Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **487,** 3644–3649 (2019)
Advance Access publication 2019 June 04

doi:10.1093/mnras/stz1537

Bayesian cross validation for gravitational-wave searches in pulsar-timing array data

Haochen Wang,[1] Stephen R. Taylor[2]★ and Michele Vallisneri[3]

[1]*Department of Physics and Astronomy, University of Southern California, Los Angeles, CA 90089, USA*
[2]*TAPIR, MC 350-17, California Institute of Technology, Pasadena, CA 91125, USA*
[3]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA*

Accepted 2019 May 28. Received 2019 May 24; in original form 2019 April 10

ABSTRACT

Gravitational-wave data analysis demands sophisticated statistical noise models in a bid to extract highly obscured signals from data. In Bayesian model comparison, we choose among a landscape of models by comparing their marginal likelihoods. However, this computation is numerically fraught and can be sensitive to arbitrary choices in the specification of parameter priors. In Bayesian cross validation, we characterize the fit and predictive power of a model by computing the Bayesian posterior of its parameters in a training data set, and then use that posterior to compute the averaged likelihood of a different testing data set. The resulting cross-validation scores are straightforward to compute; they are insensitive to prior tuning; and they penalize unnecessarily complex models that overfit the training data at the expense of predictive performance. In this article, we discuss cross validation in the context of pulsar-timing-array data analysis, and we exemplify its application to simulated pulsar data (where it successfully selects the correct spectral index of a stochastic gravitational-wave background), and to a pulsar data set from the NANOGrav 11-yr release (where it convincingly favours a model that represents a transient feature in the interstellar medium). We argue that cross validation offers a promising alternative to Bayesian model comparison, and we discuss its use for gravitational-wave detection, by selecting or refuting models that include a gravitational-wave component.

Key words: gravitational waves – methods: statistical – pulsars: general.

1 INTRODUCTION

Searches for gravitational waves (GWs) in pulsar-timing-array (PTA) data (Burke-Spolaor 2015; Lommen 2015) seek to identify weak GW signals among a plethora of other effects, including deterministic delays due to the relative motion of pulsar and observatory and to pulsar binary dynamics, stochastic delays due to the interplanetary and interstellar media, as well as intrinsic irregularities in the pulsar's period emission (Cordes 2013; Stinebring 2013). These searches are commonly formulated as Bayesian-inference problems (Gregory 2010), whereby we derive the joint posterior probability density of the GW parameters and of the noise parameters of all analysed pulsars. Choosing appropriate probabilistic models for pulsar noise is therefore crucial to reliable PTA searches (Cordes & Shannon 2010; Taylor, Gair & Lentati 2013; Lentati et al. 2016): unmodelled noise components may be interpreted as GWs, while overgenerous noise assumptions may reduce GW sensitivity. In current practice, pulsar noise models are informed by the physics of millisecond pulsars and of the interplanetary/interstellar medium, but they are largely driven by inference from PTA data sets, since these often represent the best observations to date for PTA pulsars.

2 MODEL COMPARISON

Within the data-analysis practice of the NANOGrav collaboration (Aggarwal et al. 2018; Arzoumanian et al. 2018b; NANOGrav 2019), *Bayesian model comparison* (Gregory 2010) is used to select the noise model most appropriate to each pulsar (Arzoumanian et al., in preparation). The goal is not only to improve the physical characterization of the processes affecting pulse times of arrival (TOAs), but also to isolate these processes from a putative GW signal with greater confidence. In this framework, we evaluate the fully marginalized likelihood (a.k.a. evidence) for each model M:

$$p(y|M) = \int p(y|\theta_M)p(\theta_M)\,\mathrm{d}\theta_M, \tag{1}$$

where y denotes the observed data, θ_M the parameters of model M, $p(y|\theta_M)$ the likelihood (the probability of y given θ_M), and $p(\theta_M)$ the prior probability density assigned to the parameters. We then

★ E-mail: srtaylor@caltech.edu

compare models by evaluating Bayes ratios[1] $B_{21} = p(y|M_2)/p(y|M_1)$, either directly through equation (1) (often requiring significant numerical sophistication, see Trotta 2008) or by the Monte Carlo exploration of uber-likelihoods that specialize to individual models depending on the value of an index parameter (in which case the Bayes ratio is given by the ratio of the 'time' spent in each model, see Godsill 2001; Sisson 2005). A large Bayes ratio B_{21} implies that the data favours model M_2 over M_1. However, it is difficult to give Bayes ratios a principled *quantitative* interpretation. The exception are cases where alternative models represent exclusive physical outcomes; Bayes ratios can then be calibrated in terms of statistical decision theory, by relating their sampling distribution to false-alarm and false-dismissal probabilities (see e.g. Vallisneri 2012).

