Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **503,** 484–497 (2021)
Advance Access publication 2021 February 5

doi:10.1093/mnras/stab320

Informative Bayesian model selection for RR Lyrae star classifiers

F. Pérez-Galarce ,[1]* K. Pichara,[1,2] P. Huijse,[2,3] M. Catelan [2,4,5] and D. Mery[1]

[1] *Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile, 7820436 Santiago, Chile*
[2] *Millennium Institute of Astrophysics, Av. Vicuna Mackenna 4860, 782-0436 Macul, Santiago, Chile*
[3] *Instituto de Informática, Universidad Austral de Chile, Valdivia, Chile*
[4] *Instituto de Astrofísica, Facultad de Física, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 7820436 Macul, Santiago, Chile*
[5] *Centro de Astroingeniería, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 7820436 Macul, Santiago, Chile*

Accepted 2021 January 29. Received 2021 January 28; in original form 2020 November 26

ABSTRACT

Machine learning has achieved an important role in the automatic classification of variable stars, and several classifiers have been proposed over the last decade. These classifiers have achieved impressive performance in several astronomical catalogues. However, some scientific articles have also shown that the training data therein contain multiple sources of bias. Hence, the performance of those classifiers on objects not belonging to the training data is uncertain, potentially resulting in the selection of incorrect models. Besides, it gives rise to the deployment of misleading classifiers. An example of the latter is the creation of open-source labelled catalogues with biased predictions. In this paper, we develop a method based on an informative marginal likelihood to evaluate variable star classifiers. We collect deterministic rules that are based on physical descriptors of RR Lyrae stars, and then, to mitigate the biases, we introduce those rules into the marginal likelihood estimation. We perform experiments with a set of Bayesian logistic regressions, which are trained to classify RR Lyraes, and we found that our method outperforms traditional non-informative cross-validation strategies, even when penalized models are assessed. Our methodology provides a more rigorous alternative to assess machine learning models using astronomical knowledge. From this approach, applications to other classes of variable stars and algorithmic improvements can be developed.

Key words: methods: data analysis – astronomical data bases: miscellaneous – software: data analysis – stars: variables: RR Lyrae.

1 INTRODUCTION

Machine learning has been applied intensively to the classification of variable stars in recent decades (Debosscher et al. 2007, 2009; Richards et al. 2011a; Pichara et al. 2012; Nun et al. 2015; Kim & Bailer-Jones 2016; Mackenzie, Pichara & Protopapas 2016; Benavente, Protopapas & Pichara 2017; Valenzuela & Pichara 2017; Narayan et al. 2018; Naul et al. 2018; Aguirre, Pichara & Becker 2019; Carrasco-Davis et al. 2019; Becker et al. 2020). Variable stars are considered crucial celestial objects, mainly because several of them (e.g. RR Lyrae and Cepheids) are reliable distance indicators, thus providing us with a gauge to measure topics ranging from Galactic structure to the overall expansion of the Universe.

From a machine learning perspective, the major efforts have been concentrated on developing variable star classifiers (Debosscher et al. 2007, 2009; Richards et al. 2011a; Pichara et al. 2012; Mackenzie et al. 2016; Benavente et al. 2017; Narayan et al. 2018; Aguirre et al. 2019; Carrasco-Davis et al. 2019; Becker et al. 2020) and new alternatives to represent the celestial objects (i.e. human-based and deep learning-based features; Nun et al. 2015; Kim & Bailer-Jones 2016; Valenzuela & Pichara 2017; Naul et al. 2018). However, these classifiers rely heavily on the quality of the labelled training data sets and it is challenging and highly time-consuming to generate representative training data sets to train these models. This drawback hinders the model assessment procedure and, therefore, it can impact the performance of these models when they are tested on objects beyond the labelled objects (including those from the same catalogue). More significantly, this could lead to wrong models being used to label new training data, thereby generating a cascade effect. This situation has given rise to the following questions: Do we have *confidence* in our *variable star classifiers*? Can we *improve* the *model assessment process*? To answer these questions, the metrics used to evaluate models under non-favourable conditions have become an important topic.

Several papers have discussed the impact of biases in the training data of variable stars (Debosscher et al. 2009; Richards 2012; Masci et al. 2014). However, to the best of our knowledge, no definitive solution has been proposed to more accurately assess the classifiers in this scenario. Bias in data means that there is a difference between the joint distribution of our labelled data $\mathcal{D}^S$ and that of the population $\mathcal{D}^P$, which considers all the observed objects for the astronomical project (survey). The problem arises from the fact that the current classifiers are trained with a subset of $\mathcal{D}^S$ and, subsequently, the performance is evaluated using the complement (the testing set), typically by means of a cross-validation (CV) scheme. The CV strategies assume the existence of representative training data; however, we know that data sets in astronomy are biased and, consequently, we are unable to report a realistic classification performance.

Those biases stem from several sources, the majority of which can be linked to human-related tasks and technical characteristics of the telescopes. The former is associated with the labelling process

* E-mail: fjperez10@uc.cl

since astronomers are more prone to label a class when it is easier to define. In this sense, Cabrera, Miller & Schneider (2014) presented a good discussion about this systematic bias in the astronomical labelling process. The mechanical design of receptors generates another type of bias, specifically in relation to the range in which the signal can be processed; for example, when distances increase, less luminous objects are more difficult to see (Richards 2012). Finally, the rapid development of technology accelerates the obsolescence of the models. Hence, we are unable to apply trained models to new surveys and we lack sufficient confidence in the error metrics in these newer catalogues. This problem is addressed by domain adaptation and was discussed in depth in variability surveys in Benavente et al. (2017).

To analyse the effect of these biases, they are typically divided into two categories: biases in features and biases in class representations. The existence of biases in features (e.g. period and amplitude) means that there is a difference in the joint feature distribution between $\mathcal{D}^S$ and $\mathcal{D}^P$. That is to say, zones of the feature space without labelled objects or an overrepresentation of other zones, it impacts the relevance of those zones during the training and assessment process. Bias in the representation of classes is associated with some classes of variable stars that are more/less represented in $\mathcal{D}^S$ compared to $\mathcal{D}^P$.

Notwithstanding this underlying problem, few efforts have been made to study metrics and validation strategies to evaluate the performance of light-curve classifiers. Furthermore, it is a challenge to provide more accurate metrics to assess models in a scenario in which we cannot entirely trust the data. One natural framework with which to address the aforementioned problems is Bayesian modelling, which has been increasingly used in different fields of astronomy, such as to compare astrophysical models (Ford & Gregory 2007) or make predictions on the properties of celestial objects (Sanders & Das 2018). To improve the model assessment task, we propose a novel pipeline for evaluating Bayesian Logistic Regressions (BLRs) on biased training data. The methodology used is based on Bayesian machine learning, which allows us to incorporate astronomical knowledge into the model assessment process. Our approach exploits the powerful Bayesian model selection (BMS) scheme (Murray & Ghahramani 2005), which embodies desirable properties such as Bayesian Occam's razor, consistency, and comparability (Myung & Pitt 1997).

The BMS framework is based on the marginal likelihood (also known as Bayesian evidence), which is the likelihood function weighted by a prior distribution over the range of values for its parameters. In other words, the marginal likelihood contains the expected probability of data over the parameters. However, if we do not add information to these prior distributions, even this powerful and robust metric is unable to assess the models correctly when the training data are biased. Hence, to address these biases, we contribute with a strategy that exploits expert knowledge by incorporating informative priors in the marginal likelihood estimation of RR Lyrae star classifiers. Our methodology is divided into three stages; first, we propose a method to represent the prior knowledge using deterministic rules (DRs) founded on physical-based features, such as period and amplitude. In the second stage, we generate posterior samples using these informative priors. This is a suitable approach since, by means of posterior samples, we are able to ensure zones of high value in the likelihood function and the prior distribution. Moreover, we can add astronomical knowledge through the effect of the priors in the posterior distribution. Finally, in the third phase, we estimate the marginal likelihood using an approximated sampling method.

This paper is organized as follows. Section 2 introduces the background theory of metrics and validation strategies for the assessment of models. Section 3 provides an account of related works and is divided into two subsections: first, we review machine learning models in the classification of variable stars which deal with biases; secondly, we present how Bayesian data analysis has been applied in the field of astronomy. Section 4 outlines the proposed methodology used to address the challenge. Sections 5 and 6 describe the data and the implementation, respectively. After that, Section 7 shows the results. Finally, Section 8 sets out the conclusions and future work.

2 BACKGROUND THEORY

In the machine learning and statistical learning fields, the model assessment process is a central topic and one that is typically associated with three main tasks: (i) the evaluation of a population error using the training data error, (ii) the selection of the most suitable model among a set of alternatives, and (iii) the definition of a good set of hyper-parameters. In this section, we summarize the traditional methods used to assess models as follows: Section 2.1 concentrates on the metrics for model selection and Section 2.2 focuses on validation strategies.

2.1 Metrics for evaluating classifiers

There are several metrics to evaluate the performance of classification models that have been originated from different fields such as statistical learning, information theory, and data mining. Consequently, selecting one metric or a set thereof to assess our models can become challenging. Given that, it is important to consider the following well-known basic properties when using or proposing a metric. Consistency: the size of the training data should not affect our metric, Occam's razor principle: we desire a metric that can identify whether a model has the optimal complexity required, comparison: it should allow us to compare non-nested models, reference: the metric must be independent of the validation strategy, and individuality: the metric must be able to measure any given object individually (Anderson & Burnham 2004).

We present a summary of the most frequently used metrics below. They have been divided into two groups: Section 2.1.1 presents metrics based on the confusion matrix, and Section 2.1.2 provides a scheme of methods based on BMS.

2.1.1 Metrics based on confusion matrix

Within this framework, the most intuitive metric is the accuracy, which evaluates prediction quality based on the ratio of correct predictions over the total number of observations. This metric has two critical drawbacks: first, it is not able to discriminate the type of error, and secondly, it can be easily dominated by the majority class.

