Understanding and analysing time-correlated stochastic signals in pulsar timing

Rutger van Haasteren[1,2]* and Yuri Levin[2,3]

[1]*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), D-30167 Hannover, Germany*
[2]*Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands*
[3]*School of Physics, Monash University, PO Box 27, VIC 3800, Australia*

ABSTRACT

Although it is widely understood that pulsar timing observations generally contain time-correlated stochastic signals (TCSSs; red timing noise is of this type), most data analysis techniques that have been developed make an assumption that the stochastic uncertainties in the data are uncorrelated, i.e. 'white'. Recent work has pointed out that this can introduce severe bias in the determination of timing-model parameters and that better analysis methods should be used. This paper presents a detailed investigation of timing-model fitting in the presence of TCSSs and gives closed expressions for the post-fit signals in the data. This results in a Bayesian technique to obtain timing-model parameter estimates in the presence of TCSSs, as well as computationally more efficient expressions of their marginalized posterior distribution. A new method to analyse hundreds of mock data set realizations simultaneously without significant computational overhead is presented, as well as a statistically rigorous method to check the internal consistency of the results. As a by-product of the analysis, closed expressions of the rms introduced by a stochastic background of gravitational waves in timing residuals are obtained, valid for regularly sampled data. Using T as the length of the data set and $h_c(1\,\mathrm{yr}^{-1})$ as the characteristic strain, this is $\sigma_{\mathrm{GWB}}^2 = h_c(1\,\mathrm{yr}^{-1})^2 (9\sqrt[3]{2\pi^4}\Gamma(-10/3)/8008)\,\mathrm{yr}^{-4/3} T^{10/3}$.

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

1 INTRODUCTION

Over the years, pulsar timing has proved to be a useful tool for probing a wide range of science. Prime examples include the confirmation of the emission of gravitational waves (Taylor & Weisberg 1982) and very accurate tests of general relativity (Kramer et al. 2006). An overview of techniques used in pulsar timing is given in Lorimer & Kramer (2005), and a detailed description of current treatment of timing data is given in the TEMPO2 papers (Edwards, Hobbs & Manchester 2006; Hobbs, Edwards & Manchester 2006). Much of the interesting science results from the fact that accurate measurements of the times of arrival (TOAs) of the radio pulses allow one to precisely track the pulsar trajectories relative to the Earth and that the observed TOAs can be accounted for very precisely by building a physical model of pulsar trajectory, pulse propagation and pulsar spin evolution in relativistic gravity. Such a model is referred to as the *timing model*; mathematically the parameters of the timing model are determined by the χ^2 minimization for the TOA fit.

The remaining differences between the TOAs and the timing model are called timing residuals (TRs). The physics beyond that

included in the timing model is contained in TRs. Of particular interest are the time-correlated stochastic signals (TCSSs), examples of which include

(1) the so-called pulsar timing noise, or red spin noise, a generic term for random changes in the pulsar rotational rate, thought to be possibly due to the random angular-momentum exchange between the normal and superfluid components,

(2) the time-dependent influence of the interstellar medium on the optical path length between the pulsar and the Earth and

(3) the influence of the stochastic gravitational-wave background (GWB) on the pulse TOAs.

For a recent discussion of all of these, see Cordes & Shannon (2010) and Shannon & Cordes (2010).

All of the above processes feature a red spectrum, and their contributions to the TOAs are difficult to disentangle from the variation of the deterministic timing-model parameters. The purpose of this paper is to develop a rigorous and efficient procedure for a TOA analysis which includes simultaneously the timing model and a TCSS. The two specific questions which are explicitly addressed are as follows.

(1) How does the timing-model fitting affect the statistical properties of TCSS-induced TRs?

* E-mail: haasteren@strw.leidenuniv.nl

(2) Conversely, how does the presence of a TCSS affect the uncertainties of the timing-model parameters?

Our method, as well as our answers to the above questions, is extensively tested on mock data.

The detailed plan of the paper is as follows. In Section 2 we develop a formalism that casts fitting of the timing model in the presence of red noise as a non-orthogonal projection of the covariance matrix. A connection with least-squares fitting methods is made (as used by e.g. Coles et al. 2011, CHCMV hereafter). This leads to understanding of degeneracies between the covariance matrix and the timing model. We show how to exploit these degeneracies in Section 3, which results in improved expressions for the covariance function and the marginalized posterior distribution. We describe how to obtain estimates for the timing-model parameters in Section 4. In Section 5 we introduce a computationally efficient method to analyse hundreds of mock data sets simultaneously, and we describe a powerful test based on the Kolmogorov–Smirnov (K-S) statistic that we use to check that the Bayesian analysis method produces consistent results. Finally, we compare our results with the recently proposed Cholesky method of CHCMV in Section 6.

2 TIMING-MODEL FITS AND THE COVARIANCE FUNCTION

The observed TOAs of every pulsar contain contributions from many deterministic and stochastic processes. The traditional procedure ignores any (except TOA uncertainties) stochastic or unknown deterministic contributions to the TRs (e.g. the standard weighted least-squares fit in TEMPO2; see Hobbs et al. 2006). Therefore, some of these get absorbed into the timing-model fits, which also alter the estimates of the timing-model parameters. In this section we show how to model TCSSs in combination with fitting to the timing model.

2.1 The covariance function and least-squares fitting

We describe the n TOAs of a single pulsar as an addition of a deterministic and a stochastic part:

$$t^{\mathrm{arr}} = t^{\mathrm{det}} + \delta t^{\mathrm{rgp}}, \tag{1}$$

where the n elements of t are the observed TOAs, t^{det} are the deterministic contributions to the TOAs and δt^{rgp} are the stochastic contributions to the TOAs, which in this work are TCSSs all modelled by a random Gaussian process. In practice, the pre-fit TRs are produced with first estimates β_{0i} of the m timing-model parameters β_i; this initial guess is usually precise enough so that a linear approximation of the timing model can be used further on (Edwards et al. 2006). Namely, it is a good assumption that the remaining TRs depend linearly on $\xi_a = \beta_a - \beta_{0a}$:

$$\delta t = \delta t^{\mathrm{prf}} + \mathbf{M}\xi, \tag{2}$$

where δt are the TRs in the linear approximation to the timing model, δt^{prf} is the vector of pre-fit TRs, ξ is the vector with timing-model parameters and the $(n \times m)$ matrix $\mathbf{M}$ is the so-called design matrix (see e.g. section 15.4 of Press et al. 1992; van Haasteren et al. 2009, hereafter vHLML), which describes how the TRs depend on the model parameters. Without loss of generality, we assume here and in subsequent sections that $\mathbf{M}$ has been constructed such that its columns are linearly independent. Although the distinction between deterministic and TCSS contributions here seems analogous

to equation (1), we note here that there is significant absorption of TCSSs in the fit.

The deterministic signals in t^{det} are well modelled in standard pulsar timing packages like e.g. TEMPO2 (Hobbs et al. 2006). We model the TCSS contributions as a random Gaussian process with a covariance matrix defined by

$$\langle \delta t_i^{\mathrm{rgp}} \delta t_j^{\mathrm{rgp}} \rangle = C_{ij}, \tag{3}$$

where the brackets $\langle \dots \rangle$ denote the ensemble average of the random process, and the indices i and j run from 1 to n. The covariance matrix is the numerical representation of the covariance function, and we assume here that it can be parametrized with parameters $\boldsymbol{\phi}$. We use the following convention for the Wiener–Khinchin theorem to relate the covariance function to the spectral density:

$$\mathbf{C}(\tau) = \int_0^\infty S(f) \cos(\tau f) \, \mathrm{d}f, \tag{4}$$

where $S(f)$ is the spectral density of δt^{rgp} as a function of frequency, and $\tau = 2\pi|t_1 - t_2|$ is the time difference between two observations multiplied with 2π. We emphasize here that we do not model the post-fit TRs, but the TRs δt^{rgp} prior to the fitting process.

The Bayesian likelihood of the TRs is given by (vHLML)

$$P(\delta t|\xi, \boldsymbol{\phi}) = \frac{1}{\sqrt{(2\pi)^n \det C}}$$
$$\times \exp\left(\frac{-1}{2}(\delta t - \mathbf{M}\xi)^{\mathrm{T}} \mathbf{C}^{-1}(\delta t - \mathbf{M}\xi)\right). \tag{5}$$

Provided we know the value of the parameters $\boldsymbol{\phi}$ prior to the analysis, i.e. provided we know the covariance matrix $\mathbf{C}$, we can maximize the likelihood with respect to the model parameters. This results in the generalized least-squares (GLS) estimator for the timing-model parameters:

$$\chi_\xi = \left(\mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{M}\right)^{-1} \mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\delta t^{\mathrm{prf}}$$
$$\delta t^{\mathrm{pof}} = (\mathbb{I}_n - \mathbf{M}\left(\mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{M}\right)^{-1} \mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1})\delta t^{\mathrm{prf}}$$
$$= (\mathbb{I}_n - \mathbf{B})\,\delta t^{\mathrm{pof}} = \mathbf{O}\delta t^{\mathrm{pof}}, \tag{6}$$

where χ_ξ are the best-fitting timing-model parameters, $\mathbb{I}_n$ is the n-dimensional identity matrix, δt^{pof} are the post-fit residuals, and $\mathbf{O}$ and $\mathbf{B}$ are the matrices that represent the 'removal' of the timing model.

Using an estimate for the spectral density $S(f)$ or the covariance matrix $\mathbf{C}$ to improve timing-model estimates is not new to pulsar timing. First, almost three decades ago Blandford, Romani & Narayan (1984) analytically showed what the effect of the timing noise spectrum is on the timing-model parameter estimates.[1] Their methods and conclusions are similar: they use an analytically derived orthogonal basis to project out the timing-model basis vectors (Section 3.3), and they specifically consider a power-law spectral model for the timing noise. However, their results do depend on an estimate for the noise; we advocate marginalizing over those parameters in this work, since we generally do not know the values of $\boldsymbol{\phi}$ prior to the analysis.

Secondly, the Cholesky method of CHCMV uses the GLS estimator of equation (6), combined with an estimate for the covariance matrix $\mathbf{C}$.

---

[1] We thank the anonymous referee for bringing this paper to our attention. Some of the results in their work were unknown to us when we rederived them.

Thirdly, Demorest et al. (2012) also use the GLS estimator in their efforts to constrain or detect an isotropic stochastic GWB, where they use a Bayesian approach to estimate the noise covariance matrix.

We advocate to marginalize over the parameters of the covariance matrix, but we do conclude in this work that for most of the timing-model parameters the GLS method works well.

