Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **540,** 3818–3825 (2025) https://doi.org/10.1093/mnras/staf953
Advance Access publication 2025 June 11

Generalized stepping-stone sampling: efficient marginal likelihood estimation in gravitational wave analysis of pulsar timing array data

El Mehdi Zahraoui [1]★ Patricio Maturana-Russel,[1,2] Willem van Straten [3] Renate Meyer[2] and Sergei Gulyaev[1]

[1]*Department of Mathematical Sciences, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand*
[2]*Department of Statistics, University of Auckland, Auckland 1142, New Zealand*
[3]*Manly Astrophysics, 15/41-42 East Esplanade, Manly, NSW 2095, Australia*

Accepted 2025 June 6. Received 2025 June 6; in original form 2024 November 21

ABSTRACT

Globally, pulsar timing array (PTA) experiments have revealed tentative evidence of a stochastic gravitational wave background (GWB) signal in PTA data. Apart from acquiring more observations, the sensitivity of PTA experiments can be increased by improving the accuracy of noise modelling. In PTA data analysis, noise modelling is conducted primarily using Bayesian statistics, relying on the marginal likelihoods and Bayes factors to assess evidence. We introduce generalized stepping-stone sampling (GSS) as an efficient and accurate marginal likelihood estimation method for the PTA–Bayesian framework. This method enables low-cost estimates with high accuracy, especially when comparing expensive models such as the Hellings–Downs model or the overlap reduction function model. We demonstrate the efficiency and the accuracy of GSS for model selection and evidence calculation by reevaluating the evidence of previous analyses from the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) 15-yr data set and the European Pulsar Timing Array (EPTA) second data release. We find similar evidence for the GWB compared to the one reported by the NANOGrav 15-yr data set. Compared to the evidence reported for the EPTA second data release, we find a substantial increase in evidence for a GWB signal across all data sets.

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

1 INTRODUCTION

The pulsar timing array (PTA) emerged from the concept of a single-pulsar–Earth baseline to observe gravitational waves (GWs) in the low-frequency nanohertz band (Sazhin 1978). The PTA experiments were designed as a network of pulsars that could unveil the influence of GWs on the arrival times of their pulses in the form of quadrupolar correlations (Hellings & Downs 1983). Millisecond pulsars (MSPs) became the key to achieving this objective due to their long-term stability, rivalling Earth's atomic clocks (Matsakis, Taylor & Eubanks 1997). Since the initial MSP timing analysis (Stinebring et al. 1990), the red noise (RN) in the MSPs' data set has pointed to a promising future as observational radio astronomy and computational capability constantly improve. The next step for the PTA community focused on gathering more data and refining the pulsar timing model to test and derive a realistic sensitivity threshold to observe GWs (Verbiest et al. 2009; Cordes & Shannon 2010; Hazboun, Romano & Smith 2019). Recently, major PTA collaborations such as the North American Nanohertz Observatory for Gravitational Waves (NANOGrav; McLaughlin 2013), the Australian Parkes Pulsar Timing Array (PPTA; Manchester et al. 2013), the European Pulsar Timing Array (EPTA; Ferdman et al. 2010), the Indian Pulsar Timing Array (InPTA; Joshi et al. 2018), the MeerKAT Pulsar Timing Array (MeerKAT; Miles et al. 2023), and the Chinese Pulsar Timing Array (CPTA; Xu et al. 2023) reported positive statistical support for gravitational wave background (GWB) in the latest data releases (Agazie et al. 2023; Antoniadis et al. 2023b; Reardon et al. 2023). The PTA community has established a rigorous PTA–Bayesian framework (van Haasteren et al. 2009), consistently expanding and acquiring new numerical and statistical techniques (Lentati et al. 2014; Johnson et al. 2024). This framework has enabled a simultaneous fit for different noise models and GWB signals while having the ability to compare the evidence of each different model (van Haasteren & Vallisneri 2014). Before evaluating the GWB evidence, it was proven that a common red signal (CRS) is evident across different PTA data sets: NANOGrav (Arzoumanian et al. 2020), EPTA (Chen et al. 2021), and PPTA (Goncharov et al. 2022). These findings resulted in a surge of interest in defining the nature of these CRSs and potentially revealing the buried GWB signal. Through this process, an extensive number of models were used to mitigate different sources of RN that can affect the PTA's sensitivity either from pulsar contribution such as pulsar spin noise (Verbiest et al. 2009) and dispersion measure (DM) variations (Jones et al. 2017), or other noise sources such as spatial correlations due to the Solar system ephemeris and the atomic clocks used in this experiment (Tiburzi et al. 2016). To test and account for all noise models, the full PTA–Bayesian analysis results in a large model space. Down the line, incorporating new MSP data sets and updated noise models will further expand this space significantly.

★ E-mail: elmehdi.zahraoui@autuni.ac.nz

In the Bayesian framework, hypotheses comparison is done by evaluating the *Bayes factor* (BF), which is the ratio of *marginal likelihoods* of different models. This criterion is used to assess the evidence for each hypothesis (Newton & Raftery 1994). For the PTA, model selection and evidence estimation are currently carried through various Bayesian techniques (Agazie et al. 2023; Antoniadis et al. 2023b). Among these methods, thermodynamic integration (TI; Lartillot & Philippe 2006) stands out as the cornerstone method for marginal likelihood estimation. In parallel, nested sampling (Skilling 2006) is implemented to provide not only estimates of marginal likelihood but also to generate posterior samples. Additionally, reweighting (Hourihane et al. 2023) was defined to estimate the marginal likelihood for an expensive model using an approximately matching model. Finally, the hypermodel method, implemented using the Savage–Dickey density ratio approximation as a form of product space sampling for nested models (Carlin & Chib 1995; Lodewyckx et al. 2011), is routinely used across the whole PTA analyses. Johnson et al. (2024) discuss these methods in the context of the PTA, where nested sampling is used for models with small parameter spaces. In this case, nested sampling yields less accurate estimates with a similar cost to the hypermodel method. In parallel, TI estimates become expensive for large PTA models and can lead to biased estimates if the required computational power is lacking. Consequently, TI is employed only where the hypermodel fails to provide evidence estimates. Although the hypermodel method was established as the predominant method, it is limited to nested models and requires assigning arbitrary weights to each model to avoid the Markov chain getting trapped in a particular model. Alternatively, we introduce generalized stepping-stone sampling (GSS; Fan et al. 2011) as a new method to GW astronomy and in particular the PTA–Bayesian framework. GSS is intended to bypass most of the current issues with marginal likelihood estimation and provide an accurate marginal likelihood estimate at a lower computational price. This paper is structured as follows. Section 2 outlines model selection through the BF and briefly defines the TI, stepping-stone sampling (SS), and GSS methods. Section 3 demonstrates a comparison between these methods' performance applied to a Gaussian model. Thereafter, Section 4 first discusses the implementation of GSS within the PTA framework. Then, our application to different PTA analyses will revisit the GWB evidence and will demonstrate the benefits of marginal likelihood estimation in making a robust model selection and monitoring the evidence on different CRS models. Finally, Section 5 will discuss the prospects of GSS in the PTA–Bayesian framework and the significance of different PTA application results.