In order to assess the type of error, other measures can be obtained. For example, the Recall, which represents the fraction of positive patterns that are correctly classified. Or the Precision, which corresponds to the ratio between the positive objects that are correctly predicted and the total number of predicted objects for the true class. To consider a balance between Recall and Precision, we can evaluate these two metrics in conjunction through the F1-score. This metric is the harmonic-mean between Precision and Recall, and it is more robust than accuracy when the data set has imbalanced classes.

The aforementioned metrics are the most common ones in this framework, although there are many variants in the literature. A summary of these can be found in Sokolova & Lapalme (2009). Despite the large variety of metrics, there are a number of related limitations: (i) We cannot compare the trade-off between the goodness of fit and the model complexity directly; (ii) We must use validation strategies; and (iii) due to the fact that these metrics consider a hard classification (i.e. a Boolean decision about the predicted class), we cannot consider different levels of confidence in the prediction scores.

2.1.2 Bayesian model selection

A robust alternative for selecting models is the marginal likelihood, which is denoted by $p(\mathcal{D}|m)$, where m represents a model and $\mathcal{D}$ the training data. It appears in the first level of inference in the Bayesian framework:

$$p(\theta|\mathcal{D}, m) = \frac{p(\mathcal{D}|\theta, m)p(\theta|m)}{p(\mathcal{D}|m)}, \tag{1}$$

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta, m)p(\theta|m)\mathrm{d}\theta. \tag{2}$$

We use the following traditional notation: let $p(\theta|\mathcal{D}, m)$ be the posterior distribution of the parameter given the data and a model; let $p(\mathcal{D}|\theta, m)$ denote the likelihood function; let $p(\theta|m)$ represent the prior distribution over the parameters; and finally, let $p(\mathcal{D}|m)$ be the marginal likelihood.

The marginal likelihood like a model selector was analysed in depth by MacKay (1992) and, subsequently, the links with the Occam's razor principle were emphasized in Rasmussen & Ghahramani (2001), Murray & Ghahramani (2005), and Ghahramani (2013). The idea of using the marginal likelihood in model assessment comes from the second level of inference:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m, \mathcal{D})}, \tag{3}$$

where the Bayes' theorem is used to estimate the model probability given a data set $p(\mathcal{D}|m)$.

The estimation of $p(m|\mathcal{D})$ is intractable since we cannot enumerate all possible models. However, we can apply the same criteria used in the first level of inference, avoiding the denominator estimation (constant). In this way, we can estimate the model posterior by $p(m|\mathcal{D}) \propto p(\mathcal{D}|m)p(m)$. Finally, if we assume a non-informative prior for the model, $p(m|\mathcal{D})$ is proportional to the marginal likelihood. For this reason, we use the marginal likelihood to select the most appropriate model.

Despite the fact that the marginal likelihood automatically embodies all those desired properties of good metrics, it is unable to automatically manage the biases in the training data and its estimation is a computational challenge in high-dimensional data (see equation 2). To address this challenge, we can estimate the marginal likelihood by interpreting it as an expected value and then performing Monte Carlo (MC) estimation according to the following equations:

$$p(\mathcal{D}|m) = \mathbb{E}_\theta\left[p(\mathcal{D}|\theta, m)\right], \tag{4}$$

$$\frac{1}{S}\sum_{s=1}^{N} p(\mathcal{D} \mid \theta_s, m), \theta_s \sim p(\theta). \tag{5}$$

This simple approach only performs well if the prior and likelihood have a similar shape and are strongly overlapped. If this does not hold, then misleading samples can be generated in low-valued areas of the likelihood function. Therefore, a few samples with high values in the likelihood function dominate the estimator, and this could produce a high variance in the estimation procedure (Gronau et al. 2017).

Due to these difficulties, the majority of research into BMS avoids MC sampling methods by applying approximations such as the Laplace approximation and Bayesian information criterion (Schwarz et al. 1978; Watanabe 2013) or they resort to MC methods based on posterior samples (Neal 2001; Raftery et al. 2006; Overstall & Forster 2010). Our proposal is based on the latter type of strategies but adding astronomical knowledge to those posterior samples.

2.2 Validation strategies

CV is the most common family of methods for estimating metrics. We present a summary and some drawbacks of the three most common CV-based methods. First, we review *hold-out*, which is the most basic approach. Therein, two sets of data are generated, one of which is used to train the model and the other to evaluate its quality. This approach depends heavily on one data set, and for that reason, it is a good option only when we are in possession of large quantities of data or when there is some running time limitation. Moreover, in small data sets, hold-out can generate a pessimistic estimator (Lendasse, Wertz & Verleysen 2003).

Secondly, k-fold, which is the most commonly used variant of CV, considers splitting the data $\mathcal{D}^S$ into smaller chunks $\mathcal{D}_1^S, \mathcal{D}_2^S, ...\mathcal{D}_K^S$ with the same size. We train using $\mathcal{D}^S \setminus \mathcal{D}_k^S$ chunks and evaluate using the free chunk $\mathcal{D}_k^S$. The number of folds provides the bias-variance trade-off; a small number of folds reduces the bias but increases the variance. Lastly, *leave-one-out* uses each data point as a chunk, and for each object, a model is trained to leave only this object out. It provides an unbiased estimator, although its variance can be larger, and the running time can be prohibitive (Rao, Fung & Rosales 2008).

Arlot et al. (2010) presented a survey of CV procedures for model selection. Despite the effort to develop variants, these ideas consist of a fundamental assumption. They consider that we are working with representative training data. This means that $\mathcal{D}^S$ has the same probability distribution as the data beyond the labelled objects $\mathcal{D}^\mathcal{T}$. However, in astronomy, we are often unable to generate such representative training data.

In an attempt to overcome this challenge, Sugiyama, Krauledat & MÃžller (2007) proposed a CV variant to tackle the aforementioned biases (also known as *data shift* problem) by means of an importance weighted CV (IWCV) approach. The IWCV weighs each observation i in the evaluation metric with the density ratio $p(x_i)_{\mathrm{test}}/p(x_i)_{\mathrm{train}}$. Note that IWCV is proposed to address a type of *data shift*, which is named as a covariate shift (bias in features), here, $p(x_i)_{\mathrm{train}} \neq p(x_i)_{\mathrm{test}}$, but $p(y|x)_{\mathrm{train}} = p(y|x)_{\mathrm{test}}$. If we need to work on scenarios with bias in labels (target shift), which assumes $p(y)_{\mathrm{train}} \neq p(y)_{\mathrm{test}}$, but $p(x|y)_{\mathrm{train}} = p(x|y)_{\mathrm{test}}$, we must adapt the density ratio by $p(y)_{\mathrm{train}}/p(y)_{\mathrm{test}}$. This is a clever approach, but it assumes an existing knowledge of the probability density functions for $\mathcal{D}^S$ and $\mathcal{D}^T$, which can be intractable in high dimensions.

A further commonly used method is bootstrap (Efron & Tibshirani 1997), which uses sampling with replacement, in which N samples are selected in each of the k iterations. In this approach, in a given iteration, a particular sample may appear more than once, while others might not appear at all. Although the traditional bootstrap has interesting statistical properties, it fails to select classifiers in a machine learning context because it favours overfitting classifiers (Kohavi et al. 1995).

An interesting variant is found when the bootstrap approach is analysed from a Bayesian perspective (Rubin 1981). A traditional bootstrap can be understood by modelling the probability of drawing a specific observation such as a categorical distribution, $Cat(\pi)$, where the vector $\pi = (\pi_1, \pi_2, ..., \pi_N)$ is the probability of drawing each object ($\sum_i^N \pi_i = 1$). In a traditional bootstrap we have $\pi_1 = \pi_2 = \pi_k = \pi_N = 1/N$. In a Bayesian view, π draws from a Dirichlet, $Dir(\alpha)$, where for example, the expected proportion for π_1 is based on the priors $\alpha_1/\sum_{i=1}^N \alpha_i$. However, the generation of informative priors $Dir(\alpha)$ can pose a significant challenge.

3 RELATED WORK

This section is divided into two subsections. Section 3.1 studies the state of the art of variable star classifiers, with emphasis on research addressing underlying biases and how these approaches select and compare models. Section 3.2 discusses briefly how Bayesian data analysis has been used in the field of astronomy.

3.1 Classification of variable stars under bias

Several papers on the automatic classification of variable stars have sought to address the data shift problem from different perspectives. However, none has focused on model selection strategies. Over a decade ago, Richards et al. (2011b) proposed several strategies to improve the training data. For example, they designed an active learning method and presented an importance-weighted CV method to avoid underrepresented zones of feature space. However, metrics to compare models in these contexts were not analysed in depth. Masci et al. (2014) proposed a random forest (RF) classifier, which was trained with a labelled set from the Wide-field Infrared Survey Explorer (Wright et al. 2010), within an active learning approach. This classifier was able to improve the training data and mitigate the biases. This RF approach outperformed support vector machine (SVM), K nearest neighbours (KNN), and neural networks (NN) using a CV method to estimate the accuracy.

Benavente et al. (2017) proposed a full probabilistic model to address the domain adaptation problem. This model was able to transfer knowledge (feature vectors) among different catalogues. It was able to manage the covariate shift and improve the cross-validated F1-score. A Gaussian mixture model representing each catalogue (source and target) and a mixture of linear transformations (translation, scaling, and rotation) were applied. Recently, Aguirre et al. (2019) designed a convolutional NN that was able to learn from multiple catalogues, outperforming an RF based on handcrafted features. To manage the imbalanced classes, Aguirre et al. (2019) proposed a novel data augmentation scheme that creates new light curves by modifying real objects.

Sooknunan et al. (2021) reported the relevance of a non-representative $\mathcal{D}^S$ when applying trained models on data from new telescopes. Moreover, they studied how the accuracy metric decreases (training versus real) when $\mathcal{D}^S$ is small. To create the training data, they used a few real objects and synthetic light curves generated using a Gaussian process. Experiments with the following five classes of transients were conducted: active galactic nuclei (supernovae, X-ray binaries, γ-ray bursts, and novae. The results led to the conclusion that a better performance can be obtained in new surveys if contextual information (object location) and multiwavelength information are incorporated. To encourage the use of multiwavelength information, they presented results using both the optical telescope MeerLICHT (Bloemen et al. 2016) and the radio telescope MeerKAT (Booth & Jonas 2012).