2.2 Effect of fitting on the covariance function

Irrespective of what technique is used to produce the best-fitting timing-model parameters and the post-fit TRs, the resulting post-fit TRs are not correlated according to equation (3). In order to compute the exact effect that fitting has on the post-fit correlations, we take the following approach. Let us assume that we have TR vectors of length n, defined on an interval $[-T, T]$, and some deterministic process defined by m parameters represented by $\boldsymbol{\xi}$. The effect of the deterministic process on the TRs is given by $\mathbf{M}\boldsymbol{\xi}$. Typically, a fitting procedure removes the contributions of the parameters $\boldsymbol{\xi}$ to the TRs with respect to some inner product. We define the inner product on the vector space of TRs as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_E = \boldsymbol{x}^{\mathrm{T}} \mathbf{E}^{-1} \boldsymbol{y}, \tag{7}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the vectors of TRs, $\mathbf{E}$ is a positive definite symmetric (PDS) matrix and $\langle \ldots, \ldots \rangle_E$ indicates an inner product with PDS matrix $\mathbf{E}$. Consider a fitting procedure that produces post-fit TRs that satisfy $\langle \boldsymbol{x}, \mathbf{M}\boldsymbol{\xi} \rangle_E = 0$. It is straightforward to check that this is the weighted least-squares fit when $E_{ij} = \delta_{ij}\sigma_i^2$. When $\mathbf{E}$ is a more general PDS matrix, the fitting process that removes all $\mathbf{M}\boldsymbol{\xi}$ from the TRs is the GLS of equation (6). For our purposes, the exact form of $\mathbf{E}$ is not relevant.

We use the expressions for $\mathbf{O}$ and $\mathbf{B}$ as in equation (6), but now with covariance function $\mathbf{E}$: $\mathbf{B} = \mathbf{M}(\mathbf{M}^{\mathrm{T}}\mathbf{E}^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{E}^{-1}$. Both $\mathbf{O}$ and $\mathbf{B}$ are non-orthogonal projection matrices: $\mathbf{B} = \mathbf{B}^2$, $\mathbf{B}^{\mathrm{T}} \neq \mathbf{B}$ and likewise for $\mathbf{O}$. The correlations in post-fit TRs due to a random Gaussian process with covariance matrix $\mathbf{C}$ are given by (see also Demorest et al. 2012)

$$\left\langle \delta t_i^{\mathrm{pof}} \delta t_j^{\mathrm{pof}} \right\rangle = \left\langle \left(\mathbf{O}\delta t_i^{\mathrm{pof}} \right) \left(\mathbf{O}\delta t^{\mathrm{pof}} \right)_j \right\rangle = \left(\mathbf{O}\mathbf{C}\mathbf{O}^{\mathrm{T}} \right)_{ij}. \tag{8}$$

The correlations in the post-fit TRs are thus given by the covariance matrix of the random process $\mathbf{C}$, projected with the matrix $\mathbf{O}$, where $\mathbf{O}$ removes any contribution $\mathbf{M}\boldsymbol{\xi}$ with respect to the inner product of equation (7). From here onwards, we omit the superscript 'pof' for δt, and by default we assume that we are dealing with post-fit TRs, with respect to some inner product.

3 EXPLOITING FITTING DEGENERACIES

Because the fitting process is effectively a projection of the TRs, the covariance matrix that describes the post-fit residuals is also a projection of the pre-fit covariance matrix. This post-fit covariance matrix is therefore singular, and the pre-fit covariance matrix cannot be reconstructed from the post-fit covariance matrix alone: there is a degeneracy in the processes that could have generated a single realization of post-fit TRs. In this section we use this degeneracy to derive closed expressions for the post-fit covariance function and more computationally efficient expressions for the marginalized posterior distribution.

3.1 The post-fit covariance function

In equation (8) we used the non-orthogonal projection matrices $\mathbf{O}$ and $\mathbf{B}$ to remove any contribution $\mathbf{M}\boldsymbol{\xi}$ to the TRs with respect to the inner product of equation (7). Now consider two related projection matrices:

$$\mathbf{D} = \mathbf{M} \left(\mathbf{M}^{\mathrm{T}}\mathbf{M} \right)^{-1} \mathbf{M}^{\mathrm{T}}$$

$$\mathbf{W} = \mathbb{I}_n - \mathbf{D}. \tag{9}$$

Both $\mathbf{D}$ and $\mathbf{W}$ are orthogonal projections, which satisfy $\mathbf{W}^2 = \mathbf{W}$ and $\mathbf{W} = \mathbf{W}^{\mathrm{T}}$, and likewise for $\mathbf{D}$. The relation with $\mathbf{B}$ and $\mathbf{O}$ is intuitive: if the covariance matrix $\mathbf{E}$ of the inner product of equation (7) is the identity matrix, then we have $\mathbf{W} = \mathbf{O}$ and $\mathbf{D} = \mathbf{B}$. These four projection matrices have the following interesting properties: $\mathbf{BD} = \mathbf{D}$ and $\mathbf{DB} = \mathbf{B}$, and similar expressions for $\mathbf{W}$ and $\mathbf{O}$. From this it follows that

$$\langle \delta t \delta t^{\mathrm{T}} \rangle = \mathbf{O}\mathbf{C}\mathbf{O}^{\mathrm{T}} = \mathbf{O}\mathbf{W}\mathbf{C}\mathbf{W}^{\mathrm{T}}\mathbf{O}^{\mathrm{T}}, \tag{10}$$

where the square matrix $\delta t \delta t^{\mathrm{T}}$ is the dyadic product of two vectors. This expression shows that the degeneracy in the covariance function of the post-fit TRs allows us to equivalently use $\mathbf{W}\mathbf{C}\mathbf{W}^{\mathrm{T}}$ instead of $\mathbf{C}$. In Section 3.3 we derive an analytical expression for this post-fit covariance function, see equation (18).

3.2 A simplified marginalized posterior

As shown in vHLML, it is possible to analytically marginalize the posterior distribution when a flat prior distribution is assumed for the linear parameters (in Appendix B we show how to include Gaussian priors). The marginalized posterior distribution is then equal to the likelihood function of equation (5) integrated over the linear parameters $\boldsymbol{\xi}$:

$$\int \mathrm{d}^m \boldsymbol{\xi}\, P(\delta t | \boldsymbol{\xi}, \boldsymbol{\phi}) = \frac{\sqrt{\det \left(\mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{M} \right)^{-1}}}{\sqrt{(2\pi)^{n-m} \det \mathbf{C}}}$$
$$\times \exp \left(\frac{-1}{2} \delta t^{\mathrm{T}} \mathbf{C}' \delta t \right), \tag{11}$$

with

$$\mathbf{C}' = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{M} \left(\mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{M} \right)^{-1} \mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}. \tag{12}$$

Equations (11) and (12) are the computational bottleneck for the analysis of TCSSs in pulsar timing. These equations involve non-trivial operations on large, dense matrices. Specifically, they involve one n^3 operation for the inversion (or the Cholesky decomposition of $\mathbf{C}$), an m^3 operation for the inversion of $\mathbf{M}^{\mathrm{T}}\mathbf{C}\mathbf{M}$ and a lot of vector–matrix operations that scale as n^2. In this section we seek to simplify these equations for transparency and computational efficiency.

We re-express the effect of the linear parameters $\boldsymbol{\xi}$ on the TRs in terms of an orthonormal basis. To this end we factorize the matrix $\mathbf{M}$ with a singular value decomposition:

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^*, \tag{13}$$

where $\mathbf{U}$ and $\mathbf{V}$ are, respectively, $(n \times n)$ and $(m \times m)$ orthogonal matrices, and Σ is an $(n \times m)$ diagonal matrix. For our purposes, the column space of the orthogonal matrix $\mathbf{U}$ is important. The first m columns of $\mathbf{U}$ span the column space of $\mathbf{M}$, and the last $n - m$ columns of $\mathbf{U}$ span the complement. We now construct the matrices $\mathbf{F}$ and $\mathbf{G}$ as follows. $\mathbf{U} = (\mathbf{F}\ \mathbf{G})$, where $\mathbf{F}$ is the $(n \times m)$ matrix consisting of the first m columns of $\mathbf{U}$, and $\mathbf{G}$ is the $(n \times (n - m))$ matrix consisting of the other columns of $\mathbf{U}$. The following identities

hold:

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbb{I}_m$$
$$\mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbb{I}_{n-m}$$
$$\mathbf{F}\mathbf{F}^{\mathrm{T}} + \mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{D} + \mathbf{W} = \mathbb{I}_n. \tag{14}$$

Using these expressions, it is now possible to show that the marginalized likelihood of equation (11) is equal to[2]

$$\int \mathrm{d}^m\boldsymbol{\xi}\, P(\delta\boldsymbol{t}|\boldsymbol{\xi},\boldsymbol{\phi}) = \frac{1}{\sqrt{(2\pi)^{n-m}\det\left(\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}\right)}}$$
$$\times \exp\left(\frac{-1}{2}\delta\boldsymbol{t}^{\mathrm{T}}\mathbf{G}\left(\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\delta\boldsymbol{t}\right). \tag{15}$$

More intuitively, the marginalized likelihood distribution is the likelihood function of an $(n-m)$-dimensional random Gaussian process of the data $\mathbf{G}^{\mathrm{T}}\delta\boldsymbol{t}$, with a covariance function $\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}$. This more insightful expression involves two matrix–matrix multiplications and an inversion (we ignore the vector–matrix operations, which are n^2 operations). Because $\mathbf{G}$ is block diagonal and only needs to be calculated once, equation (15) can be implemented in such a way that the main computational burden is the inversion of $\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}$. Since $\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}$ is a PDS matrix, equation (15) is best evaluated by using the Cholesky decomposition, which also directly gives access to the value of the required determinant.

The theoretically more insightful expression we arrive at in equation (15) is slightly more efficient than equation (11). Also, the dependence on the divergent low-frequency cutoff terms has been removed before the inversion (see Section 3.4), which improves numerical stability. However, the final computation still scales as n^3, so it will remain a computational bottleneck in this type of analysis.