2 MARGINAL LIKELIHOOD ESTIMATION

The Bayesian framework is based on *Bayes'* theorem, which is given by

$$p(\boldsymbol{\theta}|X, M) = \frac{L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{z(X|M)} \tag{1}$$

for a model M and data set X, where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, $p(\boldsymbol{\theta}|X, M)$ the posterior probability density of $\boldsymbol{\theta}$, $L(X|\boldsymbol{\theta}, M)$ the likelihood function, and $\pi(\boldsymbol{\theta}|M)$ the prior density. The constant $z(X|M)$, commonly denoted as z, is the *marginal likelihood*. The marginal likelihood is a multidimensional integral over the parameter space Θ defined by

$$z = \int_{\Theta} L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\,\mathrm{d}\boldsymbol{\theta}, \tag{2}$$

where $\pi(\boldsymbol{\theta}|M)$ is a proper prior, i.e. $\int_{\Theta} \pi(\boldsymbol{\theta}|M)\,\mathrm{d}\boldsymbol{\theta} = 1$. The integral z is a prior weighted average of the likelihood, which quantifies the goodness of fit of model M to data X. For a set of two models M_1 and M_2 fitted to a data set X, the marginal likelihoods $z(X|M_1)$ and $z(X|M_2)$ can be used to compare these models such that the model with the higher marginal likelihood is preferred. Hence, the BF, defined by

$$\mathrm{BF}_{(M_1/M_2)} = \frac{z(X|M_1)}{z(X|M_2)}, \tag{3}$$

is used as a Bayesian criterion to compare M_1 to M_2 and perform model selection. Specifically, the BF is equivalent to the ratio of the posterior distributions of the models when the prior probabilities for the models are equal. In contrast, alternative performance tests are used to compare models, such as the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). Although the BIC and the AIC penalize unnecessary complexity by applying a uniform penalty for each parameter, these approaches ignore the importance of priors in the Bayesian analysis. The BF allows for tailored penalties based on the chosen priors, which helps to ensure that overly complex models are penalized appropriately, reducing the risk of overfitting.

Apart from simple conjugate cases, the marginal likelihood has no analytical solution. In practice, considerable efforts have been made to numerically approximate z through Markov chain Monte Carlo (MCMC) methods such as the harmonic mean (Newton & Raftery 1994), bridge sampling (Meng & Wong 1996), and nested sampling (Skilling 2006). Path sampling (Gelman & Meng 1998) or TI (Lartillot & Philippe 2006) was the cornerstone technique of what would later be known as the method of power posterior (Friel & Pettitt 2008). This method defines a geometric path that connects the prior with the posterior through the power posterior densities:

$$p_\beta(\boldsymbol{\theta}|X, M) = \frac{L(X|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{z_\beta}, \quad \text{with} \quad 0 \leq \beta \leq 1, \tag{4}$$

where β is the inverse temperature for the power posterior density p_β. When $\beta = 0$, the density p_β is equal to the proper prior distribution, which gives $z_0 = 1$. When $\beta = 1$, the density p_β is equal to the posterior distribution, resulting in $z_1 = z$. Correspondingly, TI relies on the identity

$$\log(z) = \int_0^1 E_{p_\beta}[\log(L(X|\boldsymbol{\theta}, M))]\,\mathrm{d}\beta, \tag{5}$$

where E_{p_β} denotes the expected value with respect to p_β. For a sequence of β values, samples from each power posterior can be used to estimate the expected values by the sample averages. Thus, the integral in equation (5) can be approximated using any technique for numerical integration, e.g. the trapezoidal rule. TI can yield a good marginal likelihood estimate; nevertheless, it requires a high computational expense to produce a single estimate. This computational expense obliged modifications to the method to reduce the estimation cost. SS was proposed as an attempt to reduce TI's cost, which resulted in a far more efficient and accurate method, i.e. GSS, which is considered one of the most accurate marginal likelihood estimation methods.

2.1 Stepping-stone sampling

The approximation of a continuous integral induces a discretization bias in TI, leading to an inaccurate estimate of the marginal likelihood. Hence, SS was first proposed in the phylogenetics field (Xie et al. 2011) and later introduced to LIGO (Laser Interferometer Gravitational-Wave Observatory) GW analysis (Maturana-

Russel et al. 2019) as an alternative method to enable unbiased z estimates and improve model selection. The SS method considers an unnormalized version of the power posterior density, given in equation (4), defined by

$$q_\beta = L(X|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M). \tag{6}$$

This density is normalized by z_β yielding a normalized power posterior density:

$$p_\beta = \frac{q_\beta}{z_\beta}, \quad \text{with} \quad z_\beta = \int_\Theta q_\beta \, \mathrm{d}\boldsymbol{\theta}. \tag{7}$$

Note that z_0 represents the normalizing constant of a proper prior, which is equal to 1, whereas z_1 is the normalizing constant of the posterior, i.e. the marginal likelihood. Taking this into account, SS redefines the marginal likelihood as a product of $K-1$ intermediate ratios given by

$$z = \frac{z_1}{z_0} = \frac{z_{\beta_1}}{z_{\beta_0}} \frac{z_{\beta_2}}{z_{\beta_1}} \cdots \frac{z_{\beta_{K-1}}}{z_{\beta_{K-2}}} = \prod_{k=1}^{K-1} \frac{z_{\beta_k}}{z_{\beta_{k-1}}}, \tag{8}$$

where $\beta_0 = 0 < \cdots < \beta_{k-1} < \beta_k < \cdots < \beta_{K-1} = 1$. Importance sampling can be used to estimate each of the ratios $r_k = z_{\beta_k}/z_{\beta_{k-1}}$, where $p_{\beta_{k-1}}$ acts as an excellent importance sampling distribution when $p_{\beta_{k-1}}$ is similar to p_{β_k}, which occurs for a sufficient large K. It can be shown that the ratios r_k can be expressed as

$$r_k = E_{p_{\beta_{k-1}}} \left[(L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M))^{\beta_k - \beta_{k-1}} \right]. \tag{9}$$

