Data-driven priors via hyper-parameter posteriors of Gaussian processes

L. Martino[⋆], J. López-Santiago[†], J. Miguez[†], G. Vázquez-Vilar[†]
[†] Universidad Carlos III de Madrid, Leganés (Madrid), Spain
[⋆] Università degli Studi di Catania, Corso Italia 55, Catania, Italy

Abstract

When neither prior knowledge nor expert opinion is available, non-informative priors provide a practical alternative for conducting Bayesian inference. However, in the context of model selection, genuinely non-informative priors do not exist. In fact, diffuse priors on the parameters can drastically alter the value of the Bayesian evidence, making them effectively highly informative, while improper priors are even not allowed. Furthermore, in many real-world applications, the use of informative priors can substantially improve the computational efficiency by driving sampling algorithms toward regions of high posterior probability. In this work, we introduce a data-driven procedure for an automatic prior construction. The underlying idea is to exploit the posteriors of the hyper-parameters from non-parametric models, to construct priors for Bayesian inference in parametric models. We test the proposed scheme in four different experiments, two of which involve real astronomical data.

Keywords: automatic prior design; informative priors; Bayesian inference; Gaussian Processes; Exoplanets

1 Introduction

In Bayesian inference, the prior density plays a pivotal role as it encodes the analysts beliefs or available knowledge about the parameters before observing the data [1, 2, 26]. An appropriately chosen prior can regularize the inference, improve parameter identifiability, and prevent overfitting. Moreover, the prior can serve as a mechanism to incorporate domain-specific knowledge, physical constraints, or structural properties of the model into the analysis [8, 15, 29, 10].
In the absence of expert judgment or substantive prior knowledge, an appropriate alternative is to resort to *non-informative* priors, encompassing diffuse, reference, and, in some cases, improper priors. However, the use of non-informative priors have two main drawbacks:

- In the context of model selection, non-informative priors present relevant issues. In fact, the use of diffuse priors over the parameters can radically change the value of the Bayesian evidence (a.k.a., marginal likelihood) and, as a consequence, the result of the model selection [17, 16, 21]. Hence, diffuse/vague priors are actually *informative* for model selection. Additionally, the use of improper priors is forbidden in a model selection context. Indeed, in this case, the marginal likelihood is not completely specified due to an arbitrary constant involved in the computation [16, 21].

- In high-dimensional or complex models, the use of informative priors can substantially improve the computational efficiency by driving sampling algorithms toward regions of high posterior probability. Conversely, overly diffuse or improper priors, though sometimes used for their perceived neutrality, can exacerbate computational challenges and yield unstable inferences. Therefore, careful prior elicitation is a fundamental step in ensuring a robust Bayesian analysis.

A possible solution is the use of empirical priors, also called data-driven priors [20, 12, 33]. Note that this strategy includes the well-established *empirical Bayes* approach [14, 27, 30]. This strategy reduces the need for subjective and arbitrary prior specification but there are other important advantages. From a computational point of view, the construction of prior densities that are not in conflict with the main modes of the likelihood function helps Monte Carlo algorithms to achieve good performance. That is, this approach can improve the accuracy and robustness of the inference.

In this work, we introduce an automatic prior construction, that is a data-dependent procedure which reduces the burden of manual elicitation. We consider the inference problem of a parametric model (derived by engineering, physical and/or environment knowledge) by a previous analysis done a non-parametric model as Gaussian process (GP) technique [3, 23, 32]. More specifically, we suggest the use of posterior distributions over a GP hyper-parameters as possible prior densities. The key point is that that posterior distributions derived from non-parametric models tend to be more diffuse than those obtained from specific parametric models. This property facilitates faster and more efficient exploration of the state space by sampling algorithms, enabling them to identify high-probability regions more readily. Furthermore, we discuss different tempering strategy to further spread out the hyper-parameter GP posteriors and minimize the

impact of the double use of the data.

The proposed approach is then particularly suitable when the analyzed parametric model is highly sensitive to small variations in one or more parameters, as such sensitivity can yield a markedly concentrated posterior distribution. Inference on parameters of this nature is especially challenging and necessitates the use of informative priors to ensure accurate and reliable estimation. Four numerical experiments, two of which involve real astronomical data, demonstrate the benefits of the proposed scheme.

The rest of work is structured as follows. In Section 2, we present the required framework and the main idea. Section 3 introduce the proposed scheme. Section 4 recalls the Gaussian processes and the interpretation of the hyper-parameters in different kernel function. The range of applications is discussed in Section 5. In Section 6, we test the performance of the method on simulations and real data of different types. Some conclusions are given in Section 7.

2 Framework and main idea

2.1 Problem statement

Let assume a set of M data points formed by pairs inputs-outputs, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{M}$. We assume the following observation model,

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \qquad i = 1, ..., M, \tag{1}$$

where $f(\mathbf{x}) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ is an unknown function of the independent variable $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\epsilon_i \sim \mathcal{N}(\epsilon|0, \sigma_e^2)$ is an indepedent noise perturbation, for all $i = 1, ..., M$. The variance of the observation noise, σ_e^2 is also considered unknown and must be inferred. The observed data vector is denoted as $\mathbf{y} = [y_1, ..., y_M]^\top$ and, we also set $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_M]^\top$. The observations y_i are conditional independent given $\mathbf{x}$.

We assume that we can express $f(\mathbf{x})$ with a parametric model (e.g., derived by physical or statistical knowledge of the specific application and/or phenomenon) with parameters $\boldsymbol{\lambda} = [\lambda_1, ...\lambda_{D_\lambda}, \sigma_e]$. The variable subject to the inference is then $\boldsymbol{\lambda}$.

2.2 Main concepts

Here, we formalize the underlying idea of the proposed procedure. We actually consider two type of models to express $f(\mathbf{x})$:

- as described above, a parametric model with vector of parameters (or hyper-parameters) $\boldsymbol{\lambda} = [\lambda_1, ...\lambda_{D_\lambda}]$. This first model induces the likelihood function $\ell_1(\mathbf{y}|\boldsymbol{\lambda})$;
- a non-parametric model over $f(\mathbf{x})$ (e.g., as a GP) with hyper-parameters $\boldsymbol{\theta} = [\theta_1, ...\theta_{D_\theta}]$, which induces a likelihood function $\ell_2(\mathbf{y}|\boldsymbol{\theta})$. More precisely, $\ell_2(\mathbf{y}|\boldsymbol{\theta})$ will be the marginal likelihood in the case of a GP model [32, 23].

In this work, we focus on the scenario where certain components, specifically λ_j and θ_k, are inter-related or represent analogous physical and mathematical entities. In this scenario, it becomes feasible to merge insights garnered from one model into the structure of the other. To explain the main idea, we consider below a simplified case one-dimensional case, as example.

3 Automatic procedure for prior construction

3.1 Best-case scenario

Let $D_\lambda = D_\theta$, and let the elements contained in the vector $\boldsymbol{\lambda}$ represent exactly the same magnitude as $\boldsymbol{\theta}$, in some way that we can directly consider $\boldsymbol{\lambda} = \boldsymbol{\theta}$. We assume a prior density $p_\theta(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$. This prior $p_\theta(\boldsymbol{\theta})$ could be uninformative and improper (see remark below). In this scenario, we could build a posterior distribution for $\boldsymbol{\theta}$ from the second *non-parametric* model,

$$p_2(\boldsymbol{\theta}|\mathbf{y}) \propto \ell_2(\mathbf{y}|\boldsymbol{\theta}) p_\theta(\boldsymbol{\theta}) \tag{2}$$

$$p_2(\boldsymbol{\theta}|\mathbf{y}) = \underbrace{p_2(\boldsymbol{\lambda}|\mathbf{y})}_{\texttt{posterior}} \propto \underbrace{\ell_2(\mathbf{y}|\boldsymbol{\lambda})}_{\texttt{likehood}} \underbrace{p_\theta(\boldsymbol{\lambda})}_{\texttt{prior}}, \tag{3}$$

where we set $p_\theta(\boldsymbol{\lambda}) = p_\theta(\boldsymbol{\theta})$ since $\boldsymbol{\lambda} = \boldsymbol{\theta}$. Within this constrained framework, we can construct a posterior distribution employing the likelihood $\ell_1(\mathbf{y}|\boldsymbol{\lambda})$ induced by the parametric model, and $p_2(\boldsymbol{\lambda}|\mathbf{y})$ obtained in (3) as an *informative* prior for $\boldsymbol{\lambda}$, i.e.,

$$\underbrace{p_1(\boldsymbol{\lambda}|\mathbf{y})}_{\texttt{posterior}} \propto \underbrace{\ell_1(\mathbf{y}|\boldsymbol{\lambda})}_{\texttt{likehood}} \times \underbrace{p_2(\boldsymbol{\lambda}|\mathbf{y})}_{\texttt{used-as-prior}} . \tag{4}$$

Remark. Note that the proposed strategy requires only that the prior $p_\theta(\boldsymbol{\theta})$ yields a proper posterior distribution $p_2(\boldsymbol{\theta}|\mathbf{y})$. The impact of this choice on the results is minimal, even in the context of model selection. Indeed, improper priors may be employed if the likelihood function ℓ_2 exhibits a finite fake-evidence (i.e., the area under the likelihood curve) [21]. See also Eq. (23) below.

