posterior) are raised to a common power β and normalized, we obtain the so-called *power expected posterior prior* (PEP priors) (Fouskakis et al., 2015).

Note that most of the approaches above required that $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$ be finite, otherwise they cannot be applied. However, in this case, the problem is extended to the Level-1 of Bayesian inference since the posterior would be not proper using an improper prior.

5.3.2 Other model-based approaches for building the prior

Other ways of designing objective priors consider the use the information contained in the Fisher information matrix, i.e.,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\ell(\mathbf{y}|\boldsymbol{\theta})}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log \ell(\mathbf{y}|\boldsymbol{\theta})\right)^2\right], \tag{32}$$

where the expectation is w.r.t. $\ell(\mathbf{y}|\boldsymbol{\theta})$ (fixing $\boldsymbol{\theta}$). With a Jeffrey's approach, one takes the prior to be $g(\boldsymbol{\theta}) \propto [\mathcal{I}(\boldsymbol{\theta})]^{-\frac{1}{2}}$. This prior has the property of being invariant under change of variables (Kass & Wasserman, 1996).

The *unit information prior* (UIP) is based on the idea that the information encoded in a prior pdf should be roughly the amount of information contained in a single data (Consonni et al., 2018). The Fisher information matrix divided by the number of data, i.e., $\frac{1}{D_y}\mathcal{I}(\boldsymbol{\mu})$, is thus proposed as an estimate of this information. For instance, for a continuous parameter, $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$, we can take a Gaussian prior, i.e.,

$$g(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}\middle|\boldsymbol{\mu}, \left[\frac{1}{D_y}\mathcal{I}(\boldsymbol{\mu})\right]^{-1}\right),$$

where $\boldsymbol{\mu}$ is a prior mean. In linear models, the UIP takes the same form as the g-prior (Consonni et al., 2018). Furthermore, the use of UIP is motivated since it produces a log-Bayes factor that is asymptotically equivalent to the BIC (Kass & Wasserman, 1996; Consonni et al., 2018).

5.3.3 Posterior predictive approach

The marginal likelihood approach is not the unique approach for model selection in Bayesian statistics. Here, we discuss an alternative strategy (called *predictive model selection*), that is based on the concept of prediction (Vehtari et al., 2017, Ch. 6)(Vehtari & Ojanen, 2012; Piironen & Vehtari, 2017). 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 (Vehtari et al., 2017; Piironen & Vehtari, 2017). Hence, a key quantity in these approaches is the posterior predictive distribution of generic different data $\widetilde{\mathbf{y}}$ given $\mathbf{y}$,

$$\begin{aligned} p(\widetilde{\mathbf{y}}|\mathbf{y}) = E_{\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})}[\ell(\widetilde{\mathbf{y}}|\boldsymbol{\theta})] &= \int_{\boldsymbol{\Theta}} \ell(\widetilde{\mathbf{y}}|\boldsymbol{\theta})\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \\ &= \frac{1}{Z}\int_{\boldsymbol{\Theta}} \ell(\widetilde{\mathbf{y}}|\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned} \tag{33}$$

Considering $\widetilde{\mathbf{y}} = \mathbf{y}$, we can observe that exists a clear connection with likelihood-based priors described in Section 5.3.1. Indeed, if we assume $g(\boldsymbol{\theta}) \propto 1$ and $\widetilde{\mathbf{y}} = \mathbf{y}$, Eq. (33) becomes Eq. (20). Note that the posterior predictive distribution in Eq. (33) 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 4. With a generic $g(\boldsymbol{\theta})$ and $\widetilde{\mathbf{y}} = \mathbf{y}$, the above expression can seen as a marginal likelihood obtained using the posterior as a prior pdf, stressing even more the approach in Idea-1 described in Section 5.3.1. It can be also considered as a "posterior" Bayes factor, in the sense that the likelihood is averaged w.r.t. the posterior, rather than the prior (Aitkin, 1991). Clearly, 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) (Vehtari et al., 2017). A drawback of predictive model selection is that consistency (i.e., selecting the true model as $D_y \rightarrow \infty$) is not generally ensured (Vehtari & Ojanen, 2012).

6 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 (Gregory, 2011; Barros et al., 2016). Some related code is also provided.[3]

6.1 Experiment 1

Let us consider the following Gaussian conjugate model for θ,

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

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

$$\begin{aligned} \mu_{\text{post}} &= \frac{1}{\frac{1}{\sigma_0^2} + \frac{D_y}{\sigma^2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{D_y \bar{y}}{\sigma^2} \right) \\ \sigma^2_{\text{post}} &= \left(\frac{1}{\sigma_0^2} + \frac{D_y}{\sigma^2} \right)^{-1}, \end{aligned}$$

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

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

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

where $\sigma_n = \frac{\sigma}{\sqrt{D_y}}$ and v_y denotes the sample variance of $\mathbf{y}$. We consider a single data point ($D_y = 1$), where $\mathbf{y} = y = 2.078$. 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. Hence, the improper prior is non-informative for Level-1 of inference. 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). This result is consequence of the Jeffrey-Lindley-Bartlett paradox (Lindley, 1957; Villa & Walker, 2017). This shows that diffuse priors are very informative in Level-2 of Bayesian inference.

