A note on the area under the likelihood and the fake evidence for model selection

L. Martino* and F. Llorente$^\top$

* Universita' di Catania, Catania, Italia.

$^\top$ Stony Brook University, NewYork, USA.

Abstract

Improper priors are not allowed for the computation of the Bayesian evidence $Z = p(\mathbf{y})$ (a.k.a., marginal likelihood), since in this case Z is not completely specified due to an arbitrary constant involved in the computation. However, in this work, we remark that they can be employed in a specific type of model selection problem: when we have several (possibly infinite) models belonging to the same parametric family (i.e., for tuning parameters of a parametric model). However, the quantities involved in this type of selection cannot be considered as Bayesian evidences: we suggest to use the name "fake evidences" (or "areas under the likelihood" in the case of uniform improper priors). We also show that, in this model selection scenario, using a diffuse prior and increasing its scale parameter asymptotically to infinity, we cannot recover the value of the area under the likelihood, obtained with a uniform improper prior. We first discuss it from a general point of view. Then we provide, as an applicative example, all the details for Bayesian regression models with nonlinear bases, considering two cases: the use of a uniform improper prior and the use of a Gaussian prior, respectively. A numerical experiment is also provided confirming and checking all the previous statements.

Keywords: Bayesian evidence; marginal likelihood; improper prior; diffuse prior.

1 Introduction

Nowadays, Bayesian inference is a hot topic of research and, as a consequence, Bayesian methods are considered more and more as benchmark techniques for inferring the parameters of a model (and their uncertainties), and/or for model selection purposes. Although Bayesian inference has historically been always used (e.g. [12, 22]), Bayesian analyses are now becoming more widespread: we can find Bayesian studies in very different applied fields such as remote sensing [17, 15], astronomy [1, 8], cosmology [2, 3], or optical spectroscopy [7, 25], to name a few.

In Bayesian inference, we can distinguish (at least) two levels: the inference over the parameters (Level-1) and the model selection problem (Level 2). In order to perform Bayesian model selection (Level-2), we need to compute the so-called *Bayesian evidence*, a.k.a., *marginal likelihood* of the model, denoted in this work as Z. The choice of the prior densities over the parameters (in Level-1) affects the value of the marginal likelihood.

Vague/diffuse priors and, more extremely, improper priors are generally employed (when possible) in level-1 of inference for expressing a weak a-priori information (for this reason, they are also called non-informative priors) [9, 20]. However, in model selection (level-2), the use of vague/diffuse priors over the parameters (in level-1) can radically change the value of the evidence Z. Therefore, in this sense, vague/diffuse priors are always informative in level-2. Moreover, the use of improper priors is forbidden for computing the evidence Z since, in this case, the marginal likelihood Z is not completely specified due to an arbitrary constant involved in the computation.

In this work, we firstly try to clarify and remark the issues described above in order to avoid any sort of confusion in the literature [24, 13]. Moreover, we show that although improper priors are not allowed for the computation of the evidence Z, they can be employed in a specific type of model selection problem: when we have several (possibly infinite) models belonging to the same parametric family (i.e., for tuning parameters of a parametric model). However, in this case, we are *not* actually computing an evidence Z (that is not completely specified) [24]. For this reason, we suggest to call the calculated quantity as “fake evidence” or, in the case of uniform improper priors, as “the area under the likelihood”. Furthermore, in this scenario, if we apply a vague/diffuse prior and leave the scale parameter to increase tending to infinity, it is not possible to recover the results obtained by employing improper uniform prior (as typically happens in level-1).
We show and discuss these points firstly with generic arguments, and then more specifically within a Bayesian regression model. We provide theoretical details comparing the scenario with a uniform improper prior with the case of a Gaussian prior, checking and confirming the general statements previously discussed. Moreover, a specific example of generalized linear model is considered for providing numerical checks and related simulations.

The rest of the work is structured as follows. The main background and notation, as well as the different levels of inference and types of model selection are given in Section 2. A detailed discussion about the safe use of vague and/or improper priors is given in Section 3. The key observations of the work are described in Section 4. Section 5 provides a detailed description of a regression problem with Bayesian generalized linear models considering a uniform improper prior and a Gaussian prior. Section 6 provides related numerical results. Finally, several conclusions are given in Section 7.

2 Elements in Bayesian inference

In many applications, the goal is to make inference about a variable of interest, $\boldsymbol{\theta} = \theta_{1:D_{\boldsymbol{\theta}}} = [\theta_1, \theta_2, \dots, \theta_{D_{\boldsymbol{\theta}}}] \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{D_{\boldsymbol{\theta}}}$, where $\theta_d \in \mathbb{R}$ for all $d = 1, \dots, D_{\boldsymbol{\theta}}$, given a set of observed measurements $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$. The observed vector $\mathbf{y}$ is linked with the vector of parameters of interest $\boldsymbol{\theta}$ by an observation model denoted as $\mathcal{M}$, which induces a likelihood function denoted as $\ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$ (that is a density with respect to $\mathbf{y}$ and a non-negative function fixing $\mathbf{y}$ and varying $\boldsymbol{\theta}$).

In the Bayesian framework, a complete model $\mathcal{M}$ is formed by a likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$ and a prior probability density function (pdf) $g(\boldsymbol{\theta}|\mathcal{M})$ chosen by the practitioner. Then, all the statistical information is summarized by the posterior pdf, i.e.,

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

where

$$Z = p(\mathbf{y}|\mathcal{M}) = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})g(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}, \quad (1)$$

is called *Bayesian evidence* or *marginal likelihood* [22, 12, 14]. This quantity is important for model selection purposes, as we show below. Usually $Z = p(\mathbf{y}|\mathcal{M})$ is unknown and difficult to approximate, so that in many situations we are only able to evaluate the unnormalized target function, $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M})g(\boldsymbol{\theta}|\mathcal{M}) \propto \bar{\pi}(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M})$, so that $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}) = \frac{1}{Z}\pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M})$ and $Z = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M})d\boldsymbol{\theta}$.

2.1 Levels in Bayesian inference

Generally speaking, in Bayesian inference we can distinguish between two types of problems or levels of inference [16, Ch. 28], described below:

- **Level-1: Estimation and prediction problems.** In the first level, given the m-th model $\mathcal{M}_m$, we are interested in making inferences regarding parameter $\boldsymbol{\theta}_m$ by focusing on its posterior pdf $\bar{\pi}(\boldsymbol{\theta}_m|\mathbf{y}, \mathcal{M}_m) \propto \ell(\mathbf{y}|\boldsymbol{\theta}_m, \mathcal{M}_m)g(\boldsymbol{\theta}_m|\mathcal{M}_m)$. This is also denoted as "Level-1 of inference" in the literature. Now we drop for simplicity the dependence on the m-th model $\mathcal{M}_m$ and m, then $\bar{\pi}(\boldsymbol{\theta}_m|\mathbf{y}, \mathcal{M}_m) = \bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$.
- **Level-2: Model selection problems.** In the second type of problem, we focus on the model posterior distribution

$$p(\mathcal{M}_m|\mathbf{y}) \propto p(\mathcal{M}_m)Z_m = p(\mathcal{M}_m) \int_{\boldsymbol{\Theta}_m} \ell(\mathbf{y}|\boldsymbol{\theta}_m, \mathcal{M}_m)g(\boldsymbol{\theta}_m|\mathcal{M}_m)$$

  for all $m = 1, \ldots, M$. 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 these are the two *main* levels of inference since they are associated with the two main inference scenarios: parameter estimation and model selection. We will see that the prior choice has a different impact in each of the different levels. In this work, we focus mainly on level-2.

2.2 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.

- **Type-1 — 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 spaces. In this scenario, the parameters $\boldsymbol{\theta}_m$ of each model can have a completely different physical or statistical interpretation.
- **Type-2 — Models in the same parametric families:** tuning the parameters of a parametric model can be considered a model selection problems where different models of the same parametric families are compared. Indeed, in this case, we can apply the so-called *empirical Bayesian* approach. Let consider now that the observation model depends on some vectors of parameters $\boldsymbol{\eta}$, i.e. $\ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta})$. The marginal likelihood would depend on $\boldsymbol{\eta}$,

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

  The empirical Bayesian approach consists on tuning $\boldsymbol{\eta}$ by maximizing $p(\mathbf{y}|\boldsymbol{\eta})$ keeping fixed $\mathbf{y}$, i.e.,

  $$\boldsymbol{\eta}^* = \arg\max p(\mathbf{y}|\boldsymbol{\eta}). \tag{3}$$

  In this approach, we could also include unknown parameters of the prior density over $\boldsymbol{\eta}$, i.e., $p(\mathbf{y}|\boldsymbol{\eta}_\ell, \boldsymbol{\eta}_p) = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_\ell) g(\boldsymbol{\theta}|\boldsymbol{\eta}_p) d\boldsymbol{\theta}$.
- **Type-3 — Nested models:** Nested models are models that belong to the same parametric family but, unlike in the previous scenario, the *complexity* of the model can change, i.e., the number of parameters $|\boldsymbol{\Theta}_m| = D_{\boldsymbol{\theta}_m}$ 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. Some examples of this framework are: variable selection, order selection (in polynomial regression or ARMA models etc.), clustering (when the number of clusters are unknown) and dimension reduction problems, to name a few [5].

3 Use of vague priors and/or improper priors in Level-2

For simplicity, hereafter, whenever we focus on a single although arbitrary model $\mathcal{M}_m$, 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

$$Z = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{4}$$

We can see clearly that Z is an average of likelihood values $\ell(\mathbf{y}|\boldsymbol{\theta})$, weighted according to the prior pdf $g(\boldsymbol{\theta})$.

3.1 Diffuse/vague priors are informative for model selection

If the support $\boldsymbol{\Theta}$ is unbounded and additional information is not available, one can employ a so-called *vague* prior density, i.e., a density with probability mass spread in all the state space, with a great scale parameter. This kind of prior has different names such as *diffuse*, *flat*, etc. Let us consider now an illustrative example about the impact on the inference using a vague prior.

Illustrative example

Let us assume a likelihood function that is integrable in every subset of an unbounded $\boldsymbol{\Theta}$, 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,

$$S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty. \tag{5}$$

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

$$\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta}) \mathbf{1}_B(\boldsymbol{\theta})}{\int_B \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}, \tag{6}$$

which is the normalized likelihood restricted to the set B. We discuss what happens in Level-1 and Level-2 as $|B| \to \infty$.

- *Level-1:* as we increase the volume of B, more and more mass of the likelihood is considered. As $|B| \to \infty$, we have that $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ becomes closer and closer to

$$\bar{\pi}^*(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})}{S}. \tag{7}$$

  Namely, in the limit where $B = \boldsymbol{\Theta}$, the prior $g(\boldsymbol{\theta})$ becomes equivalent to an improper uniform prior on $\boldsymbol{\theta}$. The posterior $\bar{\pi}^*(\boldsymbol{\theta}|\mathbf{y})$ contains only the information provided by the likelihood function, and is not affected or distorted by the prior. Hence, a vague/diffuse prior (or a uniform improper prior, which is its maximal expression) can be employed for expressing the absence of additional information in the choice of the prior (at Level-1 of inference).