Remark. The key point is that posterior distributions obtained from non-parametric models are often more diffuse/vague than those arising from specific parametric models. This allows a faster and easier exploration of the state space by the Monte Carlo algorithm, finding the high-probability regions.

3.2 Handling a more general scenario

Let assume $D_\lambda > 1$, $D_\theta > 1$ and possibly $D_\lambda \neq D_\theta$. Generally, only certain components, for instance, $\theta_{k_1}, \theta_{k_2}$ are inter-related or represent analogous physical and mathematical entities with λ_1, λ_2. For sake of simplicity and without loss of generality, we assume that all the elements of $\boldsymbol{\lambda}$ are contained in $\boldsymbol{\theta}$, so that we can write

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2] = [\theta_{k_1}, \theta_{k_2}]. \tag{5}$$

In this case, we need to obtain the *marginal posterior* of θ_{k_1} and θ_{k_2},

$$p_2(\boldsymbol{\lambda}|\mathbf{y}) = p_2(\theta_{k_1}, \theta_{k_2}|\mathbf{y}) = \int_{\Theta_{-\mathbf{k}}} p_2(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-\mathbf{k}}, \tag{6}$$

where

$$\boldsymbol{\theta}_{-\mathbf{k}} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{k_1-1}, \boldsymbol{\theta}_{k_1+1}, ..., \boldsymbol{\theta}_{k_2-1}, \boldsymbol{\theta}_{k_2+1}, ..., \boldsymbol{\theta}_{D_\theta}].$$

We can use the marginal posterior $p_2(\boldsymbol{\lambda}|\mathbf{y}) = p_2(\theta_{k_1}, \theta_{k_2}|\mathbf{y})$ (or its approximation) as a prior over $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$. A simple alternative is to consider the *conditional posterior* $p_2(\boldsymbol{\lambda}|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{-\mathbf{k}}) = p_2(\theta_{k_1}, \theta_{k_2}|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{-\mathbf{k}})$ as prior for λ_1, λ_2, where $\widehat{\boldsymbol{\theta}}_{-\mathbf{k}}$ is a point-wise estimation of $\boldsymbol{\theta}_{-\mathbf{k}}$.

We can compute the integral in (6) using Monte Carlo schemes [7, 6]: we draw samples $\boldsymbol{\theta}^{(n)}$, $n = 1, ..., N$, from $p_2(\boldsymbol{\theta}|\mathbf{y})$, and then simply ignore all the components with the exception of the k_1-th and k_2-th entries in the generated samples $\boldsymbol{\theta}^{(n)}$. More precisely, we draw a vector $\boldsymbol{\theta}^{(n)}$ from $p_2(\boldsymbol{\theta}|\mathbf{y})$ (by MCMC or IS+R) and then consider only the components k_1 and k_2, i.e.,

$$\boldsymbol{\lambda}_2^{(n)} = [\lambda_{2,1}^{(n)}, \lambda_{2,2}^{(n)}] = [\theta_{k_1}^{(n)}, \theta_{k_2}^{(n)}],$$

obtaining an unweighted particle approximation $\{\boldsymbol{\lambda}_2^{(n)}\}_{n=1}^N$,

$$\widehat{p}_2(\boldsymbol{\lambda}|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}_2^{(n)}), \tag{7}$$

of $p_2(\boldsymbol{\lambda}|\mathbf{y})$. Then, in the general setting, we have different possible procedures:

P1 A kernel density estimation (KDE) can be employed obtaining an evaluable function,

$$\widetilde{p}_2(\boldsymbol{\lambda}|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}K_h(\boldsymbol{\lambda} - \boldsymbol{\lambda}_2^{(n)}), \tag{8}$$

where $K_h : \mathbb{R}^{D_\lambda} \to \mathbb{R}^+$ are normalized, positive, non-linear (kernel) function, with (a) $\int K_h(\boldsymbol{\lambda})d\boldsymbol{\lambda} = 1$, (b) $\int \boldsymbol{\lambda}K_h(\boldsymbol{\lambda})d\boldsymbol{\lambda} = 0$, and where $h > 0$ represents a bandwidth. The density $\widetilde{p}_2(\boldsymbol{\lambda}|\mathbf{y})$ is normalized and evaluable for each $\boldsymbol{\lambda}$. Furthermore, h also serves as a parameter to regulate the degree of dispersion of the prior density $\widetilde{p}_2$. A kernel density estimation is a non-parametric approach. Clearly, parametric strategies can be also used for obtaining $\widetilde{p}_2$, if desired. Some example is given in Section 6.

P2 Alternatively, an IS approach can be employed:

- To each sample $\boldsymbol{\lambda}_2^{(n)}$, associate the weights

$$w_n = \ell_1(\mathbf{y}|\boldsymbol{\lambda}_2^{(n)}), \tag{9}$$

$n = 1, ..., N$.

- Resampling N times within the set $\{\boldsymbol{\lambda}_2^{(1)}, ..., \boldsymbol{\lambda}_2^{(N)}\}$, according to the normalized weights $\bar{w}_n = \frac{w_n}{\sum_{i=1}^N w_i}$, with $i = 1, ..., N$, obtaining the se of the resampled particles $\{\bar{\boldsymbol{\lambda}}_1^{(1)}, ..., \bar{\boldsymbol{\lambda}}_1^{(N)}\}$.

Therefore, we directly obtain a particle approximation of p_1, i.e.,

$$\widehat{p}_1(\boldsymbol{\lambda}|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}_1^{(n)}), \tag{10}$$

P3 Other more specific alternatives are possible if some formulas/integrals can be computed analytically, as in the applicative example in Section 5.1.

The IS approach in P2 is theoretically justified, as it is based on the idea of using $\widehat{p}_2(\boldsymbol{\lambda}|\mathbf{y})$ not only as a prior, but also as a proposal density ("deterministically" drawing N samples from this degenerated mixture [9]), i.e., the weighting function is $w = \dfrac{\ell_1(\mathbf{y}|\boldsymbol{\lambda})\cancel{\widehat{p}_2(\boldsymbol{\lambda}|\mathbf{y})}}{\cancel{\widehat{p}_2(\boldsymbol{\lambda}|\mathbf{y})}}$.

3.3 Benefits of the proposed approach and critical assessment

The described approach offers the significant advantage of incorporating an informative prior, built from the available data, which is also expected to substantially enhance the efficiency of the Monte Carlo computation. This consideration is particularly relevant when a parametric model exhibits high sensitivity to small variations in one or more parameters, as this can result in a notably concentrated posterior distribution. The inference of this kind of parameters is particularly challenging and requires informative priors to perform a proper inference. Since posterior distributions induced from non-parametric models tend to be wider than those obtained from specific parametric models, the constructed priors can be easily analyzed by Monte Carlo algorithms, enabling them to identify high-probability regions more readily. From a purist perspective, the reuse of data in this manner may be subject to criticism [16, 25]. It is worth noting, however, that well-known and widely adopted approaches such as *empirical Bayes* and *cross-validation* (CV) are subject to the same concern [3, 14, 27, 30]. Moreover, there are several strategies to minimize the impact of it on the results. Below, we describe some possible approaches.

Artificial tempering. For simplicity, let assume the framework $\boldsymbol{\lambda} = \boldsymbol{\theta}$. We can use a tempered version of the posterior $p_2(\boldsymbol{\lambda}|\mathbf{y})$, i.e.,

$$p_1(\boldsymbol{\lambda}|\mathbf{y}) \propto \ell_1(\mathbf{y}|\boldsymbol{\lambda}) \left([p_2(\boldsymbol{\lambda}|\mathbf{y})]^{\gamma}\right., \tag{11}$$

$$\propto \ell_1(\mathbf{y}|\boldsymbol{\lambda}) \left[\ell_2(\mathbf{y}|\boldsymbol{\lambda}) p_\theta(\boldsymbol{\lambda})\right]^{\gamma}, \tag{12}$$

with $0 < \gamma \leq 1$. A possible choice of γ is $\gamma = \frac{1}{M+1}$, where M is the number of data points in the vector $\mathbf{y}$. This choice comes from the usual hypothesis of conditional independent observations, which yields likelihoods consisting of products of M terms, i.e., $\ell_2(\mathbf{y}|\boldsymbol{\lambda}) = \prod_{m=1}^{M} \ell_2(y_m|\boldsymbol{\lambda})$, and $\ell_1(\mathbf{y}|\boldsymbol{\lambda}) = \prod_{m=1}^{M} \ell_1(y_m|\boldsymbol{\lambda})$. In the extreme ideal case of having M equal observations $y_1 = y_2 ... = y_M$, then $\ell_i(\mathbf{y}|\boldsymbol{\lambda}) = [\ell_i(y_1)|\boldsymbol{\lambda})]^M$, for $i = 1, 2$. Moreover, the tempering can be included directly in the approximation of the GP hyper-parameter posterior p_2, for instance, in the construction of the kDE $\widetilde{p_2}$ in Eq. (8), increasing the bandwidth h of the kernel.