In practice, these ratios are estimated using an unbiased Monte Carlo (MC) estimator for each ratio r_k, given by

$$\widehat{r}_k = \frac{1}{n} \sum_{i=1}^{n} L(X|\boldsymbol{\theta}_{k-1,i}, M)^{\beta_k - \beta_{k-1}}. \tag{10}$$

Therefore, the SS estimate of the marginal likelihood is calculated by

$$\widehat{z} = \prod_{k=1}^{K-1} \widehat{r}_k = \prod_{k=1}^{K-1} \frac{1}{n} \sum_{i=1}^{n} L(X|\boldsymbol{\theta}_{k-1,i}, M)^{\beta_k - \beta_{k-1}}, \tag{11}$$

with $\boldsymbol{\theta}_{k-1,i}$ samples drawn from $p_{\beta_{k-1}}$ and n the number of drawn samples. To avoid numerical errors, it is preferable to work with $\log \widehat{z}$. However, this transformation introduces a bias to the $\widehat{z}$ estimate which can be overcome by increasing K, the number of β chains that corresponds to n samples from $p_\beta(\boldsymbol{\theta}|X, M)$ (Xie et al. 2011). The choice of the β values and their dispersion can enhance the efficiency of SS and even more for TI (Xie et al. 2011; Maturana-Russel et al. 2019). For β_k defined as

$$\beta_k = \frac{k}{K} 100 \text{ per cent quantile of Beta}(\alpha, 1) \tag{12}$$

for $k = 1, \ldots, K-1$, it has been shown that the efficiency can be increased with $\alpha = 0.3$ for both methods (Xie et al. 2011). In general, the prior is more diffuse than the posterior, leading to much greater differences between consecutive power posterior densities near the prior, i.e. for β close to 0. Consequently, the SS method exhibits greater differences between consecutive ratios $\widehat{r}_k$ when β is close to 0. As a result, more power posterior densities are required near the prior, meaning that more β values close to 0 are needed, which is effectively achieved by the quantiles of the Beta(0.3, 1) distribution.

2.2 Generalized stepping-stone sampling

The performance of the SS reinforced the interest in decreasing the cost of marginal likelihood estimation while maintaining high accuracy. The GSS was proposed by Fan et al. (2011) to achieve this objective. The GSS method introduces an additional distribution, referred to as the reference distribution π_0, which should be an approximation of the posterior. This constrains the parameter space explored by the MCMC method, making GSS more efficient than SS. In practice, the reference distribution can be defined as the product of convenient distributions tuned by posterior samples. GSS introduces the reference distribution π_0 in the unnormalized posterior given in equation (6), redefining it as follows:

$$q_\beta = [L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)]^\beta [\pi_0(\boldsymbol{\theta}|M)]^{1-\beta}. \tag{13}$$

Thus, the power posterior in equation (7) defines a path between the reference distribution and the posterior. Note that for $\beta = 0$, p_β is equal to the reference distribution, while for $\beta = 1$, p_β is equal to the posterior. The product of ratios that defines the SS estimator in equation (8) remains the same in the GSS estimator; however, it can be shown that the ratios are now defined as follows:

$$r_k = E_{p_{\beta_{k-1}}} \left[\left(\frac{L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{\pi_0(\boldsymbol{\theta}|M)} \right)^{\beta_k - \beta_{k-1}} \right]. \tag{14}$$

The MC estimator of the ratio r_k is given by

$$\widehat{r}_k = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{L(X|\boldsymbol{\theta}_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\pi_0(\boldsymbol{\theta}_{k-1,i}|M)} \right)^{\beta_k - \beta_{k-1}}, \tag{15}$$

with n samples $\boldsymbol{\theta}_{k-1,i}$ drawn from $p_{\beta_{k-1}}$.

Finally, the marginal likelihood is estimated by multiplying over the estimated ratios as follows:

$$\widehat{z} = \prod_{k=1}^{K-1} \widehat{r}_k = \prod_{k=1}^{K-1} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{L(X|\boldsymbol{\theta}_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\pi_0(\boldsymbol{\theta}_{k-1,i}|M)} \right)^{\beta_k - \beta_{k-1}}. \tag{16}$$

The MC estimator for each ratio r_k provides an unbiased estimate of the ratio. However, this does not imply that their product is unbiased unless the estimated ratios are independent. In practice, the interaction between Markov chains in the parallel tempering algorithm in the GSS case is unnecessary due to prior knowledge of the posterior distribution. As a result, independent Markov chains can be generated, leading to independent estimated ratios and thereby eliminating bias in the GSS estimate.

The uncertainty of the GSS estimate can be approximated using the delta method – as proposed in the SS case (Xie et al. 2011) – which relies on the assumption that the samples are independent. In practice, this assumption might not be met, resulting in a biased uncertainty estimate. Given that independent GSS estimates can be obtained at a relatively low computational cost, empirical uncertainty can be directly estimated. This method is adopted in the following applications.

GSS inherits the benefits of the SS method and is equal to SS if $\pi_0(\boldsymbol{\theta}|M) = \pi(\boldsymbol{\theta}|M)$. For GSS to be efficient, one can tailor a reference distribution π_0 that approximates the posterior as the product of independent distributions, each reflecting the marginal distribution of each parameter θ_j, which is an element of the parameter vector $\boldsymbol{\theta}$. If posterior samples are available, these samples can be used to calibrate each marginal probability density of the reference distribution for each θ_j, where each distribution can be set, for example, to $\mathcal{N}(\bar{\theta}_j, S_j^2)$, with $\bar{\theta}_j$ and S_j being the mean and standard deviation of these samples, respectively.

The GSS method bears similarity to the recently proposed weighting method (Hourihane et al. 2023). It can be shown that for a target model T and an approximate model A with equal parameter space, the reweighting method is a particular case of the GSS method. In

this case, the reference distribution π_0 is equal to likelihood $\times$ prior of the approximate model A with $K = 2$. The reweighting method is particularly useful for the PTA models when the sampling from model T is computationally expensive compared to the approximate model A.