- *Level-2:* we focus now on the marginal likelihood Z which, in this case, is given by

$$Z = \frac{\int_B \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}{|B|}. \tag{8}$$

Let us consider increasing B until we cover all parameter space, i.e.,

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

Hence,

$$\lim_{|B|\to\infty} Z = 0. \tag{9}$$

We see that the marginal likelihood of a model with a increasingly-diffuse prior becomes zero. Hence, in model selection (Level-2), actually vague/diffuse priors are highly informative, in the sense that, (if S is finite) an increasingly diffuse prior penalizes more and more the considered model, so that their use has a clear impact to the results of the model selection.

Hence, we can highlight two conclusions.

Remark 1. *In the Level-1 of inference, if $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ is finite, we can use vague prior as non-informative (or weakly-informative) choice, since the idea is to perform the minimum possible perturbation to the likelihood function and, as a consequence, a minimum impact to the inference of $\boldsymbol{\theta}$ (and, generally, we can asymptotically recover some frequentist results).*

Remark 2. *In Level-2 inference, the choice of a diffuse/vague prior is actually very informative. For instance, if $S = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ is finite, diffuse priors tend to produce smaller values of the marginal likelihood Z [6, 4].*

Hence, if $S < \infty$, 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 [4, 16, 21, 14].

3.2 Improper priors: forbidden for computing the evidence Z

Let us consider again that the domain $\boldsymbol{\Theta}$ is unbounded. An improper prior is such that

$$\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty. \tag{10}$$

Note that in this case, the prior $g(\boldsymbol{\theta}) = c \cdot h(\boldsymbol{\theta})$ (where $\int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$) is not completely specified, since h cannot be normalized, i.e., the normalization constant c does not exist, and as a consequence, the constant c is arbitrary. Let us assume that, however,

$$\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} = Z_{\ell\times h} < \infty \tag{11}$$

is finite. We call $Z_{\ell\times h}$ as *fake evidence*. In this case, trying to computing the Bayesian evidence Z, we obtian

$$\begin{aligned} Z &= \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &= c\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &= c\ Z_{\ell\times h}, \end{aligned} \tag{12}$$

i. e., the marginal likelihood Z is also not completely specified due to $c > 0$ is unknown/arbitrary. Then, we can remark below:

Remark 3. *Improper priors can not be used for computing marginal likelihood Z. Thus, generally, improper priors are not allowed for model selection (***Level-2** *of inference). However, we will see that there is an exception for Type-2 of model selection.*

On the other hand, the use of an improper prior, i.e., is allowed for **Level-1** inference when $Z_{\ell\times h}$ in Eq. (11) is finite. Indeed, in this case, the corresponding posterior is still proper,

$$\begin{aligned}\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}}, \\ &= \frac{\cancel{c}\ \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{\cancel{c}\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}}, \\ &= \frac{\ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{Z_{\ell\times h}},\end{aligned}$$

since $Z_{\ell\times h} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$ is finite.

Remark 4. *Improper priors are allowed in* **Level-1** *of inference if the fake evidence $Z_{\ell\times h}$ is finite, i.e., $Z_{\ell\times h} < \infty$ (since the corresponding posteriors are still proper).*

3.3 Uniform improper prior and the area under the likelihood

An extreme case of vague prior and the simplest example if improper prior is *the uniform improper prior*, i.e., $g(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta}$ in the unbounded support $\boldsymbol{\Theta}$. It is often employed for expressing the absence of a-priori information in the **Level-1** of inference when the *area under the likelihood* (S) is finite, i.e.,

$$S(\mathbf{y}) = Z_{\ell\times 1} = \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} < \infty. \tag{13}$$

Indeed, in this case the unnormalized posterior $\pi(\boldsymbol{\theta}|\mathbf{y}) = \ell(\mathbf{y}|\boldsymbol{\theta})$ can be normalized as

$$\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{S(\mathbf{y})}\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{S(\mathbf{y})}\ell(\mathbf{y}|\boldsymbol{\theta}). \tag{14}$$

Hence, we need $S(\mathbf{y}) < \infty$, in order to be able to use improper uniform prior in Level-1 of inference. Note also that $S(\mathbf{y}) > 0$ since $\ell(\mathbf{y}|\boldsymbol{\theta})$ must be positive in a a region with non-zero measure.

Remark 5. *The quantity in Eq.* (13)*, i.e., $S(\mathbf{y})$, cannot be interpreted as an evidence $Z = p(\mathbf{y})$, since the values of the likelihood $\ell(\mathbf{y}|\boldsymbol{\theta})$ are not weighted (by a prior density). It is a special case of fake evidence $Z_{\ell\times h}$ when $h(\boldsymbol{\theta}) = 1$.*

However, we show and remark that the area under the likelihood S or, the fake evidence $Z_{\ell\times h}$, can be still useful in Type-2 of model selection 2.2.

4 Key observations

Tuning the parameters (or hyper-parameters) in a family of models is a special scenario of model selection, i.e., Type-2 described in Section 2.2. We will see that, in this specific case, we can use improper priors since the fake evidences $Z_{\ell \times h}$ are meaningful in some sense.

4.1 Comparing models which differ for the chosen parameters

For simplicity, let us consider two models which differ only for the tuning of the parameters $\boldsymbol{\eta}_\ell$ of the likelihood function (induced by the observation model), and for $\boldsymbol{\eta}_p$ some parameters of the prior. More specifically, let say that we have $\ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_{\ell,1})$ and $\ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_{\ell,2})$, considering the use of *the same type* of improper prior for both models but with different hyper-parameters $\boldsymbol{\eta}_{p,1}$ and $\boldsymbol{\eta}_{p,2}$, i.e., $g(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,i}) = c \cdot h(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,i})$ for $i = 1, 2$, the Bayes factor is

$$\begin{aligned}
\mathrm{BF}_{12} = \frac{Z_1}{Z_2} &= \frac{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_{\ell,1}) g(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,1}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_{\ell,2}) g(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,2}) d\boldsymbol{\theta}}, \\
&= \frac{\cancel{c} \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_{\ell,1}) h(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,1}) d\boldsymbol{\theta}}{\cancel{c} \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_{\ell,2}) h(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,2}) d\boldsymbol{\theta}}, \\
&= \frac{Z_{\ell_1 \times h_1}}{Z_{\ell_2 \times h_2}},
\end{aligned}$$

i.e., the ratio of marginal likelihoods $\frac{Z_1}{Z_2}$ is well-defined in this scenario, equal to the ratio fo fake evidences $\frac{Z_{\ell_1 \times h}}{Z_{\ell_2 \times h}}$. Note that the constant c of the prior must disappear and cannot be included in the vectors $\boldsymbol{\eta}_{p,i}$.

More generally, considering M possible $\boldsymbol{\eta}_{\ell,m}$, $m = 1, ..., M$, vector of parameters (i.e., M possible models) and using the same *the same type* improper prior for all the models, $g(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,i}) = c \cdot h(\boldsymbol{\theta}|\boldsymbol{\eta}_{p,i})$, we could apply a Bayesian model averaging [11] with the following normalized weights:

$$\bar{w}_m = \frac{Z_m}{\sum_{i=1}^M Z_i} = \frac{\cancel{c} Z_{\ell_m \times h_m}}{\cancel{c} \sum_{i=1}^M Z_{\ell_i \times h_i}}, \quad m = 1, ..., M,$$

that are again well-defined, since the arbitrary value c is cancelled out. In the simplest case of a uniform improper prior, we have $\bar{w}_m = \frac{S_m}{\sum_{i=1}^M S_i}$, $m = 1, ..., M$. In other words, the improper prior here can be employed since it is shared by all the models.

Remark 6. *In the scenario of tuning some parameters $\boldsymbol{\eta}$ of the observation model, the use of unique improper prior over $\boldsymbol{\theta}$ (the same prior for all models) is allowed, and the fake evidence $Z_{\ell_m \times h}$ can be employed for comparing models.*

Hence, the fake evidence $Z_{\ell_m \times h_m}$ can be employed in Type-2 of model selection.

Remark 7. *In this sense, the statement "improper priors are not allowed for model selection" is technically wrong [13]. A more correct statement is "improper priors are not allowed for computation of the Bayesian evidence Z", but they can be used in Type-2 of model selection* *computing the fake evidence* $Z_{\ell_m \times h_m}$.

In Type-3 of model selection, i.e., the nested model scenario, an improper prior is allowed only for the first variable/parameter, common and shared to all the nested models.

4.2 On the area under the likelihood S

Let us consider now the case $h(\boldsymbol{\theta}) = 1$, i.e., $Z_{\ell_m \times 1} = S$ with $S = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty$. With respect to the use of $h(\boldsymbol{\theta}) = 1$, with respect to the fake evidence, $Z_{\ell \times h}$, with the area under the likelihood, S:

Remark 8. *Unlike the fake evidence* $Z_{\ell \times h}$ *for a generic non-uniform* $h(\boldsymbol{\theta})$*, the quantity* $S = Z_{\ell \times 1}$ *in Eq.* (13) *can be used only for tuning parameters of the likelihood function* $\boldsymbol{\eta}_\ell$*, since the only possible parameter to add in* $h(\boldsymbol{\theta}) = 1$ *is a multiplicative constant that cannot be included in* $\boldsymbol{\eta}_p$.

Moreover, S cannot be recovered as an asymptotic special case using a proper prior density (as one could expect from other results in Level-1 of inference):

Remark 9. *The value of the area under the likelihood* S *cannot be obtained starting with a diffuse prior and then increase its scale parameter to infinity (as usually done in Level-1 of inference). Namely, let us consider* $S < \infty$. *Note that* $S > 0$ *since* $\ell(\mathbf{y}|\boldsymbol{\theta}) > 0$ *in a a region with non-zero measure. As the scale of the diffuse prior grows,* $Z \to 0$*, hence* $Z \nrightarrow S$.

4.3 On the empirical Bayes and profile likelihood approaches

For simplicity, in this section, we consider $h(\boldsymbol{\theta}) = 1$ but all comments in this section are valid for the more general case. We can extend the previous considerations in order to compare a *continuous* of models, i.e., infinite models belonging to the same parametric family. In this sense, find the best model is equivalent to obtain a point-wise estimator of $\boldsymbol{\eta}$ as

$$\boldsymbol{\eta}^* = \arg\max S(\mathbf{y}|\boldsymbol{\eta}) = \arg\max \int_{\boldsymbol{\Theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}) d\boldsymbol{\theta}. \tag{15}$$

With respect to $\boldsymbol{\eta}$, if we are just interested in a pointwise estimator $\boldsymbol{\eta}^*$, we are basically employing a frequentist approach over $\boldsymbol{\eta}$ but after integrating out $\boldsymbol{\theta}$ from the likelihood. Hence, it can be interpreted as a combination of Bayesian-frequentist approaches. Indeed, it differs from the classical maximization of the complete likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta})$ (see Sections 5.3.3 and 6.1), that is a completely frequentist strategy. This is called *empirical Bayes* approach, and can be also employed to select unknown parameters of the prior.