In choosing between alternative models for a data set, we need to be wary of *overfitting*: that is, while it is always possible to improve model fit by adding parameters, the enhanced model may end up conforming to contingent noise features instead of highlighting the physical properties of interest. Correspondingly, the model loses predictive power for yet-to-observed data. Bayesian model comparison incorporates a defense against overfitting, in that the evidence integral penalizes fine tunings that restrict parameters to small regions within their prior ranges. Unfortunately, this defense creates a different, significant weakness – Bayes ratios are then sensitive to parameter-prior assignments that may be largely arbitrary, and are not testable from data (Gelman et al. 2013). Consider for instance the case of two nested models $M_1 \subset M_2$, where M_2 is obtained by adding parameter θ^* to M_1; let θ^* have uniform prior in $[-a, a]$, with a arbitrarily assigned without cogent physical grounds; finally, let the data constrain θ^* to a small range close to 0. It is then easy to see that $B_{21} \propto 1/a$.

3 CROSS VALIDATION

An alternative framework for model comparison is offered by measures of *predictive performance*, which quantify how well a model that has been fit to data set y can predict yet-to-be-observed data $\tilde{y}$ (Gelman et al. 2013). These measures penalize overfitting by construction, because a model that conforms to contingent features in y will usually do worse in fitting $\tilde{y}$. In a Bayesian setting, a commonly adopted measure of predictive performance is the *log predictive density* for the new data $\tilde{y}$, as induced by the posterior $p(\theta_M|y)$

$$\log p(\tilde{y}|y; M) = \log \int p(\tilde{y}|\theta_M)\, p(\theta_M|y)\, \mathrm{d}\theta_M . \tag{2}$$

Ideally, we would average $\log p(\tilde{y}|y; M)$ over the true distribution of future data $\tilde{y}$; doing so is however seldom possible. In practice, we can: (a) estimate *within-sample* predictive accuracy using the data y that we already have, by applying corrections for the overfitting bias, as in the various 'information criteria' (Gelman et al. 2013); (b) evaluate *out-of-sample* predictive accuracy on one or more *holdout* data sets that were not used to infer parameter posteriors. The latter approach is known as *cross validation*, and we will pursue it for PTA data in the rest of this article.

Specifically, we adopt *k-fold* cross validation as follows:

[1] Bayesian model comparison calls for the computation of *odds ratios*, which account for the prior relative probability of entire models. Since it is very difficult to attribute such priors on physical grounds, we generally work directly with Bayes ratios.

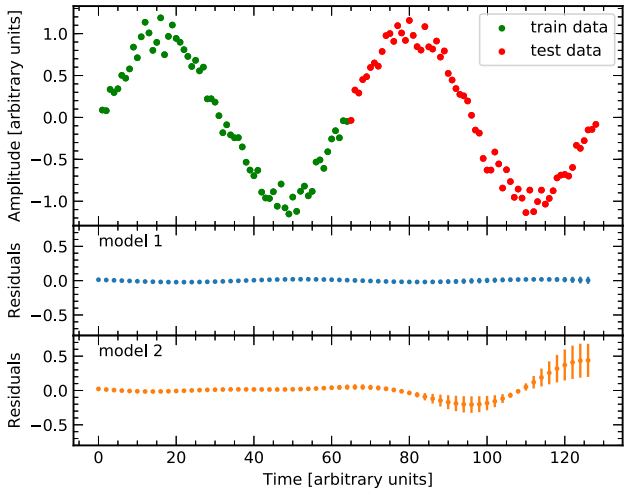

Figure 1. Cross-validation analysis of a toy model. We generate the data d (upper panel) from $f_1(A, \omega, \varphi) = A\sin(\omega t + \varphi)$ (model 1), with the addition of white noise; we then analyse d with f_1 as well as $f_2(A, \omega, \varphi, b) = A\sin((\omega + bt)t + \varphi)$ (model 2), where parameter b describes an additional frequency drift. We train both models on the first half of the data (the training set d^{train}, green in the plot), deriving the Bayesian posterior distributions $p(A, \omega, \varphi|d^{\mathrm{train}}, f_1)$ and $p(A, \omega, \varphi, b|d^{\mathrm{train}}, f_2)$. Using those posteriors with equation (3) over the second half of the data (the validation set d^{test}, red in the plot), we obtain log predictive densities -27.7 and -31.3 for models 1 and 2, respectively. The lower density for model 2 indicates that it overfits the training data, resulting in a poor fit to the validation set. The conclusion is borne out by inspection of model residuals (lower panels), represented in the plot as their averages and standard deviations over 300 posterior draws. Model-1 residuals are small and homogeneous across both training and validation sets; model-2 residuals have much larger bias and variance over the validation set, explaining the lower predictive density.