Data tempering. Recall that we are assuming for simplicity $\boldsymbol{\lambda} = \boldsymbol{\theta}$. Another possibility is to consider a data-tempering approach dividing the observations into a training data subset and a test data subset, as in a standard cross-validation (CV) procedure, using $\ell_2(\mathbf{y}_{\texttt{prior}}|\boldsymbol{\lambda})$ and $\ell_1(\mathbf{y}_{\texttt{post}}|\boldsymbol{\lambda})$

with $\mathbf{y} = [\mathbf{y}_{\texttt{prior}}^\top, \mathbf{y}_{\texttt{post}}^\top]^\top$, so that

$$p_1(\lambda|\mathbf{y}) \propto \ell_1(\mathbf{y}_{\texttt{post}}|\boldsymbol{\lambda})\ell_2(\mathbf{y}_{\texttt{prior}}|\boldsymbol{\lambda})p_\theta(\boldsymbol{\lambda}). \tag{13}$$

Using set-theoretic language, note that the vectors $\mathbf{y}_{\texttt{prior}}$ and $\mathbf{y}_{\texttt{post}}$ are *disjoint*, meaning they share no common elements. If we desire to keep all the data in the likelihood function ℓ_1 (induced by the parametric model), we could just use a subset of data $\mathbf{y}_{\texttt{subset}}$ (i.e., a smaller vector) for construction the prior density,

$$p_1(\lambda|\mathbf{y}) \propto \ell_1(\mathbf{y}|\boldsymbol{\lambda})\ell_2(\mathbf{y}_{\texttt{subset}}|\boldsymbol{\lambda})p_\theta(\boldsymbol{\lambda}). \tag{14}$$

In this case, there is a partial overlap between the data in $\mathbf{y}$ and $\mathbf{y}_{\texttt{subset}}$, meaning that the latter are effectively used twice [16]. See also the interesting discussion in [25].

4 Gaussian Processes (GPs) for regression

For a non-parametric approach, we assume that $f(\mathbf{x})$ can be represented as a realization of a Gaussian Process (GP). We defined a *kernel function* $k(\mathbf{t}, \mathbf{z}|\boldsymbol{\theta}_{\texttt{ker}}) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$, where $\boldsymbol{\theta}_{\texttt{ker}}$ is a vector of hyper-parameters of the kernel function. We also build the corresponding *kernel matrix* $[\mathbf{K}]_{ij} := k\left(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}_{\texttt{ker}}\right)$ of dimension $M \times M$ containing all kernel entries. Given a generic input $\mathbf{x}$, we also define the *kernel vector* as $\mathbf{k}(\mathbf{x}) = [k\left(\mathbf{x}, \mathbf{x}_1|\boldsymbol{\theta}_{\texttt{ker}}\right), \ldots, k\left(\mathbf{x}, \mathbf{x}_M|\boldsymbol{\theta}_{\texttt{ker}}\right)]^\top$ of dimension $M \times 1$. Given the M inputs and defining the vector $\mathbf{f} = [f(x_1), ..., f(x_M)]$, the GP prior is represented by

$$p(\mathbf{f}|\boldsymbol{\theta}_{\texttt{ker}}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}). \tag{15}$$

Denoting $\mathbf{y} = [y_1, \ldots, y_M]^\top$ the $M \times 1$ vector of observed outputs and assumed observation model in (1), the likelihood function is

$$p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}_{\texttt{ker}}, \sigma_e^2) = p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_e^2\mathbf{I}_M), \tag{16}$$

where $\mathbf{I}_M$ is the $M \times M$ identity matrix, and we have set $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\texttt{ker}}, \sigma_e^2]$. Thus, for a generic input $\mathbf{x}$ (possibly not contained in the training inputs $\mathbf{x}_{1:M}$), GPs also provide a Gaussian posterior/predictive density [32, 23]

$$p(f(\mathbf{x})|\mathbf{y}, \mathbf{x}_{1:M}, \boldsymbol{\theta}) = \mathcal{N}\left(f(\mathbf{x})\middle|\mu_{\texttt{GP}}(\mathbf{x}), \sigma_{\texttt{GP}}^2(\mathbf{x})\right), \tag{17}$$

with predictive mean $\mu_{\text{GP}}(\mathbf{x})$ and variance $\sigma^2_{\text{GP}}(\mathbf{x})$. Considering a GP prior with a zero mean and kernel function $k(\mathbf{t}, \mathbf{x})$. The predictive mean gives us the interpolating function and is given by

$$\mu_{\text{GP}}(\mathbf{x}) = \widehat{f}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma_e^2 \mathbf{I}_M)^{-1} \mathbf{y}, \tag{18}$$

$$= \mathbf{k}(\mathbf{x})^\top \boldsymbol{\alpha}, \tag{19}$$

$$= \sum_{i=1}^{M} \alpha_i k(\mathbf{x}, \mathbf{x}_i | \boldsymbol{\theta}_{\text{ker}}), \tag{20}$$

where $\widehat{f}(\mathbf{x}) = \mu_{\text{GP}}(\mathbf{x})$ is an approximation of the unknown function $f(\mathbf{x})$ after observing the data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^M$, and the weight vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]^\top$ is given by $\boldsymbol{\alpha} = (\mathbf{K} + \sigma_e^2 \mathbf{I}_M)^{-1} \mathbf{y}$, where σ_e^2 is the variance of the data. The GP formulation also provides the expression for the predictive variance,

$$\sigma^2_{\text{GP}}(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma_e^2 \mathbf{I}_M)^{-1} \mathbf{k}(\mathbf{x}). \tag{21}$$

Recall that a complete notation should be $\mu_{\text{GP}}(\mathbf{x})|\boldsymbol{\theta})$ and $\sigma^2_{\text{GP}}(\mathbf{x})|\boldsymbol{\theta})$ including the dependence on the hyper-parameters $\boldsymbol{\theta}$ of the kernel functions and the power of the noise σ_e^2.

4.1 Marginal likelihood

Recalling the assumed observation model in Eq. (1), i.e., $y_i = f(\mathbf{x}_i) + \epsilon_i$, the marginal likelihood of a GP model is analytically known [32, 17, 23],

$$\begin{aligned} p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) &= \int p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}_{\text{ker}}, \sigma_e^2) p(\mathbf{f}|\boldsymbol{\theta}_{\text{ker}}) d\mathbf{f}, \\ &= \ell_2(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_e^2 \mathbf{I}_M\right), \\ &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\text{tot}}), \end{aligned} \tag{22}$$

where we have denoted $\mathbf{K}_{\text{tot}} = \mathbf{K} + \sigma_e^2 \mathbf{I}_M$. Note that $\mathbf{K}_{\text{tot}}$ depends on the choice of the hyper-parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\text{ker}}, \sigma_e^2]$. Choosing a prior density over $\boldsymbol{\theta}$, denoted as $p(\boldsymbol{\theta})$, the complete posterior density is given by

$$p_2(\boldsymbol{\theta}|\mathbf{y}) \propto \ell_2(\mathbf{y}|\boldsymbol{\theta}) p_\theta(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\text{tot}})\, p_\theta(\boldsymbol{\theta}). \tag{23}$$

The dependence on $\boldsymbol{\theta}$ in $\ell_2(\mathbf{y}|\boldsymbol{\theta})$ is due to the fact that the covariance matrix $\mathbf{K}_{\text{tot}}$ depends on $\boldsymbol{\theta}$.

Tuning of the GP hyper-parameters. A typical approach for learning $\boldsymbol{\theta}$ is maximizing the

so-called marginal likelihood $\ell_2(\mathbf{y}|\boldsymbol{\theta})$ of the GP regression model. However, this approach does only yield point estimates. Here, we are interested in a Bayesian approach in order to obtain approximations of the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ or the marginal posterior densities $p(\theta_k|\mathbf{y})$ of the hyper-parameters [32, 17, 23].

4.2 Kernel functions and interpretation of hyper-parameters

The kernel function should encode prior knowledge or belief regarding the smoothness, correlation, and periodicity of the data. Several possible kernel functions can be used. For the purpose of this work in this section, as an example, we consider two type of kernels:

$$k(\mathbf{x}, \mathbf{z}) = v^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^\alpha}{\eta}\right), \tag{24}$$

$$k(\mathbf{x}, \mathbf{z}) = v^2 \exp\left(-\frac{2\sin\left(\frac{\pi}{P}\|\mathbf{x} - \mathbf{z}\|\right)^\alpha}{\eta}\right), \tag{25}$$

where the first one has 3 hyper-parameters $\boldsymbol{\theta}_{\texttt{ker}} = [v, \eta, \alpha]$, and the second one has 4 hyper-parameters $\boldsymbol{\theta}_{\texttt{ker}} = [v, P, \eta, \alpha]$. Considering the more sophisticated kernel, the complete vector of hyper-parameters is $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\texttt{ker}}, \sigma_e] = [v, P, \eta, \alpha, \sigma_e]$. All these hyper-parameters have a clear interpretation:

- v^2 - *a-priori signal variance*, i.e., the prior variance that the random function $f(\mathbf{x})$ (latent function) has following the user's belief, without knowing any data, i.e., $v^2 = \text{var}[f(\mathbf{x})]$. In regions where there are no data points, the posterior/predictive variance tends to be v^2.