A similar and completely frequentist approach, is the *profile likelihood* of a parameter of interest, which is defined as

$$L_p(\boldsymbol{\eta}) = \ell(\mathbf{y}|\boldsymbol{\theta}^*_{\boldsymbol{\eta}}, \boldsymbol{\eta}) = \max_{\boldsymbol{\theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}), \qquad \boldsymbol{\theta}^*_{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\theta}} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}).$$

The idea is to study this function to derive confidence regions and other statistical information (similarly with a posterior density). A normalized version can be obtained by dividing this expression by the complete maximum likelihood value, i.e., $\widetilde{L}_p(\boldsymbol{\eta}) = \frac{\max_{\boldsymbol{\theta}} \ell(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\eta})}{\max_{\boldsymbol{\theta},\boldsymbol{\eta}} \ell(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\eta})}$. However, in the profile likelihood there is not any integration and there is not any explicit, implicit prior density involved. Some advantages of integrating out $\boldsymbol{\theta}$ are shown in Section Sections 5.3.3 and 6.1.

4.4 Full-Bayesian solution with a double improper (uniform) prior

Again, here we consider $h(\boldsymbol{\theta}) = 1$ but all considerations can be extended to a generic $h(\boldsymbol{\theta})$. As an alternative approach, if $S_Z = \int_{\boldsymbol{\Theta}} S(\mathbf{y}|\boldsymbol{\eta})d\boldsymbol{\eta} < \infty$, we could consider again an improper uniform prior over $\boldsymbol{\eta}$ and the marginal posterior would be $p(\boldsymbol{\eta}|\mathbf{y}) = \frac{1}{S_Z}S(\mathbf{y}|\boldsymbol{\eta})$ (other type of improper prior can be also employed). Thus, the so-called *full-Bayesian solution* could be obtained considering improper priors twice. In this case, the full-Bayesian solution can be performed following the steps below:

1. Draw $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_R$ from $p(\boldsymbol{\eta}|\mathbf{y}) = \frac{1}{S_Z}S(\mathbf{y}|\boldsymbol{\eta})$.

2. Draw $\boldsymbol{\theta}_{r,1}, ..., \boldsymbol{\theta}_{r,N}$ from $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}_r) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\eta}_r)}{S(\mathbf{y}|\boldsymbol{\eta}_r)} \propto \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}_r)$, for $r = 1, ..., R$.

The outputs are the NR samples, $\{\boldsymbol{\theta}_{r,n}\}$, and R samples $\{\boldsymbol{\eta}_r\}$ for $r = 1, ..., R$ and $n = 1, ..., N$. With this procedure, the model parameter $\boldsymbol{\eta}$ is marginalized out. Another way to interpret this procedure is as a continuous Bayesian model averaging, i.e., consider a resulting model expressed as

$$\bar{\ell}(\mathbf{y}|\boldsymbol{\theta}) = \int \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta})p(\boldsymbol{\eta}|\mathbf{y})d\boldsymbol{\eta}, \tag{16}$$

where we are weighting the different models, $\ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta})$, according to the marginal posterior $p(\boldsymbol{\eta}|\mathbf{y})$. In the integral in Eq. (16), this marginal posterior $p(\boldsymbol{\eta}|\mathbf{y})$ is employed as an informative prior over $\boldsymbol{\eta}$, but it is an *objective prior* built after looking the data [13]. Moreover recall that, if S and S_Z are both finite, the use of the improper priors is allowed twice in this procedure. Table 1 provides a summary of the main concepts. In the next section, we check and confirm the previous statements in a Bayesian regression setting.

It is also important to remark that all the theoretical and practical considerations (in this work) are valid either when the marginal likelihood can be computed analytically, or when the marginal likelihood is intractable. In the latter scenario, we have just the additional computational problem of approximating the marginal likelihood [14].

5 Example of application to Bayesian regression models

In this section, we consider a generalized linear model for regression, considering N data points and M different non-linear bases, with $M < N$. We apply two types of priors to the vector of coefficients $\boldsymbol{\theta}$: an improper uniform prior and a Gaussian prior. In both cases, we give a complete Bayesian analysis and try to design a Level-2 of inference in order to infer a vector of parameters of the bases $\boldsymbol{\alpha}$, the noise power σ_e^2 and the rest of nuisance parameters. Firstly, the goal of this section is to show some applicative examples. Secondly, the goal is confirm some important statements provided above, from a more practical point of view. Finally, this section gives the theoretical support for the numerical example in Section 6.2.

Table 1: Summary of the main concepts.

Prior densities	**Level-1 of inference**	**Level-2 of inference**
Diffuse/vague priors	weakly informative.	informative; If $S < \infty$, then $Z \to 0$ as the prior becomes more diffuse.
Improper priors	non-informative; If $S < \infty$, they can be used. to make inference on $\boldsymbol{\theta}$.	They are not allowed for computing Z; If $S < \infty$, they can be used for Type-2 of model selection.
From diffuse → to improper uniform	We asymptotically obtain the same results using an improper uniform prior; Generally, we recover some frequentist results.	If $S < \infty$, then $Z \to 0$, hence $Z \nrightarrow S$ (since $S > 0$).

5.1 Problem statement

Let us consider the dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N$, where $\mathbf{x}_n = [x_{n,1}, \dots, x_{n,d_X}] \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ represents the inputs, $y_n \in \mathbb{R}$, denotes the outputs. The vector of outputs is then $\mathbf{y} = [y_1, \dots, y_N]^\top \in \mathbb{R}^N$. We consider the following observation probabilistic model which link the vectors $\mathbf{x}$ and $\mathbf{y}$,

$$y \;=\; f(\mathbf{x}) + e, \quad e \sim \mathcal{N}(e|0, \sigma_e^2).$$

We assume that the underlying function can have the following parametric form,

$$f(\mathbf{x}) = \sum_{m=1}^{M} \phi_m(\mathbf{x}, \boldsymbol{\alpha})\theta_m = \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\alpha})^\top \boldsymbol{\theta}, \qquad M \leq N, \tag{17}$$

where $\phi_m(\mathbf{x}, \boldsymbol{\alpha}) : \mathcal{X} \times \Omega \to \mathbb{R}$ is the m-th nonlinear function where $\boldsymbol{\alpha} \in \Omega \subseteq \mathbb{R}^{d_\theta}$ represents a vector of parameters, that the user have to tune [5]. Defining the vectors

$$\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\alpha}) \;=\; [\phi_1(\mathbf{x}, \boldsymbol{\alpha}), \phi_2(\mathbf{x}, \boldsymbol{\alpha}), \dots, \phi_M(\mathbf{x}, \boldsymbol{\alpha})]^\top, \tag{18}$$

$$\boldsymbol{\theta} \;=\; [\theta_1, \theta_2, \dots, \theta_M]^\top, \tag{19}$$

then we can rewrite the model above as

$$y = \sum_{m=1}^{M} \phi_m(\mathbf{x}, \boldsymbol{\alpha})\theta_m + e = \underbrace{\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\alpha})^\top}_{1 \times M} \underbrace{\boldsymbol{\theta}}_{M \times 1} + e. \tag{20}$$

Hereafter, for simplicity, we will remove the dependence of $\boldsymbol{\alpha}$, so that $\phi_m(\mathbf{x}) = \phi_m(\mathbf{x}, \boldsymbol{\alpha})$ and $\boldsymbol{\phi}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\alpha})$. For instance, we will write simply $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\theta}$.

Vectorial form. The model above can be written in a vectorial form as

$$\begin{aligned}\underbrace{\mathbf{y}}_{N\times 1} &= \underbrace{\mathbf{f}}_{N\times 1}+\underbrace{\mathbf{e}}_{N\times 1} &(21)\\ &= \underbrace{\mathbf{\Phi}}_{N\times M}\underbrace{\boldsymbol{\theta}}_{M\times 1}+\underbrace{\mathbf{e}}_{N\times 1}, &(22)\end{aligned}$$

where we have denoted $\mathbf{f} = \mathbf{\Phi}\boldsymbol{\theta}$, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_N)$ and we have defined $N \times M$ *design matrix* $\mathbf{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \dots, \boldsymbol{\phi}(\mathbf{x}_N)]^\top$ (that is rectangular, in general), i.e.,

$$\underbrace{\mathbf{\Phi}}_{N\times M} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_M(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \dots & \phi_M(\mathbf{x}_2) \\ \vdots & & & \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \dots & \phi_M(\mathbf{x}_N) \end{bmatrix}. \tag{23}$$

Note that, in this section, all the considerations are valid for $M \leq N$. The case $M = N$ is also included: the resulting regression method becomes non-parametric. The scenario $M > N$, since requires specific observations and analysis, is not considered here.

5.2 Likelihood function

The observation model above induce a likelihood function with respect to (w.r.t.) the coefficients $\boldsymbol{\theta}$, that is

$$\begin{aligned}\ell(\mathbf{y}|\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta}, \mathbf{\Phi}, \boldsymbol{\alpha}, \sigma_e^2) &= \left(2\pi\sigma_e^2\right)^{-\frac{N}{2}} \exp\left(-\frac{(\mathbf{y}-\mathbf{\Phi}\boldsymbol{\theta})^\top(\mathbf{y}-\mathbf{\Phi}\boldsymbol{\theta})}{2\sigma_e^2}\right)\\ &= \left(2\pi\sigma_e^2\right)^{-\frac{N}{2}} \exp\left(-\frac{||\mathbf{y}-\mathbf{\Phi}\boldsymbol{\theta}||^2}{2\sigma_e^2}\right)\\ &= \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\boldsymbol{\theta}, \sigma_e^2\mathbf{I}_N), \end{aligned}\tag{24}$$

where $\mathbf{I}_N$ is the $N\times N$ identity matrix. Clearly, a more complete notation would be $\ell(\mathbf{y}|\boldsymbol{\theta}, \mathbf{\Phi}, \boldsymbol{\alpha}, \sigma_e)$. However, we first focus on the coefficients $\boldsymbol{\theta}$ and consider, in this first stage, $\mathbf{\Phi}$ and the nonlinear bases are chosen in advance. The parameters $\boldsymbol{\alpha}$ and σ_e should be tuned and decided by the user. Then, the complete vector of hyper-parameters, denoted as $\boldsymbol{\lambda}$, is formed by $\boldsymbol{\alpha}$ and σ_e, i.e., we have $\boldsymbol{\lambda} = [\boldsymbol{\alpha}, \sigma_e]$ [5, 23]. Fixing the bases and $\boldsymbol{\alpha}$ the classical maximum likelihood approach, i.e.,

$$\left[\widehat{\boldsymbol{\theta}}_{\texttt{ML}}, \widehat{\sigma}^2_{e,\texttt{ML}}\right] = \arg\max_{\boldsymbol{\theta},\sigma_e^2} \ell(\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2), \tag{25}$$

provides the following estimators (that can be obtained analytically in this case):

$$\widehat{\boldsymbol{\theta}}_{\texttt{ML}} = (\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{y}, \qquad \widehat{\sigma}^2_{e,\texttt{ML}} = \frac{1}{N}||\mathbf{y}-\widehat{\mathbf{f}}||^2. \tag{26}$$

Note that $\widehat{\sigma}^2_{e,\texttt{ML}}$ is a consistent, but biased estimator.

5.3 Uniform improper prior over $\boldsymbol{\theta}$

In this section, we assume a uniform *improper* prior density over the weights $\boldsymbol{\theta}$, i.e., $g(\boldsymbol{\theta}) \propto 1$ for all $\boldsymbol{\theta}$.