3.3 Analytic post-fit covariance functions

In the previous section, we have shown that we are allowed to use $\mathbf{C}^{\mathrm{P}} = \mathbf{W}\mathbf{C}\mathbf{W}^{\mathrm{T}}$ instead of $\mathbf{C}$ in all our equations that describe post-fit TRs. We analytically approximate this quantity by using the inner product of equation (7),

$$\langle\boldsymbol{x},\boldsymbol{y}\rangle_E = \sum_i \frac{x(t_i)\,y(t_i)}{\sigma^2} \approx \frac{1}{\sigma^2\Delta t}\int_{-T}^{T} x(t)y(t)\,\mathrm{d}t, \tag{16}$$

where we have used $E_{ij} = \delta_{ij}\sigma^2$, with σ being the uncertainty of the TOAs, Δt is the time interval between observations, and $x(t)$ and $y(t)$ are continuous functions on the interval $[-T, T]$. On the interval $[-T, T]$, we define an orthonormal basis of quadratic functions [see van Haasteren & Levin (2010) for a similar application], $\hat{f}_1(t), \hat{f}_2(t), \hat{f}_3(t)$:

$$\hat{f}_1(t) = \frac{1}{\sqrt{2}}\sigma\sqrt{\frac{\Delta t}{T}}$$
$$\hat{f}_2(t) = \sqrt{\frac{3}{2}}\sigma\sqrt{\frac{\Delta t}{T}}\frac{t}{T}$$
$$\hat{f}_3(t) = \sqrt{\frac{45}{8}}\sigma\sqrt{\frac{\Delta t}{T}}\left[\left(\frac{t}{T}\right)^2 - \frac{1}{3}\right]. \tag{17}$$

These basis functions satisfy $\langle\hat{f}_i,\hat{f}_j\rangle_E = \delta_{ij}$. The process of fitting for quadratics can now be expressed as a projection of the covariance functions in terms of these basis functions:

$$C^{\mathrm{P}}(t_0, t_3) = S(t_0, t_1)C(t_1, t_2)S(t_2, t_3). \tag{18}$$

From here onwards, we always take inner product given by equation (16) over the repeated variables t_1 and t_2, and $S(t_k, t_l)$ is given by

$$S(t_k, t_l) = \sigma^2\Delta t\,\delta\,(t_k - t_l) - \sum_{i=1}^{3}\hat{f}_i(t_k)\hat{f}_i(t_l), \tag{19}$$

with $\delta(x)$ being the Dirac delta function. Using this formalism, it is possible to analytically derive the projected covariance function $\mathbf{C}^{\mathrm{P}}$ for TCSSs with any spectral density.

3.4 Power-law covariance function

Power-law spectra are of particular importance in PTA applications, since the stochastic GWB is expected to be a signal well described by such a power spectral density (Begelman, Blandford & Rees 1980; Phinney 2001; Jaffe & Backer 2003; Wyithe & Loeb 2003; Sesana, Vecchio & Colacino 2008). Strictly speaking, a process governed by a power-law spectral density is improper, and therefore unphysical. For PTA purposes however, the power-law behaviour of the signal is expected to hold within its expected frequency band of 0.1–10 yr^{-1}. To enforce finiteness of the covariance function, it is convenient to introduce cutoff frequencies in equation (4), see vHLML. These can be problematic in practice: the cutoff frequency needs to be low enough for the signal to represent a power-law signal, yet it must be high enough for its numerical representation not to cause numerical artefacts due to limited machine precision. In this section we show that by choosing the projection matrix to represent a fitting procedure that includes quadratic spindown, the dependence of the covariance matrix on the low cutoff frequency is explicitly removed.

We parametrize the spectral density as

$$S(f) = A^2\left(\frac{1}{1\,\mathrm{yr}^{-1}}\right)\left(\frac{f}{1\,\mathrm{yr}^{-1}}\right)^{-\gamma}, \tag{20}$$

with A being the amplitude of the signal (units time) and γ the spectral index. We require a low-frequency cutoff f_{L} if $\gamma \geq 1$. As shown in vHLML, in that case the covariance function is equal to

$$C_{ij}^{\mathrm{PL}} = A^2\left(\frac{1\,\mathrm{yr}^{-1}}{f_{\mathrm{L}}}\right)^{\gamma-1}\left\{\Gamma(1-\gamma)\sin\left(\frac{\pi\gamma}{2}\right)\left(f_{\mathrm{L}}\tau_{ij}\right)^{\gamma-1}\right.$$
$$\left. - \sum_{n=0}^{\infty}(-1)^n\frac{\left(f_{\mathrm{L}}\tau_{ij}\right)^{2n}}{(2n)!\,(2n+1-\gamma)}\right\}, \tag{21}$$

where $\tau_{ij} = 2\pi|t_i - t_j|$. In vHLML it is shown that the removal of quadratic spindown from the TRs also completely removes any dependences on the low-frequency cutoff f_{L}. In the numerical calculations, we need to choose $f_{\mathrm{L}} \ll T^{-1}$ so that $\mathbf{C}^{\mathrm{PL}}$ is PDS, and so that we can neglect the terms with $n \geq 2$ in the summation of equation (21). In practice, the diverging terms dependent on the cutoff frequency can result in numerical artefacts due to limited machine precision.

We apply equation (18) to the covariance function of a power-law signal $\mathbf{C}^{\mathrm{PL}}$ of equation (21) to obtain the projected covariance function $\mathbf{C}^{\mathrm{P}}(t_0, t_3)$ for a power-law signal. We present the details of this calculation and the explicit formulae in Appendix A. One of the key results of this calculation is that dependences on the low-frequency cutoff f_{L} are removed from the infinite summation of

---

[2] A useful identity is

$$\mathbf{U}^{\mathrm{T}}\mathbf{C}\mathbf{U} =$$
$$\begin{pmatrix} \mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G} & 0 \\ \mathbf{F}^{\mathrm{T}}\mathbf{C}\mathbf{G} & \mathbb{I}_{n-m} \end{pmatrix}\begin{pmatrix} \mathbb{I}_m & \left(\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{F} \\ 0 & \mathbf{F}^{\mathrm{T}}\mathbf{C}\mathbf{F} - \mathbf{F}^{\mathrm{T}}\mathbf{C}\mathbf{G}\left(\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}\mathbf{F} \end{pmatrix}.$$

Covariance function for a power-law signal (f$_L$ = 0.05 yr^{-1})
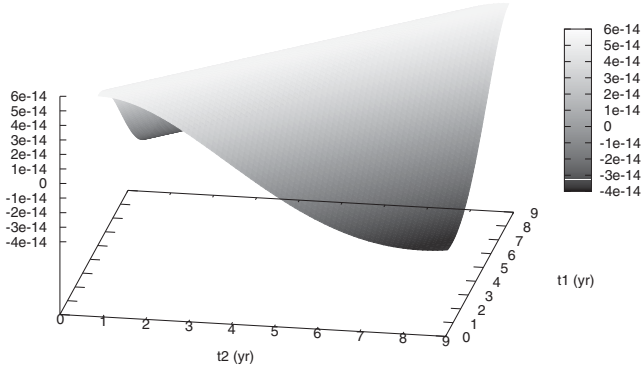
Figure 1. The covariance function $C(t_1, t_2)$ of a power-law spectrum of the form $S(f) = (A^2/1 \text{ yr}^{-1}) \times (f/\text{ yr}^{-1})^{-13/3}$, with $A = 2.9$ns. For comparison, this is equivalent to a GWB with $A_h = 10^{-15}$. Here the effect of fitting has been neglected. Note that the covariance function only depends on the value $t_1 - t_2$.

Covariance function for a power-law signal with qsd fitting

Figure 2. The covariance function $C(t_1, t_2)$ of the same TCSS as in Fig. 1. Here the effect of fitting for quadratics has been taken into account analytically with equation (A7) of Appendix A. Note that, in contrast to Fig. 1, the covariance function not only depends on the value $t_1 - t_2$, but on both t_1 and t_2.

Covariance function for a power-law signal with timing model fitting

Figure 3. The covariance function $C(t_1, t_2)$ of the same TCSS as in Fig. 1. Here the effect of fitting for the whole timing model of J1640+2224 has been taken into account numerically. Note that, in contrast to Fig. 1, the covariance function not only depends on the value $t_1 - t_2$, but on both t_1 and t_2. We use the timing-model parameters from the literature (Löhmer et al. 2005) and the timing model as used in TEMPO2, which includes the parameters: position (right ascension and declination), quadratic spindown, proper motion (right ascension and declination), eccentricity and the projected semimajor axis of the binary orbit. The sampling cadence was two weeks.

equation (21) up to $n = 2$. This ensures that the quadratic spindown fits have completely removed the sensitivity to f_L up to $\gamma < 7$. Although this result had been found before (Blandford et al. 1984), other authors (e.g. vHLML) have assumed that it was only true for $\gamma < 5$. After this work appeared as a preprint, Lee et al. (2012) also obtained both this result and equation (22).

Now that we have an analytic expression for $C^P(t_0, t_3)$ for a process with the power-law spectral density of equation (20), it is possible to derive an expression for the average rms in the post-fit TRs, valid for $1 < \gamma < 7$:

$$\sigma_{PL}^2 = \frac{1}{2T} \int_{-T}^{T} dt\, C^P(t, t)$$
$$= \frac{3(5 - \gamma)(\gamma - 3)2^\gamma (2\pi)^{\gamma-1}}{\gamma(1 + \gamma)(3 + \gamma)(5 + \gamma)}$$
$$\times A^2 \Gamma(1 - \gamma) \sin\left(\frac{\pi\gamma}{2}\right) 1 \text{ yr}^{1-\gamma} T^{\gamma-1}. \quad (22)$$

From this expression, we can derive an estimate for the rms generated by a GWB signal of the form

$$h_c(f) = A_h \left(\frac{f}{1 \text{ yr}^{-1}}\right)^{-2/3}$$
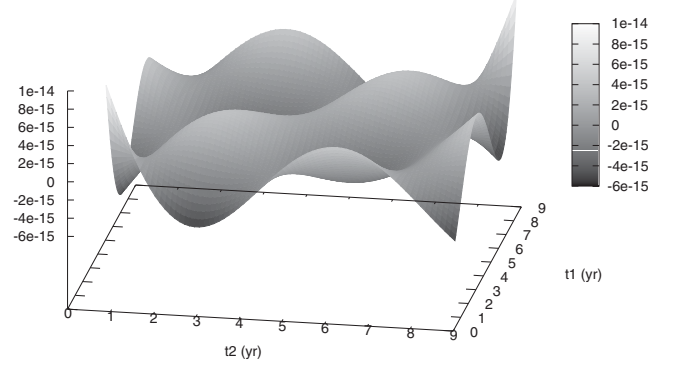$$S(f) = \frac{A_h^2}{12\pi^2} 1 \text{ yr}^3 \left(\frac{f}{1 \text{ yr}^{-1}}\right)^{-\frac{13}{3}}, \quad (23)$$

where A_h is the dimensionless amplitude of the GWB characteristic strain h_c. This results in an estimate for the rms of a GWB:

$$\sigma_{GWB}^2 = A_h^2 \left(\frac{9\sqrt[3]{4}\pi^{\frac{4}{3}}\Gamma\left(-\frac{10}{3}\right)}{1001}\right) \text{yr}^{-\frac{4}{3}} T^{\frac{10}{3}}$$
$$\sigma_{GWB} = 4.35 \times 10^{-9} \left(\frac{A_h}{10^{-15}}\right)\left(\frac{T}{\text{yr}}\right)^{\frac{5}{3}}. \quad (24)$$

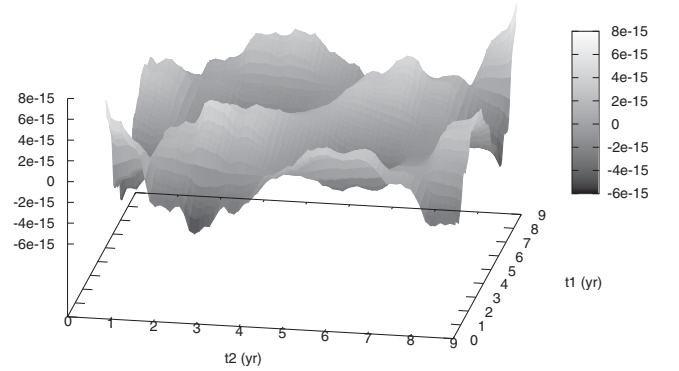Note here that $2T$ is the total duration of the experiment.