- P - *fundamental period* of the modeled phenomenon.

- η - *length-scale* of the kernel, i.e., a scale factor that controls the width of the periodic peaks. This is related to the auto-correlation of the signal $f(\mathbf{x})$. The optimal η becomes usually smaller as the number of data points grows and becoming closer and closer.

- α - *roughness*, i.e., this parameter determines the regularity/derivability and the smoothness associated to $f(\mathbf{x})$.

- σ_e^2 - *noise variance*, i.e., power of the additive perturbation variable, i.e., $\sigma_e^2 = \text{var}[\epsilon]$.

To ensure the kernel is positive definite, the following conditions are typically required: $\alpha \in (0, 2]$ (this range helps preserve positive definiteness, depending on dimensionality), and $v > 0$, $\sigma_e > 0$, $\eta > 0$, and $P > 0$. To understand the difference between v^2 and σ_e^2, we can come back to our observation model and compute the variance of each elements. For the independence, we have

$$\text{var}[y_i] = \text{var}[f(\mathbf{x}_i)] + \text{var}[\epsilon_i] = v^2 + \sigma_e^2, \tag{26}$$

since $v^2 = \text{var}[f(\mathbf{x}_i)]$ and $\sigma_e^2 = \text{var}[\epsilon_i]$. Clearly, simpler or even more complicated kernels can be employed. Some other example is provided in Eqs. (35) and (42), for instance.

5 Examples of applications

5.1 Bayesian regression with non-linear basis functions

Let assume that $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\beta}$ where $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), ..., \phi_B(\mathbf{x})]^\top$ is a vector of B non-linear functions and $\boldsymbol{\beta} = [\beta_1, ..., \beta_B]^\top$ with $B \leq M$. We also denote the $M \times B$ design matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1)..., \boldsymbol{\phi}(\mathbf{x}_M)]^\top$. The observation model is then

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{27}$$

It is common to place a (conjugate) Gaussian prior over $\boldsymbol{\beta}$, i.e.,

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\Sigma}_\beta), \tag{28}$$

with zero mean and the $B \times B$ covariance matrix $\boldsymbol{\Sigma}_\beta$. We recall that the noise vector is an independent Gaussian variable as well, more specifically, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_e^2 \mathbf{I}_M)$. In this framework, the posteriors $p(\boldsymbol{\beta}|\mathbf{y}, \sigma_e, \boldsymbol{\Sigma}_\beta)$ and $p(f(\mathbf{x})|\mathbf{y}, \sigma_e, \boldsymbol{\Sigma}_\beta)$ are both Gaussian, analytically known with mean

$$\widehat{\boldsymbol{\beta}} = \frac{1}{\sigma_e^2}\left(\frac{1}{\sigma_e^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}_\beta\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}, \tag{29}$$

and $\widehat{f}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \widehat{\boldsymbol{\beta}}$, respectively (for more details see [23, 32]). Hence, rhe regression function $\widehat{f}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \widehat{\boldsymbol{\beta}}$ is completely specified obtaining the estimator of the coefficients, but we need to choose or estimate the hyper-parameters are the matrix $\boldsymbol{\Sigma}_\beta$ and σ_e^2. Setting $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_M)]^\top$, note that

$$\text{var}[\mathbf{f}] = \boldsymbol{\Sigma}_f = \boldsymbol{\Phi}\boldsymbol{\Sigma}_\beta\boldsymbol{\Phi}^\top. \tag{30}$$

Note that a prior over $\boldsymbol{\beta}$ induces a prior over $\mathbf{f}$, that is $p(\mathbf{f}) \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, \boldsymbol{\Phi}\boldsymbol{\Sigma}_\beta\boldsymbol{\Phi}^\top)$ [23]. The two priors are related, and we can also fix $p(\mathbf{f})$ and find the corresponding $p(\boldsymbol{\beta})$. We can obtain an approximation of $\boldsymbol{\Sigma}_\beta$ by the transformation

$$\boldsymbol{\Sigma}_\beta \approx \boldsymbol{\Phi}_{\texttt{p-inv}}\boldsymbol{\Sigma}_f\boldsymbol{\Phi}_{\texttt{p-inv}}^\top, \tag{31}$$

where $\boldsymbol{\Phi}_{\texttt{p-inv}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top$ is the pseudo-inverse matrix of $\boldsymbol{\Phi}$. If we assume that $\boldsymbol{\Sigma}_f = v^2\mathbf{I}_B$, so that

$$\boldsymbol{\Sigma}_b \approx v^2\,\boldsymbol{\Phi}_{\texttt{p-inv}}^\top\boldsymbol{\Phi}_{\texttt{p-inv}}. \tag{32}$$

Therefore, building a prior over $p(v)$ (using the proposed strategy) corresponds to construct a prior over $\boldsymbol{\Sigma}_\beta$ by Eq. (32). Then, the parameters of the model are now $\lambda = [v, \sigma_e]$.

Hierarchical approach. We can consider hierarchically a prior $p(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda}$, and then approximate

$$p(\boldsymbol{\beta}|\mathbf{y}) = \int p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda})p(\boldsymbol{\lambda})d\boldsymbol{\lambda} \approx \frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda}_2^{(n)}), \quad \boldsymbol{\lambda}_2^{(n)} \sim p(\boldsymbol{\lambda}). \tag{33}$$

To automatically build the prior $p(\boldsymbol{\lambda})$, we can use a GP over $f(\mathbf{x})$ with kernel

$$k(\mathbf{x}, \mathbf{z}) = v^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{z}||^2}{\eta}\right), \tag{34}$$

and learn the posterior of the hyper-parameters $\boldsymbol{\theta} = [v, \sigma_e, \eta]$. Note that $\boldsymbol{\theta} = [\boldsymbol{\lambda}, \eta]$, i.e., the first two components of $\boldsymbol{\theta}$ coincides with the components of $\boldsymbol{\lambda}$. Applying a Monte Carlo sampling method (e.g., MCMC or an IS+R techniques), we can draw samples $\boldsymbol{\theta}^{(n)}$ from $p_2(\boldsymbol{\theta}|\mathbf{y})$ and, then, marginalizing out over η (just do not considering the last component, of the samples regarding η as described in Section 3.2). Finally, we have a particle approximation $\widehat{p}_2(\boldsymbol{\lambda}|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}_2^{(n)})$, of the bi-dimensional marginal posterior $p_2(\boldsymbol{\lambda}|\mathbf{y})$, where $\boldsymbol{\lambda}_2^{(n)} = [v^{(n)}, \sigma_e^{(n)}]$ for $n = 1, ..., N$. Note that for each $v^{(n)}$ we have a matrix $\boldsymbol{\Sigma}_b^{(n)}$ given in (32), and considering also $\sigma_e^{(n)}$, we have N vectors $\widehat{\boldsymbol{\beta}}^{(n)}$ in (29), which are the mean of the Gaussian posteriors $p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda}_2^{(n)})$ into (33) (the covariance matrices are also given and known [23]).

5.2 Parametric models involving periodicity

Several applications involving the inference about periodicity and/or parameters in frequency domain (specially in astronomy; see Section 6 for some example). In this cases, we are interested

in studying the marginal posterior of P and σ_e given the data $\mathbf{y}$ (hence, $\boldsymbol{\lambda} = [P, \sigma_e]$). In this case, P has the physical meaning of *fundamental period* of the signal.

Remark. For periodic or cyclic data, the posterior distributions of P are often highly concentrated within a narrow region of the parameter space. Consequently, the use of informative priors becomes essential to mitigate dimensionality issues and computational challenges in the inference of model parameters.

Remark. Compared to the use of Fourier series or the periodogram (which reveals the frequencies present in the signal/data) the proposed approach enables the construction of a complete prior density taking into account these frequencies. With *complete prior density*, we refer to a full characterization of the function, including its location, overall variance, uncertainty around the peaks, symmetry (skewness), tail behavior (kurtosis), and other shape-related features. Furthermore, the idea proposed in this work can be readily integrated with the preliminary use of the periodogram in several ways. For instance, the information obtained from a periodogram can be employed as a suitable initialization for the proposal densities used for sampling from $p_2(\boldsymbol{\theta}|\mathbf{y})$.
The marginal posterior of P will present peaks/modes around the frequencies contained in the signal. The main mode will correspond to the main period. However, we could also generalize the kernel given in Eq. (25) using, for example,

$$k(\mathbf{x}, \mathbf{z}) = \sum_{i}^{R} v_i^2 \exp\left(-\frac{2\sin\left(\frac{\pi}{P_i}\|\mathbf{x} - \mathbf{z}\|\right)^2}{\eta_i} \right), \tag{35}$$

which is the sum of several kernels with different hyperparameters (for instance, different periods P_i) to improve the ability of the GP posterior to capture different frequencies in the signal.

6 Numerical experiments

In this section, we test the proposed approach in two experiments with synthetic data. Furthermore, we have applied the methodology to real data from two different sources and typologies. For the first case, we chose the radial velocity data from the exoplanet host star GJ 3512 [24]. For the second case, we chose photometric data from the All-Sky Automated Survey [ASAS, 31]. We chose a noisy light curve of the binary star HD 112661. With these two examples, we checked the behaviour of the method with two common scenarios: (a) extremely concentrated posterior distributions and (b) observations with large uncertainties.