5.3.1 Posterior of the coefficients $\boldsymbol{\theta}$

Therefore, the posterior pdf of the coefficient $\boldsymbol{\theta}$ is

$$\begin{aligned}\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{p(\mathbf{y})} &\propto \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}),\\ &\propto \ell(\mathbf{y}|\boldsymbol{\theta}),\\ &\propto \left(2\pi\sigma_e^2\right)^{-\frac{N}{2}} \exp\left(-\frac{||\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta}||^2}{2\sigma_e^2}\right), \end{aligned} \tag{27}$$

i.e., proportional to the likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta})$ (we have use $g(\boldsymbol{\theta}) \propto 1$). After some algebra and rearrangements (in order to express the formula as a Gaussian density with respect to $\boldsymbol{\theta}$ instead of $\mathbf{y}$), we can express $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ as a Gaussian distribution with mean vector $\boldsymbol{\mu}_{\theta|y}$ and covariance matrix $\boldsymbol{\Sigma}_{\theta|y}$ [24], i.e.,

$$\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y}), \tag{28}$$

with

$$\boldsymbol{\mu}_{\theta|y} = \widehat{\boldsymbol{\theta}} = (\underbrace{\boldsymbol{\Phi}^\top\boldsymbol{\Phi}}_{M\times M})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}, \tag{29}$$

and

$$\boldsymbol{\Sigma}_{\theta|y} = \left(\frac{1}{\sigma_e^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1} = \sigma_e^2\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}. \tag{30}$$

Remark 10. *Note that, using a uniform improper prior, the posterior density $\bar{\pi}(\boldsymbol{\theta}|\mathbf{y})$ over the coefficient vector resembles the "frequentist" sampling distribution described in Appendix A, being both Gaussian with the same mean and covariance matrix, although they have a complete statistical different meaning (see App. A). Moreover, $\widehat{\boldsymbol{\theta}}$ coincides with the maximum likelihood estimator, i.e., $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{ML}$ (see App. A and Eq. (26)).*

5.3.2 Posteriors of the function $f(\mathbf{x})$ and vector $\mathbf{f}$

Let us recall that the assumed model is $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top\boldsymbol{\theta}$ and $\boldsymbol{\theta} \sim \bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y})$, after seeing the data. Hence, given a fixed $\mathbf{x}$, the hidden function $f(\mathbf{x})$ is a random variable with a Gaussian posterior density,

$$p(f(\mathbf{x})|\mathbf{y}) = \mathcal{N}(f(\mathbf{x})|\mu_{f|y}(\mathbf{x}), \sigma_{f|y}^2(\mathbf{x})), \tag{31}$$

with mean at $\mathbf{x}$,

$$\begin{aligned}\mu_{f|y}(\mathbf{x}) = \widehat{f}(\mathbf{x}) &= \boldsymbol{\phi}(\mathbf{x})^\top\widehat{\boldsymbol{\theta}}\\ &= \boldsymbol{\phi}(\mathbf{x})^\top(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}, \end{aligned} \tag{32}$$

and variance

$$\sigma^2_{f|y}(\mathbf{x}) = \sigma_e^2 \boldsymbol{\phi}(\mathbf{x})^\top \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \boldsymbol{\phi}(\mathbf{x}), \tag{33}$$

where we have considered the previous results regarding the posterior over the coefficients $\boldsymbol{\theta}$. We remark that the regression function is the mean solution, i.e., $\widehat{f}(\mathbf{x}) = \mu_{f|y}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{y}$ [24].

Posterior of the vector f. In the smoothing, considering only estimations at the input features, i.e., $\mathbf{f} = \mathbf{\Phi}\boldsymbol{\theta}$. Since $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \mathbf{\Sigma}_{\theta|y})$, after seeing the data, the posterior of the vector of $\mathbf{f}$ is

$$p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{f|y}, \mathbf{\Sigma}_{f|y}), \tag{34}$$

with the mean vector

$$\underbrace{\boldsymbol{\mu}_{f|y}}_{N\times 1} = \widehat{\mathbf{f}} = \mathbf{\Phi}\widehat{\boldsymbol{\theta}} = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{y}, \tag{35}$$

and with the covariance matrix

$$\underbrace{\mathbf{\Sigma}_{f|y}}_{N\times N} = \sigma_e^2 \mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top. \tag{36}$$

5.3.3 Area under the likelihood (S)

As we have remarked in the previous sections, sice we are using improper priors, we can just compute the *area under the likelihood* (S), instead of a well-defined marginal likelihood. The S can be useful in certain scenarios, for instance, performing an empirical Bayes approach. Indeed, the marginal likelihood $Z = p(\mathbf{y})$ is defined as the integral $p(\mathbf{y}) = \int_{\mathbb{R}^N} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$. However, since we are employing an improper prior $g(\boldsymbol{\theta}) \propto 1$, the marginal likelihood is not perfectly determined (a multiplicative factor is undetermined). As a consequence, we can just compute S, i.e., $S(\mathbf{y}) = \int_{\mathbb{R}^N} \ell(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$, where we miss the probability interpretation. Again, a more complete notation would be $S(\mathbf{y}|\mathbf{\Phi}, \boldsymbol{\alpha}, \sigma_e) = S(\mathbf{y}|\mathbf{\Phi}, \boldsymbol{\lambda})$. Here, we focus on the choice of the hyper-parameters $\boldsymbol{\lambda} = [\boldsymbol{\alpha}, \sigma_e]$ that we should be tuned. Then, we write

$$S(\mathbf{y}) = S(\mathbf{y}|\boldsymbol{\lambda}) = \int_{\mathbb{R}^N} \ell(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\lambda})d\boldsymbol{\theta}.$$

It is possible to show that [24]

$$S(\mathbf{y}|\boldsymbol{\lambda}) = S(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e) = \frac{(2\pi\sigma_e^2)^{-\left(\frac{N-M}{2}\right)}}{\sqrt{\det\left[\mathbf{\Phi}^\top\mathbf{\Phi}\right]}} \exp\left[-\left(\frac{\mathbf{y}^\top\mathbf{y} - \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}}{2\sigma_e^2}\right)\right], \tag{37}$$

$$= \frac{(2\pi\sigma_e^2)^{-\left(\frac{N-M}{2}\right)}}{\sqrt{\det\left[\mathbf{\Phi}^\top\mathbf{\Phi}\right]}} \exp\left[-\left(\frac{\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}}{2\sigma_e^2}\right)\right], \tag{38}$$

where we have used the equality $\widehat{\mathbf{f}} = \mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}$. Note that $\det\left[\mathbf{\Phi}^\top\mathbf{\Phi}\right] > 0$ since the matrix $\mathbf{\Phi}^\top\mathbf{\Phi}$ is symmetric, positive definite. Moreover, $S(\mathbf{y}|\boldsymbol{\lambda})$ above just depends only on $\boldsymbol{\lambda}$ (we have integrated out $\boldsymbol{\theta}$). Above, we have used the identity,

$$\widehat{\mathbf{f}}^\top\widehat{\mathbf{f}} = \mathbf{y}^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}. \tag{39}$$

Indeed, replacing the expression (35), i.e., $\widehat{\mathbf{f}} = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{y}$, in the first side of the equation above, we have

$$\begin{aligned}
\widehat{\mathbf{f}}^\top\widehat{\mathbf{f}} &= \left(\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}\right)^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}, \\
&= (\mathbf{y}^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top)\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}, \\
&= \mathbf{y}^\top\mathbf{\Phi}\cancel{\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{\Phi}}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}, \\
&= \mathbf{y}^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y},
\end{aligned} \tag{40}$$

Moreover, note also that

$$\widehat{\mathbf{f}}^\top\mathbf{y} = \mathbf{y}^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y} = \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}, \tag{41}$$

Hence, we can write to an important equality:

$$\begin{aligned}
||\mathbf{y} - \widehat{\mathbf{f}}||^2 &= \mathbf{y}^\top\mathbf{y} + \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}} - 2\widehat{\mathbf{f}}^\top\mathbf{y}, \\
&= \mathbf{y}^\top\mathbf{y} + \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}} - 2\widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}, \\
&= \mathbf{y}^\top\mathbf{y} - \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}},
\end{aligned} \tag{42}$$

where we have used $\widehat{\mathbf{f}}^\top\mathbf{y} = \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}$ in Eq. (41). Note that from (42), we have always $\mathbf{y}^\top\mathbf{y} - \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}} \geq 0$. Namely, the power of the outputs $\mathbf{y}^\top\mathbf{y}$ is always greater or equal than the power of the smoothing solution $\widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}$, i.e., $\mathbf{y}^\top\mathbf{y} \geq \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}$. This is clearly due to the *denoising* effect [24].

Then, the S can be rewritten in terms of the smoothing error $||\mathbf{y} - \widehat{\mathbf{f}}||^2$, i.e.,

$$S(\mathbf{y}|\boldsymbol{\lambda}) = \frac{(2\pi\sigma_e^2)^{-\left(\frac{N-M}{2}\right)}}{\sqrt{\det\left[\mathbf{\Phi}^\top\mathbf{\Phi}\right]}}\exp\left[-\left(\frac{||\mathbf{y} - \widehat{\mathbf{f}}||^2}{2\sigma_e^2}\right)\right]. \tag{43}$$

The negative log-S is

$$\boxed{C(\boldsymbol{\lambda}) = -\log S(\mathbf{y}|\boldsymbol{\lambda}) = \tfrac{||\mathbf{y}-\widehat{\mathbf{f}}||^2}{2\sigma_e^2} + \tfrac{N-M}{2}\log(2\pi\sigma_e^2) + \tfrac{1}{2}\log\det\left[\mathbf{\Phi}^\top\mathbf{\Phi}\right].} \tag{44}$$

We can try to minimize the cost function $C(\boldsymbol{\lambda})$ in (43) with respect to $\boldsymbol{\lambda} = [\boldsymbol{\alpha}, \sigma_e]$. Alternatively, we can try to simplify the equation above. One possibility is shown below.

Estimator of the noise variance. If we keep fixed $\boldsymbol{\alpha}$ (and hence also the matrix $\boldsymbol{\Phi}$), it is possible to show that the conditional maximum value w.r.t. σ_e (conditioned to $\boldsymbol{\alpha}$) is [18]

$$\begin{aligned}\widehat{\sigma}_e^2 &= \frac{1}{N-M}||\mathbf{y}-\widehat{\mathbf{f}}||^2, \\ &= \frac{1}{N-M}||\mathbf{y}-\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{y}||^2.\end{aligned} \qquad (45)$$

Remark 11. *This is the unbiased estimator of the noise variance. The classical maximum likelihood estimator in Eq. (26), i.e.,* $\widehat{\sigma}_e^2 = \frac{1}{N}||\mathbf{y}-\widehat{\mathbf{f}}||^2$, *is instead biased. Hence, this is an example of advantage of computing and maximizing the area under the likelihood* $S(\mathbf{y}|\boldsymbol{\lambda})$ *(i.e., the empirical Bayes approach), with respect to the classical maximum likelihood estimator.*

If we replace Eq. (45) into (44), we obtain

$$\log S(\mathbf{y}|\boldsymbol{\alpha}) = -\frac{N-M}{2} - \frac{N-M}{2}\log\left(2\pi\frac{1}{N-M}||\mathbf{y}-\widehat{\mathbf{f}}||^2\right) - \frac{1}{2}\log\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right] + const. \qquad (46)$$

Considering the value $M < N$ fixed and chosen by the user, we can write a cost function as

$$\boxed{C(\boldsymbol{\alpha}) = -\log S(\mathbf{y}|\boldsymbol{\alpha}) = \underbrace{\frac{N-M}{2}\log\left(||\mathbf{y}-\widehat{\mathbf{f}}||^2\right)}_{\text{fitting term}} + \underbrace{\frac{1}{2}\log\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]}_{\text{penalty term}} + const,} \qquad (47)$$

which is just function of $\boldsymbol{\alpha}$. We desire to minimize the cost function $C(\boldsymbol{\alpha}) = -\log S(\mathbf{y}|\boldsymbol{\alpha})$, in term of $\boldsymbol{\alpha}$. We can clearly identify two parts:

- A *fitting term*, $\frac{N-M}{2}\log\left(||\mathbf{y}-\widehat{\mathbf{f}}||^2\right)$, which decreases to $-\infty$ at $\mathbf{y} = \widehat{\mathbf{f}}$ (maximum overfitting). Bigger errors $||\mathbf{y}-\widehat{\mathbf{f}}||^2$ correspond to more positive values of this term. Recall that $\widehat{\mathbf{f}}$ depends on $\boldsymbol{\Phi}$, in fact, $\widehat{\mathbf{f}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}$.