Naul et al. (2018) proposed the use of a recurrent autoencoder to learn a variable star embedding. The measurement error in observations is used for weighting the reconstruction metric in the loss function so that those observations with large measurement error were less important. Subsequently, this embedding is used to classify by means of an RF classifier. The new representation is compared with two baseline sets of handcrafted features (Richards et al. 2011a; Kim & Bailer-Jones 2016), being competitive with traditional approaches when folded light curves were used. It outperformed or was similar to the baselines in the LIncoln Near-Earth Asteroid Research survey (Sesar et al. 2013) and the MAssive Compact Halo Object catalogue (Alcock et al. 1997).

Recently, Becker et al. (2020) presented a scalable recurrent NN that was capable of learning a representation without human support. The researchers obtained a competitive accuracy in shorter running time than an RF that was based on handcrafted features. Furthermore, they provided a comparison between biases affecting handcrafted features and those based on deep-learning features, thereby supporting the line of thought that deep learning models are capable of learning features that are less biased when working in specific surveys.

Table 1 provides a summary of model assessment strategies for variable star classifiers. We conclude that the majority of the papers analysed herein have applied metrics based on the confusion matrix and have primarily utilized k-fold for the validation thereof.

3.2 Bayesian data analysis in astronomy

In recent decades, several astronomical papers have proposed the application of a Bayesian analysis. For example, pioneering research was conducted by Gregory & Loredo (1992) and Saha & Williams (1994) on the parameter estimation of astrophysical models. The research field most heavily influenced by these developments has been probably that of cosmological parameter estimation (Christensen & Meyer 1998; Christensen et al. 2001). Accordingly, Trotta (2008) provided a comprehensive review of Bayesian statistics with an emphasis on cosmology. Sharma (2017) produced a literature review that focuses on the Monte Carlo Markov Chain (MCMC) for Bayesian analysis in astronomy, providing an extensive overview of several MCMC methods, while also emphasizing how astronomers have used Bayesian data analyses in the past and how such approaches should, in fact, be used more commonly in the present. Furthermore, Sharma (2017) exemplified a number of basic concepts for model selection in a Bayesian approach. Subsequently, Hogg & Foreman-Mackey (2018) provided a pedagogical overview of MCMC in astronomical contexts and discussed its foundations, highlighting certain aspects to consider to avoid obtaining misleading results from applications of this otherwise powerful technique. Moreover, several papers have shown the advantages of the BMS approach (Parviainen, Deeg & Belmonte 2013; Ruffio et al. 2018) in astrophysical model selection.

Weinberg (2013) presented a software package to apply Bayesian statistics in astronomy, including methods for estimating the posterior distribution and managing the model selection. This paper also provides a comprehensive introduction to Bayesian inference. Moreover, Weinberg (2013) included two applications where the system performance on astrophysical models (semi-analytic galaxy formation model and Galaxy photometric attributes) is evidenced.

Budavári, Szalay & Loredo (2017) designed an incremental Bayesian method to decide whether observations correspond to faint objects or noise from the data set (multiepoch data collection). To classify each object, thought a Bayes factor scheme the marginal likelihoods of competing hypotheses (object or no object), at each epoch, are compared. In order to define these hypotheses, expert knowledge of the flux of each alternative is included.

Table 1. Summary of strategies for selecting variable star classifiers. The bold letters in the classifiers column represent the best model according to the papers listed in the final column. GMM = Gaussian mixture model classifier, BN = Bayesian network, BAANN = Bayesian average of artificial neural networks, SVM = support vector machine, CNN = convolutional neural network, RNN = recurrent neural network.

Classifiers	Metrics	Validation	Reference
GMMC, BN, BAANN, SVM	Accuracy	10-fold	(Debosscher et al. 2007, 2009)
CART, random forest,	Error rate	10-fold	(Richards et al. 2011a)
Boosted trees, C4.5, SVM	–	–	–
RF	Error rate	*k*-fold	(Bloom et al. 2012)
Boosted RF, regular RF, SVM	F1-score	10-fold	(Pichara et al. 2012)
RF+BN	F1-score	*k*-fold	(Nun et al. 2014)
NN, RF, SVM, KNN	Accuracy, ROC	Hold-out	(Masci et al. 2014)
Meta Classifier (RF)	Precision-F1-score-recall	*k*-fold	(Pichara, Protopapas & León 2016)
SVM, RF	F1-score	10-fold	(Mackenzie et al. 2016)
LR, RF, CART, SBoost, AdaBoost	Precision, Recall, F1, AUC	10-fold	(Elorrieta et al. 2016)
SVM, LASSO, NN, DNN	–	–	–
RF, SVM	F1-score	CV	(Benavente et al. 2017)
Decision tree	F1-score	Bootstrap	(Castro, Protopapas & Pichara 2017)
Recurrent CNN, RF	Accuracy, Av. recall	Hold-out	(Carrasco-Davis et al. 2019)
RF	Accuracy	3-fold	(Sooknunan et al. 2021)
CNN	Recall, F1-score, MC	Repetitive hold-out	(Mahabal et al. 2017)
RF	OOB - Accuracy- ROC	*k*-fold	(Narayan et al. 2018)
AE-RNN+RF	Accuracy	5-fold	(Naul et al. 2018)
CNN, RF	Accuracy	10-fold	(Aguirre et al. 2019)

In spite of the fact that several papers have applied BMS to astronomy, to the best of our knowledge, our proposal is the first approach that adds physical information during the assessment process of machine learning classifiers for variable stars.

4 INFORMATIVE BAYESIAN MODEL SELECTION

This section provides a comprehensive description of our method to add human knowledge to the assessment and selection of RR Lyrae star classifiers. The methodology assumes that we have a set of models $\{m_1, m_2, .., m_i, ..m_n\} \in \mathcal{M}$ and a biased set of labelled objects (variable stars) to train them. Our goal is to rank these models to obtain a good performance in a shifted data set (testing set).

The method can be divided into three main steps. Section 4.1 focuses on obtaining priors from DRs. Section 4.2 considers the generation of posterior samples running an MCMC algorithm. Section 4.3 presents the mechanism to add informative posterior samples to the marginal likelihood estimation procedure.

Fig. 1 shows a diagram of our method, in which the output for each step is highlighted. The final output is a ranking of models based on an informative estimation of the marginal likelihood. We propose to mitigate biases through this informative marginal likelihood.

4.1 Obtaining informative priors

In the Bayesian framework, informative priors offer a great opportunity to add expert knowledge to machine learning models; however, the majority of Bayesian approaches use non-informative priors, and hence, they rely completely on the likelihood function (Gelman, Simpson & Betancourt 2017). The use of non-informative priors, if there is expert knowledge, can be controversial (Gelman et al. 2008; Golchi 2019), and it is valid for both levels of inference: the first level, when we make inference on parameters, and the second level, when we make inference on models.

For some models, the addition of human knowledge can be less complex, since it can be transferred from the space of features to the space of parameters directly; this is the case of Bayesian GMM or Bayesian Naive Bayes. However, in models such as BLRs or Bayesian neuronal networks, it is not direct. Proposing informative priors for BLR can be a great challenge (Hanson et al. 2014), despite the fact that some alternatives have been proposed to add expert knowledge when it is available. Gelman et al. (2008) proposed weakly informative priors (Cauchy priors) that are also useful to solve the complete separation problem (Zorn 2005). Hanson et al. (2014) provided an informative *g*-priors approach; this scheme is suitable if there is information about the probability of each class. In spite of these proposals, and to the best of our knowledge, information about the relationship between classes and features cannot easily be incorporated. To face this challenge, we propose a novel methodology to obtain informative Gaussian priors for BLR classifiers.

We propose to obtain astronomical knowledge through DRs. DRs can be used to filter celestial objects without resorting to machine learning methods. The DRs are based on physical features such as period, mean magnitude, and amplitude. To design these rules for RR Lyrae stars, we can use literature in the field to define physical features that may be particularly relevant in characterizing this class of variable stars.

DRs can be understood as a relationship between an antecedent (if) and a consequent (then). To define a rule, we use a standard notation, $A{\Rightarrow}B$, where A represents a physical condition (antecedent) and B represents a class of variable stars (consequent). Some examples of DRs for pulsating stars include

(i) (period ∈ [0.2, 1.0] days) ⇒ RR Lyrae
(ii) (amplitude ∈ [0.3 − 1.2] in *V*-band) ⇒ RR Lyrae
(iii) (amplitude ∈ [0.2 − 0.8] *I*-band) ⇒ RR Lyrae
(iv) (period ∈ [1, 100] days) ⇒ Classical Cepheid
(v) (period ∈ [0.75, 30] days) ⇒ Type II Cepheid
(vi) (period ∈ [0.5, 8.0] hours) ⇒ Dwarf Cepheid

Note that some physical conditions can be valid for more than one variable star class; however, when applying a chain of several DRs, this drawback is reduced. Despite that, we recommend