6.2 Experiment 2: Normal linear regression

Let us consider the normal linear regression setting with two models for the observations $\mathbf{y} = \{y_i\}_{i=1}^{D_y}$,

$$\mathcal{M}_0 : y_i = \beta_0 + \epsilon_i,$$
$$\mathcal{M}_1 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\{x_i\}_{i=1}^{D_y}$ are fixed/known and $\epsilon_i \sim \mathcal{N}(0, \sigma_{\text{like}}^2)$ with σ_{like} known. Hence, model $\mathcal{M}_0$ has parameter $\boldsymbol{\theta}_0 = \beta_0$, and model $\mathcal{M}_1$ has parameter $\boldsymbol{\theta}_1 = (\beta_0, \beta_1)$. We set Gaussian priors for both models,

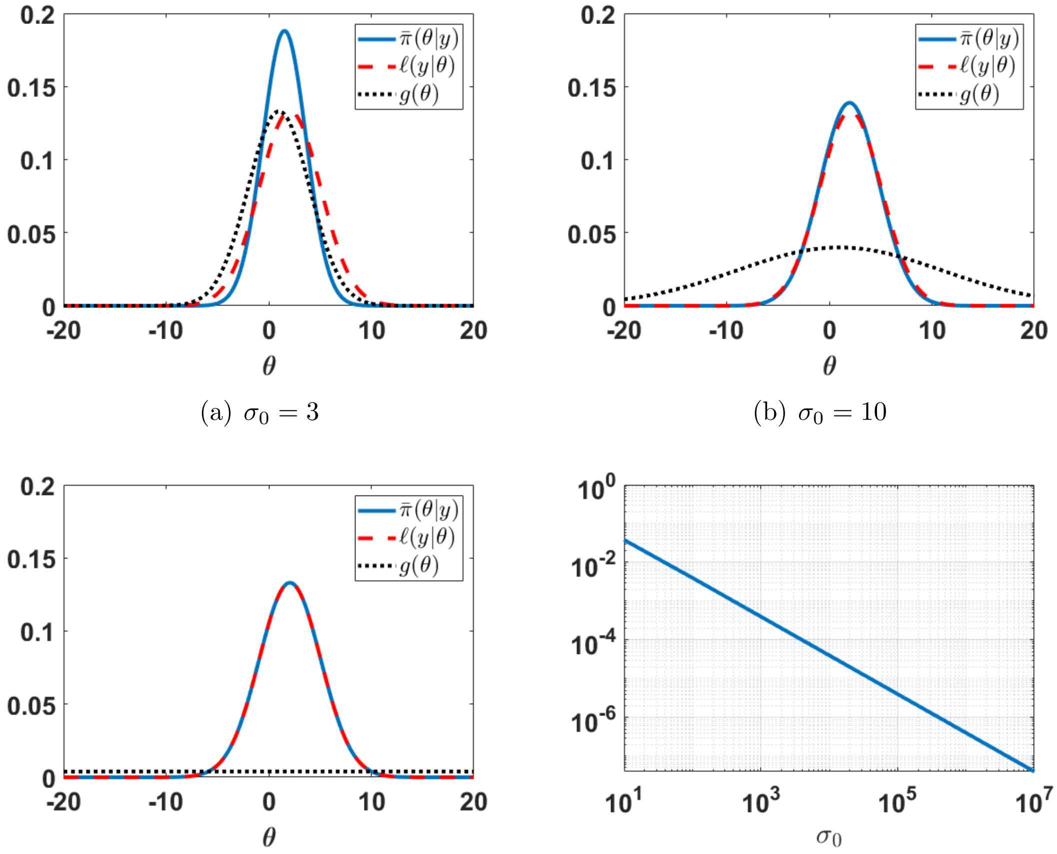
$$g_0(\beta_0) = \mathcal{N}(\beta_0|0, \sigma_0^2) \quad \text{and} \quad g_1(\beta_0, \beta_1) = g_0(\beta_0)\mathcal{N}(\beta_1|0, \sigma_1^2). \tag{34}$$

We aim to analyze the sensitivity of the Bayes factor BF_{01}, given by

$$\text{BF}_{01} = \frac{Z_0}{Z_1} = \frac{\int \ell(\mathbf{y}|\beta_0) g_0(\beta_0) d\beta_0}{\int \ell(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1) g_1(\beta_0, \beta_1) d\beta_0 d\beta_1}, \tag{35}$$

where $g_1(\beta_0, \beta_1) = g_0(\beta_0)\mathcal{N}(\beta_1|0, \sigma_1^2)$, and when we vary different features such as the dispersions σ_0 and σ_1.