- A *model complexity penalization*, $\frac{1}{2}\log\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]$, that penalizes the overfitting. It fosters smaller values of $\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]$ (this usually happens when $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ tends to be a full matrix), and penalizes greater values of $\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]$ (this usually happens when $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ becomes more similar to a diagonal matrix). Note that the matrix $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ is always symmetric, and positive semi-definite [24].

Minimizing Eq. (47) can be employed to tune the vector of parameters $\boldsymbol{\alpha}$ [24, 13, 14]. Note also that

$$\begin{aligned}\exp\left(-C(\boldsymbol{\alpha})\right) &= \frac{1}{\sqrt{\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]}}\left[||\mathbf{y}-\widehat{\mathbf{f}}||^2\right]^{-\left(\frac{N-M}{2}\right)}, \\ &= \frac{1}{\sqrt{\det\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]}}\left[||\mathbf{y}-\widehat{\mathbf{f}}||\right]^{-(N-M)},\end{aligned} \qquad (48)$$

that resembles the form of a t-student density.

Remark 12. *Both expressions in Eqs.* (44)-(47) *seems to be adequate for tuning the parameters of the model (Type-2 of model selection). We test them numerically in Section 6.2.*

Remark 13. *It is possible an alternative derivation: the expressions* (47)-(48) *could be also obtained assuming an improper Jeffreys prior, $h(\theta) = 1/\sigma_e$, and integrating out σ_e from* (43) *[24, Chapter 2]. This also confirms again that the use of improper priors is allowed in Type-2 of model selection.*

The expressions (47)-(48) can be used for tuning $\boldsymbol{\alpha}$, i.e., we are tuning the bases ϕ_m's, in terms of location and scale parameters, for instance.

5.4 Gaussian prior over $\boldsymbol{\theta}$

In the previous section, we assume a improper uniform prior over $\boldsymbol{\theta}$. Now, let us consider a Gaussian prior density over $\boldsymbol{\theta}$; more specifically, we assume

$$g(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \underbrace{\boldsymbol{\Sigma}_\theta}_{M\times M}),$$

as a prior, where $\boldsymbol{\Sigma}_\theta$ is a $M \times M$ covariance matrix decided and/or tuned by the user. This prior is related to the so called *Tikhonov's regularization* in least squares problems. We recall that the likelihood function does not change and it is again given in Eq. (24), i.e., $\ell(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi\theta}, \sigma_e^2\mathbf{I}_N)$.

5.4.1 Posterior of $\boldsymbol{\theta}$

It is possible to show that the posterior density of $\boldsymbol{\theta}$ is distributed again as Gaussian density [5, 24],

$$\bar{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y}), \tag{49}$$

where the mean vector is

$$\begin{aligned}\widehat{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\theta|y} &= \frac{1}{\sigma_e^2}\left(\frac{1}{\sigma_e^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{y},\\ &= \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{y},\\ &= \boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top\left(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right)^{-1}\mathbf{y}, \qquad (50)\end{aligned}$$

and the covariance matrix is

$$\begin{aligned}\boldsymbol{\Sigma}_{\theta|y} &= \left(\frac{1}{\sigma_e^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1},\\ &= \sigma_e^2\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{\Sigma}_\theta^{-1}\right)^{-1},\\ &= \boldsymbol{\Sigma}_\theta - \boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top\left(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta, \qquad (51)\end{aligned}$$

where in the last expression we have used the matrix identity in [5].

5.4.2 Posterior of $f(\mathbf{x})$

Recall that the assumed model is $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\theta}$ and we consider $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y})$, after seeing the data. For a fixed $\mathbf{x}$, the hidden function $f(\mathbf{x})$ is again a random variable with a Gaussian posterior density,

$$p(f(\mathbf{x})|\mathbf{y}) = \mathcal{N}(f(\mathbf{x})|\mu_{f|y}(\mathbf{x}), \sigma^2_{f|y}(\mathbf{x})), \tag{52}$$

with mean at $\mathbf{x}$,

$$\begin{aligned}
\mu_{f|y}(\mathbf{x}) = \widehat{f}(\mathbf{x}) &= \boldsymbol{\phi}(\mathbf{x})^\top \widehat{\boldsymbol{\theta}} \\
&= \boldsymbol{\phi}(\mathbf{x})^\top (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \sigma_e^2 \boldsymbol{\Sigma}_\theta^{-1})^{-1} \boldsymbol{\Phi}^\top \mathbf{y}, \\
&= \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}^\top \left(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2 \mathbf{I}_N\right)^{-1} \mathbf{y},
\end{aligned} \tag{53}$$

and variance

$$\begin{aligned}
\sigma^2_{f|y}(\mathbf{x}) &= \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\theta|y} \boldsymbol{\phi}(\mathbf{x}), \\
&= \boldsymbol{\phi}(\mathbf{x})^\top \left(\frac{1}{\sigma_e^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1} \boldsymbol{\phi}(\mathbf{x}), \\
&= \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}^\top \left(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2 \mathbf{I}_N\right)^{-1} \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta \boldsymbol{\phi}(\mathbf{x}).
\end{aligned} \tag{54}$$

where we have considered $\boldsymbol{\theta}$ is distributed as its posterior, $p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y})$ and the matrix identities given in [5, 24]. The regression function is the mean solution, i.e., $\widehat{f}(\mathbf{x}) = \mu_{f|y}(\mathbf{x})$ [19].

Posterior of the vector f. In the smoothing case, we have $\mathbf{f} = \boldsymbol{\Phi}\boldsymbol{\theta}$. Moreover, recall that the posterior of $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y})$, hence we obtain that

$$p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{f|y}, \boldsymbol{\Sigma}_{f|y}), \tag{55}$$

where the mean vector is

$$\begin{aligned}
\widehat{\mathbf{f}} = \boldsymbol{\mu}_{f|y} &= \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}, \\
&= \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{y}, \\
&= \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top\left(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right)^{-1}\mathbf{y},
\end{aligned} \tag{56}$$

and the covariance matrix is

$$\begin{aligned}
\boldsymbol{\Sigma}_{f|y} &= \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{\Phi}^\top, \\
&= \boldsymbol{\Phi}\left(\frac{1}{\sigma_e^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\boldsymbol{\Phi}^\top, \\
&= \left[\left(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top\right)^{-1} + \left(\sigma_e^2\mathbf{I}_N\right)^{-1}\right]^{-1}, \\
&= \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top - \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top\left(\sigma_e^2\mathbf{I}_N + \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top,
\end{aligned} \tag{57}$$

where in the last expression we have used the matrix identity in [5].

5.4.3 Marginal likelihood

Since we have used a proper prior density $g(\boldsymbol{\theta})$, we can compute the integral $\int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$ without any arbitrary constant, and it can be interpreted as a marginal likelihood $Z = p(\mathbf{y})$. Due to the assumed observation model, $\mathbf{y} = \mathbf{f} + \mathbf{e}$, we can observe that the vector $\mathbf{y}$ is the sum of two independent Gaussian vectors, $\mathbf{f}$ and $\mathbf{e}$, where

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top) \quad \text{and} \quad p(\mathbf{e}) = \mathcal{N}(\mathbf{e}|\mathbf{0}, \sigma_e^2\mathbf{I}_N). \tag{58}$$

The first density $p(\mathbf{f})$ is induced by the prior over $\boldsymbol{\theta}$, the second density $p(\mathbf{e})$ is given by assumption. Thus, $\mathbf{y}$ is also distributed as a Gaussian density with mean the sums of the means, and covariance matrix the sum of the two covariance matrices, i.e.,

$$Z = p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N), \tag{59}$$

where we recall that a complete notation would be $p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\Phi}, \boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta)$, but considering fixed the bases (hence $\boldsymbol{\Phi}$), we study the marginal likelihood as function of the hyper-parameters $\boldsymbol{\alpha}$, σ_e and $\boldsymbol{\Sigma}_\theta$. Therefore, the minus log-marginal likelihood $-\log Z$ is

$$\begin{aligned} -\log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta) &= \\ &= \frac{1}{2}\mathbf{y}^\top(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N)^{-1}\mathbf{y} + \frac{1}{2}\log\det\left[\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right] + \frac{N}{2}\log 2\pi, \end{aligned}$$

and finally

$$\boxed{-\log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta) = \underbrace{\frac{1}{2}\mathbf{y}^\top(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N)^{-1}\mathbf{y}}_{\text{fitting term}} + \underbrace{\frac{1}{2}\log\det\left[\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right]}_{\text{penalty term}} + const.,} \tag{60}$$

The factor $\mathbf{y}^\top(\boldsymbol{\Phi}\boldsymbol{\Sigma}_\theta\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N)^{-1}\mathbf{y}$ is a *fitting term*. The second term is a *penalization of the model complexity*, as we have already discussed above. Generally, one can try to maximize $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta)$, or minimize $-\log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta)$, in order to learn $\boldsymbol{\alpha}$, σ_e and $\boldsymbol{\Sigma}_\theta$ [14, 13, 23]. In this case, the covariance matrix $\boldsymbol{\Sigma}_\theta$ is a parameter of the prior density over $\boldsymbol{\theta}$.