(i) We divide the data set y randomly in k exclusive subsets $y^{(k)}$ (the *testing* data sets);

(ii) For each k, we derive the posterior $p(\theta_M|y^{(-k)})$, where $y^{(-k)} = \cup_{j\neq k} y^{(j)}$ is the *training* data set corresponding by omitting $y^{(k)}$. We represent posteriors as sequences $\{\theta^{(-k)}_{M,i}\}$ of N quasi-independent samples, obtained by Monte Carlo methods;

(iii) For each k, we evaluate the log predictive density $\log p(y^{(k)}|y^{(-k)})$, given by

$$\log \int p(y^{(k)}|\theta_M)\, p(\theta_M|y^{(-k)})\, \mathrm{d}\theta \simeq \log \frac{1}{N}\sum_{i=1}^{N} p(y^{(k)}|\theta^{(-k)}_{M,i}); \tag{3}$$

(iv) We repeat this procedure for every model under consideration, and then compare the respective log predictive densities, averaged over the k repetitions. The variance of the densities is a measure of their statistical uncertainty. The choice of number of repetitions k will be somewhat dependent on the time-scale of processes in a given model under consideration. PTA experiments are interested in signals and processes that span time-scales comparable to the duration of the data, for which we need a large subset of the data to initially train our models. Thus we employ small k ($\sim$2–3) in the following.

In Fig. 1, we exemplify this process with a simple toy problem: a sinusoidal signal parametrized by amplitude, frequency, and phase, as modelled by that very model and by an expanded model that includes a linear frequency drift. The more complicated model re-

sults in significantly lower log predictive density, demonstrating that cross validation can recognize and reject overfitting. Furthermore, cross validation avoids the problem of arbitrary parameter prior assignments by integrating the likelihood of the testing data over the informative parameter posterior distribution under the trained model, rather than over an uninformative prior distribution (as in the case of Bayes ratios).

4 CROSS VALIDATION FOR SINGLE-PULSAR NOISE MODELING

In current practice (see van Haasteren & Vallisneri 2014 for a recent review), probabilistic noise models for pulsars are built as the sum of a number of Gaussian processes (GPs; Rasmussen & Williams 2006) representing all sources of correlated noise: errors in the parameters of the deterministic timing model, pulsar-rotation irregularities, dispersion-measure (DM) variations along the pulse propagation path, jitter-like noise in multifrequency observations, and more. The TOAs are also subject to 'white' radiometer measurement noise, conceptualized as independent and heteroskedastic normal variates. Both the GPs and measurement noise are governed by a set of hyperparameters (e.g. the amplitude and spectral slope of delays due to rotation irregularities) that are estimated from PTA data sets.

In keeping with basis–kernel duality for GPs, the model likelihood can be written in two complementary ways. In *hierarchical form*, the likelihood is given by

$$p(y|\eta_N, \eta_{\rm GP}, c_{\rm GP}) = p(y|\eta_N, c_{\rm GP}) \times p(c_{\rm GP}|\eta_{\rm GP}) = \frac{e^{-(y-F_{\rm GP}c_{\rm GP})^T N^{-1}(y-F_{\rm GP}c_{\rm GP})/2}}{\sqrt{(2\pi)^n |N|}} \times \frac{e^{-c_{\rm GP}^T \Phi_{\rm GP}^{-1} c_{\rm GP}/2}}{\sqrt{(2\pi)^m |\Phi|}}, \quad (4)$$

where y is the vector of n *timing residuals* obtained by subtracting the best-fittng timing model from the observed pulse TOA; the $n \times m$ matrix $F_{\rm GP}$ collects the m GP basis vectors, and the $c_{\rm GP}$ are the corresponding weights (or coefficients); N (a function of the hyperparameters η_N) is a diagonal matrix expressing measurement-noise variance; and $\Phi_{\rm GP}$ (a function of the hyperparameters $\eta_{\rm GP}$) represents the normal priors for the GP weights. In *marginalized form*, we eliminate the dependence on the GP weights by integrating over them (van Haasteren & Vallisneri 2014):

$$p(y|\eta_N, \eta_{\rm GP}) = \int p(y|\eta_N, c_{\rm GP}, \eta_{\rm GP})\, {\rm d}c_{\rm GP} = \frac{e^{-y^T (N + F_{\rm GP}\Phi_{\rm GP}F_{\rm GP}^T)^{-1} y/2}}{\sqrt{(2\pi)^n |N + F_{\rm GP}\Phi_{\rm GP}F_{\rm GP}^T|}}. \quad (5)$$

The marginalized form is usually employed for the Monte Carlo exploration of hyperparameter posteriors. The GP weights can still be characterized by way of their conditional posterior given the data and the hyperparameters, which follows the jointly normal distribution $p(c_{\rm GP}|y; \eta_N, \eta_{\rm GP}) = \mathcal{N}(\bar{c}, \Sigma)$ with mean

$$\bar{c}(y; \eta_N, \eta_{\rm GP}) = \Sigma F_{\rm GP}^T N^{-1} y \quad (6)$$

and covariance

$$\Sigma(\eta_N, \eta_{\rm GP}) = (\Phi_{\rm GP}^{-1} + F_{\rm GP}^T N^{-1} F_{\rm GP})^{-1}. \quad (7)$$

Armed with equations (4)–(7), we perform steps 1–3 of k-fold cross validation for a single-pulsar data set as follows:

(i) We partition the timing residuals y into exclusive testing data sets $y^{(k)}$, making sure that each training data set $y^{(-k)}$ depends on all weights and hyperparameters. For instance, for DM variations described as piecewise-constant 'DMX' functions, each DMX epoch must be populated by at least one residual in every $y^{(k)}$; likewise, for 'EFAC' measurement noise that is rescaled differently in each radio backend, each backend must be represented in every $y^{(k)}$;

(ii) We sample the marginalized hyperparameter posterior $p(\eta_N, \eta_{\rm GP}|y^{(-k)})$ (proportional to equation 5 times prior $p(\eta_N, \eta_{\rm GP})$, using the PTA data-analysis package ENTERPRISE[2] and the Markov Chain Monte Carlo sampler PTMCMCSAMPLER;[3]

(iii) For each of the N quasi-independent $(\eta_{N,i}^{(-k)}, \eta_{{\rm GP},i}^{(-k)})$ obtained at step 2, we draw P weight vectors $\{c_{{\rm GP},ij}^{(-k)}\}$ from their conditional distribution (equations 6–7), then we evaluate the log predictive density by averaging the hierarchical likelihood (4) over the $N \times P$ triples $(\eta_{N,i}^{(-k)}, \eta_{{\rm GP},i}^{(-k)}, c_{{\rm GP},ij}^{(-k)})$.

The 'representation' condition imposed in step 1 is necessary so that the parameters for which we derive posteriors at step 2 fully specify the model's prediction for each testing data set; this prediction is used in step 3 to evaluate the predictive density. Another important technical subtlety is that the GP weights must conserve the same identity across all $y^{(-k)}$ (for example, Fourier coefficients for the same set of frequencies in the case of correlated spin noise); by contrast, the GP basis vectors would change in value, because they refer to different time-of-arrival measurements (continuing our example, the basis elements would be sines and cosines of the same frequencies for all subdatasets, but would be evaluated at different times).

We note also that in step 2 we could have used the hierarchical likelihood to sample the full parameter set $(\eta_N, \eta_{\rm GP}, c_{\rm GP})$, avoiding the conditional $c_{\rm GP}$ draws at step 3. However, the hierarchical likelihood is considerably harder to explore stochastically. Also, in step 3 the sum over the $c_{{\rm GP},ij}^{(-k)}$ for each i could be replaced by analytical integration in terms of $\bar{c}$ and Σ, at the cost of some algebraic complication (see below for a related development).

5 RESULTS

To demonstrate how cross validation can be applied to PTA noise-model selection, we first consider the simulated TOA residuals for pulsar J1713+0747 from the 1st International Pulsar Timing Array (IPTA) Data Challenge (Hazboun, Mingarelli & Lee 2018).[4] These residuals (130 values at 14-d cadence) are generated from a simplified deterministic timing model and a simple noise model: the timing model parameters describe the intrinsic-spin, astrometry, and binary-orbit properties of the pulsar; the noise model includes white measurement noise (described by 'EFAC' and 'EQUAD' parameters) and red correlated spin noise, specified by a power-law spectrum (Arzoumanian et al. 2018b)

$$P(f) = A_{\rm red}^2 \left(\frac{f}{f_{\rm yr}}\right)^{-\gamma_{\rm red}}, \quad (8)$$

where $A_{\rm red}$ is the amplitude of the red-noise process in units of $\mu s \times {\rm yr}^{1/2}$, $\gamma_{\rm red}$ is its spectral index (set to 13/3 to generate the data), and $f_{\rm yr} = 1\,{\rm yr}^{-1}$. We compare power-law red-noise models with different $\gamma_{\rm red}$ by evaluating their respective cross-validation predictive densities, shown in Fig. 2 for $\gamma_{\rm red}$ ranging from 1 to 6.

[2] https://github.com/nanograv/enterprise
[3] https://github.com/jellis18/PTMCMCSampler
[4] http://ipta4gw.org/data-challenge