6.1 First Experiment

We consider a non-linear basis model

$$y_i = \boldsymbol{\phi}(x_i)^\top \boldsymbol{\beta} + \epsilon_i, \tag{36}$$

$$= \beta_1 \exp\left(-|x_i - 2|\right) + \beta_2 \exp\left(-\frac{1}{2}|x_i - 7|\right) + \epsilon_i, \tag{37}$$

where $\boldsymbol{\phi}(x_i) = [\exp\left(-|x_i - 2|\right), \exp\left(-|x_i - 7|\right)]^\top$, $\boldsymbol{\beta} = [\beta_1, \beta_2]^\top$ and $\epsilon_i \sim \mathcal{N}(\epsilon|0, \sigma_e^2)$. We generate $M = 34$ observations y_i, $i = 1, ..., 34$, according to the model above with $\sigma_e = 0.3$. The M inputs x_i are obtained a uniform grid from 0 to 10 with step 0.3. The data vector is then $\mathbf{y} = [y_1, ..., y_{34}]^\top$. We desire to perfom Bayesian inference on $\boldsymbol{\beta} = [\beta_1, \beta_2]^\top$ and σ_e^2. The idea is to build automatically a data-driven prior density over σ_e^2. Moreover, with respect to $\boldsymbol{\beta}$ we consider the conjugate prior density,

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\Sigma}_\beta), \tag{38}$$

where

$$\boldsymbol{\Sigma}_\beta = \begin{bmatrix} \sigma_{\beta,1}^2 & \rho_\beta \\ \rho_\beta & \sigma_{\beta,2}^2 \end{bmatrix}, \quad [\boldsymbol{\Sigma}_\beta]_{i,i} = \sigma_{\beta,i}^2, \quad [\boldsymbol{\Sigma}_\beta]_{1,2} = [\boldsymbol{\Sigma}_\beta]_{2,1} = \rho_\beta,$$

is a 2×2 covariance matrix. We desire to construct an automatic prior for the matrix $\boldsymbol{\Sigma}_\beta$, i.e., over the three scalar variables $\sigma_{\beta,1}^2$, $\sigma_{\beta,2}^2$, and ρ_β. For this purpose, we apply a GP with the kernel in Eq. (34), i.e. $k(x, z) = v^2 \exp\left(-\frac{\|x-z\|^2}{\eta}\right)$. Hence, the hyper-parameters of the GP are $\boldsymbol{\theta} = [v^2, \sigma_e, \eta]$. For simplicity, we consider uniform priors over the elements of $\boldsymbol{\theta}$, i.e., $\mathcal{U}([0, \text{var}(y_i)])$ for v^2 and $\mathcal{U}([0, \text{std}(y_i)])$ for σ_e and $\mathcal{U}([0, 10])$ for η. We consider the empirical variance for computing $\text{var}(y_i)$ and $\text{std}(y_i)$.

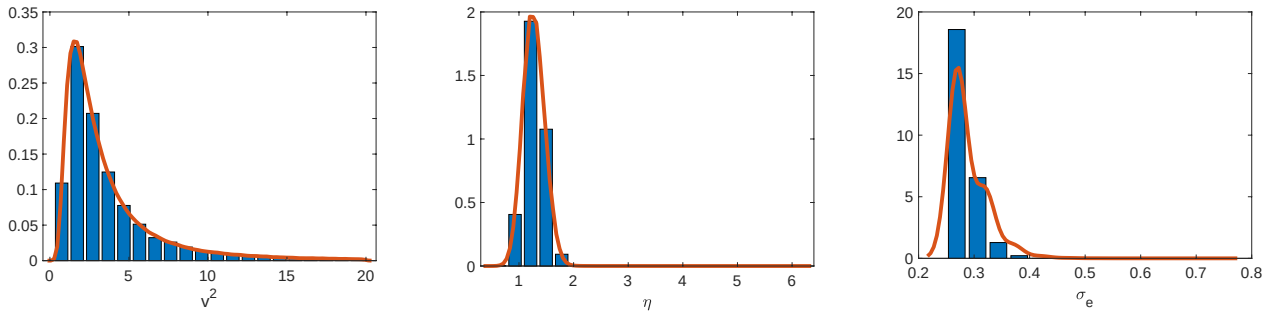
We use an importance sampler with $N = 10^6$ particles (using the prior over $\boldsymbol{\theta}$ as proposal density), for drawing samples from the posterior $p_2(\boldsymbol{\theta}|\mathbf{y})$. The approximations of the marginal posteriors of the components of $\boldsymbol{\theta} = [v^2, \sigma_e, \eta]$ are depicted in Figure 1. Note that the marginal posterior $p_2(\sigma_e|\mathbf{y})$ (or a more diffuse, spread-out versions) can be used as an informative prior over σ_e. Some bidimensional conditional posteriors are also shown in Fig. 2.

Furthermore, given the design matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}(x_1), ..., \boldsymbol{\phi}(x_{34})]^\top$ and the expressions (30)-(31)-(32), we can obtain samples of the matrix $\boldsymbol{\Sigma}_b \approx v^2 \, \boldsymbol{\Phi}_{\texttt{p-inv}}^\top \boldsymbol{\Phi}_{\texttt{p-inv}}$, having samples of v^2. Thus, the approximations of the marginal densities of the elements into $\boldsymbol{\Sigma}_b$, i.e., $\sigma_{\beta,1}^2$, $\sigma_{\beta,2}^2$, and ρ_β are depicted in Fig. 3. Clearly, at the same time, we have also obtained the approximation of the

joint density. This joint distribution over the matrix $\mathbf{\Sigma}_b$ must be employed as prior to ensure the matrix to be definite positive. The mean $\widehat{\mathbf{\Sigma}}_b$ and a credible interval (CI) of 90% are given below:

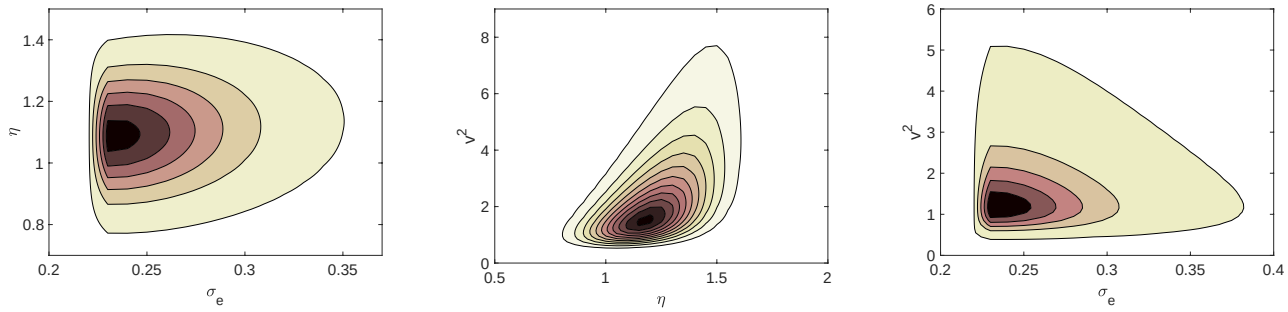
$$\widehat{\mathbf{\Sigma}}_\beta = \begin{bmatrix} 1.14 & -0.12 \\ -0.12 & 0.57 \end{bmatrix}, \quad \text{CI} = \begin{bmatrix} [0.29, 3.20] & [-0.031, -0.34] \\ [-0.031, -0.34] & [0.15, 1.6] \end{bmatrix}.$$

If desired, a more diffuse prior over $\mathbf{\Sigma}_\beta$ can be obtained by adopting a more dispersed density for v^2. This can be forced, for example, by employing a larger bandwidth in the KDE estimation (or, in some cases, by applying a more diffuse prior over $\boldsymbol{\theta}$).



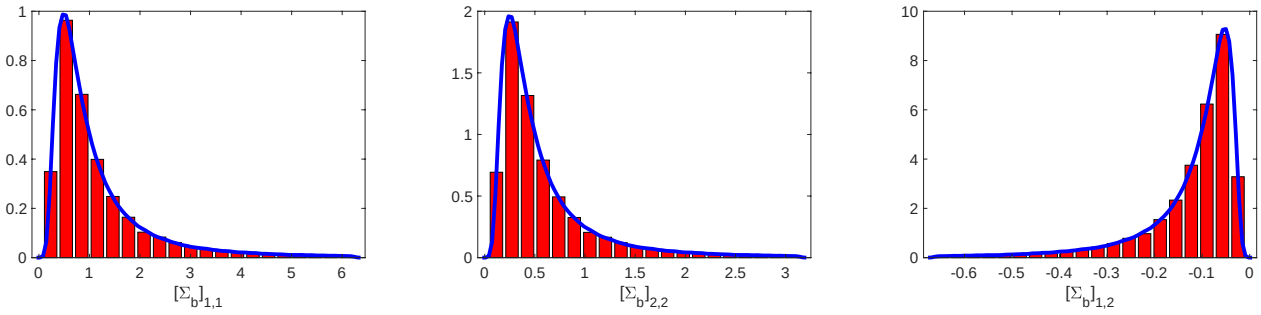
(a) (b) (c)

Figure 1: Approximation of the marginal posteriors of the hyper-parameters of the employed GP: **(a)** of v^2, **(b)** of η, and **(c)** of σ_e. Note that the marginal posterior $p_2(\sigma_e|\mathbf{y})$ (or more diffuse, spread-out versions) can be used as a prior over σ_e.