5.4.4 Trying to come back to the case of the improper uniform prior

For simplicity, we assume that the covariance matrix of the prior of $\boldsymbol{\theta}$ is diagonal, i.e., $\boldsymbol{\Sigma}_\theta = \sigma_p^2\mathbf{I}_N$ so that, replacing in Eq. (60),

$$\begin{aligned} -\log Z = -\log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e, \boldsymbol{\Sigma}_\theta) &= \\ &= \frac{1}{2}\mathbf{y}^\top\frac{1}{\sigma_p^2}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \frac{\sigma_e^2}{\sigma_p^2}\mathbf{I}_N\right)^{-1}\mathbf{y} + \frac{1}{2}\log\det\left[\sigma_p^2\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right] + const., \end{aligned}$$

First, we focus on the first term. Applying the following Woodbury matrix identity [10],

$$\left(\mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^\top\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1} + \mathbf{C}^\top\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^\top\mathbf{A}^{-1},$$

to the first term, i.e.,

$$\left(a\mathbf{I}_N + \mathbf{\Phi}\mathbf{\Phi}^\top\right)^{-1},$$

where $a = \frac{\sigma_e^2}{\sigma_p^2}$, $\mathbf{A} = a\mathbf{I}_N$, $\mathbf{B} = \mathbf{I}_N$ and $\mathbf{C} = \mathbf{\Phi}$, then we obtain

$$\begin{aligned}
&\frac{1}{2\sigma_p^2}\mathbf{y}^\top \left(a\mathbf{I}_N + \mathbf{\Phi}\mathbf{\Phi}^\top\right)^{-1}\mathbf{y} = \\
&\qquad = \frac{1}{2\sigma_p^2}\mathbf{y}^\top \left(a^{-1}\mathbf{I}_N - a^{-1}\mathbf{\Phi}\left(\mathbf{I}_N + a^{-1}\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top a^{-1}\right)\mathbf{y} \\
&\qquad = \frac{1}{2\sigma_p^2}\mathbf{y}^\top \left(a^{-1}\mathbf{I}_N - a^{-1}\mathbf{\Phi}\left(a\mathbf{I}_N + \mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\right)\mathbf{y} \\
&\qquad = \frac{1}{2\sigma_p^2}a^{-1}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi}\left(a\mathbf{I}_N + \mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}\right)
\end{aligned}$$

Replacing $a = \frac{\sigma_e^2}{\sigma_p^2}$, we have

$$\frac{1}{2\sigma_p^2}\mathbf{y}^\top \left(a\mathbf{I}_N + \mathbf{\Phi}\mathbf{\Phi}^\top\right)^{-1}\mathbf{y} = \frac{1}{2\sigma_e^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi}\left(\frac{\sigma_e^2}{\sigma_p^2}\mathbf{I}_N + \mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}\right). \tag{61}$$

Finally, for $\sigma_p^2 \to \infty$, we get

$$\begin{aligned}
\lim_{\sigma_p^2\to\infty} \quad \frac{1}{2\sigma_e^2}\mathbf{y}^\top\left(\frac{\sigma_e^2}{\sigma_p^2}\mathbf{I}_N + \mathbf{\Phi}\mathbf{\Phi}^\top\right)^{-1}\mathbf{y} \quad &= \frac{1}{2\sigma_e^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi}\left(0\cdot\mathbf{I}_N + \mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}\right), \\
&= \frac{1}{2\sigma_e^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\top\mathbf{y}\right) \\
&= \frac{1}{2\sigma_e^2}\left(\mathbf{y}^\top\mathbf{y} - \widehat{\mathbf{f}}^\top\widehat{\mathbf{f}}\right), \\
&= \frac{1}{2\sigma_e^2}||\mathbf{y} - \widehat{\mathbf{f}}||^2, \qquad (62)
\end{aligned}$$

where we have employed that $\widehat{\mathbf{f}} = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{y}$ as in Eq. (56) when $\sigma_p^2 \to \infty$, becoming equal to Eq. (35), and we have used the expression in (39)-(40). Moreover, in the last equality, we have used Eq. (42). Hence, as $\sigma_p^2 \to \infty$, we are able to recover the first term in Eq. (44).

Let consider now the second term as $\sigma_p^2 \to \infty$. We obtain

$$\lim_{\sigma_p^2\to\infty} \frac{1}{2}\log\det\left[\sigma_p^2\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma_e^2\mathbf{I}_N\right] = \infty, \tag{63}$$

so that, for the complete expression in (60), we have $-\log Z \to \infty$ and hence $Z \to 0$.

Remark 14. *Namely, as discussed in Section 3.1, making the Gaussian prior more diffuse penalizes more the corresponding model and its marginal likelihood decreases to 0 (since $S < \infty$ always in this model).*

Remark 15. *Moreover, we are not able to recover completely the improper prior case. We are just able to recover the fitting term of the $-\log S$ in Eq.* (44) *as asymptotic case of $-\log Z$ in Eq.* (60)*, as shown in Eq.* (62)*. This confirms the fact that improper priors are not allowed for computing the evidence Z, and that the the area under the likelihood S cannot be considered an asymptotical special case of a marginal likelihood.*

6 Numerical examples

The aim of this section is to clarify the use and the meaning of the generic equations and concepts described previously. We provide two examples that can help the interested reader and practitioner to apply the expressions and ideas of the work. For instance, the first example deals with a very simple scenario of estimation of the mean of a Gaussian density given N data, in a Bayesian setting. However, even this simple example can serve as a guide for a proper use of the more general formulas given above, and to provide some interesting theoretical insights. The second example addresses a more general regression problem.

6.1 First numerical example

In order to clarify the contents in Sections 3.1 and 5.4.4, we consider a simple univariate linear model with a Gaussian prior over θ. This is also a very interesting special case of all the general expressions for the Bayesian regression. Indeed, le us consider

$$y_n = \theta + e_n, \qquad n = 1, ..., N, \tag{64}$$

where $e_n \sim \mathcal{N}(\epsilon|0, \sigma_e^2)$, so that the likelihood of only one observation is

$$p(y_n|\theta, \sigma_e) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{(y_n - \theta)^2}{2\sigma_e^2}\right), \quad \text{and} \tag{65}$$

$$g(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2}\right), \tag{66}$$

is the prior over θ with hyper-parameters μ_θ and σ_θ^2, hence, here $\mathbf{\Sigma}_\theta = \sigma_\theta^2 \mathbf{I}_N$. Let us consider to observe N i.i.d. data y_i and let recall $\mathbf{y} = [y_1, ..., y_N]^\top$. The complete likelihood is

$$\ell(\mathbf{y}|\theta, \sigma_e) = \prod_{i=1}^{N} p(y_n|\theta, \sigma_e). \tag{67}$$

In a vectorial form, we can write

$$\mathbf{y} = \mathbf{1}_N \theta + \mathbf{e}, \tag{68}$$

where $\mathbf{1}_N = [1, ..., 1]^\top$ is $N \times 1$ vector of ones. If we compare Eq. (68) with the Eq. (22), that we rewrite below,

$$\underbrace{\mathbf{y}}_{N\times 1} = \underbrace{\mathbf{\Phi}}_{N\times M} \underbrace{\boldsymbol{\theta}}_{M\times 1} + \underbrace{\mathbf{e}}_{N\times 1},$$

we can see that $M = 1$ and $\mathbf{\Phi} = \mathbf{1}_N$, in this example. Hence, the marginal likelihood in Eq. (59) becomes

$$Z = p(\mathbf{y}) = p(\mathbf{y}|\sigma_p, \mu_p, \sigma_e) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma_p^2 \mathbf{1}_N \mathbf{1}_N^\top + \sigma_e^2 \mathbf{I}_N), \tag{69}$$

where $\mathbf{1}_N \mathbf{1}_N^\top$ is a $N \times N$ matrix of 1's in all the entries. This has an important statistical meaning: the observed data y_n's are *conditional independent* given θ, but they are not *independent*. Indeed, we have that $\mathbf{\Sigma}_{yy} = \sigma_p^2 \mathbf{1}_N \mathbf{1}_N^\top + \sigma_e^2 \mathbf{I}_N$, hence

$$\begin{aligned} \text{cov}[y_n, y_j] &= \sigma_p^2, \quad \text{for } n \neq j, \\ \text{cov}[y_n, y_n] &= \text{var}(y_n) = \sigma_p^2 + \sigma_e^2. \end{aligned}$$

Namely, the covariance matrix of $\mathbf{y}$ is

$$\mathbf{\Sigma}_{yy} = \sigma_p^2 \mathbf{1}_N \mathbf{1}_N^\top + \sigma_e^2 \mathbf{I}_N = \begin{bmatrix} \sigma_p^2 + \sigma_e^2 & \sigma_p^2 & \dots & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 + \sigma_e^2 & \dots & \sigma_p^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_p^2 & \sigma_p^2 & \dots & \sigma_p^2 + \sigma_e^2 \end{bmatrix}. \tag{70}$$

The log-marginal likelihood in Eq. (60) (with the difference that here the prior has non-zero mean, i.e., $\mu_p \neq 0$) can be rewritten in this scenario as

$$\log Z = \log p(\mathbf{y}) = -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_p)^\top \mathbf{\Sigma}_{yy}^{-1} (\mathbf{y} - \boldsymbol{\mu}_p) - \frac{1}{2} \log[\det \mathbf{\Sigma}_{yy}] - \frac{N}{2} \log(2\pi), \tag{71}$$

where we set $\boldsymbol{\mu}_p = [\mu_p, \mu_p]^\top$. Recall that actually $p(\mathbf{y}) = p(\mathbf{y}|\sigma_p, \mu_p, \sigma_e)$.

6.1.1 Analysis of the area under the likelihood $S(\mathbf{y}|\sigma_e)$

First of all, note that in this example the only auxiliary parameter in the observation model (hence the likelihood) is σ_e^2. Clearly, there is also the main variable object of the inference, i.e., θ. In a frequentist approach, we could find

$$\left[\widehat{\theta}_{\texttt{ML}}, \widehat{\sigma}_{e\texttt{ML}}\right] = \arg\max_{\theta, \sigma_e} \ell(\mathbf{y}|\theta, \sigma_e). \tag{72}$$

In this specific scenario, we have the analytical expressions of the two estimators in close form,

$$\widehat{\theta}_{\texttt{ML}} = \frac{1}{N} \sum_{n=1}^{N} y_n, \quad \widehat{\sigma}_{\texttt{ML}}^2 = \frac{1}{N} \sum_{n=1}^{N} \left(y_n - \widehat{\theta}_{\texttt{ML}}\right)^2. \tag{73}$$

Note that $\frac{1}{N} \sum_{n=1}^{N} \left(y_n - \widehat{\theta}_{\texttt{ML}}\right)^2$ is a *biased* estimator of the variance σ_e^2. However, if the area under the likelihood w.r.t. θ is finite, in a more Bayesian fashion, we could also study

$$S(\mathbf{y}|\sigma_e) = \int_\Theta \ell(\mathbf{y}|\theta, \sigma_e) d\theta. \tag{74}$$

If $S(\mathbf{y}|\sigma_e) < \infty$ for each σ_e, we can maximize this function, for instance.[1] Note that a symmetric approach could be also employed w.r.t. θ, i.e., analyzing $S(\mathbf{y}|\theta) = \int \ell(\mathbf{y}|\theta, \sigma_e) d\sigma_e$, if finite for each value of $\theta \in \Theta$. It is possible to show that maximizing $S(\mathbf{y}|\sigma_e)$, we obtain

$$\widehat{\sigma}_{\texttt{S}}^2 = \arg\max_{\sigma_e} S(\mathbf{y}|\sigma_e) = \frac{1}{N-1} \sum_{n=1}^{N} (y_n - \widehat{\theta}_{\texttt{ML}})^2. \tag{75}$$

that is the *unbiased* estimator of the variance (see also Eq. (45)). This shows a clear benefit in integrating out θ in (74). We have checked and tested numerically this result with different vectors $\mathbf{y}$ of $N = 2$ data (we have considered symmetric data only, for simplicity; in this case $\widehat{\theta}_{\texttt{ML}} = 0$). See Table 2.