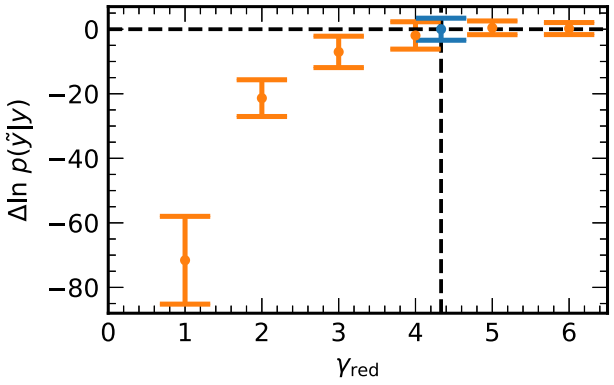

Figure 2. Cross-validation analysis of power-law models for pulsar red noise, as demonstrated in the J1713+0747 data set from the 1st IPTA Mock Data Challenge. The data set includes GW-like correlated noise with spectral slope $\gamma_{\rm red} = 13/3$. We perform two-fold cross validation using power-law models with integer $\gamma_{\rm red} \in [1, 6]$, as well as the correct $\gamma_{\rm red} = 13/3$. For each $\gamma_{\rm red}$, we plot the average and standard deviation of the log predictive density over five random shuffles of the data into training and validation subsets. We adopt the standard deviation as a proxy for the uncertainty of the predictive density. All values are shown relative to the $\gamma_{\rm red} = 13/3$ result (plotted in blue). Lower values of the spectral slope are clearly disfavoured, while the data set cannot discriminate among slopes $\gamma_{\rm red} > 4$, for which the characteristic correlation time-scale of red noise exceeds the span of the measurements.

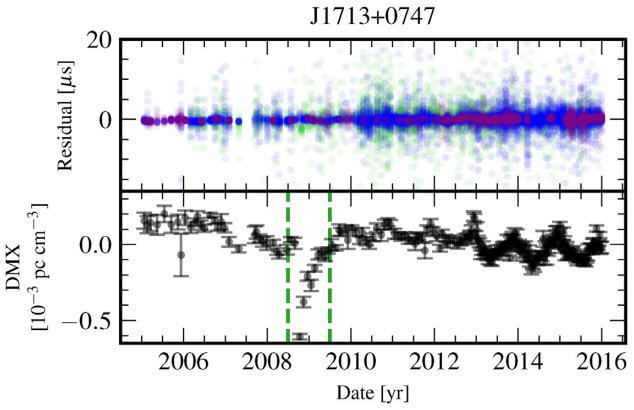

Figure 3. Residuals and dispersion-measure variation (DMX) for pulsar J1713+0747 in the NANOGrav 11-yr data release (Arzoumanian et al. 2018a). DMX shows a dramatic dip around year 2009. The dashed green vertical lines on the lower panel indicate a one-year window centred around 2009. Colours in the upper panel indicate observations taken at different radio frequencies: Blue: 1.4 GHz; Purple: 2.1 GHz; Green: 820 MHz.

It is clear that a value larger than 4 (and consistent with 13/3) is preferred.

Moving on to real PTA data sets, we perform two-fold cross validation on the TOA residuals of pulsar J1713+0747 from the NANOGrav 11-yr data release (Arzoumanian et al. 2018a). As apparent in Fig. 3 (from Arzoumanian et al. 2018a), around 2009 the residuals underwent a DM dip (i.e. an apparent decrease in the electron density experienced by radio pulses traveling to Earth, resulting in reduced 'fanning' across frequencies). We compare two noise models: the first (*GP + dip*) represents DM with a Gaussian process, but it includes also a transient feature with exponential

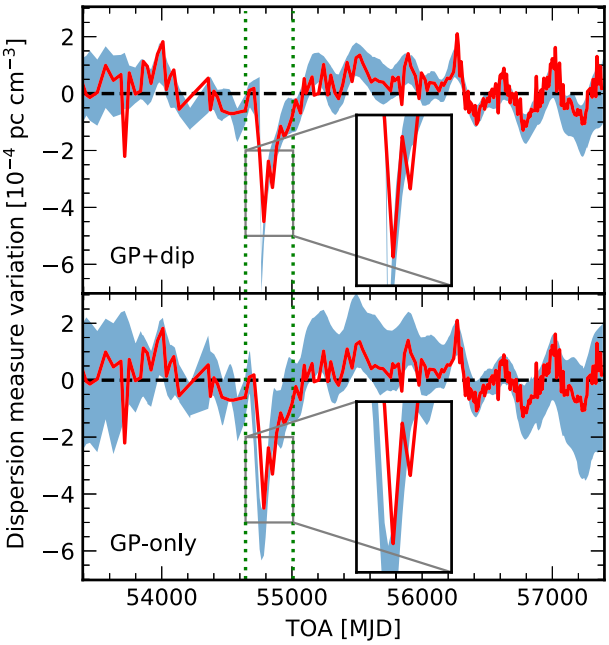

Figure 4. Reconstructed dispersion measure according to the GP + dip model (upper panel) and the GP-only model (lower panel), shown as the 5–95 per cent posterior interquantile range (grey bands) in a single cross-validation run. 'True' DM values, as obtained by fitting independent 'DMX' parameters at each epoch, are plotted in red. The dashed vertical lines delimit the one-year dip window. As apparent in the plot insets, the GP + dip model captures the transient feature more accurately. Furthermore, the GP-only model pays the flexibility required to fit the dip with higher variance across the entire data set. Both circumstances lead to higher predictive density for the GP + dip model; in this particular run, the delta log density is 39.

decay to represent the dip; the second (*GP-only*) represents DM with the Gaussian process alone. In keeping with the representation condition introduced above, when we select the training data set we need to make sure that it includes a sufficient number of residuals around the dip. To achieve this, we build the training data set by randomly selecting half of the residuals within a one-year window centred around 2009 (see Fig. 3), as well as half of the residuals outside the window.