As the covariance function is a function of two variables, it is instructive to inspect the results of this section visually. First, consider the pre-fit covariance matrix $\mathbf{C}^{PL}$ of equation (21). This is a function of only the difference between the two parameters $|t_1 - t_2|$, since it describes a stationary random process. This is illustrated in Fig. 1, where lines of equal covariance are lines of equal $|t_1 - t_2|$. Secondly, we demonstrate what the effect of fitting for quadratics is

on the covariance function in Fig. 2. The symmetry due to the time stationarity of the random process that is present in Fig. 1 is broken, and we see prominent cubical features at the edges of the plot. Thirdly, we have included the effect of fitting to the entire timing model of an example pulsar in Fig. 3. This figure is similar to Fig. 2, except that there are some small-scale features on top of the general structure. This demonstrates that the quadratic spindown fitting is the effect that most prominently affects the covariance matrix; all other contributions are minor in comparison.

Because the unweighted least-squares fit is unlikely to be optimal for any realistic data set, it may seem that Figs 2 and 3 do not represent a post-fit TCSS if a more appropriate fitting procedure is used (e.g. the Cholesky method). One may therefore argue that equation (24) is not a good measure of the amount of detectable GWB signal in the data. However, the equivalence of $\mathbf{C}$ and $\mathbf{WCW}^T$ in equations (10) and (15) shows that the extra rms in the post-fit TRs

that may result from an improved fitting routine does not increase the sensitivity to a TCSS, since the marginalized posterior distribution for the TCSS parameters is always the same. The GWB rms quoted in equation (15) is the part of the TCSS that cannot be absorbed in the timing model: the averaged trace of any post-fit covariance function is greater than or equal to this value. The quoted GWB rms is therefore a measure of how much 'detectable' signal is in the data.

4 TIMING-MODEL ANALYSIS

When doing a Markov Chain Monte Carlo simulation (MCMC), we analytically marginalize over the timing-model parameters. We would like to retain the information about the timing-model parameters, without adding these dimensions to the MCMC. In this section we show how to do this efficiently.

The marginalized posterior of equation (15) allows one to numerically marginalize over all the stochastic parameters of the model, while analytically marginalizing over the timing-model parameters with a flat prior. Using that equation it is impossible to obtain best estimates for the timing-model parameters. Here we show what extra steps are needed to be taken in order to infer the timing-model parameters. We rewrite the likelihood of equation (5) as follows, using the same notation as in equation (15):

$$
P(\delta t|\boldsymbol{\xi}, \boldsymbol{\phi}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}}
$$
$$
\times \exp\left(\frac{-1}{2}(\boldsymbol{\xi} - \boldsymbol{\chi})^\mathrm{T} \Sigma^{-1} (\boldsymbol{\xi} - \boldsymbol{\chi})\right)
$$
$$
\times \exp\left(\frac{-1}{2}\delta t^\mathrm{T} \mathbf{G} \left(\mathbf{G}^\mathrm{T}\mathbf{C}\mathbf{G}\right)^{-1} \mathbf{G}^\mathrm{T}\delta t\right), \quad (25)
$$

where $\boldsymbol{\chi} = (\mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\mathbf{M})^{-1}\mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\delta t$, $\Sigma^{-1} = \mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\mathbf{M}$, and as before, the stochastic parameters are stored in the l-dimensional vector $\boldsymbol{\phi}$. In Appendix B we show how to include timing-model parameters with Gaussian priors in a similar manner. We are interested in recovering the timing-model parameters

$$
\boldsymbol{\xi} = \sum_{i=1}^{m} \xi_i \hat{e}_i, \quad (26)
$$

where the $\hat{e}_i$ denote the basis vectors of the timing-model parameter space.

The main idea is that in equation (25), which is based on a linear approximation to the timing model, the likelihood function is a multivariable Gaussian with respect to the timing-model parameters. The full posterior distribution can be reconstructed without numerically exploring the timing-model parameters as follows.

(1) Constructing a Markov chain, using the vHLML posterior distribution of equation (15) that faithfully samples the stochastic parameters.

(2) Arithmetically averaging the Gaussian timing-model posteriors from each point of the chain.

The full timing-model posterior distribution can then be used to obtain marginalized posterior distributions for any combination of timing-model parameters. Importantly, as will be demonstrated shortly, all this is done without any extra significant computational or memory cost on top of what is already used in building a chain in the stochastic parameter space.

Operationally, as in vHLML, the MCMC is performed using equation (15), where we analytically marginalize over the timing-

model parameters. However, at each step of the Markov chain, we save the following quantities:

(i) $\boldsymbol{\phi}$
(ii) $P(\delta t|\boldsymbol{\phi})$
(iii) $\boldsymbol{\chi} = (\mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\mathbf{M})^{-1}\mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\delta t$
(iv) $\Sigma^{-1} = \mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\mathbf{M}$.

This does not require any additional calculations in the MCMC, and for each MCMC step the amount of data that have to be saved is of the order of m^2, which is not expected to be a bottleneck in terms of storage space in modern workstations. We store these quantities for each step of the Markov chain, which has been run in the l-dimensional parameter space of $\vec{\boldsymbol{\phi}}$, and we thus have enough information to fully characterize the $(l+m)$-dimensional posterior distribution function. Just as in vHLML, the marginalized posterior for the stochastic parameters $\vec{\boldsymbol{\phi}}$ can be calculated as usual from the MCMC.

We assume here that we are interested in calculating the two-dimensional marginalized posterior as a function two timing-model parameters, say ξ_k and ξ_l, with $1 \leq k, l \leq m$, but the generalization to a dimensionality other than two is straightforward. The evaluation of the two-dimensional marginalized posterior consists of numerically integrating over the stochastic parameters $\vec{\boldsymbol{\phi}}$ (summing over the MCMC samples) and analytically integrating over all but two timing-model parameters along the lines of equation (15). Details of this calculation are given in Appendix B; here we give the result

$$
P(\xi_k, \xi_l|\delta t) = \left\langle \frac{\exp\left(\frac{-1}{2}\Delta\boldsymbol{\xi}^\mathrm{T}\mathbf{L}_G \left(\mathbf{L}_G^\mathrm{T}\Sigma\mathbf{L}_G\right)^{-1}\mathbf{L}_G^T\Delta\boldsymbol{\xi}\right)}{\sqrt{(2\pi)^2 \det\left(\mathbf{L}_G^T\Sigma\mathbf{L}_G\right)}} \right\rangle, \quad (27)
$$

where we use $\dots\rangle$ to average over all MCMC samples, the ($m \times 2$) matrix $\mathbf{L}_G = (\hat{e}_k \quad \hat{e}_l)$, with $\hat{e}_i$ being the ith basis vector for $\mathbb{R}^m$, and

$$
\Delta\boldsymbol{\xi} = \begin{pmatrix} \xi_k - \chi_k \\ \xi_l - \chi_l \end{pmatrix}. \quad (28)
$$

Equation (27) allows one to correctly infer the parameters of the timing model, while taking into account the effect of red timing noise.

5 TESTS ON AN ENSEMBLE OF MOCK DATA SETS

We test the procedures we describe in this work with mock TOAs. The TOAs are simulated observations of a pulsar with known timing-model parameters and added noise with known statistical parameters. In previous studies, analysing just one data set with an MCMC was a computational challenge (vHLML; van Haasteren et al. 2011). Extensive statistical studies of the behaviour of the data analysis method have therefore not been carried out in those studies. In this section we show that the Bayesian data analysis method has the desired statistical properties by introducing and applying a method with which a whole ensemble of mock data sets can be analysed simultaneously without much computational overhead.

5.1 MCMC and importance sampling

Doing a full analysis of a single data set is a computationally challenging task because we have to do non-trivial matrix algebra at each step of the MCMC. This makes a straightforward analysis of a whole ensemble of data sets, say $k = 1000$ data sets, computationally prohibitive. We seek to overcome this problem by analysing a

whole ensemble of data sets simultaneously when doing only one MCMC simulation. At each sample of the chain we efficiently evaluate the likelihood for each data set, which in the end can be used to construct the respective marginalized posterior distributions.

A necessary requirement for doing this is that equation (25) can be evaluated for each data set without redoing all the matrix algebra. This is possible if all data sets are different realizations of the same process, which is not a restriction for the purposes of this work. To ensure realistic simulations, we model our mock data after the data for pulsar J1713+0747 as published in van Haasteren et al. (2011). This model for our data sets has irregular sampling, greatly varying error bars for different TOAs and an unknown jump in the middle of the data set. The n TOAs of each data set are generated as perfect realizations of the published timing model, observed at the same MJDs, combined with a TCSS modelled as a random Gaussian process with a red spectral density and a flat high-frequency tail.

We simulate the contributions of the random Gaussian process to the TOAs by, for each data set, appropriately transforming a vector of pseudo-random numbers ζ with entries drawn from a normal distribution with mean 0 and width 1. The simulated TRs are then constructed as $\delta t = \mathbf{J}\zeta$, with $\mathbf{J}$ being the lower diagonal Cholesky decomposition of the covariance matrix $\mathbf{C}$ of equation (4) of the random Gaussian process, defined by $\mathbf{C} = \mathbf{J}\mathbf{J}^{\mathrm{T}}$. We generate all data sets in the ensemble this way, an example of which is shown in Fig. 7. All data sets in the ensemble are generated with the same input parameters.