(a) $p_2(\sigma_e, \eta|\mathbf{y}, v^2 = 1)$. (b) $p_2(\eta, v^2|\mathbf{y}, \sigma_e = 0.3)$. (c) $p_2(\sigma_e, v^2|\mathbf{y}, \eta = 1)$.

Figure 2: Conditional bidimensional posteriors, **(a)** $p_2(\eta, \sigma_e|\mathbf{y}, v^2 = 1)$, **(b)** $p_2(\eta, v^2|\mathbf{y}, \sigma_e = 0.3)$, and **(c)** $p_2(\sigma_e, v^2|\mathbf{y}, \eta = 1)$.



(a) $\sigma^2_{\beta,1}$ (b) $\sigma^2_{\beta,2}$ (c) ρ_β

Figure 3: Marginal densities over the elements of the matrix $\mathbf{\Sigma}_b$. The joint distribution over the matrix $\mathbf{\Sigma}_b$ must be employed to ensure to be definite positive.

6.2 Second Experiment

In this section, we consider the following model,

$$y_i = b + A \sin\left(\frac{2\pi}{P} x_i + \phi\right) + e_i, \tag{39}$$

where again e_i is a normally distributed random variable with standard deviation σ_e and zero mean, i.e. $e_i \sim \mathcal{N}(0, \sigma_e^2)$. We generate $M = 15$ observations y_i according to the model above (hence $\mathbf{y} = [y_1, ..., y_{15}]^\top$), setting $\sigma_e = 0.5$, $P = 2\pi \approx 6.28$, $A = 1$, $\phi = 0$, and $b = 0$. We randomly selected $M = 15$ points x_i, uniformly in the interval $x \in [-4, 4]$.

6.2.1 Prior construction

In this experiment, we focus on the period parameter P. Our goal is to construct an informative prior distribution for P making inference on the GP hyper-parameters $\boldsymbol{\theta}_{\mathtt{ker}} = [v^2, P, \eta]$ of a periodic kernel in Eq. (25). The complete vector of hyper-parameters would be $\boldsymbol{\theta} = [v^2, P, \eta, \sigma_e]$. However, for simplicity, in this example, σ_e is kept fixed to the true value so that $\boldsymbol{\theta} = [v^2, P, \eta]$. The posterior distributions of the period P for periodic data are often concentrated in a small region of the parameter space, so there is a need to use informative priors to avoid dimensionality issues and computational problems when inferring model parameters. An adaptive importance sampling (AIS) scheme [7] has been employed to sample from the posterior of the GP hyper-parameters $p_2(\boldsymbol{\theta}|\mathbf{y})$, with uniform priors $\mathcal{U}([0, 100])$ over the elements of $\boldsymbol{\theta}$. Gaussian proposal density is adopted for all components of $\boldsymbol{\theta}$, with its parameters being adaptively updated throughout the iterations of the AIS scheme. The AIS algorithm provides weighted samples: after a

resampling step, we obtain the approximations by histograms of the marginal posteriors of v^2, η and P, shown in Fig. 4(a). The continuous, black line denotes the mean of the samples. The dashed lines are the 90% credibility intervals. Finally, the continuous red lines represent the maxima a posteriori of the marginal posteriors: 0.42 for v^2, 0.84 for η and 5.47 for P, respectively. The marginal distribution of the hyper-parameter P shown in Fig. 4(a) (bottom, right panel) can be used as the prior for the period of the generic sine curve in Eq. (39). We employ the procedure P1 in Section 3.2, we apply a KDE with Gaussian kernel with the optimal bandwidth suggested in [4]. The resulting approximated density, $\widetilde{p_2}$, is depicted in Figure 4(b). A more vague prior over P can be obtained by employing a larger bandwidth in the KDE estimation.

6.2.2 Bayesian inference for the parametric model

For the sake of simplicity, we consider $\sigma_e = 0.5$ fixed, and we assume uniform priors $\mathcal{U}([0, 100])$ over the parameters A, ϕ, and b. For the period P, we consider the constructed prior in Figure 4(b). We run the AIS with $N = 10^5$ and 10^3 iterations, for approximating the posterior $p_1(\boldsymbol{\lambda}|\mathbf{y})$ where $\boldsymbol{\lambda} = [A, \phi, b, P]$. Figure 5 shows the particle approximations of the marginal posterior densities for each model parameter. Note that the posterior distribution of the period P is more concentrated, compared to that used as the prior. All the marginal posteriors are located around the true values of the parameters A, ϕ, b and P.

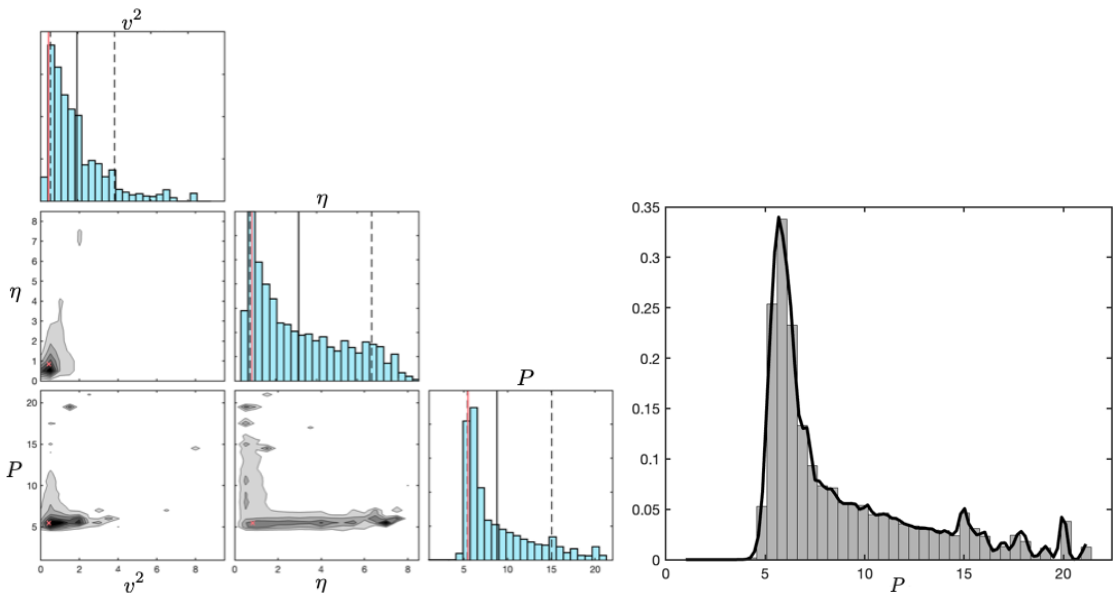
6.3 Radial velocity model with real data

.

For this example, we consider the radial velocity data from [24]. The data correspond to radial velocity measurements of the star GJ 3512, which hosts an exoplanetary system [5, 11]. They were acquired with the CARMENES instrument, a high resolution, optical and near-infrared spectrograph mounted on the 3.5m telescope at the Calar Alto Observatory [34]. The 158 observations expand for 900 days. In the present work, only optical data are utilised, without any loss of generality. The mean value for RV uncertainties is $\sim 2\ \mathrm{m\,s^{-1}}$ for the optical channel of the instrument.

Radial velocity (RV) model. The observations can be succinctly modeled as

$$y_i = v_i + e_i, \tag{40}$$



(a) (b)

Figure 4: **(a)** Approximations of the marginal posterior distribution for the GP hyper-parameters v^2, η and P. The mean and the 90% credibility intervals are depicted with continuous black line and dashed lines, respectively. The red continuous line represents the estimated maximum a posteriori (MAP). **(b)** KDE approximation $\widetilde{p}_2$ (continuous line) for the marginal posterior of the hyperparameter P (histogram).