We perform two-fold cross validation 32 times with different random data partitions. The GP + dip model yields consistently higher predictive densities than the GP-only model, with delta log density $54.8^{+584.2}_{-19.9}$ (quoted as median augmented by the 90 per cent interquantile range). The stronger predictive performance of the GP + dip model (the testing data are e^{39} times more probable under this model) offers statistical evidence that the dip is a real physical feature. This is confirmed by inspecting the reconstructed DM. The grey bands in Fig. 4 show the 90 per cent interquantile range of reconstructed DM over 100 posterior draws of the model parameters in a single cross-validation run, while the red line traces the 'true' DM at each epoch, as obtained by fitting independent dispersion values to multifrequency data.[5] It is apparent that the GP + dip model captures the DM transient more accurately, and that it follows the overall evolution of DM with smaller variance. Both conditions result in higher predictive density.

[5]In PTA jargon, the red line shows the best-fittng 'DMX' parameters.

6 CROSS VALIDATION FOR GW DETECTION IN MULTIPULSAR DATA SETS

To perform cross validation on multipulsar data sets and models (the latter possibly including a common GP describing the correlated delays induced by GWs for each pulsar), we may proceed without change *if* we satisfy the representation condition: that is, if we partition the multipulsar residual vector, Y, into testing data sets $Y^{(k)}$ in such a way that all parameters of every pulsar are represented in every training data set $Y^{(-k)}$. However, it seems natural to partition Y instead into subsets that correspond to individual pulsars, or groups of individual pulsars. Such an arrangement may allow us to identify pulsars contaminated by pathological observations, pulsars that are poorly described by the noise model chosen for them, or pulsars that are too noisy to contribute to GW inference. We next discuss how to proceed for this more general partitioning. We describe separately the case of deterministic and stochastic GWs.

For GWs described by *deterministic* models (e.g. isolated supermassive black hole binaries), individual pulsars are described by the likelihoods of equations (4) and (5), with the replacement $y \to Y^{(a)} - d^{(a)}(\theta_{\rm GW})$, where $Y^{(a)}$ is the vector of timing residuals for pulsar a, where the $\theta_{\rm GW}$ describe the GW parameters (a common set for all pulsars), and where the $d^{(a)}(\theta_{\rm GW})$ are the delays induced by the GWs on pulsar a. In this case, cross validation would proceed as follows: (a) for each $Y^{(-k)}$ we would sample the joint posterior distribution of the $\theta_{\rm GW}$ and of the noise hyperparameters $\eta^{(-k)}_{N,{\rm GP}}$ describing the pulsars represented in $Y^{(-k)}$; (b) for each corresponding $Y^{(k)}$, we would evaluate the log *marginalized* predictive likelihood

$$\log \int p(Y^{(k)}|\eta^{(k)}, \theta_{\rm GW}) p(\theta_{\rm GW}|Y^{(k)}) p(\eta^{(k)})\, {\rm d}\eta^{(k)}\, {\rm d}\theta_{\rm GW}$$
$$\simeq \log \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \int p(Y^{(k)}|\eta^{(k)}, \theta^{(-k)}_{{\rm GW},i}) p(\eta^{(k)}) {\rm d}\eta^{(k)}, \qquad (9)$$

where the $\theta^{(-k)}_{{\rm GW},i}$ are Markov Chain (sub-)samples from the training posterior $p(\theta_{\rm GW}, \eta^{(-k)}_{N,{\rm GP}}|Y^{(-k)})$, and where we have dropped the $\eta_{N,\,{\rm GP}}$ suffix for compactness. From an implementation standpoint, the nested sum/integral in equation (9) may require a dedicated stochastic algorithm similar to those employed to evaluate the Bayesian evidence (Trotta 2008).

It is important to notice that equation (9) depends directly on the noise-hyperparameter priors $p(\eta^{(k)})$, which invalidates some of our motivation for computing predictive likelihoods in the first place. In practice, we may worry that we cannot compare predictive likelihoods for different $Y^{(k)}$ because they have different prior 'calibrations'. With respect to this objection, it seems then natural to consider the ratio of equation (9) to the noise-only evidence $\int p(Y^{(k)}|\eta^{(k)}) p(\eta^{(k)})\, {\rm d}\eta^{(k)}$ (which in fact factorizes over the pulsars in $Y^{(k)}$). We leave to future work the exploration of marginalized predictive likelihoods as GW detection statistics, as well as the development of an efficient sampling method for equation (9).