We evaluate the likelihood function for each data set i and for each MCMC sample j, where i runs from 1 to the number of MCMC samples N, and j runs from 1 to n. The samples in which we evaluate the likelihood values L_{ij} of all data sets come from an MCMC chain that we call the kernel chain. We postpone the details of how we have constructed this kernel chain for the next section, for now we assume that we have found a suitable kernel, where for each sample we have access to the kernel likelihood L_{0j}, and the values of the parameters. For a canonical MCMC simulation, a marginalized posterior distribution $p(\theta)$ can be calculated as

$$
\begin{aligned}
p(x) &= \int_{\theta_i = x} L(\boldsymbol{\theta}) P_0(\boldsymbol{\theta})\, \mathrm{d}^{m-1}\theta \\
&\approx \langle 1 \rangle_{\theta_i = x},
\end{aligned}
\tag{29}
$$

where $p(\theta_i)$ is the marginalized posterior, θ_i is the ith component of the m-dimensional parameter vector $\boldsymbol{\theta}$, $L(\boldsymbol{\theta})$ is the likelihood function, $P_0(\boldsymbol{\theta})$ is the prior distribution and $\langle \cdots \rangle_{\theta_i = x}$ indicates an ensemble average over the MCMC samples over all samples with ith parameter equal to θ_i. This expression assumes that the samples in the MCMC are sampled with a probability proportional to $L(\boldsymbol{\theta}) P_0(\boldsymbol{\theta})$. In our case, however, the MCMC samples are taken with a probability proportional to the kernel likelihood $L_0(\boldsymbol{\theta})$. We adjust equation (29) to suit this new situation:

$$
\begin{aligned}
\int_{\theta_i = x} L(\boldsymbol{\theta}) P_0(\boldsymbol{\theta})\, \mathrm{d}^{m-1}\theta &= \int L_0(\boldsymbol{\theta}) \frac{L(\boldsymbol{\theta}) P_0(\boldsymbol{\theta})}{L_0(\boldsymbol{\theta})}\, \mathrm{d}^{m-1}\theta \\
&\approx \left\langle \frac{L(\boldsymbol{\theta}) P_0(\boldsymbol{\theta})}{L_0(\boldsymbol{\theta})} \right\rangle_{\theta_i = x}.
\end{aligned}
\tag{30}
$$

This approach where one samples not from the true distribution, but from a distribution similar to the true distribution (the kernel distribution) is called importance sampling (Newman & Barkema 1999), where our samples are reused from the MCMC on the kernel.

5.2 Choosing a suitable kernel

With equation (30), we can efficiently produce marginalized posterior distributions for many data sets at a time. Provided there are enough samples in the MCMC, this expression is valid for any kernel likelihood function $L_0(\boldsymbol{\theta})$. However, with importance sampling, the efficiency of the MCMC is highly dependent on the choice of the kernel likelihood function, with this only being practical if the kernel likelihood function is similar to the likelihood functions of the data sets. Our data sets satisfy this condition, because they are all realizations of the same processes. We therefore take the following approach to the construction of a suitable kernel data set, which is then used to form the kernel likelihood function.

(1) We produce a realization of data, which we call the kernel data set, in an identical manner to how we produced the k data sets.

(2) We randomly delete 4/5 of the data points in the kernel data set to ensure that the kernel data set has a broader posterior distribution for all parameters than the mock data sets.

(3) The likelihood function that belongs to the kernel data set is the kernel likelihood.

(4) We make sure that for all the parameters that vary during the MCMC, the true value of each parameter is inside the 1σ region of the kernel likelihood. If not, we discard this chain, and start at (1) again to form a new kernel data set. This step makes sure that our particular realization is not a so-called outlier for our model parameters.

By constructing a kernel likelihood like this, we are ensured that our kernel likelihood distribution covers all the high probability density regions of all the likelihood functions, which allows for faster convergence of equation (30). We note though that convergence is ensured for *any* kernel data set.

5.3 Statistical properties of the ensemble

We use the method outlined above to test $k = 1000$ data sets. The random Gaussian process is a summation of several components:

(1) the error bars of the individual data points of J1713+0747, as described in van Haasteren et al. (2011);

(2) an extra component of noise that is added in quadrature to all error bars. This parameter is the same for all data points and represents the pulse phase jitter (usually referred to as 'EQUAD'). This random pulse jitter is expected to be one of the fundamental limits to pulsar timing precision (Cordes & Shannon 2010); and

(3) a red timing noise TCSS, described by a power-law spectrum of the form $S(f) = N_r^2 (1/1\,\mathrm{yr}^{-1})(f/1\,\mathrm{yr}^{-1})^{-\gamma_r}$, with N_r being the noise amplitude and γ_r the spectral index that describes the 'redness' of the timing noise, with a low-frequency cutoff of $f_L = 0.03\,\mathrm{yr}^{-1}$. As part of our model, we keep f_L fixed during our analysis.

For each data set, we have three parameters that vary during the MCMC, and we have 12 timing-model parameters that we analytically marginalize over during the MCMC. We show the marginalized posterior distribution of the red timing noise parameters here as an example in Fig. 4. Our reweighting scheme of equation (30) has modified the kernel likelihood correctly to match the true posterior distribution. The sample in the MCMC chain that has the highest likelihood value is a good estimator for the maximum likelihood in an MCMC with this few dimensions. In Fig. 5 we present the maximum likelihood estimators for the red timing noise parameters that we obtain in this way. The collection of estimators of the
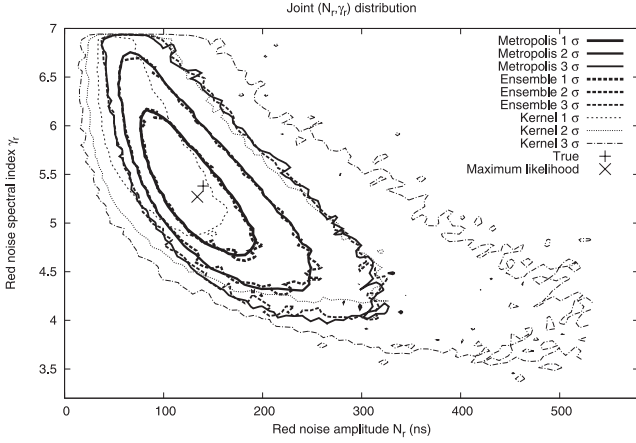
Figure 4. Analysis of the TOAs of Fig. 7 with two methods: a regular MCMC and the importance sampling method of equation (30). The MCMC contours are marked 'Metropolis', and the importance sampling contours are marked 'Ensemble'. Also, the contours of the kernel set that has been used are shown, marked 'Kernel'. In all cases, the 68, 95 and 99.9 per cent contours are shown. The true values for this simulation were $N_r = 145$ ns and $\gamma_r = 5.4$. The maximum likelihood values were $N_r^{\mathrm{ml}} = 134$ ns and $\gamma_r^{\mathrm{ml}} = 5.27$.
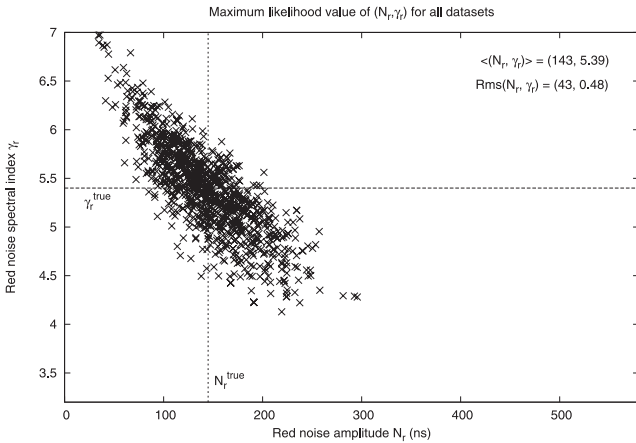
Figure 5. The maximum likelihood values for the parameters N_r and γ_r, for the $k = 1000$ data sets of Section 5.3. The maximum likelihood values are taken to be the values of the parameters of the MCMC sample with the highest likelihood. This collection of estimators seems to display the same characteristics as the marginalized posterior of Fig. 4. As in Fig. 4, the true values for this simulation were $N_r = 145$ ns and $\gamma_r = 5.4$. The mean values were $\langle (N_r, \gamma_r) \rangle = (143, 5.39)$, and the standard deviations were $\mathrm{rms}(N_r, \gamma_r) = (43, 0.48)$.

ensemble seems to follow a distribution with the same shape as the marginalized posterior of Fig. 4, the maximum likelihood value of which is also shown to be close to the centre of the distribution of maximum likelihood estimators.

The ensemble analysis has resulted in $k = 1000$ distribution functions in 15 dimensions. In order to keep our presentation of the results transparent, we restrict our discussion to the $15k$ one-dimensional marginalized posterior distribution functions that follow from this analysis. For each of the $15k$ distributions, we have access to the true value that we gave as an input to our simulations. A basic check would be to verify that for 68 per cent of the data sets, the true value lies within the inner 68 per cent of the marginalized posterior distribution. We generalize this type of basic check to a

more extensive test of both the width and the shape of all the $15k$ distributions.

Provided that our model is correct, the posterior distribution gives the probability that the true value of a parameter has a certain value. Since we have done many trials, we can count how many times the true value θ_i^{true} of the parameter θ_i lies within the most likely x per cent of the posterior distribution. By definition of the posterior, for large number of trials this number approaches x per cent of the total number of trials. More formally, we define the inner high-probability region (HPR) of the one-dimensional marginalized posterior as

$$\int_W p(\theta_i)\, \mathrm{d}\theta_i = a,$$
$$W = \{\theta_i \in \mathbb{R} : P(\theta_i) > L_a\}, \tag{31}$$

where L_a is some value >0 unique for each a, where a is a probability with $0 \le a \le 1$. For each parameter, we define a threshold value $L_t = P(\theta_i^{\mathrm{true}})$. The true value of the parameter lies within the HPR of the marginalized posterior distribution when $L_t > L_a$. By definition of the posterior distribution, the probability that the true value lies within the HPR is given by $\mathrm{Pr}(L_t > L_a) = a$, where we use Pr to denote probabilities. We define the empirical distribution function (EDF) as (Vaart 2000)

$$F_{i,k}(a) = \frac{1}{k} \sum_{j=1}^{k} \Theta(L_t - L_a), \tag{32}$$

where $\Theta(x)$ is the Heaviside function, here used as an indicator. The term in the summation of equation (32) is the indicator for the event $L_t > L_a$. For a fixed L_a, this is a Bernoulli random variable with probability a. Hence, $F_{i,k}(a)$ is a binomial random variable with mean a and variance $a(1 - a)/k$. Therefore, by the law of large numbers,

$$\lim_{k \to \infty} F_{i,k}(a) = a. \tag{33}$$

The Glivenko–Cantelli theorem (Glivenko 1933; Cantelli 1933) states that this convergence happens uniformly over a. In Fig. 6 we present the EDF for all timing-model parameters in our simulation.