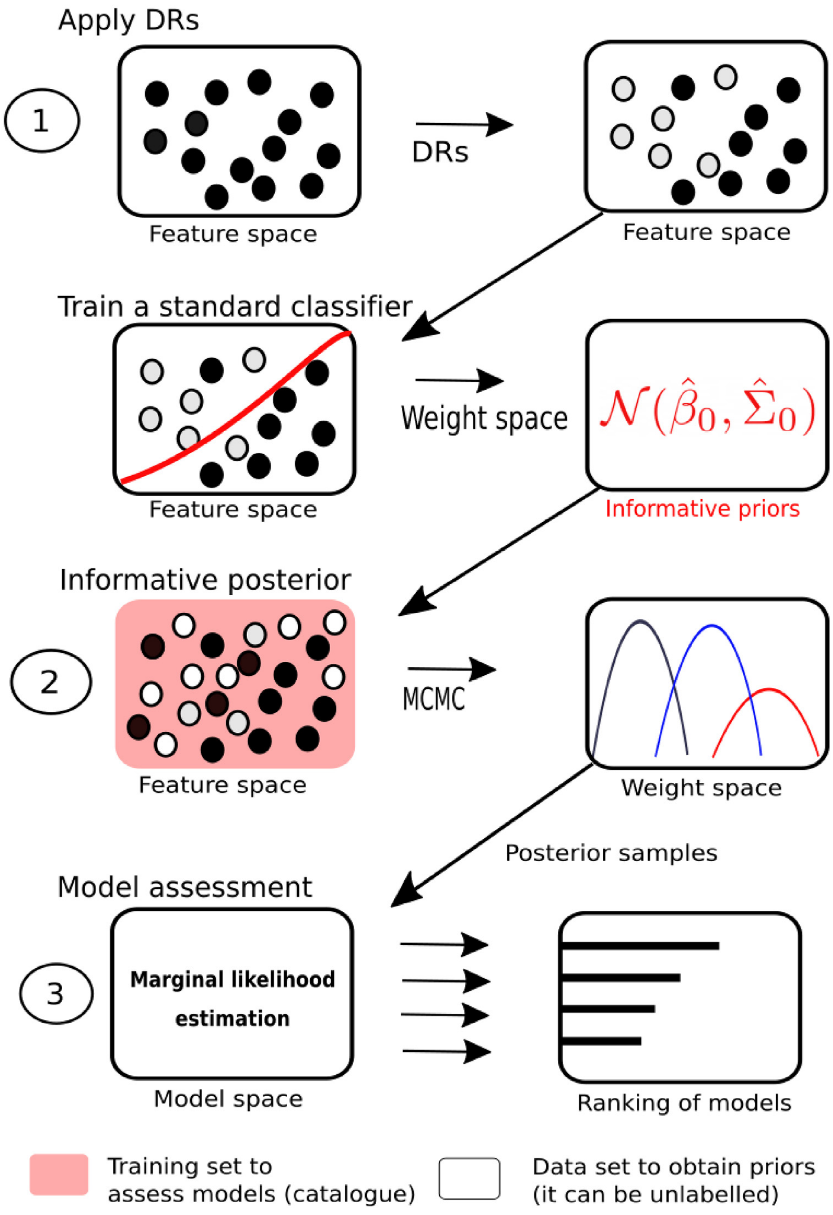


Figure 1. Proposed method overview. (1) First, we label a variable star data set using the DRs. (2) Then, we train a standard logistic regression to obtain its weights. (3) After that, the mean and variance of those weights are used in an MCMC frame to generate posterior samples. (4) The marginal likelihood is estimated using these informative posterior samples. Lastly, the estimated marginal likelihoods are used to rank the models.

mitigating this possible overlap using not only various DRs but also DRs based on features that do not vary across different surveys (invariant features), e.g. period and amplitude (Catelan & Smith 2015).

Once we have obtained the DRs, we propose algorithm 1 to obtain informative priors. This algorithm identifies priors in a binary classification scheme; thus, we must use a set of rules for each class of variable stars.

The priors $\hat{\theta}$ are generated by fitting a standard (non-Bayesian) logistic regression. The training data with which to fit this model becomes critical at this stage. This because depending on the training set, our DRs can find a different distribution of objects for both the true class and the false class. It is possible to use an entire survey, a subset of a survey, or even an improved set (data augmentation, adversarial examples, down-sampling, or oversampling).

This method allows transferring astronomical knowledge from the space of physical features to the space of model parameters through the collected DRs based on physical features of RR Lyrae stars. In particular, we define the mean estimator vector, $\hat{\theta}$, and the variance estimators, $\mathrm{Var}(\hat{\theta})$, for a normal prior. $\mathrm{Var}(\hat{\theta})$ is defined by the diagonal of the inverse Fisher Information matrix $\mathbf{I}(\theta)$. To avoid very small values for the estimated prior of variance $\mathrm{Var}(\hat{\theta})$, we add a small constant ϵ (for example, ϵ =0.1) after applying algorithm 1.

4.2 Posterior samples generation

Our path for transferring human knowledge is by means of posterior samples since these contain both prior knowledge and data information. In this step, for each $m \in \mathcal{M}$, we train a BLR with priors obtained using algorithm 1.

Algorithm 1 Procedure to obtain priors from physical based features.

```
Input: Data $\mathcal{D}_{\mathcal{X}}$, classifier $m$, DRs
Output: weights for the classifier $m$, $\beta$
  $\mathcal{D}_{\mathcal{Y}} = \mathbf{1}$
  for $r \in$ Rules do
    for $d \in \mathcal{D}_{\mathcal{X}}$ do
      state ← r.applyDR($d$)
      if state == False then
        $\mathcal{D}_{\mathcal{Y}}[d] = 0$
      end if
    end for
  end for
  $\hat{\theta} \leftarrow$ m.fit($\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}}$)
  Var($\hat{\theta}$) ← diag($\mathbf{I}(\theta)^{-1}$)
  return $\hat{\theta}$, Var($\hat{\theta}$)
```

To estimate the posterior $p(\theta|\mathcal{D}_{\mathcal{X}}, m)$, we propose to use standard MCMC techniques, such as Metropolis–Hastings or Hamiltonian MC algorithms. Moreover, we use the Gelman–Rubin test to validate the sample convergence in each dimension (Gelman et al. 1992). Lastly, to manage imbalanced classes, we downsample the data sets.

This step is time-consuming; hence, more efficient sampling strategies could speed up our strategy. We did not consider variational inference since the samples from this approach can be biased and our approach requires precise and unbiased samples (Blei, Kucukelbir & McAuliffe 2017).

4.3 Informative marginal likelihood estimation

The marginal likelihood has been widely studied to compare and select machine learning models, despite the fact that its estimation represents a significant computational challenge. Comprehensive references for the study of estimation methods can be found in Gronau et al. (2017) and Wang, Chen & Kuo (2018).

We propose addressing an informative estimation of the marginal likelihood using a bridge sampling approach (Overstall & Forster 2010; Gronau et al. 2017). Unlike standard MC estimators (importance sampling or harmonic mean estimator), bridge sampling allows us to avoid dealing with typical constraints of standard MC methods in relation to the shape of a proposal probability distributions. Indeed, this method has suitable properties in our context, mainly due to the following reasons: (i) it does not waste resources by generating samples in low-value zones, and (ii) it allows us to incorporate astronomical knowledge in order to reduce the impact of biases in the training data.

The bridge sampling estimator is based on a ratio of two expected values as follows:

$$p(\mathcal{D}) = \frac{\mathbb{E}_{g(\theta)}\left[p(\mathcal{D}|\theta)p(\theta)h(\theta)\right]}{\mathbb{E}_{p(\theta|\mathcal{D})}\left[h(\theta)g(\theta)\right]}. \tag{6}$$

To estimate $\mathbb{E}_{g(\theta)}\left[p(\mathcal{D}|\theta)p(\theta)h(\theta)\right]$, we use samples from a proposal distribution, $g(\theta)$, and to estimate $\mathbb{E}_{p(\theta|\mathcal{D})}\left[h(\theta)g(\theta)\right]$ we need posterior samples, $p(\theta|\mathcal{D})$, that contain astronomical knowledge.

The desired match between the samples from the proposal and those from the posterior is managed through a function, which is named bridge function,

$$h(\theta) = C\frac{1}{s_1 p(\mathcal{D}|\theta)p(\theta) + s_2 p(\mathcal{D})g(\theta)}, \tag{7}$$

which plays a central role in the bridge sampling estimator (Meng & Wong 1996). When the bridge function is introduced to the estimator, the function depends recursively on $p(\mathcal{D})$; hence, for estimating it, it is solved iteratively by

Table 2. Class distribution of OGLE labelled set.

Class	Abbreviation	Number of objects
Long-period variable	lpv	323 999
RR Lyrae	rrlyr	42 751
Eclipsing binary	ecl	41 787
Cepheids	cep	7952
Delta Scuti	dsct	2807
Type II Cepheid	t2cep	589
Double Periodic Variable	dpv	135
Anomalous Cepheid	acep	81
Dwarf nova	dn	35
R CrB variable	rcb	22

$$\hat{p}(\mathcal{D})^{t+1} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\frac{p(\mathcal{D}|\theta_i)p(\theta_i)}{s_1 p(\mathcal{D}|\theta_i)p(\theta_i)+s_2\hat{p}(\mathcal{D})^t g_(\theta_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1}\frac{g(\theta_j)}{s_1 p(\mathcal{D}|\theta_j)p(\theta_j)+s_2\hat{p}(\mathcal{D})^t g_(\theta_j)}},$$

$$\theta_j \sim p(\theta|\mathcal{D}); \theta_i \sim g(\theta). \tag{8}$$

Through this estimator, astronomical knowledge is incorporated into the assessment process. Using an informative prior, we can reduce the effect of biases in the training sets on the posterior. A proof of this estimator is presented in Appendix A (Gronau et al. 2017).

5 DATA AND CLASSIFIERS

This section presents the inputs used to validate our methodology. Section 5.1 describes the Optical Gravitational Lensing Experiment (OGLE) catalogue. Section 5.2 describes how we obtain the final training set from the raw light curves. In Section 5.3, the procedure to obtain a ground truth is explained. Lastly, in Section 5.4, we present a set of models that are assessed through our method.

5.1 OGLE-III catalogue of variable stars

For testing purposes, we use the OGLE-III variable star catalogue, which corresponds to the third phase of the OGLE project (Udalski et al. 2008). The main goal of OGLE has been to identify microlensing events and transiting planets in four fields: the Galactic bulge, the Large and Small Magellanic Clouds, and the constellation of Carina. We use light curves with at least 25 observations in the I band. The final number of labelled light curves is 420,126. In Table 2, we present the number of objects per class.

To estimate the informative priors (step 1 in Fig. 1), we also use the OGLE-III catalogue. When applying the DRs to this data set, we obtained a subset with ∼75 per cent of RR Lyraes, ∼20 per cent of eclipsing binaries, and ∼ 5 per cent distributed amongst the remaining classes.