Last, for *stochastic GWs* described by their spectrum and by their correlations across pulsars, we need to account not only for the common GW parameters $\theta_{\rm GW}$, but also for the correlations between the GW GP weights in each pulsar. To sketch the mathematical structure of the problem with more readable notation, let us consider the case of training on pulsar 1 and testing on pulsar 2; formulaes generalize readily to k-fold validation testing pairs $(Y^{(-k)}, Y^{(k)})$. We do as follows: We first obtain samples $(\eta^{(1)}_i, \theta_{{\rm GW},i})$ from the posterior $p(\eta^{(1)}, \theta_{\rm GW}|Y^{(1)})$; we then evaluate the log marginalized predictive likelihood in the form

$$\log \sum_{i=1}^{N} \int p(Y^{(2)}|\eta^{(2)}, c^{(2)}_{\rm GW}) p(c^{(2)}_{\rm GW}|c^{(1)}_{\rm GW}, \theta_{{\rm GW},i}) \times$$
$$p(c^{(1)}_{\rm GW}|Y^{(1)}; \eta^{(1)}_i, \theta_{{\rm GW},i})\, {\rm d}\eta^{(2)}\, {\rm d}c^{(2)}_{\rm GW}\, {\rm d}c^{(1)}_{\rm GW}. \qquad (10)$$

The integral over the $c^{(1)}_{\rm GW}$ can be performed analytically, and the resulting conditional prior for the $c^{(1)}_{\rm GW}$ expressed as

$$p(c^{(2)}_{\rm GW}|Y^{(1)}; \eta^{(1)}_i, \theta_{{\rm GW},i}) = \mathcal{N}(\bar{c}^{(2)|(1)}, \Sigma^{(2)|(1)}) \qquad (11)$$

with

$$\bar{c}^{(2)|(1)}(Y^{(1)}; \eta^{(1)}_i, \theta_{{\rm GW},i}) = \Phi_{21} \Phi_{11}^{-1} \bar{c}^{(1)}, \qquad (12)$$

and

$$\Sigma^{(2)|(1)}(\eta^{(1)}_i, \theta_{{\rm GW},i})$$
$$= \Phi_{22} - \Phi_{21}(\Phi_{11}^{-1} - \Phi_{11}^{-1} \Sigma^{(1)} \Phi_{11}^{-1}) \Phi_{21}, \qquad (13)$$

where $\bar{c}^{(1)}(Y^{(1)}; \eta^{(1)}_i, \theta_{{\rm GW},i})$ and $\Sigma^{(1)}(\eta^{(1)}_i, \theta_{{\rm GW},i})$ are given in equations (6) and (7), and the $\Phi_{ij}(\theta_{{\rm GW},\,i})$ denote the blocks of the joint normal prior for the GW GP weights. The integral over the $c^{(2)}_{\rm GW}$ in equation (10) may also be performed analytically by way of equation (5), with the replacement $y \to Y^{(2)} - F^{(2)}_{\rm GW} \bar{c}^{(2)|(1)}$ and $\Phi_{\rm GW} \to \Sigma^{(2)|(1)}$ (in this notation, $F^{(2)}_{\rm GW}$ and $\Phi_{\rm GW}$ represent the GW blocks of $F_{\rm GP}$ and $\Phi_{\rm GP}$). Again, we leave to future work the investigation of marginalized predictive likelihoods as detection statistics for stochastic GWs in PTA data.

7 DISCUSSION

The development of sophisticated noise models is of paramount importance to ongoing PTA searches for nanohertz GWs, especially so because much of the PTA 'instrument' was not engineered by humans. Contrast LIGO's precisely fabricated arms with a PTA Earth–pulsar 'arm', which comprises a radiotelescope, a distant (∼kpc) pulsar, and the expanse of spacetime in between. The telescope admits some experimental control, since we can test its signal response and mitigate noise in the receiver, but we still have to contend with poorly constrained physical processes in pulsar interiors and emission regions (Lasky et al. 2015), not to mention dispersive effects as radio pulses propagate through the ionized interstellar medium to the Earth (Cordes & Shannon 2010; Lam et al. 2016, 2018, 2019). Thus PTA noise models must be well motivated physically, and yet flexible enough to accommodate unknown unknowns, if they are to allow the identification of subtle GW-induced delays.

Current techniques to test the relative aptness and robustness of noise models include basic checks (i.e. evaluating χ^2 fit residuals under different models), frequentist approaches (computing receiver operating characteristic curves to maximize the signal detection probability at a given false-alarm probability), as well as Bayesian model comparison (Taylor et al. 2013, 2017; Taylor, Ellis & Gair 2014; Sampson, Cornish & McWilliams 2015; Cornish & Sampson 2016). In this last technique, we marginalize likelihoods for different models over their respective prior volumes, and use the ratios of marginal likelihoods (usually under the assumption of equal prior probabilities for each model) to make statements about the posterior odds with which one model is favoured over the other. This approach is perfectly valid, but it can be troubled by practical issues such as accurately integrating into the tails of the likelihood distribution, as well as assigning the ranges of parameter priors in the first place.