We compare our EDF $F_{i,k}(a)$ to the EDF as used in the K-S test in statistics, where one can test for the equality of a sampled distribution function to a reference distribution function. Given a number of samples, the K-S test statistic quantifies how much the distribution of the samples and the reference distribution are alike. Although we have not one single, but many reference distributions, we can define a similar K-S statistic for our EDF:

$$D_{i,k} = \sup_a \left| F_{i,k}(a) - a \right|. \tag{34}$$

This K-S statistic is our quantitative test whether or not the null hypothesis – our data analysis method is consistent – should be rejected. In the canonical application of the K-S statistic, one chooses as threshold the quantity $\sqrt{k} D_{i,k}$, which is expected to follow a Kolmogorov distribution $P_k(K)$:

$$\sqrt{k} D_{i,k} > K_\alpha, \tag{35}$$

where our significance α is determined by $P_k(K \le K_\alpha) = 1 - \alpha$. Two commonly used values are $\alpha = 0.05$ with $K_\alpha = 1.36$ and $\alpha = 0.01$ with $K_\alpha = 1.63$. We choose our significance level as $\alpha = 0.01$, which together with $k = 1000$ implies that we should reject the null hypothesis that our analysis method is consistent when
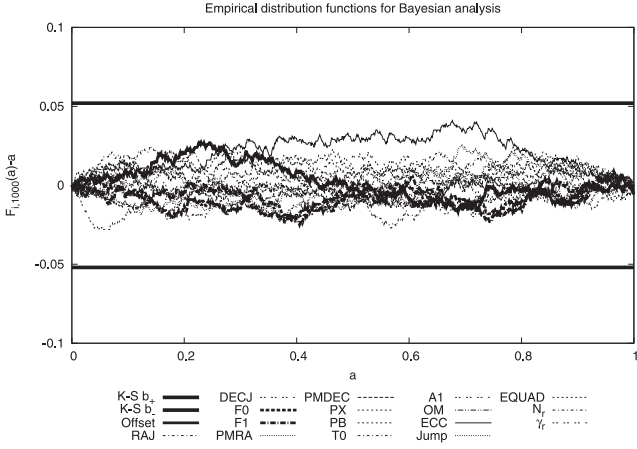
Figure 6. EDF $F_{i,k}(a) - a$ for all parameters i, with $k = 1000$, for the Bayesian analysis. We used mock data of J1713+0747 of Section 5.3, with red noise modelled with a power-law spectral density. The K-S boundaries with a significance level of $\alpha = 0.01$ are displayed as the (b_+, b_-) lines. The TEMPO2 parameter identifiers are as follows. Offset: unknown absolute phase offset; RAJ: right ascension of the pulsar; DECJ: declination of the pulsar; F0: pulse frequency; F1: pulse frequency derivative; PMRAJ: proper motion in right ascension; PMDEC: proper motion in declination; PX: parallax; PB: orbital period; T0: epoch of periastron; A1: projected semi-major axis of the orbit; OM: longitude of periastron; ECC: eccentricity of the orbit; Jump: random phase jump; EQUAD: random pulse phase jitter.

$D_{i,k} > 0.052$. We can see in Fig. 6 that we do not need to reject the null hypothesis.

6 COMPARISON WITH THE CHOLESKY METHOD

Recently, CHCMV have proposed a new method to include TCSSs in the analysis of pulsar timing observations: the Cholesky method. The Cholesky method describes the problem of fitting to the timing model as a whitening problem, where both the data and the description of the timing model need to be whitened with a Cholesky decomposition matrix. This approach is identical to a GLS fit to the timing model given by equation (6). This requires prior knowledge of the covariance matrix **C**, which CHCMV substitute with a best estimator for the power spectral density of the TRs, produced with an advanced spectral analysis method. This spectral analysis method is implemented in the form of a TEMPO2 plugin called SPECTRALMODEL.

The algorithm implemented in SPECTRALMODEL allows the determination of the power spectral density of the post-fit TRs, assuming that the power spectral density has some specific form. For steep red TCSSs as used in this work, the power spectral density is modelled as a power law with a so-called corner frequency f_c:

$$P(f) = \frac{A^2}{1 \, \mathrm{yr}^{-1} \left(1 + (f/f_c)^2\right)^{\alpha/2}}, \qquad (36)$$

where A is the amplitude of the TCSS, f is the frequency and α is the spectral index. The difference with a pure power law as we use in equation (20) is that this power spectral density does not diverge for $f \to 0$. The user needs to provide estimates for α and f_c, and check that these are correct; fitting to the data is only done for the amplitude A.

A direct comparison between the Cholesky method of CHCMV and the inference of timing-model parameters with a Bayesian analysis as done in this work can be made. Both methods take into account the fact that the TOAs may contain a TCSS with a red power

spectrum, of which estimates can be obtained. And both methods obtain improved estimates of the timing-model parameters due to the incorporation of the TCSS contribution to the TOAs. However, several important differences should be highlighted.[3]

First, the modelling of the observations is different. We model the TCSS as a stationary random Gaussian process that is added to the TOAs, prior to the fitting procedure. In the Cholesky method, the TCSS is modelled as a stationary signal in the post-fit TRs. As we have shown in Figs 1–3, this stationarity breaks down in the fitting process. We believe that this raises a question about the spectral estimation method of CHCMV, since the post-fit TRs cannot be described by a stationary TCSS with a mathematically defined spectral density.

Secondly, CHCMV do not fully account for the covariance between the TCSS and the timing-model parameters. The use of an optimal spectral estimate in a parameter estimation technique analogous to equation (6) is not completely appropriate: the covariance matrix of the TCSS is itself covariant with the timing-model parameters, which results in an incorrect covariance matrix for the timing-model parameter estimates and incorrect uncertainties in the spectral estimates. CHCMV show that the incorrectness of the uncertainties is significant for the quadratic spindown parameters, while it is less of a problem for all the other timing-model parameters.

In Fig. 7 we present one realization of mock data of the ensemble of data sets we generated for J1713+0747. Besides the true residuals as generated by the random Gaussian process, we also present three reconstructions of the TRs:

(1) the input TRs to all analysis methods. These were not the true TRs,[4] but the TRs after a weighted least-squares fit was subtracted from the TRs with TEMPO2;

(2) Cholesky TRs. We used the SPECTRALMODEL plugin for TEMPO2 to produce an estimate for the covariance matrix of the post-fit TRs. The Cholesky TRs are constructed using this estimate and equation (6); and

(3) ML TRs. We used the maximum likelihood of equation (15) for the stochastic parameters $\boldsymbol{\phi}$ to produce a best estimator for the covariance matrix **C**. This results in the maximum likelihood estimator TRs through equation (6).

One can see that the maximum likelihood TRs approximate the true TRs slightly better than the Cholesky TRs: the Cholesky TRs deviate slightly more at the sides, with the true residuals not everywhere inside the 1σ bound of the pulse frequency, indicating an error in the low-frequency behaviour. The 1σ bounds of the pulse frequency (and, not shown in the figure, for the pulse frequency derivative) are smaller for the Cholesky method than those for the maximum likelihood TRs. These statements were generally true for all the realizations of the ensemble of mock data sets. Besides due to the issues raised above, this may also be due to the difference in modelling of the power spectral density: the true TRs have been generated with a strict power law and a (fixed) low-frequency cut-off. However, the results here seem to be consistent with table 4 of CHCMV.

---

[3] We emphasize that we only refer to the theoretical description of CHCMV. This does not include the practical implementation of the Cholesky method, SPECTRALMODEL, which in itself has very useful general features such as robust spectral estimation.

[4] We actually worked with TOAs. The residuals plotted in Fig. 7 are produced using different TEMPO2 '.par' files.
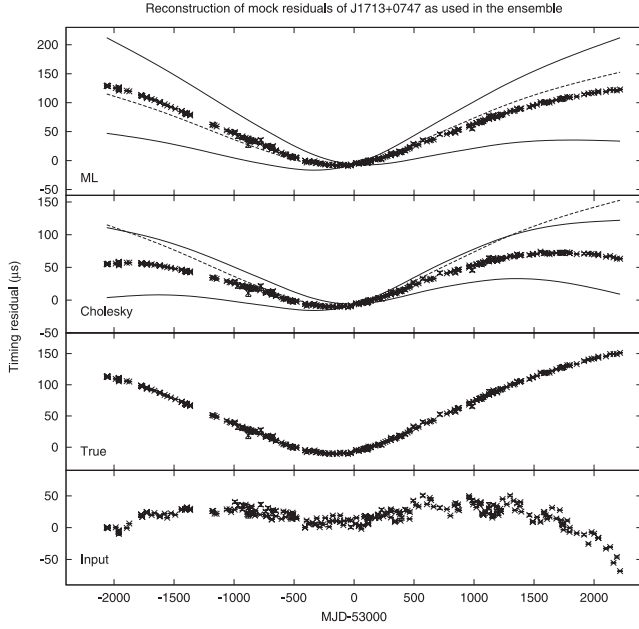
Figure 7. Example of the mock TRs analysed in the ensemble and their reconstruction with various algorithms, all offset from each other for clarity. The mock TRs are based on the observing scheme and timing model of J1713+0747 as used in van Haasteren et al. (2011). The error bars in the figure are mostly too small to see in this figure, and they vary between different observations. The TCSS in these residuals comes from a source with the following spectral density components. (1) The error bars of the individual observations. (2) A white noise component describing the pulse phase jitter (EQUAD), with rms 200 ns. (3) A power-law red noise component $S(f) = N_r^2(1/1 \, \mathrm{yr}^{-1})(f/1 \, \mathrm{yr}^{-1})^{-\gamma_r}$, with amplitude $N_r = 145$ ns, and $\gamma_r = 5.4$. In the figure, four reconstructions of the same realization are shown: True: the true TRs as generated by the TCSS. ML: the TRs as reconstructed using the maximum likelihood values for all parameters: both stochastic parameters and timing-model parameters. Cholesky: the TRs, reconstructed using a covariance matrix produced with the TEMPO2 plugin SPECTRALMODEL, which is an implementation of the Cholesky method of CHCMV. Input: the TRs as produced by TEMPO2 after a normal weighted least-squares fit. This 'Input' set is used as the input TRs for all methods. In the ML and Cholesky reconstructions, we have marked the true TRs as a dashed line, and we have marked the pulse period 1σ boundaries with a solid line. These solid lines demonstrate what the residuals would look like if we changed the pulse period, F0, by $\pm 1\sigma$, and therefore give an impression of how well this parameter is determined from the data.