5.2 Processing of light curves

To extract features from the light curves, we use the Feature Analysis for Time Series (FATS) library (Nun et al. 2015), thus obtaining a 420 126 × 63 matrix, where 63 stands for the number of features included in our analysis. Subsequently, to manage both the high dimensionality and multicollinearity, we apply principal component analysis (PCA). Spyroglou et al. (2018) used a similar strategy that combines BLR and PCA to avoid multicollinearity among features.

Table 3. Number of objects in the training and testing sets for each class. TC represents the true class.

$\mathcal{D}$	Training	Testing	TC training	TC testing
rrlyrae-1	389 364	30 762	27 240 (6.9%)	15 269 (49.6%)
rrlyrae-2	402 787	17 339	34 233 (8.5%)	8500 (49.0%)
rrlyrae-3	335 721	84 405	34 001(10.1%)	8732(10.3%)

5.3 Shifted training and testing sets

To evaluate the performance of our approach, we simulate some challenging scenarios, where the training objects are shifted from the testing objects. To create this scenario, we propose a procedure (algorithm 2) for splitting a labelled catalogue (OGLE-III in our case) into two shifted (biased) data sets.

Algorithm 2 Procedure to introduce bias in the distribution of objects from a catalogue.

```
Input:  Data $\mathcal{D} = (\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}})$, classifier $m$, bias control $(T)$
Output:  Biased Data $\hat{\mathcal{D}} = (\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{test}})$

  $m$.fit$(\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}})$
  for $(d_x, d_y) \in (\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}})$ do
      $P_A, P_B \leftarrow m$.softPredict$(d_x)$
      $h = 1 - (P_A^2 + P_B^2)$
      $p = e^{-h/T}$
      $r$ = uniform(0, 1).sample()
      if $r \leq p$ then
          $\mathcal{D}_{\mathrm{train}}$.add$((d_x, d_y))$
      else
          $\mathcal{D}_{\mathrm{test}}$.add$((d_x, d_y))$
      end if
  end for
  return $\hat{\mathcal{D}}$
```

First, we fit a binary classifier (m) that is trained with the entire catalogue. We use an RF classifier (m) to obtain a soft prediction (probability) for each star, and then, we use these predictions to split the data set $\mathcal{D}$. To split the objects, we define a threshold to assess whether an object can be easily classified or not; to measure that, we use the following metric for each object $i \in \mathcal{D}$: $h_i = 1 - (P(A)_i^2 + P(B)_i^2)$. This is based on the Gini impurity index (Raileanu & Stoffel 2004), where $P(A)_i^2$ is a soft prediction for the true class (RR Lyrae) and $P(B)_i^2 = 1 - P(A)_i^2$ a prediction for the false class.

To avoid a hard threshold when deciding the set (training or testing) for each object, we add a random selection, which is tuned by a constant T. This is based on the annealing principle (Van Laarhoven & Aarts 1987) and allows us to provide a probabilistic selection of objects, assigning difficult objects [$P(A)_i$ close to 0.5] more frequently to the testing set. As higher values for T are defined a less shifted sets is generated.

We apply algorithm 2 to obtain data sets with different levels of bias for the RR Lyrae class. The bias was managed by the parameter T, and we obtained the data sets rrlyrae-1, rrlyrae-2, and rrlyrae-3 for $T \in \{1, 2, 4\}$. These three configurations allow us to evaluate our proposal under different bias scenarios. Table 3 provides a summary of different biased data sets.

Fig. 2 shows the hardness distribution (classification difficulty) for objects in the training and testing sets. As we said before, we assume that objects whose prediction scores are close to 0.5 are more difficult to classify than whose predictions scores are close to 1 or 0. According to this definition, in the training sets in Figs 2(a), (c), and (e), we can observe that the training sets have a higher frequency of easier objects than the testing sets in Figs 2(b), (d), and (f). The relative frequency of objects at different levels of hardness can be visualized in both type plots, in the histograms and those bars on the top of each figure.

Fig. 3 presents the resulting amplitude versus period distribution, also known as *Bailey diagram*, obtained using algorithm 2, in the space of features for data set rrlyrae-1. Figs 3(a) and (b) show a clear shift in the joint distribution of period and amplitude for RR Lyrae from the Small Magellanic Cloud between the training and test sets. Figs 3(c) and (d) show a similar behaviour for RR Lyrae of the Galactic disc. We note that the bimodal distributions that are seen in these Bailey diagrams are similar to those typically found for RR Lyrae stars (e.g. Catelan & Smith 2015, and references therein). In particular, stars in the sequence with the longest periods at any given amplitude are fundamental-mode pulsators, also known as ab-type RR Lyrae stars. Conversely, stars located in the sequence with relative short periods and small amplitudes are first-overtone pulsators, or c-type RR Lyrae stars (RRc). Double-mode RR Lyrae, which pulsate simultaneously in the fundamental and first-overtone modes, also exist, and are commonly denoted as RRd. Their position in the Bailey diagram will depend on which mode is selected as the dominant one. We note that c-type and d-type RR Lyraes stars are mainly assigned to testing sets [see Figs 3(b) and 3(d)]. In other words, when we trained the RR Lyrae classifier (m) in algorithm 2 these types (RRc and RRd) were more difficult to classify.

5.4 Classifiers

As mentioned before, we focus on assessing and ranking a set of BLR classifiers. We compare rankings provided by our method with the accuracy-based rankings in a CV framework, considering two traditional logistic regression variants. Below we present a brief description of each of these models.

5.4.1 Standard logistic regression:

The standard LR classifier models the success probability of a binary dependent variable, $y \in \{0, 1\}$, by means of a Bernoulli distribution:

$$p(y|\mathbf{x}, \theta) = \mathrm{Ber}(y|s(\mathbf{x}, \theta)). \tag{9}$$

In this model, a sigmoid function,

$$s(\mathbf{x}, \theta) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}} = \frac{e^{\theta^T \mathbf{x}}}{1 + e^{\theta^T \mathbf{x}}}, \tag{10}$$

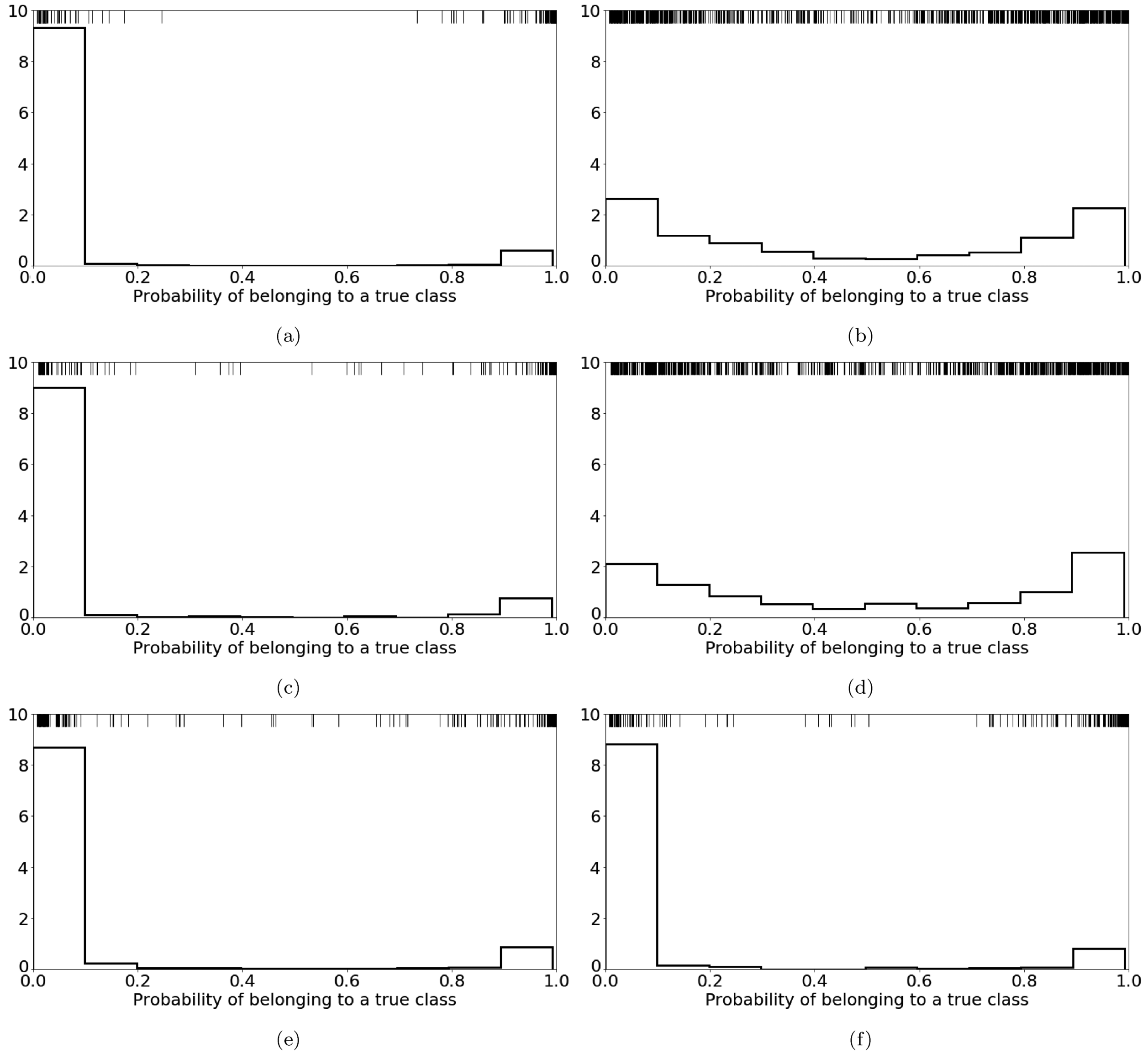
of input ($\mathbf{X}$) and parameters (θ) is used to model the Bernoulli parameter ($p = s(\mathbf{x}, \theta)$). The Likelihood function,

$$p(y|\mathbf{x}, \theta) = s(\mathbf{x}, \theta)^y (1 - s(\mathbf{x}, \theta))^{1-y}, \tag{11}$$

is optimized, giving rise to the maximum likelihood estimator.