3 SIMULATION STUDY

In this simulation study, we will implement the GSS method and compare its performance to that of SS and TI. Since the marginal likelihood z of real cases generally does not have an analytical solution, we resort to a simple Gaussian model where the marginal likelihood is known (Lartillot & Philippe 2006). This Gaussian model is parametrized by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ with a null data vector X, where the prior on $\boldsymbol{\theta}$ is a product of standard normal distributions $\theta_j \sim \mathcal{N}(0,\ 1)$, and the likelihood is defined by

$$L(X|\boldsymbol{\theta}) = \prod_{j=1}^{d} \exp\left(-\frac{\theta_j^2}{2v}\right), \tag{17}$$

with d the number of dimensions, and a known variance v. This set-up enables access to the analytical form for each component of the Bayesian model. The power posterior density can be expressed as a product of d normal distributions $\mathcal{N}(0,\ v/(v+\beta))$. The density p_β is equal to the posterior distribution when $\beta = 1$, yielding a marginal likelihood equal to

$$z = \left(\frac{v}{1+v}\right)^{d/2}. \tag{18}$$

Since the analytical power posteriors are available, we sample directly from these densities to avoid MCMC sampling. For SS and TI, n independent samples are generated for each β_k chain, and then K, the number of chains, is increased until the numerical estimate reaches the true value. In contrast, the GSS estimator is executed in two steps. First, independent samples from the posterior distribution (calibration samples $N_{\rm cal}$) are used to tune the reference distribution, then n samples are generated for each β_k from the power density $\mathcal{N}_{\beta_k,j}(\mu_{\beta_k,j},\ \sigma^2_{\beta_k,j})$ with

$$\mu_{\beta_k,j} = \frac{\bar{\theta}_j(1-\beta_k)}{S_j^2\beta_k\left(\frac{1+v}{v}\right) + (1-\beta_k)} \text{ and}$$

$$\sigma^2_{\beta_k,j} = \left(\beta_k\left(\frac{1+v}{v}\right) + \frac{(1-\beta_k)}{S_j^2}\right)^{-1}, \tag{19}$$

where $\bar{\theta}_j$ and S_j^2 are, respectively, the mean and the variance of the calibration samples from the jth dimension.

For this comparison, we set the number of dimensions to $d = 50$ and the variance to $v = 0.01$, resulting in a $\log(z) = -115.38$. With β_k uniformly dispersed along the quantiles of Beta(0.3, 1), we fix $n = 1000$ independent samples, $K \in [4, 8, 16, 32, 64]$ number of β chains for each method. For the GSS estimator, the same set-up is used; however, we set $N_{\rm cal} = 1000$ and $n = 10$. To evaluate the uncertainty of each method's estimate, 1000 independent estimates are generated for each K. Fig. 1 shows that TI requires more than 32 chains and 16 chains for SS to converge to the true $\log(z)$, while the GSS estimator yields an accurate estimate of $\log(\widehat{z}) = -115.37$ with $K = 4$. This comparison demonstrates how the $\log(z)$ estimation using SS and TI can be biased even when error constraints are available. Thus, only increasing K will ensure convergence and an accurate $\log(z)$ estimate.

The GSS estimator's performance remains consistent with a higher number of dimensions as shown in Fig. 2 where the number of dimensions $d = 2000$ and $N_{\rm cal} = 500$. Fig. 2 shows that one can either use fewer samples and increase K or use few β chains and increase n the number of samples per chain. In practice, numerical experiments suggest that the computational cost should be allocated to the number of β chains since a higher number of K increases the similarity between the consecutive importance distributions (Maturana-Russel 2017).

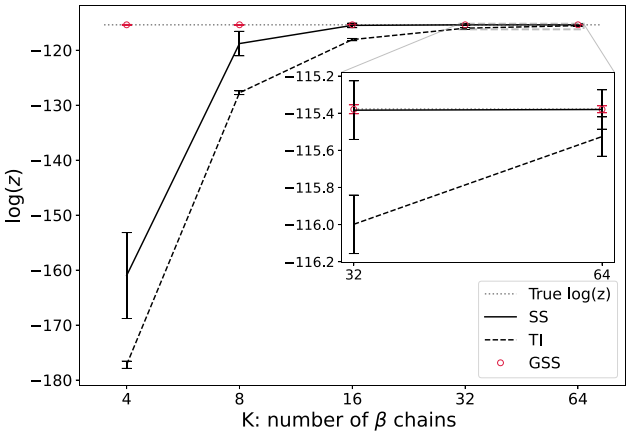

Figure 1. Comparison of log-marginal likelihood estimated as a function of K number of β chains for the Gaussian model with TI, SS, and GSS methods.

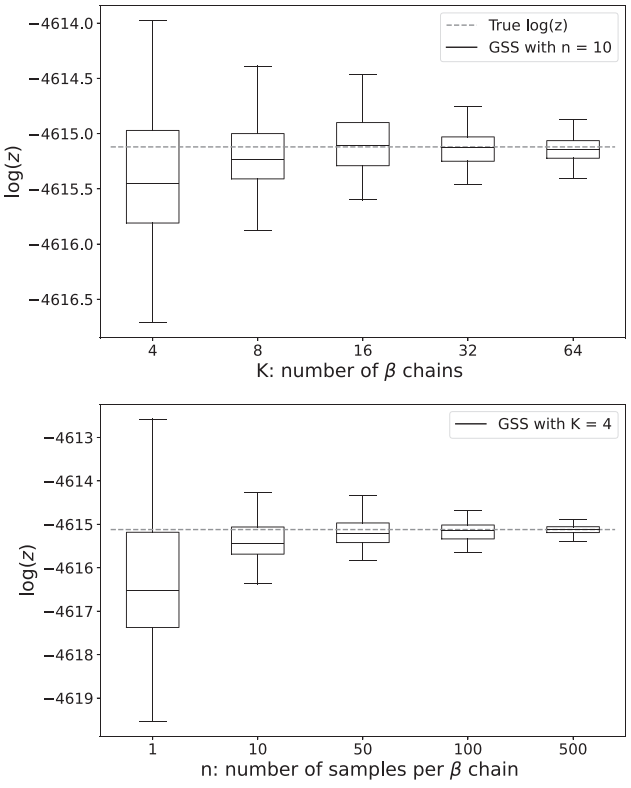

Figure 2. Convergence of the marginal likelihood estimated using GSS for the Gaussian model with $d = 2000$ and $v = 0.01$. Top panel: Comparison of log-marginal likelihood estimated as a function of K number of β chains with $n = 10$. Bottom panel: Comparison of log-marginal likelihood estimated as a function of n number of independent samples per β chains with $K = 4$.

Table 1. Mean and standard deviation of 100 independent log-marginal likelihood estimates for NANOGrav single-pulsar noise models obtained via GSS with $K = 8$.