Sensitivity w.r.t. the choice of σ_0. We generate $D_y = 4$ observations from model $\mathcal{M}_1$ with $\beta_0^{\text{true}} = \beta_1^{\text{true}} = 1$. We consider $\sigma_1 = 1$ fixed and we compute BF_{01} for a sequence of increasing values of σ_0. We show the Bayes factor BF_{01} versus σ_0 in Figure 2(a). We see that BF_{01} is much lower than 1 for every σ_0, indicating that $\mathcal{M}_1$ is the preferred model. As expected, BF_{01} is stable



(a) $\sigma_0 = 3$

(b) $\sigma_0 = 10$

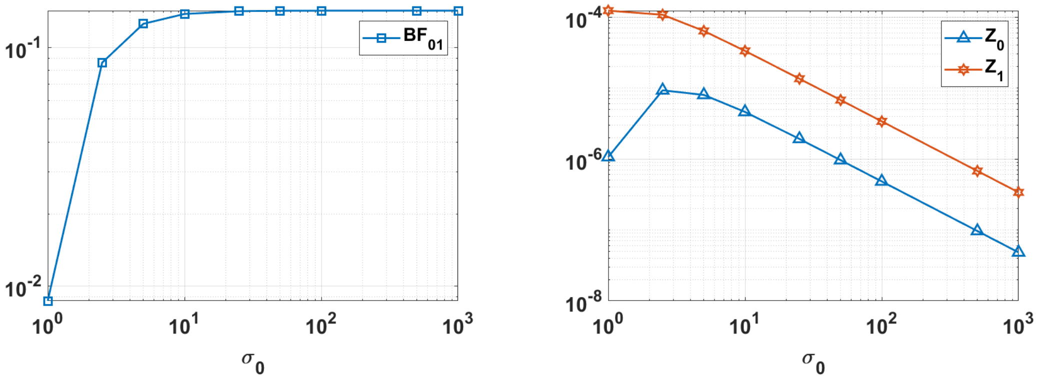
(c) $\sigma_0 = 100$

(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.

under increasing σ_0, reaching a plateau at $\sigma_0 = 10$ and becoming constant from there on. This is a well-known fact: the choice of prior for the common parameter β_0 does not affect much the comparison. This is a consequence of choosing the same prior $g_0(\beta_0)$ for both models. In Figure 2(b), we see that increasing σ_0 reduces the marginal likelihood of both models simultaneously, hence the pitfalls of using increasingly diffuse priors are solved when we compute the quotient.

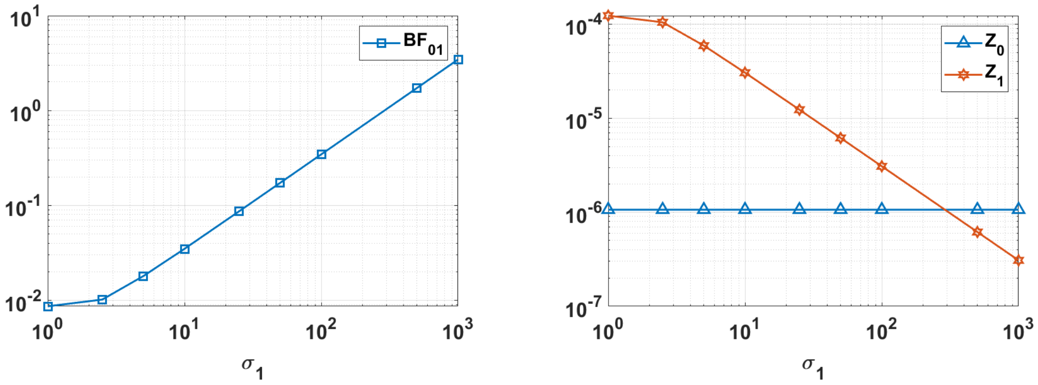
Sensitivity w.r.t. the choice of σ_1. We repeat the experiment but considering a fixed $\sigma_0 = 1$, and we compute BF_{01} for a sequence of increasing values of σ_1. We show the Bayes factor BF_{01} and both marginal likelihoods Z_0, Z_1 versus σ_1 in Figure 3. Opposite to the previous case, this time we see that BF_{01} is greater than 1 when $\sigma_1 > 500$. Indeed, in Figure 3(b), we see that only Z_1 decreases as σ_1. This is due to we are only varying the dispersion of the prior in model $\mathcal{M}_1$ not $\mathcal{M}_0$. As a consequence, increasing the dispersion of the prior on β_1 makes us eventually choose the wrong model $\mathcal{M}_0$ (again, this is the Lindley-Bartlett paradox).



(a) Bayes factor

(b) Z_0 and Z_1

Figure 2: In **(a)** Bayes factor versus σ_0 in log-scale; In **(b)** Marginal likelihoods of models $\mathcal{M}_0$ and $\mathcal{M}_1$ versus σ_0 in log-scale. We consider $\sigma_1 = 1$ is fixed.



(a) Bayes factor

(b) Z_0 and Z_1

Figure 3: In **(a)** Bayes factor versus σ_1 in log-scale; In **(b)** Marginal likelihoods of models $\mathcal{M}_0$ and $\mathcal{M}_1$ versus σ_1 in log-scale. We consider $\sigma_0 = 1$ is fixed.