Bayesian cross validation addresses some of these issues by partitioning data sets into training and testing samples. Models are conditioned on the training set, producing posterior probability distributions for the parameters, over which we average the likelihood of the testing set. The cross-validation score is then the probability of the test data under the trained model. For reasonably informative training data, posteriors will be more compact than the priors, shielding the integration from ad hoc prior choices. To evaluate the integrals, it is convenient to average the likelihood over posterior samples from a conventional MCMC analysis of the training set.

In this article we argued for the power of Bayesian cross validation as applied to PTA data. Our case studies included the spectral characterization of a GW background, and the modelling of dispersive noise from the interstellar medium. For the former, we performed two-fold cross validation on a data set from the 1st IPTA mock data challenge (Hazboun et al. 2018), showing that the data favoured a correlated process consistent with a stochastic GW background from supermassive binary black holes. For the latter, two-fold cross validation on 11 yr of NANOGrav data for pulsar J1713+0747 illustrated the necessity to include a transient dispersive noise feature around the year 2009, consistent with a void in the electron density along the line of sight (Lentati et al. 2016). We also introduced a formalism for multipulsar cross validation, where GW models conditioned on training data from a subarray are assessed for predictive performance on left-out pulsars. In future work we will investigate this approach as a tool to validate claims of GW detections in real PTA data sets.

We expect cross validation to be similarly useful in analysing data from other GW detectors. For instance, within the global network of ground-based GW interferometers, support for a GW signal in one detector could be validated using data from a different widely separated detector. Furthermore, signals such as GW150914, which could have been observed by a LISA-like detector years before it was seen by LIGO (Sesana 2016), raise the prospect of multiband, multidetector cross validation of GW signals.

ACKNOWLEDGEMENTS

We thank Joseph Simon and Michael Lam for discussions regarding the dispersion-measure variation of PSR J1713+0747. This research was performed in part using the Zwicky computer cluster at Caltech supported by the National Science Foundation under MRI-R2 award No. PHY0960291 and by the Sherman Fairchild Foundation. Portions of this research were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This work was supported in part by National Science Foundation Grant No. PHYS-1066293 and by the hospitality of the Aspen Center for Physics. MV was supported by the Jet Propulsion Laboratory RTD program. SRT was supported by the NANOGrav National Science Foundation Physics Frontier Center, award number 1430284. Parts of this work were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract to the National Aeronautics and Space Administration. Copyright 2019 California Institute of Technology. Government sponsorship acknowledged.

REFERENCES

Aggarwal K. et al., 2018, preprint (arXiv:1812.11585)
Arzoumanian Z. et al., 2018a, ApJS, 235, 37
Arzoumanian Z. et al., 2018b, ApJ, 859, 47
Burke-Spolaor S., 2015, preprint (arXiv:1511.07869)
Cordes J. M., 2013, Class. Quantum Gravity, 30, 224002
Cordes J. M., Shannon R. M., 2010, preprint (arXiv:1010.3785)
Cornish N. J., Sampson L., 2016, Phys. Rev. D, 93, 104047
Gelman A., Stern H. S., Carlin J. B., Dunson D. B., Vehtari A., Rubin D. B., 2013, Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton, FL USA
Godsill S. J., 2001, J. Comp. Graph. Stat., 10, 230
Gregory P., 2010, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge U. Press, Cambridge, UK
Hazboun J. S., Mingarelli C. M. F., Lee K., 2018, preprint (arXiv:1810.10527)
Lam M. T. et al., 2018, ApJ, 861, 132
Lam M. T., Cordes J. M., Chatterjee S., Jones M. L., McLaughlin M. A., Armstrong J. W., 2016, ApJ, 821, 66
Lam M. T., Lazio T. J. W., Dolch T., Jones M. L., McLaughlin M. A., Stinebring D. R., Surnis M., 2019, preprint (arXiv:1903.00426)
Lasky P. D., Melatos A., Ravi V., Hobbs G., 2015, MNRAS, 449, 3293
Lentati L. et al., 2016, MNRAS, 458, 2161
Lommen A. N., 2015, Rep. Prog. Phys., 78, 124901
NANOGrav, 2019, North American Nanohertz Observatory for Gravitational Waves, Available at: http://nanograv.org.
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA
Sampson L., Cornish N. J., McWilliams S. T., 2015, Phys. Rev. D, 91, 084055
Sesana A., 2016, Phys. Rev. Lett., 116, 231102
Sisson S. A., 2005, J. Am. Stat. Ass., 100, 1077
Stinebring D., 2013, Class. Quantum Gravity, 30, 224006
Taylor S., Ellis J., Gair J., 2014, Phys. Rev. D, 90, 104028
Taylor S. R., Gair J. R., Lentati L., 2013, Phys. Rev. D, 87, 044035
Taylor S. R., Lentati L., Babak S., Brem P., Gair J. R., Sesana A., Vecchio A., 2017, Phys. Rev. D, 95, 042002
Trotta R., 2008, Contemp. Phys., 49, 71
Vallisneri M., 2012, Phys. Rev. D, 86, 082001
van Haasteren R., Vallisneri M., 2014, Phys. Rev. D, 90, 104012

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