5.4.2 Penalized logistic regression (l_2-LR-C):

In Bayesian terms, penalized LRs (l_2 and l_1) embody a prior distribution over θ, and subsequently, the maximum value for the resulting distribution (Maximum a posteriori or MAP) is selected. In particular, l_2-LR is equivalent to a vague Gaussian prior centred at the origin. Let $1/C$ be the penalization factor; hence, if C is small, we obtain a stronger regularization. This approach does not use human knowledge to define the shape of priors.



(a) (b) (c) (d) (e) (f)

Figure 2. Histogram for the probability of belonging to the true class in rrlyrae-1 (a–b), rrlyrae-2 (c–d) and rrlyrae-3 (e–f). (a), (c), and (d) are training sets and (b), (e), and (f) are testing sets. The bars on the top of each figure represent objects. To create these plots, a sample of 10 000 objects was used.

5.4.3 Bayesian logistic regression:

BLR focuses on estimating and using the posterior of the distribution of the weights $p(\theta|\mathcal{D})$ in the LR. In our proposal, the informative priors $\sim\mathcal{N}(\hat{\theta}, \hat{\sigma})$ are estimated using the method laid out in Section 4. For these experiments, we consider DRs for period and amplitude, both of which were estimated with the FATS library (Nun et al. 2015).

Each of these models (LR, l_2-LR, and BLR) represents a family of models ($\mathcal{M}$), which are defined by transformations over their input matrices. Regarding these transformations, first, we apply a linear transformation using PCA retaining the r most important principal components, where $r \in \{2, 4, 6, 8, 10, 12\}$, and after that, we also apply polynomial transformations over each component $p \in \{1, 2\}$. The interactions among the components were not considered.

Let ($\mathbf{X}^{n\times q}$) be the final matrix, where n are the objects in the training set and q the product between the polynomial degree and the number of components used in each model, $m \in \mathcal{M}$. This processing allows us to control the complexity of the model by increasing the number of PCA components or by increasing the degree of the polynomial transformation. Due to convergence problems, the model $m(2, 1)$ was not considered in our experiments ($|\mathcal{M}| = 11$). Downsampled data were used to deal with imbalanced classes.

In LR and l_2-LR, the models are sorted by their cross-validated accuracy in training. The BLR models are ordered according to our method (informative marginal likelihood). In the following section, we compare the rankings obtained and show empirical results in which the marginal likelihood (BLR case) can provide improved rankings with respect to CV (LR and l_2-LR cases).

6 IMPLEMENTATION

Our methodology was implemented using PYTHON 3.7. The most important libraries in our code are presented below:

PYMC: probabilistic modelling framework and posterior sampling algorithms (Salvatier, Wiecki & Fonnesbeck 2016).

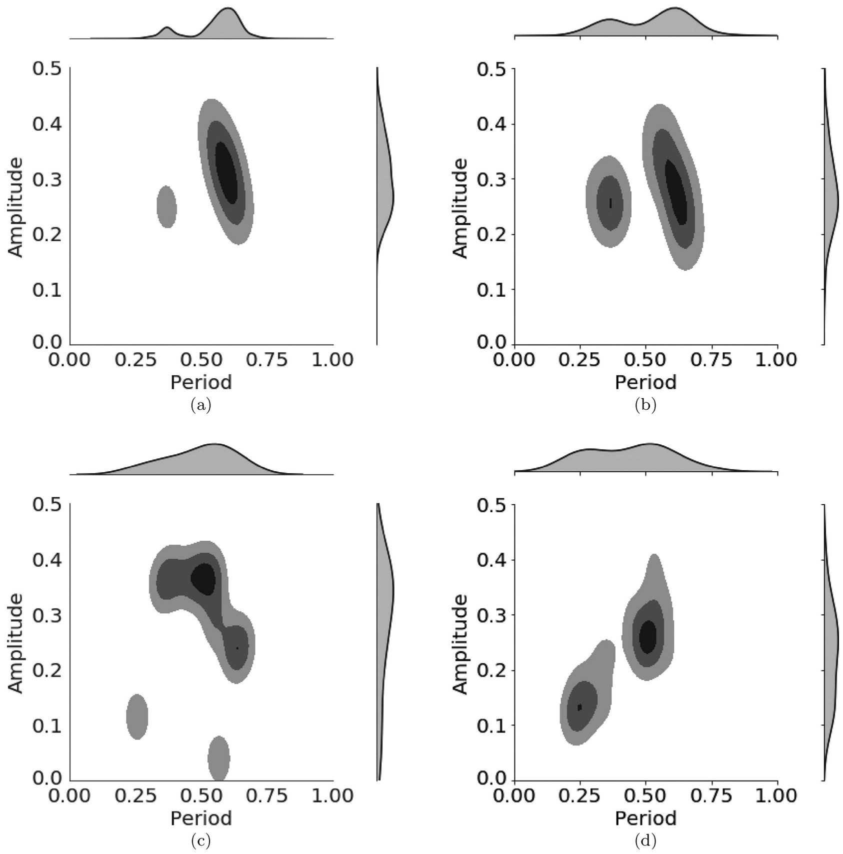


Figure 3. Density plots for RR Lyrae variable stars in rrlyrae-1 data set. (a) Small Magellanic Cloud – Training. (b) Small Magellanic Cloud – Testing. (c) Galactic disc – Training. (d) Galactic disc – Testing.

SCIKIT-LEARN: preprocessing, traditional machine learning models (e.g. Logistic Regression, RF, and PCA), CV methods and metrics for assessing models (Pedregosa et al. 2011).

PANDAS: methods for reading and managing data sets (McKinney et al. 2011).

SEABORN: visualization methods, e.g. scatter plots, histograms, and density plots (Waskom et al. 2014).

We also use a PYTHON implementation of bridge sampling, which was developed by Grunwald (2004). Lastly, the code source is available at https://github.com/frperezgalarce/vsbms.

7 RESULTS

Fig. 4 presents two examples of rankings that were generated by different strategies for selecting models: Fig. 4(a) provides a ranking of models from our proposed method, while Fig. 4(c) shows a ranking using a k-fold cross-validated ($k = 10$) accuracy. In these simple examples, we can note that the marginal likelihood strategies [Figs 4(a) and 4(b)] provide a better ranking coherence compared to the cross-validated accuracy. In fact, according to Fig. 4(c), the cross-validated accuracy selects the worst model.

To obtain a more rigorous comparison of rankings among methods, we define a set of metrics to quantify several viewpoints thereof. The selection of metrics used to compare models is presented below.

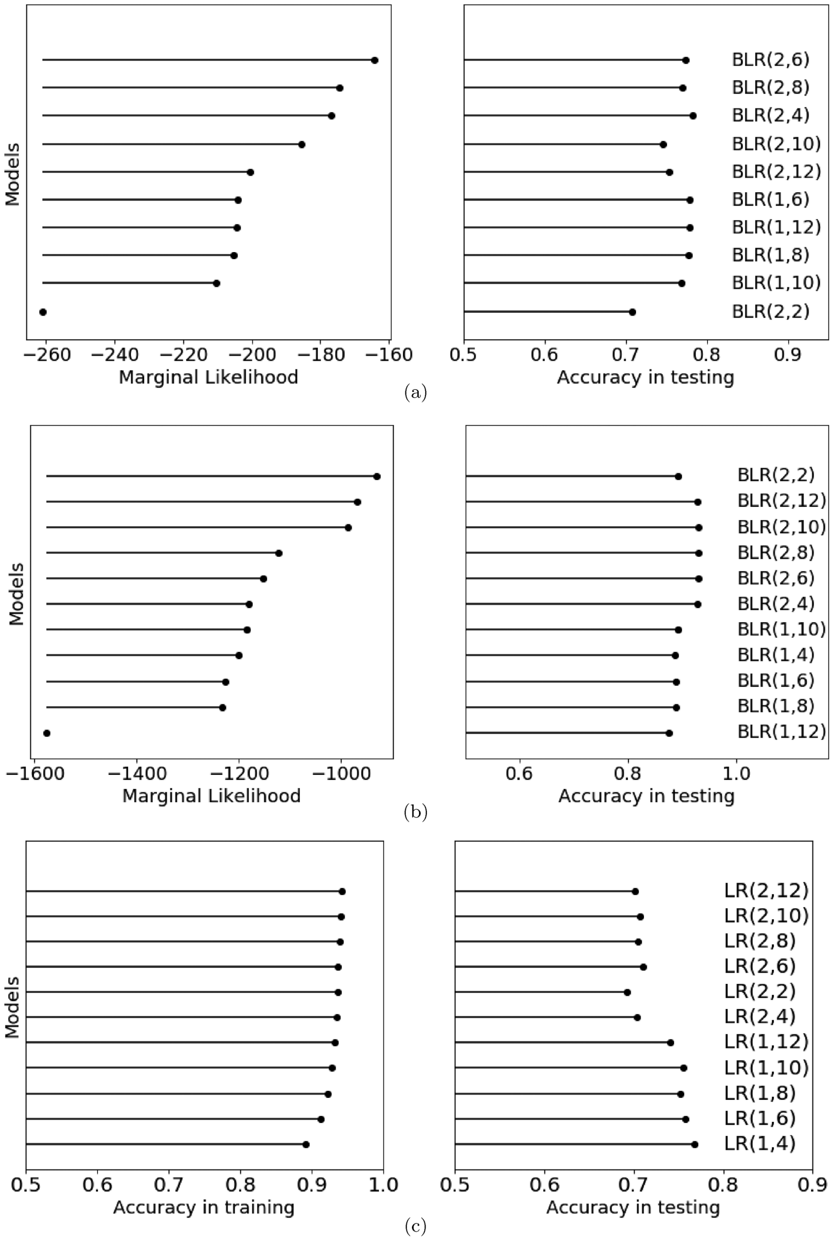
(i) **Kendall-tau (τ):** this metric is estimated by $\frac{n_\mathrm{c} - n_\mathrm{d}}{\frac{1}{2}n(n-1)}$, where n_c represents the number of concordant models and n_d is the number of discordant models in the ranking. It identifies the coincidences between training and testing rankings. Both rankings (training and testing) are concordant when the selection is correct. In preliminary experiments, similar results were obtained using the Spearman's rank coefficient.