Table 2: Numerical maximization of $\arg\max_{\theta,\sigma_e} \ell(\mathbf{y}|\theta, \sigma_e)$ and $\arg\max_{\sigma_e} S(\mathbf{y}|\sigma_e)$. The results coincide with $\widehat{\sigma}_{\texttt{ML}}^2$ and $\widehat{\sigma}_{\texttt{S}}^2$ in Eqs. (73)-(75).

Estimator	**unbiased**	$\mathbf{y} = [0, 0]^\top$	$\mathbf{y} = [-1, 1]^\top$	$\mathbf{y} = [-2, 2]^\top$	$\mathbf{y} = [-3, 3]^\top$
$\widehat{\sigma}_{\texttt{ML}}^2$	✗	0	1	4	9
$\widehat{\sigma}_{\texttt{S}}^2$	✓	$10^{-6} (\approx 0)$	1.988 (≈ 2)	8.008 (≈ 8)	17.977(≈ 18)

6.1.2 Using a proper diffuse prior: asymptotic analysis $\sigma_p \to \infty$

For the sake simplicity, first we keep fixed $\sigma_e^2 = 1$, $\mu_p = 2$ and consider $\mathbf{y} = [-2, 2]^\top$. In Figure 1(a), we can see

$$\log Z = \log p(\mathbf{y}|\sigma_p, \sigma_e = 1, \mu_p = 2) = p(\mathbf{y}|\sigma_p),$$

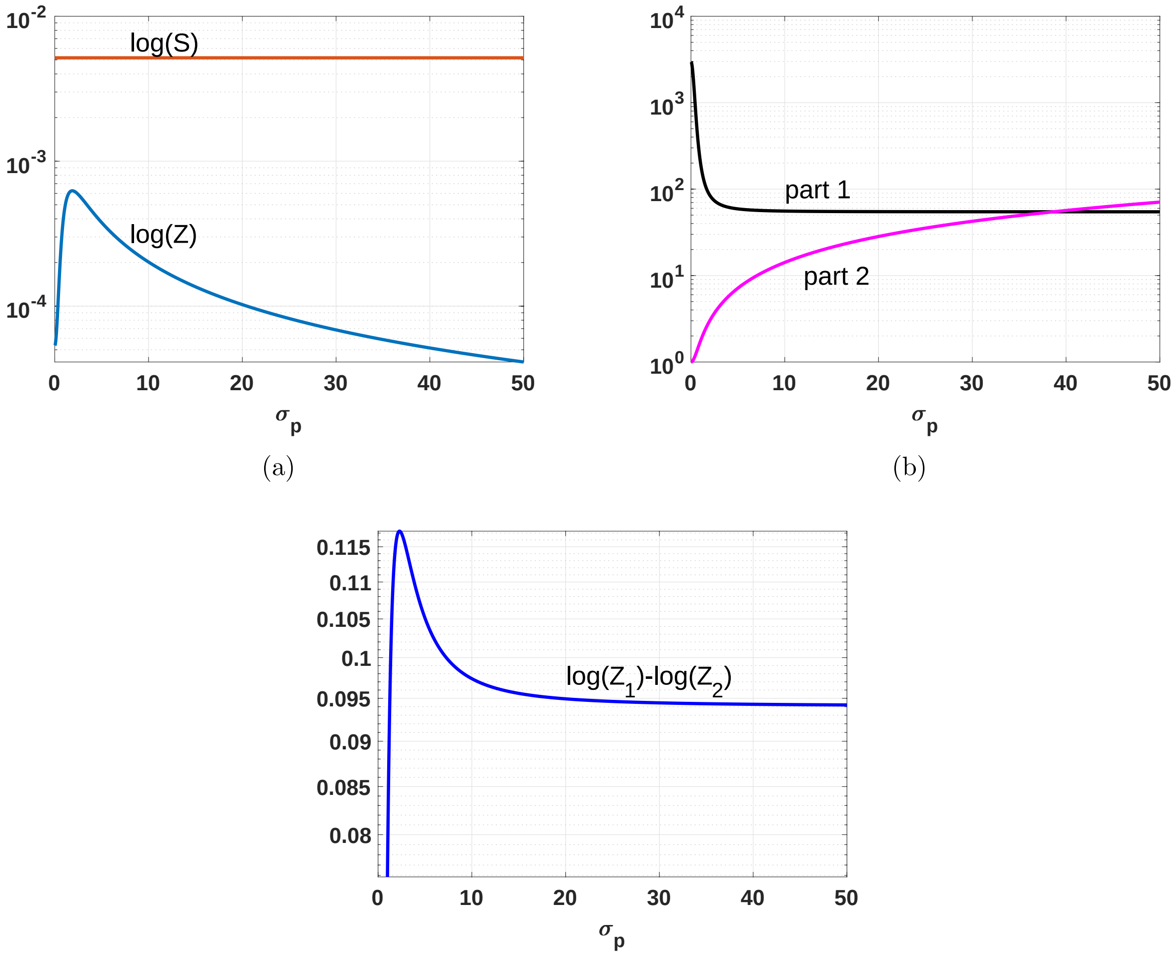
in Eq. (71) as function σ_p and the area under the likelihood $S = S(\mathbf{y}|\sigma_e = 1)$. We can observe that $\log Z \to -\infty$ hence $Z \to 0$ (hence a diffuse prior, when S is finite, penalizes more and more the model) and $Z \nrightarrow S$, i.e., S cannot be recovered with asymptotical arguments. Moreover, rewriting the negative log-marginal likelihood,

$$-\log Z = -\log p(\mathbf{y}|\sigma_p) = \underbrace{\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_{yy}^{-1} (\mathbf{y} - \boldsymbol{\mu}_p)}_{\text{part 1}} + \underbrace{\frac{1}{2} \log[\det \boldsymbol{\Sigma}_{yy}]}_{\text{part 2}} + const, \tag{76}$$

we can study the behavior of the first and second part. See Figure 1(b). As $\sigma_p \to \infty$, we can observe numerically that the first part converges to the term $\frac{1}{2\sigma_e} ||\mathbf{y} - \widehat{\mathbf{f}}||^2$ where here $\widehat{\mathbf{f}} = \widehat{\theta} = \frac{1}{N} \sum_{n=1}^{N} y_n$. This result coincides with Eq. (62). The second part diverges to ∞, as $\sigma_p \to \infty$, as shown in (63). As a consequence, we have $-\log Z \to \infty$ and $Z \to 0$. Finally, in Figure

[1] Recall that, technically, $S(\mathbf{y}|\sigma_e)$ cannot be called marginal likelihood since has been not obtained averaging likelihood values according to a proper prior. Indeed, we can see that its value does not belong the marginal likelihood values as depicted in Figure 1(a).

1(c), we show the log-difference between two marginal likelihood $Z_1 = p(\mathbf{y}|\sigma_p, \sigma_e = 1, \mu_p = 2)$ and $Z_2 = p(\mathbf{y}|\sigma_p, \sigma_e = 4, \mu_p = 2)$, that converges to a constant (horizontal asymptote) as $\sigma_p \to \infty$. This also confirm that the Bayes factor $\frac{Z_1}{Z_2}$ still contains useful statical information even when $\sigma_p \to \infty$.



(a)

(b)

(c)

Figure 1: **(a)** The area under the likelihood $S = S(\mathbf{y}|\sigma_e = 1)$ and Z in log-domain as function of σ_p. **(b)** Part 1 in $-\log Z$ converges to a constant, whereas part 2 in $-\log Z$ diverges. **(c)** The difference $\log Z_1 - \log Z_2$ converges to a constant (horizontal asymptote); $Z_1 = p(\mathbf{y}|\sigma_p, \sigma_e = 1, \mu_p = 2)$ corresponds to $\sigma_e = 1$ and $Z_2 = p(\mathbf{y}|\sigma_p, \sigma_e = 4, \mu_p = 2)$ corresponds to $\sigma_e = 4$.

6.2 Second numerical example

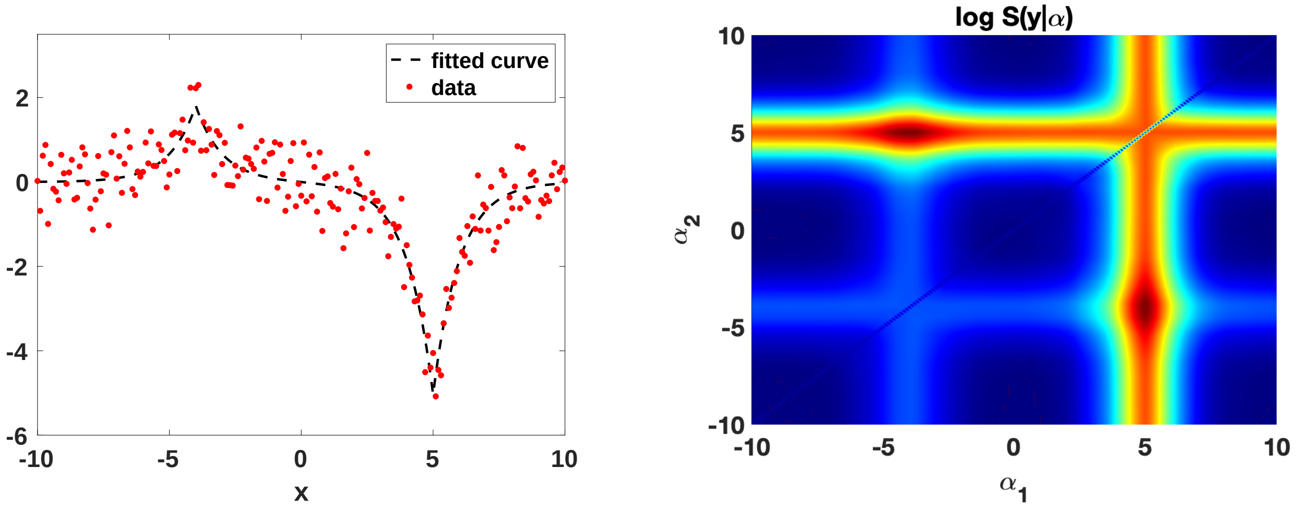
Let us consider to observe the data points $\{x_i, y_i\}_{i=1}^N$ generated by the following model,

$$y_i = \theta_1 \exp\left(|x_i - \alpha_1|\right) + \theta_2 \exp\left(|x_i - \alpha_2|\right) + e_i, \text{ with } \alpha_1 < \alpha_2, \tag{77}$$

where $e_i \sim \mathcal{N}(e|0, \sigma_e^2)$. We set $\sigma_e^2 = 0.5$, $\boldsymbol{\theta}_{\texttt{true}} = [\theta_1 = 2, \theta_2 = -5]^\top$ and $\boldsymbol{\alpha}_{\texttt{true}} = [\alpha_1 = -4, \alpha_2 = 6]^\top$. We consider a vector $\mathbf{y} = [y_1, ..., y_N]^\top$ of $N = 200$, generated by the model above considering $\boldsymbol{\theta}_{\texttt{true}}$, $\boldsymbol{\alpha}_{\texttt{true}}$ and equispaced values of x_i, from -10 to 10.