6.3 Experiment 3

6.3.1 First analysis

Let us consider the problem of selecting between two models, $\mathcal{M}_1 = \{\ell_1(y|\theta) = \theta^y e^{-\theta}/y!,\ g_1(\theta)\}$ and $\mathcal{M}_2 = \{\ell_2(y|\phi) = \phi(1-\phi)^y,\ g_2(\phi)\}$, namely a Poisson and a geometric distribution (Lindley, 1957). We use a uniform prior $g_2(\phi) = 1$ for the proportion $\phi \in [0,1]$, and also a uniform prior $g_1(\theta) = \frac{1}{L}$ for $\theta \in [0, L]$. We generate D_y independent data $\mathbf{y} = (y_1, \ldots, y_{D_y})$ from $\mathcal{M}_1$ with $\theta_{\text{true}} = 2$. The goal of this example is to show empirically the sensitivity of the Bayes factor to increasing L (i.e., $g_1(\theta)$ becomes more diffuse), and the number of data D_y. For doing this, for each pair of values (L, D_y), we study the average number of *errors* in model selection (i.e., the number of times $\text{BF}_{12} < 1$) in 100 independent simulated datasets of size D_y.

First, we compute the number of errors as we increase L for two fixed sample sizes, $D_y = 30$ and $D_y = 100$. Table 1 shows the results when $D_y = 30$ for the values $L = 10^\alpha$ ($\alpha = 1, \dots, 6$). Specifically, we show the maximum and minimum values of BF_{12}, obtained in the 100 simulations, along with the number of errors. As expected, 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 $\mathcal{M}_1$ (i.e., the Lindley-Bartlett paradox). Table 2 shows results when $D_y = 100$. On the contrary, we observe here that the number of errors is very low even for large L, namely, having more data compensates the potential drawbacks of using a very diffuse prior. In addition, in Figure 4(a), 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 rapidly reduces the number of times we choose the wrong model. Figure 4(b) shows the average number of errors as a function of both L and D_y. We can see again that for fixed L, the number of errors is very sensitive to increasing D_y. Namely, a small increase in sample size produces a large reduction in the average number of errors (i.e. the results are consistent). On the other hand, the number of errors is rather insensitive to increasing L, as compared to D_y. In fact, for $D_y > 50$, the number of errors remains constant and close to 0 for all the considered values of L (up to $L = 10^4$). Although increasing L eventually gives the wrong results, this effect is noticeable only when the sample size is small enough.
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 4). 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 (Llorente et al., 2020; Bos, 2002).

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

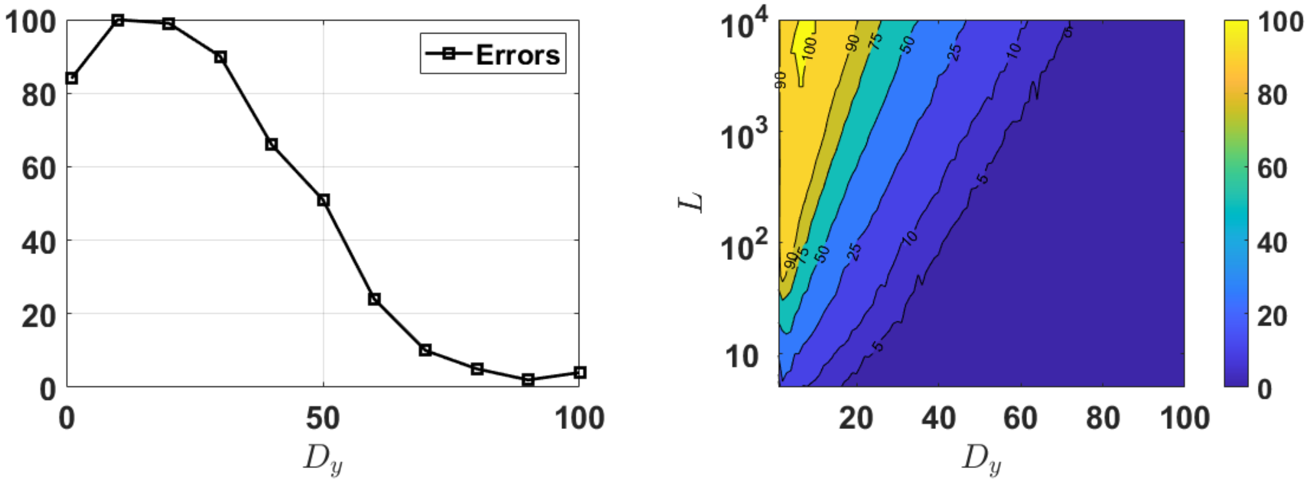
True model = $\mathcal{M}_1$ (with $\theta_{\text{true}} = 2$)			
L	min	max	Errors in model choice, over 100 simulations
10	0.094	4.77×10^5	3
10^2	0.059	2.49×10^4	15
10^3	0.0012	1.46×10^3	31
10^4	1.06×10^{-4}	339.86	67
10^5	1.02×10^{-4}	41.05	84
10^6	1.59×10^{-6}	0.7080	100

6.3.2 Using partial and intrinsic BFs

Consider again the comparison between a Poisson distribution and a geometric distribution. Previously in Section 6.3, 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 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 1 and

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



(a)

(b)

Figure 4: In **(a)** 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. **(b)** The average number of errors for different values of L and D_y.