(ii) **Average top-3-Accuracy (A):** In order to discriminate beyond the rankings, we also use the average accuracy (in test) over the three best models for each ranking. Thus, we can identify the quality of the selected models. This metric provides a perspective about how good are the models prioritized by each strategy. Note that the best model accuracy can be a harsh metric in our context (ranking of models). On the other hand, the average performance over a family of models is a poorly informative measure.

(iii) **Average top-3-F_1-score (F_1):** As was explained in Section 2.1.1, this metric is a better option than A in case of having unbalanced classes. To assess the performance of each method, we estimate the average F_1-score in testing for the three foremost models in the ranking.

(iv) **Delta training/testing (Δ_T):** This metric seeks to evaluate how far, on average, the predicted accuracy for the testing set is with respect to the training set. We report the average distance over the set of models in each family.

In Tables 4–6, we summarize the metrics for each data set (rrlyrae-1, rrlyrae-2, and rrlyrae-3), considering three subset sizes s (for training). As was commented in Section 5.4, we run three baseline strategies to rank models in addition to three approaches based on the marginal likelihood. The baseline approaches (LR, l_2-LR-1, and l_2-LR-100) are based on k-fold CV with $k = 10$. The rankings based on the marginal likelihood consider flat priors (BLR-FP),



(a)

(b)

(c)

Figure 4. Comparison of model rankings with 1000 samples on rrlyrae-3 set. (a) Models sorted by the marginal likelihood **BLR-IP**. (b) Models sorted by the marginal likelihood **BLR-IP**($\sigma = 10$). (c) Models sorted by a cross-validated ($k = 10$) accuracy for the l_2-**LR**-1 family of models. Let **BLR**(p, c) and **LR**(p, c) be classifiers that is defined by an $n \times (p \times c)$ input matrix, where r represents the retained principal components and p the polynomial transformation degree.

informative priors for the mean and fixed variance (BLR-IP $\sigma =$ 10), and informative priors on both the mean and variance (BLR-IP).

From Tables 4–6, we can also observe that τ values are greater for marginal-likelihood-based rankings than those based on k-fold CV. These results demonstrate empirically that the marginal likelihood is more robust than the cross-validated A for assessing and prioritizing RR Lyrae star classifiers under different levels of bias. In fact, according to τ values, when looking at rrlyrae-1 and rrlyrae-2 results, the best rankings were provided by BLR-IP (2000), while the best ranking for rrlyrae-3 was obtained using BLR-IP (1000).

Concerning F_1-score and A for the three sets, the best three informative Bayesian models obtain a better performance than the best three likelihood-based models. This means that the predictive performance of posterior samples (posterior mean) is also improved when we add prior knowledge from DRs. When looking at experiments with rrlyrae-1, it is worth noting that the difference is more significant in the A metric.

When looking the impact of bias (managed by T in Algorithm 2) on A and Δ_T, we can note that in highly biased data sets (rrlyrae-1 and rrlyrae-2) our proposal is the best alternative, but even BMS is competitive in the less biased set (rrlyrae-3).

Table 6 shows that, given a smaller bias, the six alternatives obtain better A. However, in this set, this metric is poorly informative due to the imbalance problem in this testing set (10 per cent of RR Lyraes, see Table 3). In spite of that, it also shows a better F1-score value for those models selected by the informative marginal likelihood (BLR-IP). Figs 4(a), (b) and (c) show the rankings provided by **BLR-IP** (σ=10), **BLR-IP** and l_2-**LR**-1 for this experiment (data set rrlyrae-3

Table 4. Evaluation of rankings of models in rrlyrae-1. τ is the Kendall's tau rank correlation; A and F1 are the mean Accuracy and the mean F1-score, respectively; of the three foremost models, Δ_T is the average difference between the accuracy in training and testing. The bold numbers represent the best strategy for model selection by each metric.

$\mathcal{M}$	s	τ	F_1	A	Δ_T
		k-fold CV			
LR	1000	0.11	0.62	0.62	0.35
	2000	0.09	0.62	0.63	0.34
	4000	0.26	0.63	0.65	0.32
l_2-LR-100	1000	−0.31	0.64	0.64	0.35
	2000	−0.09	0.64	0.63	0.35
	4000	0.24	0.68	0.66	0.32
l_2-LR-1	1000	−0.16	0.64	0.64	0.35
	2000	−0.02	0.63	0.63	0.34
	4000	0.49	0.69	0.66	0.32
		Marginal likelihood			
BLR-FP	1000	0.70	0.70	0.69	0.32
	2000	0.82	0.70	0.69	0.32
	4000	0.85	0.70	0.69	0.31
BLR-IP (σ=10)	1000	0.31	0.68	0.69	0.32
	2000	0.56	0.70	0.69	0.32
	4000	0.75	0.70	0.69	0.31
BLR-IP	1000	0.60	0.50	0.61	0.24
	2000	0.85	0.65	0.66	0.34
	4000	0.71	0.69	0.68	0.32

Table 5. As in Table 4, but for the case of rrlyrae-2.

$\mathcal{M}$	s	τ	F_1	A	Δ_T
		k-fold CV			
LR	1000	0.13	0.67	0.64	0.32
	2000	0.38	0.66	0.66	0.32
	4000	0.38	0.63	0.65	0.32
l_2-LR-100	1000	0.27	0.65	0.64	0.32
	2000	0.24	0.65	0.63	0.33
	4000	0.45	0.65	0.63	0.34
l_2-LR-1	1000	0.42	0.66	0.65	0.32
	2000	0.27	0.65	0.63	0.33
	4000	0.45	0.65	0.63	0.34
		Marginal likelihood			
BLR-FP	1000	0.71	0.68	0.66	0.32
	2000	0.64	0.68	0.66	0.33
	4000	0.76	0.69	0.67	0.32
BLR-IP (σ=10)	1000	0.20	0.68	0.67	0.33
	2000	0.61	0.68	0.66	0.33
	4000	0.73	0.68	0.67	0.32
BLR-IP	1000	0.60	0.49	0.59	0.24
	2000	0.82	0.68	0.66	0.33
	4000	0.78	0.68	0.67	0.33

Table 6. As in Table 4, but for the case of rrlyrae-3.

$\mathcal{M}$	s	τ	F_1	A	Δ_T
		k-fold CV			
LR	1000	−0.75	0.46	0.76	0.19
	2000	−0.53	0.45	0.75	0.20
	4000	−0.05	0.57	0.85	0.12
l_2-LR-100	1000	−0.75	0.46	0.76	0.20
	2000	−0.78	0.44	0.74	0.22
	4000	−0.42	0.51	0.80	0.16
l_2-LR-1	1000	−0.71	0.46	0.76	0.20
	2000	−0.78	0.44	0.74	0.22
	4000	−0.49	0.51	0.80	0.17
		Marginal likelihood			
BLR-FP	1000	−0.16	0.49	0.78	0.19
	2000	−0.36	0.46	0.76	0.22
	4000	−0.39	0.46	0.76	0.21
BLR-IP (σ=10)	1000	0.09	0.49	0.79	0.18
	2000	−0.30	0.47	0.76	0.22
	4000	−0.33	0.46	0.76	0.21
BLR-IP	1000	0.56	0.71	0.93	−0.25
	2000	−0.53	0.47	0.76	0.22
	4000	−0.53	0.46	0.76	0.22

and s = 1000). When looking at Fig. 4(c), we can observe that the cross-validated Accuracy is unable to prioritize models correctly; in fact, this approach selects the worst model in this case.

To sum up, when there are biased labelled objects and we have expert knowledge, our scheme provides an excellent alternative to select models. Note that BLR-IP (s = 1000) obtains the best Δ_T in the three sets; this means that the reported A is more reliable. It comes from informative priors on small training sets can penalize the performance highly in training since this expert knowledge helps to limit the likelihood function (based on data), but when we use them, the selected models are more robust to biases.

8 CONCLUSIONS

We have presented a novel approach to assess and sort models considering expert knowledge. The method is based on the design of informative priors using deterministic physical rules that allow us to estimate an informative marginal likelihood, which includes well-known properties like a model selector. The method offers a good alternative for selecting variable star classifier with biased (and/or small) sets of labelled objects. This gives rise to an original and simple methodology to add prior knowledge in the model assessment process of RR Lyrae star classifiers without having to undergo a time-consuming adaptation process.

For evaluation purposes, we have designed a method capable of introducing bias to a data set according to the classification difficulty of each object. This allows us to test different strategies to assess models under three conditions of bias. The results show that the informative marginal likelihood is able to identify more suitable models than non-informative cross-validated metrics.

Future work can consider extensions such as (i) the use of other types of informative priors, e.g. priors over the proportion of classes or heavy-tailed distributions over BLR's weights; (ii) analysis of other time-series survey data sets; (iii) and the application of this approach to other classes (or subtypes) of variable stars.

ACKNOWLEDGEMENTS

We acknowledge the support from CONICYT-Chile, through the FONDECYT Regular project number 1180054. F. Pérez-Galarce acknowledges the support from National Agency for Research and Development (ANID), through Scholarship Program/Doctorado Nacional/2017-21171036. Support for M. Catelan is provided by ANID's Millennium Science Initiative through grant ICN12_120009, awarded to the Millennium Institute of Astrophysics (MAS); by Proyecto Basal AFB-170002; and by FONDECYT grant #1171273. P. Huijse acknowledges financial support from project ANID's project PAI 79170017.

DATA AVAILABILITY STATEMENT

The list of objects in each set (rrlyrae-1, rrlyrae-2, and rrlyrae-3) and their descriptors (63 features from FATS Nun et al. (2015)) will be shared on reasonable request to the corresponding author.