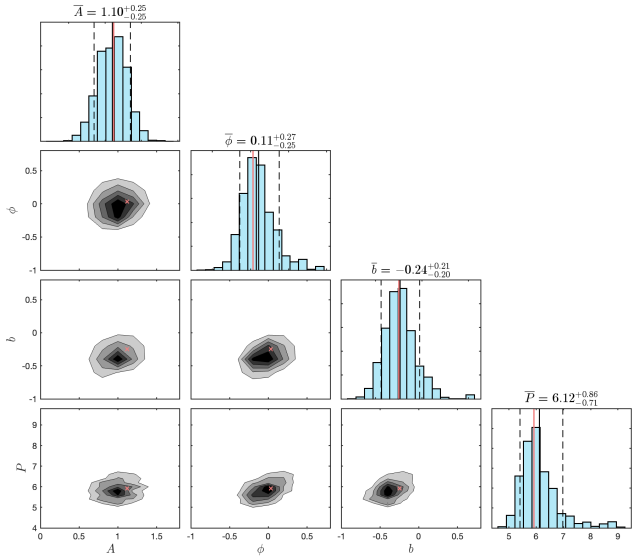


Figure 5: Approximations of the marginal posteriors for the parameters A, ϕ, b and P of the parametric model in Eq. (39), using the constructed prior in Figure 4(b) for P. The value over each histogram is the mean and the 90% credibility level (continuous black line and dashed lines, respectively). The continuous red line represents the estimated MAP.

where y_i is the observation of the star's radial velocity at time i, which combines the intrinsic velocity component v_i and an additive noise term ξ_i. The parametric model is defined by Eqs. 40 and

$$v_i = V_0 + K\left[\cos\left(u_i + \omega\right) + e\cos\left(\omega\right)\right], \quad i = 1, \ldots, M. \tag{41}$$

where M is the number of observations, V_0 is the mean radial velocity of the star, K is the amplitude of the curve, u is the true anomaly, ω is the periastron angle and e is the orbit eccentricity. It is noteworthy that a period parameter P is encompassed within the parameter u in Eq. (41). See [18], for a detailed definition of the parametric model. For the standard deviation of the random variable e_i, we used $\sigma_e = 3.33\ \mathrm{m\,s^{-1}}$. This value was determined in [18].

Prior construction. The kernel we used for this application is a composition of the periodic kernel from Sect. 6.2 and a squared exponential kernel,

$$k(x, z|\boldsymbol{\theta}_{\mathtt{ker}}) = \quad v_1^2 \exp\left(-\frac{2\sin((\pi/P)\|x-z\|)^2}{\eta_1^2}\right) + v_2^2 \exp\left(\frac{(\|x-z\|)^2}{2\eta_2^2}\right), \tag{42}$$

where $\boldsymbol{\theta}_{\mathtt{ker}} = [v_1, v_2, \eta_1, \eta_2, P]$. For simplicity, again we keep fixed σ_e (out of the inference), so that we can write $\boldsymbol{\theta}_{\mathtt{ker}} = \boldsymbol{\theta}$. Considering a product uniform priors from 0 to 100 (one for each hyper-parameter) for building $p_\theta(\boldsymbol{\theta})$, we apply an AIS scheme for drawing from $p_2(\boldsymbol{\theta}|\mathbf{y})$ [22, 19]. Figure 6 shows the results of the GP regression considering the MAP estimation $\widehat{\boldsymbol{\theta}}_{MAP}$, and the full Bayesian solution obtained considering all the weighted particles generated by the AIS algorithm. Generally, the mean regression function are very similar, but the variance of the full Bayesian solution is larger in some regions. The approximation the the marginal posterior $p_2(P|\mathbf{y})$ is given in Figure 7(a). Using the generated samples, we also fit a Laplace density for obtaining $\widetilde{p}_2$ (i.e., we use a parametric P1 strategy, described in Section 3.2). We use this Laplace density as prior for the subsequent inference of the orbital period of the planet within the parametric model.

Inference of the RV model. We use the Laplace approximation of the marginal posterior $p_2(P|\mathbf{y})$ in Figure 7(a), as prior for the period P in the RV model in Eqs. (40)-(41). The rest of priors are set as in [18]. We apply again the AIS scheme for sampling from $p_1(\boldsymbol{\lambda}|\mathbf{y})$ [22, 19]. Figure 7(b) shows the approximated marginal posterior distribution of P (using the aforementioned Laplacian-derived prior). The mean of the posterior distribution is 204.2 days, which can be compared with 204.5 days obtained in [18] for a single planet model and 203.6 days obtained

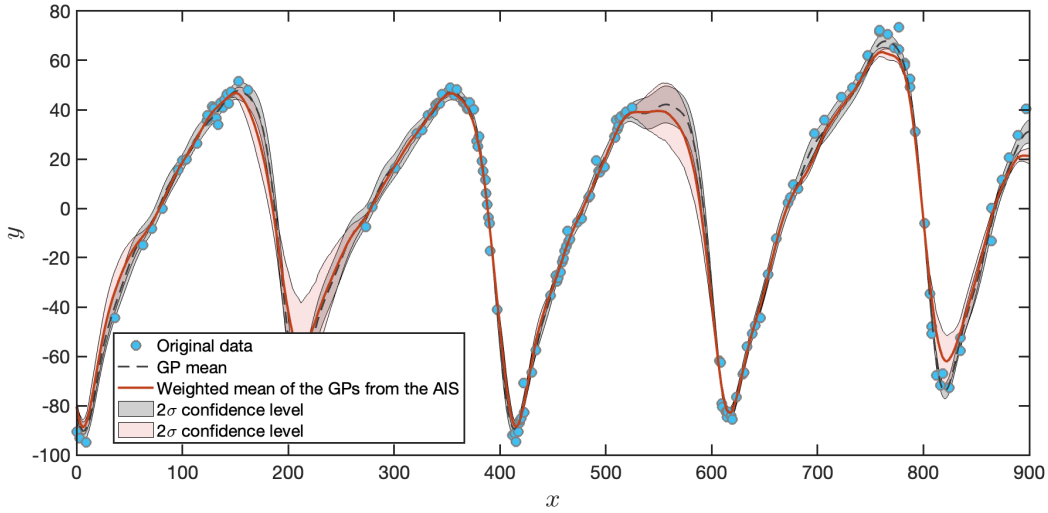


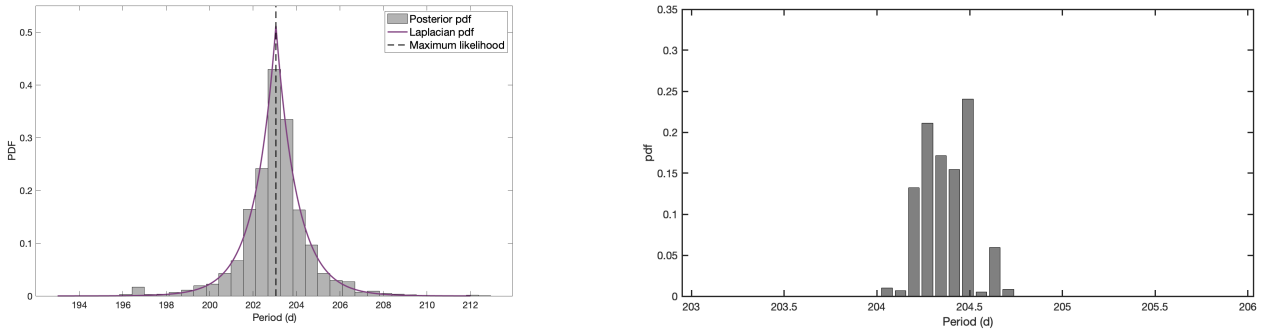
Figure 6: GP regression obtained by using $\widehat{\boldsymbol{\theta}}_{MAP}$ in dashed line and the full Bayesian GP solution (considering all the weighted particles from the AIS) with solid red line [23]. We show also two uncertainty intervals, one corresponding to the GP with $\widehat{\boldsymbol{\theta}}_{MAP}$ (grey color) and the other one corresponding to the full Bayesian solution (pink color) [23].The blue points represent the real data from [24].

in [24]. The histogram represents the posterior density of this parameter, which is concentrated within a narrow interval. In contrast to the findings reported in [18], our approach enabled an accurate approximation of the marginal posterior for the orbital period (computationally speaking, we obtain more different particles with higher weights). This outcome remarks the importance of constraining the prior for the parameter P, as such constraints contribute to a more precise estimation of its posterior distribution.

6.4 Fitting a photometric light curve with real data

In this section, we analyze data from the ASAS source id. 125922-6217.3. This source is the counterpart of HD 112661, an evolved B star classified as B0/1 III/IV by [13] and B2 IVn by [28]. The observations expand during approximately nine years for a total of $M = 600$ data points. The uncertainties in the magnitude are of the same order of the amplitude of the low amplitude variations in the light curve.

We consider the same model in Eq. (39) and the procedure described in Sect. 6.2, i.e., the model



(a) (b)

Figure 7: **(a)** Approximated marginal posterior $p_2(P|\mathbf{y})$ of the GP hyper-parameter P in Eq. (42). We also obtain a parametric approximation $\widetilde{p}_2$ fitting a Laplace density. **(b)** Marginal posterior distribution of the period in the model obtained from Bayesian inference using the distribution in the left panel as a prior.

is

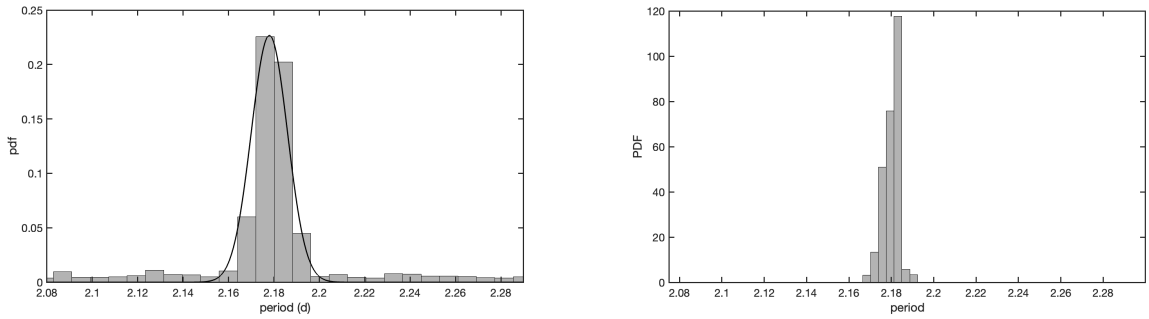
$$y_i = b + A \sin\left(\frac{2\pi}{P} x_i + \phi\right) + e_i. \tag{43}$$

$i = 1, ..., M$. First, a GP with a periodic kernel as in Eq. (25) was used to construct the informative prior over the period parameter in the parametric model. This prior distribution was then used for fitting the model in Eq. (43) to the data. The priors over the elements $\boldsymbol{\theta}$ are uniform from 0 to 200. Figure 8(a) depicts the approximation of marginal posterior distribution of the GP kernel hyper-parameter P, obtained by an AIS scheme. We employ the strategy P1, considering a parametric estimation of the density fitting a normal distribution (shown in solid line). We used this normal distribution as the prior of the period of the parametric model. Figure 8(b) shows the marginal posterior approximated for the parameter P for the light curve model. As in the previous example, the support of the marginal posterior distribution for the period in the parametric model is more concentrated than that of the same kernel hyper-parameter.