PSR	WN		WN+DMGP		WN+RN		WN+RN+DMGP	
	$\log(\widehat{z})$	Std	$\log(\widehat{z})$	Std	$\log(\widehat{z})$	Std	$\log(\widehat{z})$	Std
J1012+5307	275 708.30	0.20	275 708.35	0.25	275 744.84	0.48	275 745.76	0.61
J1600−3053	261 778.81	0.57	261 778.87	0.65	261 784.37	0.81	261 784.82	0.87
J1705−1903	111 504.39	0.29	111 600.51	0.32	111 562.63	0.44	111 633.47	0.41
J1713+0747	750 942.80	3.24	750 946.68	1.18	751 285.52	0.78	751 284.93	2.13

4 APPLICATION OF GSS TO THE PTA

In this section, we will illustrate how the GSS integrates into the PTA–Bayesian framework, and demonstrate the different benefits of estimating the *marginal likelihood* in a PTA analysis. For this application, a Metropolis–Hastings algorithm (Chib & Greenberg 1995) was implemented to perform the sampling for the GSS method, where MPI for PYTHON (Dalcín, Paz & Storti 2005) was employed to parallelize the sampling of β chains. This implementation was designed to adequately fit into the Bayesian framework used in PTA analyses (ENTERPRISE; Ellis et al. 2019; Taylor et al. 2021). In the following, the performance of GSS is assessed based on K, the number of β chains, and $n_{\rm ESS}$, the effective sample size (ESS) per β chain. We will use the GSS method to produce multiple independent estimates for each analysis. These multiple estimates allow the direct comparison of all models by constructing a distribution of the BF of the compared models. In practice, single random draws of $\log(\widehat{z})$ from the marginal likelihood distributions for model M_1 and M_2 are used to compute each $\log(\mathrm{BF}_{M_1/M_2})$, the standard deviation of the resulting $\log(\mathrm{BF}_{M_1/M_2})$ distribution is used to measure uncertainty.

4.1 Application to NANOGrav

This section demonstrates the application of the GSS to estimate the marginal likelihood for different models used in the PTA single-pulsar noise analysis and GWB analysis. We reproduce the analysis and the public NANOGrav 15-yr data release and the NANOGrav tutorials (Agazie et al. 2023) using ENTERPRISE. The consistency of the reproduced Bayesian analysis is compared with the results discussed in the NANOGrav 15-yr data set.

4.1.1 Model selection for single-pulsar analysis

The following four pulsars were selected from NANOGrav (Agazie et al. 2023) as candidates for single-pulsar noise analysis: PSRs J1012+5307, J1600−3053, J1705−1903, and J1713+0747. Around 100 000 posterior samples are generated for each combination of the following noise models: white noise (WN), dispersion measure Gaussian process (DMGP), and RN. The 20 000 samples of the burn-in period are discarded from the posterior samples used to calibrate the reference distribution π_0. In this case, the posterior samples are reasonably unimodal; therefore, for each model, a normal distribution is used to approximate each parameter's marginal posterior distribution. The product of these distributions defines the reference distribution π_0, as it usually provides a sufficiently accurate approximation to the posterior for the GSS method. The mean and the variance of each parameter's set of samples are used to calibrate π_0, then a mean ESS $n_{\rm ESS} \approx 10$ is generated for each β chain with $K = 8$ for each single estimate. Table 1 reports the mean and standard deviation of 100 independent $\log(\widehat{z})$ for the combination of the three noise models. These values of $\log(\widehat{z})$ are used to compute the $\log(\widehat{\mathrm{BF}})$ for comparing models' relative performance.

For each pulsar, the WN model is used as a baseline for model comparison; the best-performing model compared to WN is reported with their $\log(\widehat{\mathrm{BF}}_{\mathrm{M/WN}}) \pm 1\sigma$. Across all four pulsars, strong statistical evidence is observed for the RN model. For PSR J1012+5307, WN+RN is favoured with $\log(\widehat{\mathrm{BF}}) = 36.47 \pm 0.56$. When combined with the other two models, the DMGP model has weak statistical evidence since $\log(\widehat{\mathrm{BF}}) < 1$. For PSR J1705−1903 and PSR J1713+0747, the model comparison shows high statistical significance between each fitted model, where the most significant model is WN+RN+DMGP with $\log(\widehat{\mathrm{BF}}) = 129.07 \pm 0.49$ for PSR J1705−1903, and WN+RN with $\log(\widehat{\mathrm{BF}}) = 342.58 \pm 3.29$ for PSR J1713+0747. PSR J1600−3053 demonstrates the least RN significance compared to the other three pulsars with $\log(\widehat{\mathrm{BF}}) = 5.56 \pm 0.99$. The evidence of the excess noise becomes more significant for the model WN+RN+DMGP with $\log(\widehat{\mathrm{BF}}) = 6.00 \pm 1.05$.

One can further examine the performance of each model by inspecting the influence of its parameters $\boldsymbol{\delta}$ when included. This influence can be demonstrated for the PTA models using the previous combinations of RN models. This evaluation can be conducted using the *inclusion Bayes factor* (IBF; Hinne et al. 2020), which examines the proportional change of the marginal likelihood when a set of parameters $\boldsymbol{\delta}$ is included in a set of models. Let $\mathcal{M}_{\delta^1}$ and $\mathcal{M}_{\delta^0}$ be, respectively, the sets of models that include and exclude $\boldsymbol{\delta}$ from their respective parameter vector. The IBF is given by

$$\mathrm{IBF} = \frac{\sum_{M \in \mathcal{M}_{\delta^1}} z(X|M)}{\sum_{M \in \mathcal{M}_{\delta^0}} z(X|M)}, \tag{20}$$

which denotes the sum of marginal likelihoods of models including $\boldsymbol{\delta}$ divided by the sum of marginal likelihoods of models excluding it. Table 2 shows the IBF to compare the overall performance of RN and DMGP parameters. IBF reinforces the strong statistical evidence of RN across all pulsars, while including DMGP parameters delivers a better performance only for PSR J1705−1903 and PSR J1713+0747. Since pulsar customized noise models are used for the full GWB analysis, the IBF can provide additional statistical evidence and help with a robust model selection where decisions are critical.

Table 2. Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{IBF}})$ estimates for the parameter inclusion of DMGP model and RN model in NANOGrav single-pulsar noise models.