2 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 (O'Hagan, 1995, 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 (Berger & Pericchi, 1996), called the *intrinsic* Bayes factor (IBF). Let $\mathbf{y}_{-i}$ denote the vector of all D_y data 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,

$$\mathrm{IBF}_{12} = \frac{1}{D_y}\sum_{i=1}^{D_y} \frac{\int_0^\infty \ell_1(\mathbf{y}_{-i}|\theta)\bar{\pi}_1(\theta|y_i)d\theta}{\int_0^1 \ell_2(\mathbf{y}|\phi)d\phi} = \frac{1}{D_y}\sum_{i=1}^{D_y} \frac{\int_0^\infty \ell_1(\mathbf{y}|\theta)d\theta / \int_0^\infty \ell_1(y_i|\theta)d\theta}{\int_0^1 \ell_2(\mathbf{y}|\phi)d\phi}. \tag{36}$$

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. 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[4]. 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 $\mathcal{M}_1$ 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 and intrinsic 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. This results in using the following intrinsic Bayes factor

$$\mathrm{IBF}_{12} = \frac{1}{D_y}\sum_{i=1}^{D_y} \frac{\int_0^\infty \ell_1(\mathbf{y}_{-i}|\theta)\bar{\pi}_1(\theta|y_i)d\theta}{\int_0^1 \ell_2(\mathbf{y}_{-i}|\phi)\bar{\pi}_2(\phi|y_i)d\phi} = \frac{1}{D_y}\sum_{i=1}^{D_y} \frac{\int_0^\infty \ell_1(\mathbf{y}|\theta)d\theta / \int_0^\infty \ell_1(y_i|\theta)d\theta}{\int_0^1 \ell_2(\mathbf{y}|\phi)d\phi / \int_0^1 \ell_2(y_i|\phi)d\phi}. \tag{37}$$

We run 100 simulations employing this procedure and observed that the number of errors in detecting the model $\mathcal{M}_2$ when $\phi_{\text{true}} \in \{0.5, 0.8\}$ gets reduced to, respectively, 18 and 16 when $D_y = 30$.

6.4 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 (Gregory, 2011; Barros et al., 2016; Affer et al., 2019; Trifonov et al., 2019). The problem consists in fitting a dynamical model 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

[4] 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

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.

6.4.1 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

$$y_t = V_0 + \sum_{i=1}^{S} K_i \left[\cos\left(u_{i,t} + \omega_i\right) + e_i \cos\left(\omega_i\right)\right] + \xi_t, \tag{38}$$

with $t = 1, \dots, T$.[5] 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. The angle $u_{i,t}$ is the true anomaly of the planet i and it can be determined by solving a differential equation and applying an iterative procedure, as a function of P_i, τ_i and e_i (see (Martino, Llorente, et al., 2021; López-Santiago et al., 2021) for more details). For certain sets of parameters, this iterative procedure can be particularly slow and the computation of the likelihood becomes quite costly. The variable of interest $\boldsymbol{\theta}$ is the vector of dimension $D_{\boldsymbol{\theta}} = 1 + 5S$ (where S is the number of planets),

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

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

Table 5: Description of parameters in Eq. (38).

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	$\mathrm{m\,s^{-1}}$

The meaning of each parameter is given in Table 5. 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 $D_{\boldsymbol{\theta}} = 11$, etc. The Eq. (38) induces a likelihood function, i.e.,

$$\ell(\mathbf{y}|\boldsymbol{\theta}, \sigma_e) = \prod_{t=1}^{T} \ell(y_t|\boldsymbol{\theta}, \sigma_e),$$

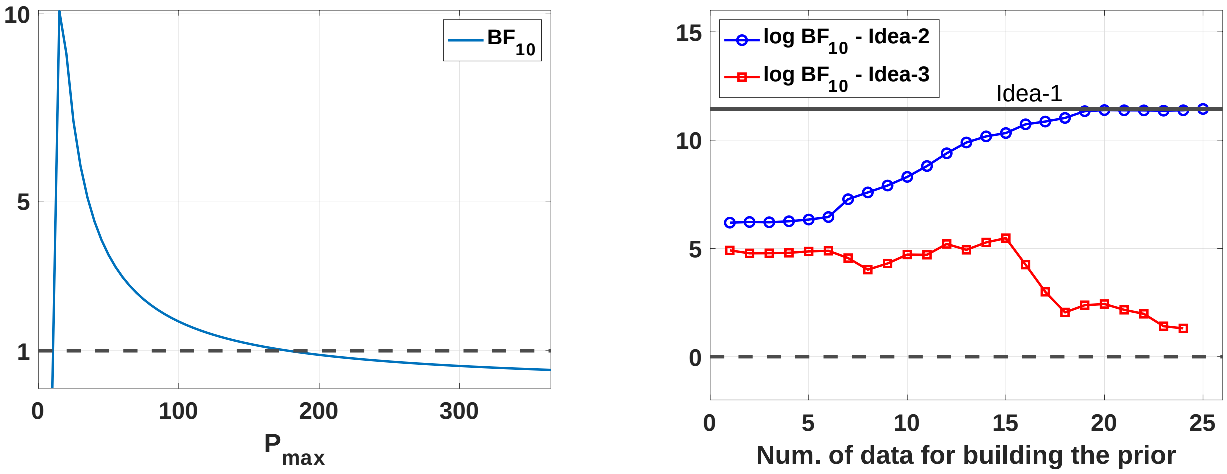
where $\mathbf{y} = \{y_1, \ldots, y_T\}$. Our goal is to infer the number S of planets in the system. For this purpose, given prior densities $g_i(\boldsymbol{\theta}_i)$ for each model, we have to approximate the model evidences

$$Z_i = \int_{\Theta_i} \ell(\mathbf{y}|\boldsymbol{\theta}_i, \sigma_e) g_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i.$$

For simplicity, we consider the noise variance σ_e^2 is given.