The fitted curve is shown in Fig. 9. The data are plotted in phase for clarity. The estimation for the model parameters is given in Table 1, joinlty with the 90% credibility levels.

Table 1: Estimated model parameters for the light curve of HD 112661.

A	ϕ	b	P
$0.03^{+0.01}_{-0.01}$	$-9.51^{+0.27}_{-0.38}$	$9.23^{+0.01}_{-0.01}$	$2.18^{+0.01}_{-0.01}$



(a) (b)

Figure 8: **(a)** Approximated marginal posterior of the hyper-parameter P of the GP regression for HD-112661. **(b)** Approximated marginal posterior of the period P in the photometric light curve model.

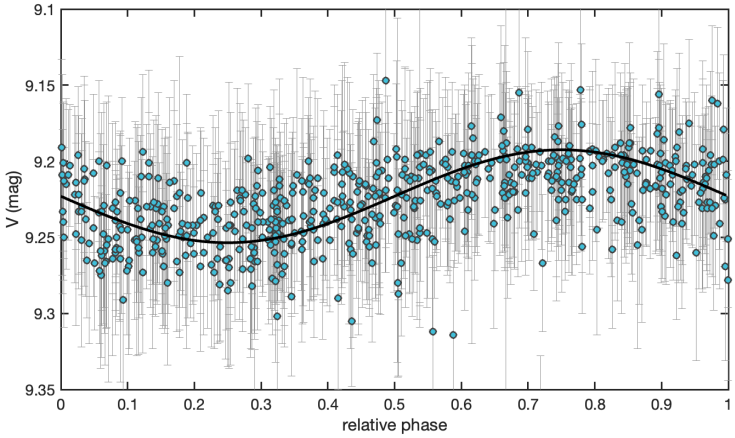


Figure 9: Resulting curve fitted for the ASAS data of HD 112661. The light curve is plotted in phase, for clarity.

7 Conclusions

In this work, we have described a novel methodology to build data-driven prior densities. The proposed approach constructs automatically informative priors that are particularly necessary for parameters highly concentrated posterior distributions. Indeed, in these cases, the use of informative priors becomes essential to mitigate dimensionality issues and computational challenges in the inference of model parameters. Namely, it helps substantially the computational inference algorithm employed to approximate the posterior, correctly exploring the parameter space. Furthermore, the constructed prior is particularly well-suited for model selection purposes (being

both informative, since data-driven, and proper). The novel procedure use posteriors over the hyper-parameters of a GPs as priors for parameters in parametric physical models. We have provided several numerical examples, including two astrophysical experiments involving real data, showing the benefit of the proposed approach.

References

[1] K. M. Banner, K. M Irvine, and T. J. Rodhouse. The use of Bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889, 2020.

[2] J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith. Highly informative priors. In *Bayesian Statistics 2*, pages 329–360. Citeseer, 1985.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[4] A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

[5] G. L. Bretthorst. Generalizing the lomb-scargle periodogram. In *AIP Conference Proceedings*, volume 568, pages 241–245. American Institute of Physics, 2001.

[6] Johannes Buchner. Ultranest: Pythonic nested sampling development framework and ultranest. *Astrophysics Source Code Library*, pages ascl–1611, 2016.

[7] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

[8] B. Clarke. Implications of reference priors for prior information and for sample size. *Journal of the American Statistical Association*, 91(433):173–184, 1996.

[9] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized Multiple Importance Sampling. *Statistical Science*, 34(1):129 – 155, 2019.

[10] H. Finch and J. Miller. The use of incorrect informative priors in the estimation of mimic model parameters with small sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, 26:1–12, 01 2019.

[11] Philip C Gregory and Debra A Fischer. A bayesian periodogram finds evidence for three planets in 47 ursae majoris. *Monthly Notices of the Royal Astronomical Society*, 403(2):731–747, 2010.

[12] M. Holden, M. Pereyra, and K. C. Zygalakis. Bayesian imaging with data-driven priors encoded by neural networks. *SIAM Journal on Imaging Sciences*, 15(2):892–924, 2022.

[13] N. Houk and A. P. Cowley. *University of Michigan Catalogue of two-dimensional spectral types for the HD stars. Volume I. Declinations -90_ to -53_f0.* 1975.

[14] I. Klebanov, A. Sikorski, C. Schutte, and S. Roblitz. Objective priors in the empirical bayes framework. *arXiv:1612.00064*, 2020.

[15] P. Lenk and B. Orme. The value of informative priors in bayesian inference with sparse data. *Journal of Marketing Research*, 46(6):832–845, 2009.

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

[17] 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.

[18] J. Lopez-Santiago, L. Martino, J. Míguez, and M. A. Vázquez. A likely magnetic activity cycle for the exoplanet host m dwarf gj 3512. *The Astronomical Journal*, 160(6):273, 2020.

[19] J. Lpez-Santiago, L. Martino, M. A. Vzquez, and J. Miguez. A Bayesian inference and model selection algorithm with an optimization scheme to infer the model noise power. *Monthly Notices of the Royal Astronomical Society*, 507(3):3351–3361, 08 2021.

[20] R. Martin and S. G. Walker. Data-driven priors and their posterior concentration rates. *Electronic Journal of Statistics*, 13(2):3049 – 3081, 2019.

[21] L. Martino and F. Llorente. A note on the area under the likelihood and the fake evidence for model selection. *Computational Statistics*, pages 1–24, 2025.

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

[23] 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, 2021.

[24] J. C. Morales, A. J. Mustill, I. Ribas, M. B. Davies, A. Reiners, F. F. Bauer, D. Kossakowski, E. Herrero, E. Rodríguez, M. J. López-González, et al. A giant exoplanet orbiting a very-low-mass star challenges planet formation models. *Science*, 365(6460):1441–1445, 2019.

[25] G. E. Moran, D. M. Blei, and R. Ranganath. Holdout predictive checks for Bayesian model criticism. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):194–214, 09 2023.

[26] Satoshi Morita, Peter F Thall, and Peter Müller. Evaluating the impact of prior assumptions in bayesian biostatistics. *Statistics in biosciences*, 2:1–17, 2010.

[27] S. Nabi, H. Nassif, J. Hong, H. Mamani, and G. Imbens. Bayesian meta-prior learning using empirical Bayes. *Management Science*, 68(3):1737–1755, 2022.

[28] M. Pantaleoni González, J. Maíz Apellániz, R. H. Barbá, and B. C. Reed. The alma catalogue of ob stars–ii. a cross-match with gaia dr2 and an updated map of the solar neighbourhood. *Monthly Notices of the Royal Astronomical Society*, 504(2):2968–2982, 2021.

[29] C. Pedroza, W. Han, V. T. Thanh, C. Green, and J. E. Tyson. Performance of informative priors skeptical of large treatment effects in clinical trials: A simulation study. *Statistical Methods in Medical Research*, 27(1):79–96, 2018.

[30] S. PETRONE, J. ROUSSEAU, and C. SCRICCIOLO. Bayes and empirical Bayes: do they merge? *Biometrika*, 101(2):285–302, 2014.

[31] G. Pojmanski and G. Maciejewski. The all sky automated survey. the catalog of variable stars. iii. 12h-18h quarter of the southern hemisphere. *arXiv preprint astro-ph/0406256*, 2004.

[32] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[33] M. Ravasi, J. Romero, M. Corrales, N. Luiken, and C. Birnie. Striking a balance: Seismic inversion with model-and data-driven priors. In *Developments in structural geology and tectonics*, volume 6, pages 153–200. Elsevier, 2024.

[34] S. Stock, E. Nagel, J. Kemmer, V. M. Passegger, S. Reffert, A. Quirrenbach, J. A. Caballero, S. Czesla, V. J. S. Béjar, C. Cardona, et al. The carmenes search for exoplanets around m dwarfs-three temperate-to-warm super-earths. *Astronomy & Astrophysics*, 643:A112, 2020.