PSR	DMGP		RN	
	$\log(\widehat{\mathrm{IBF}})$	Std	$\log(\widehat{\mathrm{IBF}})$	Std
J1012+5307	0.99	0.87	73.94	0.82
J1600−3053	0.52	1.41	11.42	1.43
J1705−1903	166.92	0.75	91.20	0.74
J1713+0747	3.46	4.14	680.88	4.17

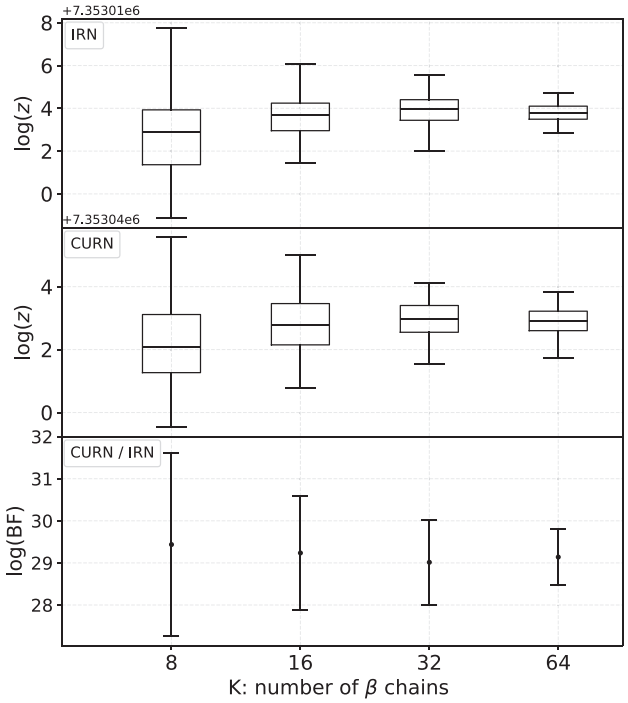

Figure 3. Estimations as a function of K number of β chains for the NANOGrav data set via GSS. Top panel: Log-marginal likelihood estimated for the IRN model. Middle panel: Log-marginal likelihood estimated for the CURN model. Bottom panel: $\log(\widehat{\mathrm{BF}}_{\mathrm{CURN/IRN}})$ comparing CURN to IRN.

4.1.2 Evidence estimation for the GWB analysis

In parallel with our previous analysis, the GSS estimator can also be used to investigate the evidence of GWB in the PTA. We reproduce the posterior samples used in the NANOGrav analysis where all CRS models share the log-spectral amplitude $\log(A_M)$ and the spectral index γ_M of the power law as free parameters for the four following models: intrinsic red noise (IRN), common uncorrelated red noise (CURN), spline overlap reduction function (Spl ORF), and the Hellings–Downs (HD) serving as the probe for GWB. The default set-up and codes provided by the NANOGrav public data release are used in this demonstration. We refer to Agazie et al. (2023) for extended details on the set-up and the models. In this case, around 120 000 samples are generated for each of the four models. Before estimating $\log(z)$, the posterior samples are compared with the previous analyses for consistency. The means and the standard deviations of the generated samples are in agreement with the reported results in Agazie et al. (2023).

To configure the GSS, the 20 000 samples of the burn-in period are discarded from the posterior samples used to tune the reference distribution π_0. By tuning the reference distribution and using it as a sampling proposal distribution, the GSS estimator enables us to skip the new burn-in period in the GSS sampling. In Fig. 3, each $\log(z)$ estimate is computed using $n_{\mathrm{ESS}} \approx 50$ per β chain, while the number of parallel chains grows with $K \in [8, 16, 32, 64]$. A higher number of n_{ESS} is used to cross-check the consistency of the $\log(\widehat{z})$ estimates. As demonstrated in Fig. 3, the GSS estimator produces a consistent median over 100 $\log(\widehat{z})$ starting from $K = 16$ while constraining the uncertainty as the number of chains increases, a higher number of n_{ESS} can improve the estimate of the median $\log(\widehat{z})$ for $K = 8$ as shown in Fig. 2. The $\log(\widehat{\mathrm{BF}})$ is consistent through all K and the slight variation would not affect the inference.

Table 3. Mean and standard deviation of 100 independent log-marginal likelihood estimates with $K = 16$ for different common red process models in the NANOGrav data set.

	$\log(\widehat{z})$	Std
IRN	7353 013.52	1.07
CURN	7353 042.72	0.84
Spl ORF	7353 037.75	1.18
HD	7353 047.80	0.91

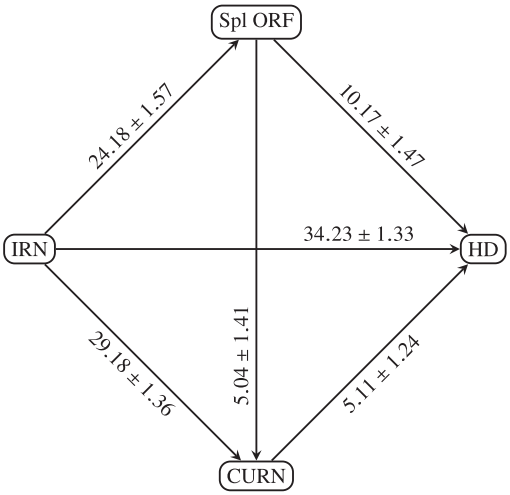

Figure 4. Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{BF}}_{M_1/M_2})$ estimates comparing different common red process models in the NANOGrav data set where the arrow starts at model M_2 and ends at model M_1.

To evaluate the GWB evidence, 100 independent estimates were computed for each of the four models. Table 3 reports the mean and the standard deviation of 100 estimates of $\log(\widehat{z})$. Fig. 4 reports the mean and standard deviation where 1000 estimates of $\log(\widehat{\mathrm{BF}})$ were derived to compare the models. Compared to IRN, the other three models are highly supported, reinforcing the strong evidence of an extra red process in the PTA data set. Furthermore, with a $\log(\widehat{\mathrm{BF}}_{\mathrm{CURN/IRN}})$ of 29.18 ± 1.36, a CURN process is evident in the data set reconfirming the previous PTA analyses (Arzoumanian et al. 2020; Chen et al. 2021; Goncharov et al. 2022). Currently, CURN is favoured over the Spl ORF model with $\log(\widehat{\mathrm{BF}}_{\mathrm{CURN/Spl\,ORF}}) = 5.04 \pm 1.41$. For the GWB evidence, $\log(\widehat{\mathrm{BF}}_{\mathrm{HD/CURN}}) = 5.11 \pm 1.24$ is consistent with the NANOGrav reported estimation of 5.42 ± 4.25, whereas the uncertainty is more constrained by the GSS method.