We would like to perform a K-S test on the results of the Cholesky method to check for consistency. However, this comparison on the ensemble of data sets presented in this work would not be fair because the spectral model used by the SPECTRALMODEL plugin would then be incorrect. Also, we believe that some of the issues with the Cholesky method that we raised above can be overcome. We therefore perform a K-S test on the maximum likelihood equivalent of the Cholesky method: we use equation (25) in conjunction with the maximum likelihood of equation (15) as an estimator for the covariance matrix $\mathbf{C}$. This is equivalent to the Cholesky method when using an 'optimal' estimate for $\mathbf{C}$. Because this estimator does take into account the non-stationary nature of the post-fit residuals, and because the modelling is the same as in the marginalized posteriors, this is effectively a limit on how well any whitening method can perform. Applying this method to the same ensemble of mock data as in Section 5.3 yields Fig. 8. We see that the quadratic spindown parameters are slightly rejected by the K-S test, which shows that at least some of the discrepancy found by CHCMV for the quadratic
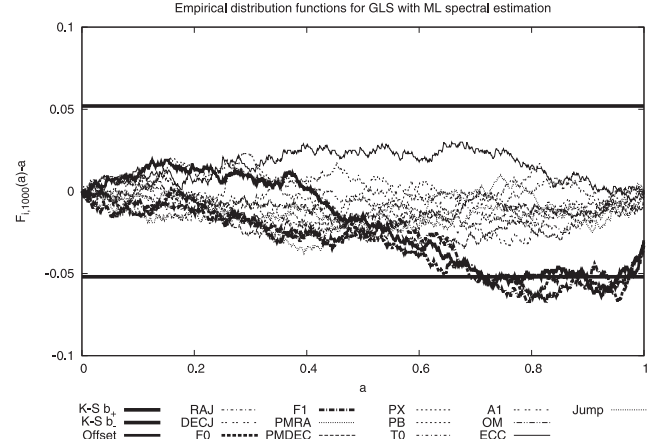
Figure 8. A similar plot as Fig. 6, but now for a method that combines a GLS fit with a maximum likelihood spectral estimator (similar to the Cholesky method). The same mock data sets as in Fig. 6 are used. We see that such a method performs well, with results similar to the Bayesian analysis. Only the quadratic spindown parameters [the average (offset), the pulse period F0 and the period derivative F1] are slightly outside the K-S boundaries. This means, assuming Gaussian probability distributions, that the rms of the parameter estimates was at least 1.11 times the estimated uncertainty. The parameter labels have the same meaning as in Fig. 6.

spindown parameters is due to the covariance of the TCSS with the quadratic spindown parameters. The width of all the marginalized posterior distributions was similar between the approach of Figs 6 and 8, except for the quadratic spindown parameters. For the quadratic spindown parameters, the width of the marginalized posterior distributions was smaller for the maximum likelihood estimates of Fig. 8 than for the full Bayesian method of Fig. 6.

For the quadratic spindown parameters to be rejected by a K-S test of this magnitude means that, assuming Gaussian probability distribution functions, the rms of the parameter estimates was at least a factor of 1.11 times larger than the estimated uncertainty. Table 4 of CHCMV shows that their estimates for the pulse frequency and frequency derivative were over a factor of 3 too large, which corresponds to $\sup_a |F_{i,k}(a) - a| > 0.43$. With the 100 realizations of mock data they used, the K-S bound would be 0.16. This would be a firm rejection, more so than our maximum likelihood estimate. This is at least partially due to the underestimated uncertainties for the quadratic spindown parameters. Whether or not there is also a bias in the estimates of the Cholesky method for these parameters due to incorrect modelling of the covariance function is not clear from the current analysis. We agree with CHCMV that the Cholesky method can be further improved to give more reliable spectral estimates at the very low frequencies and that the Cholesky method performs well for the other timing-model parameters. One possible way to improve the Cholesky method is to use a maximum likelihood estimator for the covariance matrix as we have done here, which models the non-stationarity, and which by design does not suffer from spectral leakage since it does not rely on a periodogram.

7 CONCLUSIONS

We investigate TCSSs in pulsar timing data analysis. TCSSs are significantly influenced by fitting procedures that solve for timing-model parameters, and timing-model parameter estimates can be biased due to absorption of power of the TCSS. We formally analyse the covariance between the timing model and TCSSs, and obtain closed expressions describing the behaviour of the TCSSs when

fitting to the timing model. New results we derive in our analysis are as follows.

(1) Proof that the results of the Bayesian analysis are unaffected by the use of different fitting methods [e.g. (un)weighted least-squares], provided that the timing solution has converged.

(2) Closed expressions for the post-fit correlations of signals with known power spectra.

(3) Analytical closed expressions for the post-fit covariance function of power-law signals with quadratic spindown fitting. This includes proof that the low-frequency cutoff is removed up to spectral indices of $\gamma = 7$, corresponding to $\alpha = 3$ for the GWB.

(4) More computationally efficient expressions for the marginalized posterior distribution of vHLML.

(5) An analytical expression of the post-fit rms induced by a stochastic GWB.

(6) Equations on how to extract the timing-model parameters from Bayesian MCMC simulations.

(7) A new method to analyse hundreds of mock data sets simultaneously with a Bayesian analysis, without significant computational overhead.

(8) A powerful test to check whether any data analysis method produces consistent results, based on the K-S test.

We test our method on many realizations of mock data and find that the shape, width and position of the posterior distributions are consistent with the input values of the parameters. We compare our results to methods that use a spectral estimate to whiten the TRs, like CHCMV, and find that an optimal whitening method performs equally well as our own method, except for the quadratic spindown parameters, in which case the Bayesian analysis produces more consistent results.

REFERENCES

Begelman M. C., Blandford R. D., Rees M. J., 1980, Nat, 287, 307
Blandford R., Romani R. W., Narayan R., 1984, JA&A, 5, 369
Cantelli F. P., 1933, Giorn. Ist. Ital. Attuari., 4, 421
Coles W., Hobbs G., Champion D. J., Manchester R. N., Verbiest J. P. W., 2011, MNRAS, 418, 561 (CHCMV)
Cordes J. M., Shannon R. M., 2010, arXiv:1010.3785C
Demorest P. B., Ferdman R. D., Gonzalez M. E., Nice D., Ransom S., Stairs I. H., Arzoumanian Z., Brazier A., 2012, arXiv:1201.6641D
Edwards R. T., Hobbs G. B., Manchester R. N., 2006, MNRAS, 372, 1549
Glivenko V., 1933, Giorn. Ist. Ital. Attuari., 4, 92
Hobbs G. B., Edwards R. T., Manchester R. N., 2006, MNRAS, 369, 655
Jaffe A., Backer D., 2003, ApJ, 583, 616
Kramer M. et al., 2006, Sci, 314, 97
Lee K. J., Bassa C. G., Janssen G. H., Karuppusamy R., Kramer M., Smits R., Stappers B. W., 2012, MNRAS, 423, 2642
Löhmer O., Lewandowski W., Wolszczan A., Wielebinski R., 2005, ApJ, 621, 388
Lorimer D. R., Kramer M., 2005, Handbook of Pulsar Astronomy. Cambridge Univ. Press, Cambridge
Newman M., Barkema G., 1999, Monte Carlo Methods in Statistical Physics. Oxford Univ. Press, Oxford, p. 31
Phinney E. S., 2001, Astrophysics, astro-ph/0108028
Press W., Teukolsky S., Vetterling W., Flannery B., 1992, Numerical Recipes in C, 2nd edn. Cambridge Univ. Press, Cambridge
Sesana A., Vecchio A., Colacino C. N., 2008, MNRAS, 390, 192
Shannon R. M., Cordes J. M., 2010, ApJ, 725, 1607
Taylor J. H., Weisberg J. M., 1982, ApJ, 253, 908
Vaart A., 2000, Asymptotic Statistics, Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge Univ. Press, Cambridge
van Haasteren R., Levin Y., 2010, MNRAS, 401, 2372
van Haasteren R., Levin Y., McDonald P., Lu T., 2009, MNRAS, 395, 1005 (vHLML)
van Haasteren R. et al., 2011, MNRAS, 414, 3117
Wyithe J., Loeb A., 2003, ApJ, 595, 614

APPENDIX A: POWER-LAW COVARIANCE FUNCTIONS

In this appendix, we analytically derive the post-fit covariance function $\mathbf{W}\mathbf{C}^{\mathrm{PL}}\mathbf{W}^{\mathrm{T}}$ from equation (10) of the main text. We rewrite the relevant expressions for the basis functions and the projection operators here for convenience, with the same notation as in Section 3.3:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_E \approx \frac{1}{\sigma^2 \Delta t} \int_{-T}^{T} x(t)y(t) \, \mathrm{d}t$$

$$\hat{f}_1(t) = \frac{1}{\sqrt{2}} \sigma \sqrt{\frac{\Delta t}{T}}$$

$$\hat{f}_2(t) = \sqrt{\frac{3}{2}} \sigma \sqrt{\frac{\Delta t}{T}} \frac{t}{T}$$

$$\hat{f}_2(t) = \sqrt{\frac{45}{8}} \sigma \sqrt{\frac{\Delta t}{T}} \left[\left(\frac{t}{T}\right)^2 - \frac{1}{3} \right].$$

$$C^{\mathrm{P}}(t_0, t_3) = S(t_0, t_1)C^{\mathrm{PL}}(t_1, t_2)S(t_2, t_3)$$

$$S(t_k, t_l) = \sigma^2 \Delta t \delta(t_k - t_l) - \sum_{i=1}^{3} \hat{f}_i(t_k)\hat{f}_i(t_l), \qquad \text{(A1)}$$

hereafter we always sum over the repeated indices t_1 and t_2. Because the pre-fit covariance function of a power-law spectral density depends only on $\tau = 2\pi|t_0 - t_3|$, we first calculate the quantity

$$Z_\zeta^{\mathrm{P}}(t_0, t_3) = S(t_0, t_1)|t_1 - t_2|^\zeta S(t_2, t_3). \qquad \text{(A2)}$$

We can then construct C^{P} with several Z_ζ^{P} terms.[5] We write the resulting Z_ζ^{P} in the following form:

$$Z_\zeta^{\mathrm{P}}(t_0, t_3) = |t_0 - t_3|^\zeta - \sum_{i=1}^{3} Z_i(t_0, t_3) + \sum_{i,j=1}^{3} Z_{ij}(t_0, t_3). \qquad \text{(A3)}$$

The Z_{ij} terms are symmetric in i and j, and after evaluation of the (somewhat tedious) integrals we find[6]

$$Z_{ij}(t_0, t_3) = \left[\hat{f}_i(t_1)|t_1 - t_2|^\zeta \hat{f}_j(t_2) \right] \hat{f}_i(t_0)\hat{f}_j(t_3)$$

$$U_{11}(t_0, t_3) = \frac{1}{2(1+\zeta)(2+\zeta)}$$

$$U_{12}(t_0, t_3) = 0$$

$$U_{13}(t_0, t_3) = \frac{15\zeta\left(\left(\frac{t_0}{T}\right)^2 + \left(\frac{t_3}{T}\right)^2 - \frac{2}{3}\right)}{4(2+\zeta)(3+\zeta)(4+\zeta)}$$

$$U_{22}(t_0, t_3) = -\frac{9\zeta \frac{t_0 t_3}{T^2}}{2(1+\zeta)(2+\zeta)(4+\zeta)}$$

---

[5] We note that in general $c_{ij} = |t_i - t_j|^\zeta$ is not a PDS matrix, and it therefore does not correspond to a physical stochastic process.