REFERENCES

Aguirre C., Pichara K., Becker I., 2019, MNRAS, 482, 50
Alcock C. et al., 1997, ApJ, 486, 697
Anderson D., Burnham K., 2004, Second NY: Springer-Verlag, 63, 10
Arlot S. et al., 2010, Stat. Surv., 4, 40
Becker I., Pichara K., Catelan M., Protopapas P., Aguirre C., Nikzat F., 2020, MNRAS, 493, 2981
Benavente P., Protopapas P., Pichara K., 2017, ApJ, 845, 147
Blei D. M., Kucukelbir A., McAuliffe J. D., 2017, J. Am. Stat. Assoc., 112, 859
Bloemen S. et al., 2016, Proc. SPIE , 9906, 990664
Bloom J. et al., 2012, PASP, 124, 1175
Booth R., Jonas J., 2012, Afr. Skies, 16, 101
Budavári T., Szalay A., Loredo T., 2017, ApJ, 838, 52
Cabrera G. F., Miller C. J., Schneider J., 2014, in Helen J., ed., Proc. SPIE, 22nd International Conference on Pattern Recognition. SPIE, Bellingham, p. 4417
Carrasco-Davis R. et al., 2019, PASP, 131, 108006
Castro N., Protopapas P., Pichara K., 2017, AJ, 155, 16
Catelan M., Smith H., 2015, Pulsating Stars. Wiley-VCH, Weinheim
Christensen N., Meyer R., 1998, Phys. Rev. D, 58, 082001
Christensen N., Meyer R., Knox L., Luey B., 2001, Class. Quantum Gravity, 18, 2677
Debosscher J. et al., 2009, A&A, 506, 519
Debosscher J., Sarro L., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Efron B., Tibshirani R., 1997, J. Am. Stat. Assoc., 92, 548
Elorrieta F. et al., 2016, A&A, 595, A82
Ford E., Gregory P., 2007, in Babu G. J., Feigelson E. D., eds, ASP Conference Series, Vol. 371, Statistical Challenges in Modern Astronomy IV. Pennsylvania State University, Pennsylvania, USA
Gelman A., Rubin D., 1992, Stat. Sci., 7, 457
Gelman A. et al., 2008, Ann. Appl. Stat., 2, 1360
Gelman A., Simpson D., Betancourt M., 2017, Entropy, 19, 555
Ghahramani Z., 2013, Phil. Trans. R. Soc. A, 371, 20110553
Golchi S., 2019, Stat. Anal. Data Mining: The ASA Data Sci. J., 12, 45
Gregory P., Loredo T., 1992, ApJ, 398, 146
Gronau Q. et al., 2017, J. Math. Psychol., 81, 80
Grunwald P., 2004, preprint (arXiv:math/0406077)
Hanson T. E. et al., 2014, Bayesian Anal., 9, 597
Hogg D., Foreman-Mackey D., 2018, ApJS, 236, 11
Kim D., Bailer-Jones C., 2016, A&A, 587, A18
Kohavi R. et al., 1995, Ijcai. No. 2 in 14. Morgan Kaufmann Publishers Inc., San Francisco, CA, p. 1137
Lendasse A., Wertz V., Verleysen M., 2003, in Artificial Neural Networks and Neural Information Processing-ICANN/ICONIP 2003. Springer-Verlag, Berlin, p. 573
MacKay D., 1992, PhD thesis. California Institute of Technology
Mackenzie C., Pichara K., Protopapas P., 2016, ApJ, 820, 138
Mahabal A., Sheth K., Gieseke F., Pai A., Djorgovski S., Drake A., Graham M., 2017, Symposium Series on Computational Intelligence (SSCI). IEEE, p. 1
Masci F., Hoffman D., Grillmair C., Cutri R., 2014, AJ, 148, 21
McKinney W. et al., 2011, Python for High Performance and Scientific Computing, International Conference for High Performance Computing, Networking, Storage, and Analysis. University of Rochester, New York
Meng X., Wong W., 1996, Stat. Sin., 831
Murray I., Ghahramani Z., 2005, A Note On The Evidence And Bayesian Occam'S Razor, Gatsby Unit
Myung I. J., Pitt M. A., 1997, Psychonomic Bull. Rev., 4, 79
Narayan G. et al., 2018, ApJS, 236, 9
Naul B., Bloom J. S., Pérez F., van der Walt S., 2018, Nat. Astron., 2, 151
Neal R., 2001, Stat. Comput., 11, 125
Nun I., Pichara K., Protopapas P., Kim D., 2014, ApJ, 793, 23
Nun I., Protopapas P., Sim B., Zhu M., Dave R., Castro N., Pichara K., 2015, preprint (arXiv:1506.00010)
Overstall A., Forster J., 2010, Comput. Stat. Data Anal., 54, 3269
Parviainen H., Deeg H., Belmonte J., 2013, A&A, 550, A67
Pedregosa F. et al., 2011, J. Mach. Learning Res., 12, 2825
Pichara K., Protopapas P., Kim D., Marquette J., Tisserand P., 2012, MNRAS, 427, 1284
Pichara K., Protopapas P., León D., 2016, ApJ, 819, 18
Raftery A. E., Satagopan J, Krivitsk P, Newton M, 2006, Estimating the Integrated Likelihood via Posterior Simulation Using The Harmonic Mean Identity, Bayesian statistics, 8, 1
Raileanu L. E., Stoffel K., 2004, Ann. Math. Artif. Intell., 41, 77
Rao R., Fung G., Rosales R., 2008, in Apte Ch, Park H, Wang K, Zaki M, eds, Proceedings of the 2008 SIAM International Conference on Data Mining, On the Dangers of Cross-validation. An Experimental Evaluation. p. 588
Rasmussen C., Ghahramani Z., 2001, Advances in Neural Information Processing Systems, Occam's razor. NIPS, Denver, US, p. 294
Richards J. et al., 2011a, ApJ, 733, 10
Richards J. et al., 2011b, ApJ, 744, 192
Richards J., 2012, Astrostatistics and Data Mining. Springer-Verlag, Berlin, p. 213
Rubin D., 1981, The Annals of Statistics, 9, 30
Ruffio J. et al., 2018, AJ, 156, 196
Saha P., Williams T., 1994, AJ, 107, 1295
Salvatier J., Wiecki T. V., Fonnesbeck C., 2016, Peer J. Comput. Sci., 2, e55
Sanders J. L., Das P., 2018, MNRAS, 481, 4093
Schwarz G. et al., 1978, Ann. Stat., 6, 461
Sesar B. et al., 2013, AJ, 146, 21
Sharma S., 2017, ARA&A, 55, 213
Sokolova M., Lapalme G., 2009, Inf. Process. Manag., 45, 427
Sooknunan K. et al., 2021, MNRAS, 502, 206
Spyroglou I., Spöck G., Chatzimichail E., Rigas A., Paraskakis E., 2018, A Bayesian Logistic Regression approach in Asthma Persistence Prediction, Epidemiology, Biostatistics and Public Health, p. 15
Sugiyama M., Krauledat M., MÃžller K., 2007, J. Mach. Learn. Res., 8, 985
Trotta R., 2008, Contemporary Phys., 49, 71
Udalski A., Szymanski M., Soszynski I., Poleski R., 2008, Acta Astron., 58, 69
Valenzuela L., Pichara K., 2017, MNRAS, 474, 3259
Van Laarhoven P. J., Aarts E. H., 1987, Simulated Annealing: Theory and Applications. Springer-Verlag, Berlin, p. 7
Wang Y., Chen M., Kuo Land Lewis P., 2018, Bayesian Anal., 13, 311
Waskom M. et al., 2014, Availablle at: https://seaborn.pydata
Watanabe S., 2013, J. Mach. Learn. Res., 14, 867
Weinberg M., 2013, MNRAS, 434, 1736
Wright E. L. et al., 2010, AJ, 140, 1868
Zorn C., 2005, Political Anal., 13, 157

APPENDIX: PROOF OF BRIDGE SAMPLING ESTIMATOR

The proposed bridge sampling approach (Gronau et al. 2017) begins with the following identity:

$$1 = \frac{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)g(\theta)\mathrm{d}\theta}{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)g(\theta)\mathrm{d}\theta}, \tag{A1}$$

where $g(\theta)$ is the proposal distribution. Subsequently, it is multiplied by the marginal likelihood on both sides. Then, we obtain the following equation:

$$p(\mathcal{D}) = \frac{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)g(\theta)\mathrm{d}\theta}{\int \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}h(\theta)g(\theta)\mathrm{d}\theta}. \tag{A2}$$

Note that the posterior distribution appears on the right side's denominator. After that, by means of

$$p(\mathcal{D}) = \frac{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)}{\int h(\theta)g(\theta)}\frac{g(\theta)\mathrm{d}\theta}{p(\theta|\mathcal{D})\mathrm{d}\theta}, \tag{A3}$$

the right-hand side of A2 is separated into two ratios, and consequently, we can obtain the expected values in the denominator and numerator as follows:

$$p(\mathcal{D}) = \frac{\mathbb{E}_{g(\theta)}\left[p(\mathcal{D}|\theta)p(\theta)h(\theta)\right]}{\mathbb{E}_{p(\theta|\mathcal{D})}\left[h(\theta)g(\theta)\right]}. \tag{A4}$$

Finally, we use the definition of optimal bridge function provided by Meng & Wong (1996):

$$h(\theta) = C\frac{1}{s_1 p(\mathcal{D}|\theta)p(\theta) + s_2 p(\mathcal{D})g(\theta)}. \tag{A5}$$

Due to the obtained estimator depends recursively on the marginal likelihood, the iterative scheme presented below is applied:

$$\hat{p}(\mathcal{D})^{t+1} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\frac{p(\mathcal{D}|\theta_i)p(\theta_i)}{s_1 p(\mathcal{D}|\theta_i)p(\theta_i)+s_2\hat{p}(\mathcal{D})^t g(\theta_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1}\frac{g(\theta_j)}{s_1 p(\mathcal{D}|\theta_j)p(\theta_j)+s_2\hat{p}(\mathcal{D})^t g(\theta_j)}}. \tag{A6}$$

This paper has been typeset from a TEX/LATEX file prepared by the author.