6.4.2 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 $D_y = 25$ 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 $\mathrm{BF}_{10} = \frac{Z_1}{Z_0}$, 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, $\boldsymbol{\theta}_0 = V_0$ and we choose a uniform 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. We use the same prior for V_1 in $\mathcal{M}_1$. For the period P_1, we use $\mathcal{U}([0, P_{\max}])$ with $P_{\max} > 0$. Namely, we use a uniform prior with varying width. When $P_{\max} = 365$, we are considering a uniform prior over all the possible values of P. 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 diffuse) 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 5(a), we show the Bayes factor as a function of $P_{\max}$. For $P_{\max}$ greater than 200, we have $\mathrm{BF}_{10} < 1$, that is, we wrongly choose the model with zero planets. This illustrates again the problematic with the use of vague priors.



(a)

(b)

Figure 5: **(a)** 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 diffuse) eventually produces $\mathrm{BF}_{10} < 1$. Note also that, when $P_{\max}$ is small (lower than $P_{\max} = 15$), we have $\mathrm{BF}_{10} < 1$, preferring the model with zero planet $\mathcal{M}_0$. **(b)** The log-BF_{10} obtained using likelihood-based priors in both models (specifically, the ideas 2 and 3 in Section 5.3.1) adding sequentially data in the prior construction. Note that log-$\mathrm{BF}_{10} > 0$ preferring always (and correctly) the model with one planet. Figure (b) also shows Idea-1 as limit of Idea-2, which provides an upper-bound for the rest of values.

Hierarchical solution. Let us denote as $Z_1(P_{\max})$, 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 $g_h(P_{\max}) = \mathcal{U}([10, 365])$, hence the new marginal likelihood is

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

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. Note that $g_h(P_{\max}) = \mathcal{U}([10, 365])$ is virtually the more diffuse hyper-prior that we can use in this experiment, since the parameter P represents a period of rotation (measure in “days”), so it varies between 0 and 365.

Likelihood-based priors. Another possible solution is to employ likelihood-based priors. We apply the Idea-2 and Idea-3 given in Section 5.3.1 to both models. In Idea-2, a subset of data is used twice (for building the prior and in the likelihood as well) whereas, in Idea-3, we split the data in training (for building the prior) and test (used only in the likelihood). Note that, if we use all the data ($D_y = 25$) for building the prior, Idea-2 becomes Idea-1 in Section 5.3.1. We start building the prior with only one data (the first one), and compute the corresponding BF_{10}. Then, we add sequentially the rest of data starting from the second one, until the 25-th data for Idea-2, and 24-th data for Idea-3. The log-BF_{10} is given in Figure 5(b). In this case, we always choose the true model. As expected, Idea-2 tends to promote more the more-complex model with respect to Idea-3. Again as expected, Idea-1 provides an upper bound for the BF_{10} obtained by Idea-2 and Idea-3.

7 Conclusions

In this work, we have highlighted some important considerations regarding the computation of marginal likelihoods, which are fundamental quantities for Bayesian model selection. We have discussed the dependence on the choice of the prior density and shown some comforting asymptotical results. Moreover, we have clarified that the use of improper priors is not suitable for model selection. More generally, we have also discussed the use of diffuse priors, whether proper (vague priors) or improper, are actually very informative for the model selection procedure (Level-2 of inference). We have shown by means of illustrative examples the potential pitfalls of using vague priors, and we have provided and discussed several possible solutions for all these scenario, such as the construction of likelihood-based or model-based priors, and partial/fractional Bayes factors. 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 detecting the number of objects orbiting a star.

Acknowledgments

The authors would like to thank the two anonymous referees for their detailed comments and suggestions. This work has been supported by Spanish government via grant FPU19/00815 and by Agencia Estatal de Investigación AEI (project SP-GRAPH, ref. num. PID2019-105032GB-I00).

References

Affer, L., et al. (2019, February). HADES RV program with HARPS-N at the TNG. IX. A super-Earth around the M dwarf Gl 686. *arXiv:1901.05338*, *622*, A193.

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(1), 111–128.

Anfinogentov, S. A., Nakariakov, V. M., Pascoe, D. J., & Goddard, C. R. (2021, January). Solar Bayesian Analysis Toolkit—A New Markov Chain Monte Carlo IDL Code for Bayesian Parameter Inference. *Astrophysical Journal Supplement Series*, *252*(1), 11. doi: 10.3847/1538-4365/abc5c1

Ashton, G., & Talbot, C. (2021, October). BILBY-MCMC: an MCMC sampler for gravitational-wave inference. *Monthly Notices of the Royal Astronomical Society*, *507*(2), 2037-2051. doi: 10.1093/mnras/stab2236

Ayuso, I., Lazkoz, R., & Salzano, V. (2021, March). Observational constraints on cosmological solutions of f (Q) theories. *Physical review d*, *103*(6), 063505. doi: 10.1103/PhysRevD.103.063505

Barros, S. C. C., et al. (2016). WASP-113b and WASP-114b, two inflated hot Jupiters with contrasting densities. *Astronomy and Aastrophysics*, *593*, A113. doi: 10.1051/0004-6361/201526517

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of statistics*, *40*(3), 1550–1577.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122.