4.2 Application to EPTA+InPTA

The EPTA+InPTA collaboration has publicly released its second data set with the analysis for GW evidence (Antoniadis et al. 2023b). This analysis combines the data set of 25 pulsars into four subsets: DR2full covers 24.7 yr of EPTA data; DR2new is limited to the last 10.3 yr of observations with the upgraded radio telescopes; DR2full+ and DR2new+ add 0.7 yr of InPTA data to the previous subsets. The EPTA+InPTA study uses customized pulsar noise models reported in Antoniadis et al. (2023a). We will reproduce the analysis using the four subsets for the following models: CURN, GWB (referred to as HD in Section 4.1.2), and binned ORF by using the default set-up provided in the second data release. Around 120 000 posterior samples are generated for each model, and then their means and standard deviations are examined for consistency with the reported

Table 4. Mean and standard deviation of 100 log-marginal likelihood estimates for each common red process model applied to the EPTA+InPTA data sets with $K = 16$.

	DR2new		DR2new+		DR2full		DR2full+	
	$\log(\widehat{z})$	Std	$\log(\widehat{z})$	Std	$\log(\widehat{z})$	Std	$\log(\widehat{z})$	Std
PSRN+CURN	493 684.31	1.02	525 786.67	1.08	607 134.56	1.20	639 232.25	1.03
PSRN+binned ORF	493 681.86	1.30	525 783.84	1.36	607 130.75	1.31	639 228.67	1.05
PSRN+GWB	493 692.55	1.14	525 794.58	0.98	607 140.21	1.12	639 238.08	1.01

Table 5. Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{BF}}_{M/\mathrm{CURN}})$ estimates comparing binned ORF and GWB model to CURN model across EPTA+InPTA data sets.

Model	DR2new		DR2new+		DR2full		DR2full+	
	$\log(\widehat{\mathrm{BF}})$	Std	$\log(\widehat{\mathrm{BF}})$	Std	$\log(\widehat{\mathrm{BF}})$	Std	$\log(\widehat{\mathrm{BF}})$	Std
PSRN+binned ORF	−2.35	1.65	−2.81	1.76	−3.62	1.42	−3.80	1.81
PSRN+GWB	8.24	1.47	7.89	1.48	5.79	1.42	5.70	1.57

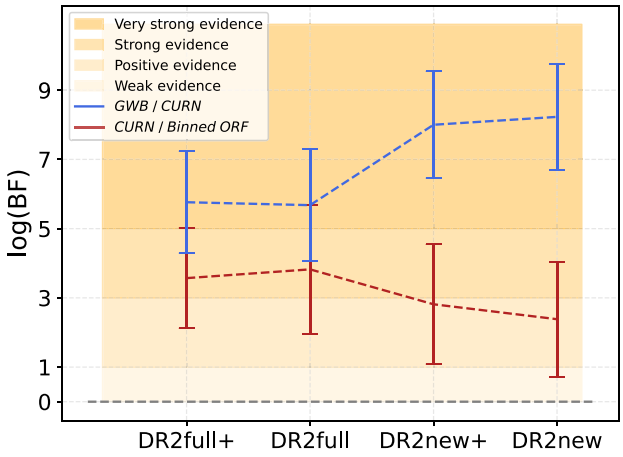

Figure 5. Evolution of $\log(\widehat{\mathrm{BF}}_{M_1/M_2})$ comparing different common red process models as a function of the data set size of EPTA+InPTA data set with DR2full+, the full data set, and DR2new, the shortest subset.

ones in Antoniadis et al. (2023b). A burn-in period of the first 20 000 samples was discarded from the posterior samples used to calibrate the reference distribution π_0. A 100 $\log(\widehat{z})$ per model are estimated with $K = 16$ and $n_{\mathrm{ESS}} \approx 20$ samples per β chain. To assess the stability of the estimation, n_{ESS} and K are increased and $\log(\widehat{z})$ estimates are cross-checked for consistency.Table 4 reports the mean and the standard deviation of 100 $\log(\widehat{z})$ estimates for the three models across all EPTA+InPTA data sets. To compare each model's performance, Table 5 reports the mean and the standard deviation of 1000 $\log(\mathrm{BF}_{M/\mathrm{CURN}})$ independent estimates of the other two models compared to the CURN model. For DR2full and DR2full+, the $\log(\mathrm{BF}_{\mathrm{GWB/CURN}})$ supports the GWB model over the CURN model. This support is further reinforced when fitting only to the new generation data set DR2new. However, the CURN model is favoured over the binned ORF model, and this evidence decreases for DR2new. The addition of the InPTA data set in DR2full+ and DR2new+ slightly decreases the evidence on the GWB model and slightly increases the evidence favouring CURN over binned ORF. Fig. 5 summarizes this evolution of evidence in the EPTA+InPTA data set. Although the InPTA data set is recent, the slight decrease in evidence of GWB might arise from adding 1 yr of data set asymmetrically to only 12 pulsars out of 25 candidates, which will directly affect the correlations between pulsars.

5 DISCUSSION

In the simulation study, we highlighted the differences between methods of power posterior and showed their relative performance. The GSS method has outperformed both TI and SS and proved efficient for higher dimensional cases. GSS can benefit from the previous efforts usually done in the parameter estimation phase. The success of a cheaper GSS estimation relies on having a good posterior approximation, which shortens the path to an accurate estimate. Without such an approximation, the GSS estimator will require more effort (but still less than SS and TI) to achieve an accurate estimate. These $\log(z)$ estimates result in an accurate evaluation of the log(BF) and a robust interpretation of evidence supporting different hypotheses. This evidence can further be tested using the IBF criteria, which examine the influence of a particular set of parameters on the evidence.

Because of its lower computational cost, the GSS method enabled multiple $\log(z)$ estimates with high accuracy for each PTA–Bayesian analysis in this study. For single-noise analysis, GSS has shown its ability to offer model selection without constraints, compared to the hypermodel, which can fail to provide evidence evaluation in cases such as estimation for RN cases (Smarra et al. 2023). In the NANOGrav data set, the four pulsars have shown different levels of significance for RN and DMGP signals, where PSR J1713+0747 showed extreme evidence for RN. Note that PSR J1713+0747 contributed the most to the common RN signal in the EPTA+InPTA data set (Antoniadis et al. 2023b). For NANOGrav GWB evidence, the GSS method has provided a consistent and more accurate estimate than the hypermodel method's estimate of evidence supporting a GWB signal compared to a common RN signal. Furthermore, GSS enabled a direct comparison of expensive models, efficiently eliminating the arbitrary weights required by the hypermodel method. The EPTA GWB analysis consistently reconfirms the evidence supporting a GWB signal in the PTA data set. This evidence becomes more significant for the new-generation data set as shown in Fig. 5.