The goal is to make inference regarding $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. We consider a Bayesian approach for inferring $\boldsymbol{\theta}$ considering an improper uniform prior. For tuning $\boldsymbol{\alpha}$, we consider the maximization of the area under the likelihood $S(\mathbf{y}|\boldsymbol{\alpha})$ given in Section 5.3.3. We use Eq. (29) as estimator of $\boldsymbol{\theta}$, Moreover, considering known the noise power $\sigma_e^2 = 0.5$, we numerically maximize $S(\mathbf{y}|\boldsymbol{\alpha})$ in Eq. (43), or equivalently minimize (44) (or Eq. (47)). We compute the mean square error (MSE) in estimation for both $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. We repeat the procedure in 2000 different independent runs.

The MSEs averaged over 2000 independent runs are 0.0271 and 0.0317 with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, respectively. An example of data realization and the corresponding curve fitting is given in Figure 2(a). An example of log area under the likelihood $\log S(\mathbf{y}|\boldsymbol{\alpha})$ for a given realization is shown in Figure 2(b). The histograms of the estimated values of each component of the vector of $\boldsymbol{\alpha}$ (in different realizations) are given in Figure 3.
Therefore, by this numerical example, we can confirm that the maximization of the area under the likelihood $\log S(\mathbf{y}|\boldsymbol{\alpha})$ can be employed for tuning parameters of the observation model.



(a)

(b)

Figure 2: **(a)** One realization of the data vector $\mathbf{y}$ and a corresponding fitted curve according to the observation model. **(b)** Example of the log area under the likelihood $\log S(\mathbf{y}|\boldsymbol{\alpha})$ in one realization of the data vector $\mathbf{y}$. We can see that the maxima are localized around approximately $[-4, 5]$ and $[5, -4]$ (just $[-4, 5]$ is admissible since $\alpha_1 < \alpha_2$).

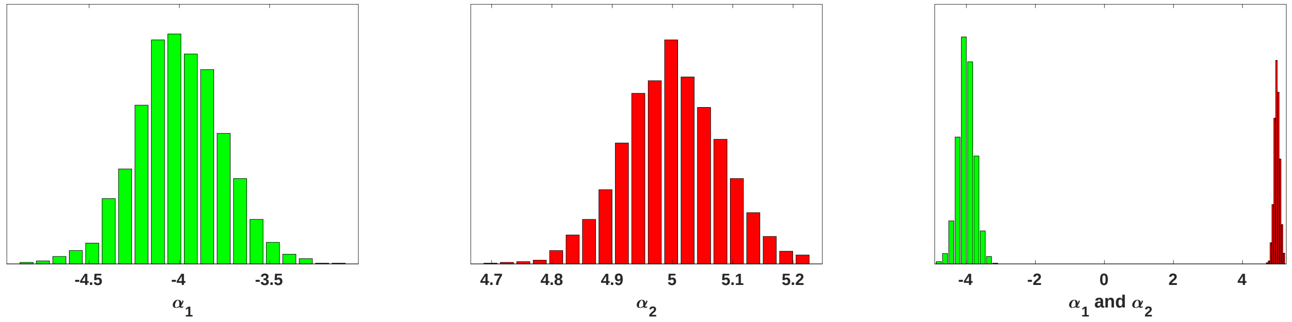


Figure 3: Histograms of each component of estimated vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^\top$ (maximizing $S(\mathbf{y}|\boldsymbol{\alpha})$) over 2000 independent realizations. We can observe that bias is virtually zero and the variance in bigger for α_1 with respect to α_2. This is reasonable looking the realization of data in Figure 2(a) where the (negative) pick at $x = 5$ seems much more clear/evident, than the first pick at $x = -4$.

7 Conclusions

In this work, we have remarked some relevant points regarding the use of diffuse priors and improper priors in level-1 and level-2 of Bayesian inference. We have stressed their impact into the inference and their possible use with a complete discussion and by illustrative examples and numerical experiments. The relationship with the profile likelihood approach has been also described. As a summary, we have pointed out the following important statements:

- Diffuse priors are uninformative in the Level-1 of inference, but can be *very informative* in Level-2 (i.e., for Bayesian model selection). Indeed, if $S < \infty$, a more diffuse prior penalizes more the model providing a smaller marginal likelihood Z (and asymptotically $Z \to 0$).

- An improper prior are not allowed for the computation of the evidence Z. However, we can compute the quantity $Z_{\ell \times h}$ which contains statistical information useful for Type-2 of model selection. Namely, the improper priors can be used when we have several models belonging to the same parametric family (i.e., for tuning parameters of a parametric model) or, more generally, for tuning parameters that are shared by different models.

- We suggest to call the computed quantity $Z_{\ell \times h}$ as "fake evidence" or, in the case of uniform improper priors, as "the area under the likelihood" $S = Z_{\ell \times 1}$.

- Another interesting aspect is that the area under the likelihood S cannot be obtained as a special asymptotic case of marginal likelihood Z, applying a diffuse prior and increase its scale parameter to infinity. Namely, a diffuse prior can become asymptotically a uniform improper prior. In Level-1 of inference, we can recover asymptotically the results obtained using a uniform improper prior, starting from a diffuse prior. In Level-2, this is not true and $Z \nrightarrow S$, i.e., we cannot recover the area under the likelihood S by a sequence of marginal likelihoods correspond to different diffuse priors (obtained increasing their scale parameter).

- Let assume $S < \infty$. Even if the evidence Z is not defined with an improper prior (and $Z \to 0$ using a diffuse prior), the ratio of two evidences Z_1/Z_2 (Bayes factor) corresponding a two different values of a shared parameter is still well-defined and can be computed by the ratio of the fake-evidences $Z_{\ell_1 \times h}/Z_{\ell_2 \times h}$ (applying the same improper prior to both models).

We have discussed all these aspects firstly in a general way, and then more specifically within a Bayesian regression model, considering a uniform improper prior and a Gaussian prior. Moreover, two numerical experiments, one involving an interesting special case and a specific example of generalized linear model, have been also provided performing clarifications, checks and additional remarks by numerical simulations.

References

[1] Sergey A. Anfinogentov, Valery M. Nakariakov, David J. Pascoe, and Christopher R. Goddard. Solar Bayesian Analysis Toolkit—A New Markov Chain Monte Carlo IDL Code for Bayesian Parameter Inference. *Astrophysical Journal Supplement Series*, 252(1):11, January 2021.

[2] G. Ashton and C. Talbot. BILBY-MCMC: an MCMC sampler for gravitational-wave inference. *Monthly Notices of the Royal Astronomical Society*, 507(2):2037–2051, October 2021.

[3] Ismael Ayuso, Ruth Lazkoz, and Vincenzo Salzano. Observational constraints on cosmological solutions of f (Q) theories. *Physical review d*, 103(6):063505, March 2021.

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

[5] C. M Bishop. *Pattern recognition and machine learning.* springer, 2006.

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

[7] J. Emmert, S. J. Grauer, S. Wagner, and K. J. Daun. Efficient Bayesian inference of absorbance spectra from transmitted intensity spectra. *Opt. Express*, 27(19):26893–26909, 2019.

[8] Farhan Feroz, Michael P. Hobson, Ewan Cameron, and Anthony N. Pettitt. Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, 2(1):10, November 2019.

[9] Philip C. Gregory. Bayesian re-analysis of the Gliese 581 exoplanet system. *Monthly Notices of the Royal Astronomical Society*, 415(3):2523–2545, August 2011.

[10] W. W. Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):221–239, 1989.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

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

[13] F. Llorente, L. Martino, E. Curbelo, J. Lopez-Santiago, and D. Delgado. On the safe use of prior densities for Bayesian model selection. *WIREs Computational Statistics*, 15(1):e1595, 2023.

[14] F. Llorente, L. Martino, D. Delgado, and J. López-Santiago. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review*, 65(1):3–58, 2023.

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

[16] D. J. C. MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

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

[18] L. Martino, F. Llorente, E. Curbelo, J. Lopez-Santiago, and J. Miguez. Automatic tempered posterior distributions for Bayesian inversion problems. *Mathematics*, 9(7), 2021.

[19] L. Martino and J. Read. A joint introduction to Gaussian Processes and Relevance Vector Machines with connections to Kalman filtering and other kernel smoothers. *Information Fusion*, 74:17–38, Oct 2021.

[20] D. J. Pascoe, A. Smyrli, T. Van Doorsselaere, and A. M. Broomhall. Bayesian Analysis of Quasi-periodic Pulsations in Stellar Flares. *Astrophysical Journal*, 905(1):70, December 2020.

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

[22] C. Robert and G. Casella. *Monte Carlo statistical methods.* Springer, 2004.

[23] C. P. Robert. *The Bayesian choice: a decision-theoretic motivation.* Springer-Verlag, 1994.

[24] J. K. O Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Springer, 1996.

[25] U. Von Toussaint. Bayesian inference in physics. *Rev. Mod. Phys.*, 83:943–999, 2011.

A Maximum likelihood estimator

The maximum of the likelihood function is reached at

$$\widehat{\boldsymbol{\theta}}_{\texttt{ML}} = (\underbrace{\boldsymbol{\Phi}^\top \boldsymbol{\Phi}}_{M \times M})^{-1} \boldsymbol{\Phi}^\top \mathbf{y}. \tag{78}$$

Moreover, the covariance matrix of the estimator above is

$$\boldsymbol{\Sigma}_{\widehat{\theta}} = \sigma_e^2 (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \tag{79}$$

and we can also write

$$\widehat{\boldsymbol{\theta}}_{\texttt{ML}} \sim \mathcal{N}(\widehat{\boldsymbol{\theta}} | \boldsymbol{\theta}_{\texttt{true}}, \boldsymbol{\Sigma}_{\widehat{\theta}}), \tag{80}$$

where $\boldsymbol{\theta}_{\texttt{true}}$ above represents the true vector generating the observations $\mathbf{y}$ following the model in Eq.(22), and it is also the mean of the Gaussian density above. This density is the *sampling distribution* of the estimator $\widehat{\theta}_{\texttt{ML}}$.

Note that the sampling distribution of the estimator, $\mathcal{N}(\widehat{\theta}_{\texttt{ML}} | \boldsymbol{\theta}_{\texttt{true}}, \boldsymbol{\Sigma}_{\widehat{\theta}})$, is philosophically completely different from a posterior distribution over the vector $\boldsymbol{\theta}$, given in Section 5.3.1. Indeed, the sampling distribution is a probability density that describes the probabilities with which the possible values for the estimator $\widehat{\theta}_{\texttt{ML}} = \widehat{\theta}_{\texttt{ML}}(\mathbf{y})$ occur, when different realizations of the data $\mathbf{y}$ are given. Namely, for a different realization $\mathbf{y}'$ we have a different estimator $\widehat{\theta}_{\texttt{ML}}(\mathbf{y}')$. If we have another realization $\mathbf{y}''$, we have a different estimator $\widehat{\theta}_{\texttt{ML}}(\mathbf{y}'')$. These vectors, $\widehat{\theta}_{\texttt{ML}}(\mathbf{y}')$ and $\widehat{\theta}_{\texttt{ML}}(\mathbf{y}'')$, are samples distributed according to the sampling distribution of the estimator. On the other hand, in a Bayesian framework, the data $\mathbf{y}$ are considered fixed and given (conceptually just one realization of data is considered). The posterior distribution over $\boldsymbol{\theta}$ expresses all the statistical knowledge about $\boldsymbol{\theta}$ after observing the data (and considering the prior information and/or beliefs) [5, 23].