Bernardo, J. M., & Smith, A. F. M. (1994a). *Bayesian Theory.* Wiley.

Bernardo, J. M., & Smith, A. F. M. (1994b). *Bayesian theory.* Wiley & sons.

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, *128*, 1–58.

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. In *Compstat* (pp. 111–116).

Cameron, E., & Pettitt, A. (2014). Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis. *Statistical Science*, *29*(3), 397–419.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*(453), 270–281.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679.

Dawid, A. P. (2011). Posterior model probabilities. In *Philosophy of statistics* (pp. 607–630). Elsevier.

Emmert, J., Grauer, S. J., Wagner, S., & Daun, K. J. (2019). Efficient Bayesian inference of absorbance spectra from transmitted intensity spectra. *Opt. Express*, *27*(19), 26893-26909.

Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. (2019, November). Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, *2*(1), 10. doi: 10.21105/astro.1306.2144

Fouskakis, D., Ntzoufras, I., & Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, *10*(1), 75–107.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis.* CRC press.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 967–1033.

Gregory, P. C. (2011, August). Bayesian re-analysis of the Gliese 581 exoplanet system. *Monthly Notices of the Royal Astronomical Society*, *415*(3), 2523-2545. doi: 10.1111/j.1365-2966.2011.18877.x

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 190-195.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*(4), 382-417.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, *34*(28), 3724–3749.

Jeffreys, H. (1998). *The theory of probability.* OUP Oxford.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American statistical Association*, *91*(435), 1343–1370.

Knuth, K. H., Habeck, M., Malakar, N. K., Mubeen, A. M., & Placek, B. (2015). Bayesian evidence and model selection. *Digital Signal Processing*, *47*, 50–67.

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling.* Springer Science & Business Media.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & B., J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing.* Springer.

Llorente, F., Martino, L., Delgado, D., & Lopez-Santiago, J. (2020). Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *(to appear) SIAM Review, - extended version in arXiv:2005.08334*.

Llorente, F., Martino, L., Delgado-Gomez, D., & Camps-Valls, G. (2021). Deep importance sampling based on regression for model inversion and emulation. *Digital Signal Processing*, *116*, 103104.

López-Santiago, J., Martino, L., Vázquez, M., & Miguez, J. (2021). A Bayesian inference and model selection algorithm with an optimization scheme to infer the model noise power. *Monthly Notices of the Royal Astronomical Society*, *507*(3), 3351-3361.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Martino, L., Elvira, V., Lopez-Santiago, J., & Camps-Valls, G. (2021). Compressed particle methods for expensive models with application in astronomy and remote sensing. *IEEE Transactions on Aerospace and Electronic Systems*, 1-15. doi: 10.1109/TAES.2021.3061791

Martino, L., Llorente, F., Cuberlo, E., López-Santiago, J., & Míguez, J. (2021). Automatic tempered posterior distributions for Bayesian inversion problems. *Mathematics*, *9*(7), 784.

Martino, L., & Read, J. (2021). A joint introduction to 0aussian Processes and Relevance Vector Machines with connections to Kalman filtering and other kernel smoothers. *Information Fusion*, *74*, 17–38.

Martino, L., Read, J., Elvira, V., & Louzada, F. (2017). Cooperative parallel particle filters for on-line model selection and applications to urban mobility. *Digital Signal Processing*, *60*, 172-185.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., ... others (2021). Prior knowledge elicitation: The past, present, and future. *arXiv preprint arXiv:2112.01380*.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118.

Pascoe, D. J., Smyrli, A., Van Doorsselaere, T., & Broomhall, A. M. (2020, December). Bayesian Analysis of Quasi-periodic Pulsations in Stellar Flares. *Astrophysical Journal*, *905*(1), 70. doi: 10.3847/1538-4357/abc69d

Pérez, J. M., & Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, *89*(3), 491–512.

Petrone, S., Rizzelli, S., Rousseau, J., & Scricciolo, C. (2014). Empirical Bayes methods in classical and Bayesian inference. *Metron*, *72*(2), 201–215.

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735.

R Oaks, J., A. Cobb, K., N Minin, V., & D. Leaché, A. (2019). Marginal likelihoods in phylogenetics: a review of methods and applications. *Systematic biology*, *68*(5), 681–697.

Robert, C. P. (2014). On the Jeffreys–Lindley paradox. *Philosophy of Science*, *81*(2), 216–232.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods.* Springer.

Rossell, D., & Rubio, F. J. (2021). Balancing Sparsity and Power: Likelihoods, Priors, and Misspecification. In *Handbook of bayesian variable selection* (pp. 371–394). Chapman and Hall/CRC.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Spiegelhalter, D., Best, N. G., Carlin, B. P., & der Linde, A. V. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*, 583-616.