The PTA experiments rely on model selection to tune each part of the Bayesian noise model, which increases the number of Bayesian analyses required to achieve better sensitivity. The current GSS set-up substantially minimizes the cost of a single $\log(z)$ estimate for both light and expensive models, and would further benefit in the future

from better samplers. The efficiency of the GSS method will help reduce the cost of evidence estimation for the continuously growing model portfolio for GWs and new physics in the PTA (Afzal et al. 2023; Smarra et al. 2023).

ACKNOWLEDGEMENTS

EMZ, PM-R, WvS, and RM gratefully acknowledge support by the Marsden Fund Council grant MFP-UOA2131 from New Zealand Government funding, managed by the Royal Society Te Apārangi. This work was performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR programme receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government, and from the Victorian Higher Education State Investment Fund (VHESIF) provided by the Victorian Government.

Software used: ENTERPRISE (Ellis et al. 2019; Taylor et al. 2021), ENTERPRISE_EXTENSION (Taylor et al. 2021), PTMCMC (Ellis & van Haasteren 2017), MPI4PY (Dalcín et al. 2005), JUPYTER (Kluyver et al. 2016), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), and ARVIZ (Kumar et al. 2019).

DATA AVAILABILITY

Data used in this paper are available on GSS-estimator GitHub.[1]

REFERENCES

Afzal A. et al., 2023, ApJ, 951, L11
Agazie G. et al., 2023, ApJ, 951, L8
Antoniadis J. et al., 2023a, A&A, 678, A49
Antoniadis J. et al., 2023b, A&A, 678, A50
Arzoumanian Z. et al., 2020, ApJ, 905, L34
Carlin B. P., Chib S., 1995, J. R. Stat. Soc. B, 57, 473
Chen S. et al., 2021, MNRAS, 508, 4970
Chib S., Greenberg E., 1995, Am. Stat., 49, 327
Cordes J. M., Shannon R. M., 2010, preprint (arXiv:1010.3785)
Dalcín L., Paz R., Storti M., 2005, J. Parallel Distrib. Comput., 65, 1108
Ellis J., van Haasteren R., 2017, jellis18/PTMCMCSampler: Official Release (1.0.0). Zenodo (https://doi.org/10.5281/zenodo.1037579)
Ellis J. A., Vallisneri M., Taylor S. R., Baker P. T., 2019, Astrophysics Source Code Library, record ascl:1912.015
Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P. O., 2011, Mol. Biol. Evol., 28, 523
Ferdman R. D. et al., 2010, Class. Quantum Grav., 27, 084014
Friel N., Pettitt A. N., 2008, J. R. Stat. Soc. B, 70, 589
Gelman A., Meng X.-L., 1998, Stat. Sci., 13, 163
Goncharov B. et al., 2022, ApJ, 932, L22
Harris C. R. et al., 2020, Nature, 585, 357
Hazboun J. S., Romano J. D., Smith T. L., 2019, Phys. Rev. D, 100, 104028
Hellings R., Downs G., 1983, ApJ, 265, L39
Hinne M., Gronau Q. F., van den Bergh D., Wagenmakers E.-J., 2020, Adv. Methods Pract. Psychol. Sci., 3, 200
Hourihane S., Meyers P., Johnson A., Chatziioannou K., Vallisneri M., 2023, Phys. Rev. D, 107, 084045
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Johnson A. D. et al., 2024, Phys. Rev. D, 109, 103012
Jones M. L. et al., 2017, ApJ, 841, 125
Joshi B. C. et al., 2018, J. Astrophys. Astron., 39, 51
Kluyver T. et al., 2016, in Loizides F., Schmidt B., eds, Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, Amsterdam, p. 87
Kumar R., Carroll C., Hartikainen A., Martin O., 2019, J. Open Source Softw., 4, 1143
Lartillot N., Philippe H., 2006, Syst. Biol., 55, 195
Lentati L., Alexander P., Hobson M. P., Feroz F., van Haasteren R., Lee K., Shannon R. M., 2014, MNRAS, 437, 3004
Lodewyckx T., Kim W., Lee M. D., Tuerlinckx F., Kuppens P., Wagenmakers E.-J., 2011, J. Math. Psychol., 55, 331
McLaughlin M. A., 2013, Class. Quantum Grav., 30, 224008
Manchester R. et al., 2013, Publ. Astron. Soc. Aust., 30, e017
Matsakis D. N., Taylor J. H., Eubanks T. M., 1997, A&A, 326, 924
Maturana-Russel P. A., 2017, PhD thesis, University of Auckland, Auckland.
Maturana-Russel P., Meyer R., Veitch J., Christensen N., 2019, Phys. Rev. D, 99, 084006
Meng X.-L., Wong W. H., 1996, Stat. Sin., 6, 831
Miles M. T. et al., 2023, MNRAS, 519, 3976
Newton M. A., Raftery A. E., 1994, J. R. Stat. Soc. B, 56, 3
Reardon D. J. et al., 2023, ApJ, 951, L6
Sazhin M. V., 1978, Sov. Astron., 22, 36
Skilling J., 2006, Bayesian Anal., 1, 833
Smarra C. et al., 2023, Phys. Rev. Lett., 131, 171001
Stinebring D. R., Ryba M. F., Taylor J. H., Romani R. W., 1990, Phys. Rev. Lett., 65, 285
Taylor S. R., Baker P. T., Hazboun J. S., Simon J., Vigeland S. J., 2021, enterprise_extensions v2.4.3.
Tiburzi C. et al., 2016, MNRAS, 455, 4339
van Haasteren R., Vallisneri M., 2014, Phys. Rev. D, 90, 104012
van Haasteren R., Levin Y., McDonald P., Lu T., 2009, MNRAS, 395, 1005
Verbiest J. et al., 2009, MNRAS, 400, 951
Virtanen P. et al., 2020, scipy/scipy: SciPy 1.6.0 (v1.6.0). Zenodo Available at: (https://doi.org/10.5281/zenodo.4406806)
Xie W., Lewis P. O., Fan Y., Kuo L., Chen M.-H., 2011, Syst. Biol., 60, 150
Xu H. et al., 2023, Res. Astron. Astrophys., 23, 075024

[1] https://github.com/nz-gravity/GSS-estimator

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