Spiegelhalter, D. J., & Smith, A. F. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(3), 377–387.

Trifonov, T., Stock, S., Henning, T., Reffert, S., Kürster, M., Lee, M. H., ... Vogt, S. S. (2019, March). Two Jovian Planets around the Giant Star HD 202696: A Growing Population of Packed Massive Planetary Pairs around Massive Stars? *The Astronomical Journal*, *157*(3), 93. doi: 10.3847/1538-3881/aafa11

Urteaga, I., Bugallo, M. F., & Djurić, P. M. (2016). Sequential Monte Carlo methods under model uncertainty. In *2016 ieee statistical signal processing workshop (ssp)* (p. 1-5).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413–1432.

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

Villa, C., & Walker, S. (2017). On the mathematics of the Jeffreys–Lindley paradox. *Communications in Statistics-Theory and Methods*, *46*(24), 12290–12298.

Von Toussaint, U. (2011). Bayesian inference in physics. *Rev. Mod. Phys.*, *83*, 943–999.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

Appendices

A Implicit model penalization contained in Z

The illustrative example in Section 4.3.1 allows us to show that the marginal likelihood Z contains an implicit model penalization (MacKay, 2003, Ch. 28). In that example, we consider the uniform prior $g(\boldsymbol{\theta}) = \frac{1}{|B|}\mathbf{1}_B(\boldsymbol{\theta})$, where $|B|$ represents the volume of B. Without loss of generality, let us consider the case of B being a hypercube centered at the origin with side length δ, i.e, $B = [-\delta/2, \delta/2]^{D_{\boldsymbol{\theta}}} \subseteq \boldsymbol{\Theta}$, with volume $|B| = \delta^{D_{\boldsymbol{\theta}}}$. From Eq. (12), we have

$$\begin{aligned}\log Z &= \log \int_B \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} - \log |B|, \\ &= \log \int_B \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} - D_{\boldsymbol{\theta}} \log \delta. \end{aligned} \tag{39}$$

Note that both terms depend on the size δ and the dimensionality $D_{\boldsymbol{\theta}}$.[6] For a fixed $D_{\boldsymbol{\theta}}$, increasing δ affects both the fitting and penalty terms. Both terms grows as δ increases. However, note that while the first term is bounded by $S_{D_{\boldsymbol{\theta}}} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$,[7] and the second term can grow indefinitely in δ. Hence, we have the following upper bound for $\log Z$, that is

$$\log Z \leq \underbrace{\log S_{D_{\boldsymbol{\theta}}}}_{\texttt{fitting}} \underbrace{- D_{\boldsymbol{\theta}} \log \delta}_{\texttt{penalty}}, \tag{40}$$

where we can interpret the first term in the above equation as a fitting term, and the second term as a penalty term over the model complexity/order (MacKay, 2003, Ch. 28).

Remark 10. *This penalty term can also be interpreted as an implicit log-prior term over the corresponding model.*

Moreover, for $\delta \to \infty$, we have $\log Z \to -\infty$ (keeping fixed $D_{\boldsymbol{\theta}}$). Similar considerations and the connection with information criteria are also given in the following Appendix B.

B Marginal likelihood Z and information criteria

The marginal likelihood can be expressed as

$$Z = \ell_{\max} W, \tag{41}$$

where $W \in [0, 1]$ is the *Occam factor* (Knuth et al., 2015, Sect. 3). More specifically, the Occam factor is defined as

$$W = \frac{1}{\ell_{\max}} \int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{42}$$

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.

[6] B depends on both δ and $D_{\boldsymbol{\theta}}$, whereas the $\ell(\mathbf{y}|\boldsymbol{\theta})$ depends on $D_{\boldsymbol{\theta}}$.

[7] B and $\boldsymbol{\Theta}$ depend both on the parameter dimension $D_{\boldsymbol{\theta}}$. Hence, also S depends on $D_{\boldsymbol{\theta}}$. For this reason, here we use the more proper notation $S_{D_{\boldsymbol{\theta}}} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$.

Considering the expression (41) and taking the logarithm, we obtain

$$\log Z = \log \ell_{\max} + \log W \tag{43}$$

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 (Schwarz et al., 1978; Hannan & Quinn, 1979; D. Spiegelhalter et al., 2002). Most of the criteria, suggested in the literature, can be expressed as

$$C = \underbrace{-2 \log \ell_{\max}}_{\text{fitting}} \underbrace{+2\eta D_\theta}_{\text{penalization}} , \tag{44}$$

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 11. *Note that the expression of C is similar to*

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

considering Eq. (43)*, where $-2 \log W$ plays the role of the second factor $2\eta D_\theta$ in Eq.* (44)*.*

The expression (44) encompasses several well-known information criteria proposed in the literature and shown in Table 6, which differ for the choice of η.

Remark 12. *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) (Schwarz et al., 1978)	$\frac{1}{2} \log D_y$
Akaike information criterion (AIC) (D. Spiegelhalter et al., 2002)	1
Hannan-Quinn information criterion (HQIC) (Hannan & Quinn, 1979)	$\log(\log(D_y))$