[6] The calculations are available from the authors by request.

$$U_{23}(t_0, t_3) = 0$$

$$U_{33}(t_0, t_3) = \frac{225\zeta(\zeta - 2)\left(\left(\frac{t_0}{T}\right)^2 - \frac{1}{3}\right)\left(\left(\frac{t_3}{T}\right)^2 - \frac{1}{3}\right)}{8(1 + \zeta)(2 + \zeta)(4 + \zeta)(6 + \zeta)}, \qquad (A4)$$

where

$$U_{ij} = \begin{cases} \frac{T^2}{(2T)^{2+\zeta}}\left(Z_{ij} + Z_{ji}\right) & \text{if } i \neq j \\ \frac{T^2}{(2T)^{2+\zeta}}Z_{ij} & \text{if } i = j \end{cases}. \qquad (A5)$$

We find for the Z_i terms

$$Z_i(t_0, t_3) = \hat{f}_i(t_0)\,\hat{f}_i(t_1)\,|t_1 - t_3|^\zeta$$
$$+ |t_0 - t_2|^\zeta\,\hat{f}_i(t_2)\,\hat{f}_i(t_3)$$

$$Z_1(t_0, t_3) = \frac{(T + t_0)^{\zeta+1} + (T - t_0)^{\zeta+1}}{2T(1 + \zeta)}$$
$$+ \frac{(T + t_3)^{\zeta+1} + (T - t_3)^{\zeta+1}}{2T(1 + \zeta)}$$

$$Z_2(t_0, t_3) = \frac{3\left(-\frac{t_3}{T}(T + t_0)^{\zeta+1} + \frac{t_3}{T}(T - t_0)^{\zeta+1}\right)}{2T(1 + \zeta)}$$
$$+ \frac{3\left(-\frac{t_0}{T}(T + t_3)^{\zeta+1} + \frac{t_0}{T}(T - t_3)^{\zeta+1}\right)}{2T(1 + \zeta)}$$
$$+ \frac{3\left(\frac{t_0}{T}(T + t_3)^{\zeta+2} - \frac{t_0}{T}(T - t_3)^{\zeta+2}\right)}{2T^2(1 + \zeta)(2 + \zeta)}$$
$$+ \frac{3\left(\frac{t_3}{T}(T + t_0)^{\zeta+2} - \frac{t_3}{T}(T - t_0)^{\zeta+2}\right)}{2T^2(1 + \zeta)(2 + \zeta)}$$

$$Z_3(t_0, t_3) = \frac{15\left(\left(\frac{t_0}{T}\right)^2 - \frac{1}{3}\right)\left((T + t_3)^{\zeta+1} + (T - t_3)^{\zeta+1}\right)}{4T(1 + \zeta)}$$
$$+ \frac{15\left(\left(\frac{t_3}{T}\right)^2 - \frac{1}{3}\right)\left((T + t_0)^{\zeta+1} + (T - t_0)^{\zeta+1}\right)}{4T(1 + \zeta)}$$
$$- \frac{45\left(\left(\frac{t_0}{T}\right)^2 - \frac{1}{3}\right)\left((T + t_3)^{\zeta+2} + (T - t_3)^{\zeta+2}\right)}{4T^2(1 + \zeta)(2 + \zeta)}$$
$$- \frac{45\left(\left(\frac{t_3}{T}\right)^2 - \frac{1}{3}\right)\left((T + t_0)^{\zeta+2} + (T - t_0)^{\zeta+2}\right)}{4T^2(1 + \zeta)(2 + \zeta)}$$
$$+ \frac{45\left(\left(\frac{t_0}{T}\right)^2 - \frac{1}{3}\right)\left((T + t_3)^{\zeta+3} + (T - t_3)^{\zeta+3}\right)}{4T^3(1 + \zeta)(2 + \zeta)(3 + \zeta)}$$
$$+ \frac{45\left(\left(\frac{t_3}{T}\right)^2 - \frac{1}{3}\right)\left((T + t_0)^{\zeta+3} + (T - t_0)^{\zeta+3}\right)}{4T^3(1 + \zeta)(2 + \zeta)(3 + \zeta)}. \qquad (A6)$$

Then $\mathbf{C}^P$ are obtained by substituting the above expressions into the following:

$$C^P = A^2\left(\frac{1\,\text{yr}^{-1}}{f_L}\right)^{\gamma-1}\left\{\Gamma(1 - \gamma)\sin\left(\frac{\pi\gamma}{2}\right)(f_L 2\pi)^{\gamma-1}Z_{\gamma-1}^P \right.$$
$$\left. - \sum_{n=0}^\infty (-1)^n\frac{(f_L 2\pi)^{2n}}{(2n)!\,(2n + 1 - \gamma)}Z_{2n}^P\right\}. \qquad (A7)$$

Interestingly, $Z_0^P = Z_2^P = Z_4^P = 0$. This means that for $\gamma < 7$, all the f_L-dependent terms in $\mathbf{C}^P$ vanish due to the removal of quadratic spindown.

In the calculations of the rms of signals, we also need the following integral, valid for $\zeta > 0$:

$$\frac{1}{2T}\int_{-T}^T Z_\zeta^P(t, t)\,\mathrm{d}t = \frac{3(4 - \zeta)(\zeta - 2)2^{1+\zeta}T^\zeta}{(1 + \zeta)(2 + \zeta)(4 + \zeta)(6 + \zeta)}. \qquad (A8)$$

This result does not contradict $Z_0^P = 0$, since for $\zeta = 0$ the integral does not exist: for $\gamma \leq 1$ we also need a high-frequency cutoff for the power spectral density.

APPENDIX B: GAUSSIAN PRIORS AND TIMING-MODEL ANALYSIS

In this appendix we show how to include Gaussian priors for the timing-model parameters efficiently. We also present a derivation of equation (27).

B1 Gaussian priors

Besides with flat priors, analytically marginalizing over timing-model parameters is also possible with Gaussian priors for the timing-model parameters. We define Gaussian priors for the m timing-model parameters $\boldsymbol{\xi}$ as

$$P_0(\boldsymbol{\xi}) = \frac{\exp\left(\frac{-1}{2}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_0\right)^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\xi} - \boldsymbol{\xi}_0\right)\right)}{\sqrt{(2\pi)^m \det \boldsymbol{\Sigma}_0}}, \qquad (B1)$$

where $\boldsymbol{\xi}_0$ are the maxima of the prior probabilities and $\boldsymbol{\Sigma}_0$ is the $(m \times m)$ prior covariance matrix of the timing-model parameters. We now proceed with equation (5) multiplied with this prior and rewrite this analogous to what we did in equation (25):

$$P(\boldsymbol{\phi}, \boldsymbol{\xi}|\delta t) = \frac{\exp\left(\frac{-1}{2}\left[\delta t^T \mathbf{C}^{-1}\delta t + \boldsymbol{\chi}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\chi} + \boldsymbol{\xi}_0^T \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\xi}_0\right]\right)}{\sqrt{(2\pi)^{n+m}\det\boldsymbol{\Sigma}_0 \det C}}$$
$$\times \exp\left(\frac{-1}{2}(\boldsymbol{\xi} - \boldsymbol{\chi})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi} - \boldsymbol{\chi})^T\right), \qquad (B2)$$

with

$$\boldsymbol{\chi} = \left(\mathbf{M}^T \mathbf{C}^{-1}\mathbf{M} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{M}^T \mathbf{C}^{-1}\delta t + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\xi}_0\right)$$
$$\boldsymbol{\Sigma}^{-1} = \mathbf{M}^T \mathbf{C}^{-1}\mathbf{M} + \boldsymbol{\Sigma}_0^{-1}. \qquad (B3)$$

Up to a normalization constant due to the inclusion of the prior, these expressions reduce to equation (25) if we take $\Sigma_0^{-1} = 0$ and $\boldsymbol{\xi}_0 = 0$. Marginalizing equation (B2) over the timing-model parameters gives

$$P(\boldsymbol{\phi}|\delta t) = \frac{\sqrt{\det\boldsymbol{\Sigma}}}{\sqrt{(2\pi)^n}\det\boldsymbol{\Sigma}_0 \det\mathbf{C}}$$
$$\times \exp\left(\frac{-1}{2}\left[\delta t^T \mathbf{C}^{-1}\delta t + \boldsymbol{\chi}^T \Sigma^{-1}\boldsymbol{\chi} + \boldsymbol{\xi}_0^T \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\xi}_0\right]\right). \qquad (B4)$$

B2 Posteriors for the timing-model parameters

The MCMC samples are drawn from $P(\boldsymbol{\phi}|\delta t)$ of equation (B4), which is $P(\boldsymbol{\phi}, \boldsymbol{\xi}|\delta t)$ of equation (B2) marginalized over $\boldsymbol{\xi}$. However, we are interested in $P(\boldsymbol{\xi}|\delta t)$, which is $P(\boldsymbol{\phi}, \boldsymbol{\xi}|\delta t)$ marginalized over all stochastic parameters $\boldsymbol{\phi}$. Using an importance sampling approach, we approximate the full posterior distribution with the MCMC samples as

$$P(\boldsymbol{\xi}|\delta t) \approx \left\langle\frac{P(\boldsymbol{\phi}, \boldsymbol{\xi}|\delta t)}{P(\boldsymbol{\phi}|\delta t)}\right\rangle$$
$$= \left\langle\frac{\exp\left(\frac{-1}{2}(\boldsymbol{\xi} - \boldsymbol{\chi})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi} - \boldsymbol{\chi})^T\right)}{\sqrt{(2\pi)^m \det\boldsymbol{\Sigma}}}\right\rangle, \qquad (B5)$$

where we use $\langle \ldots \rangle$ to average over all MCMC samples.

We assume that we would like to obtain the two-dimensional marginalized posterior as a function of the parameters ξ_k and ξ_l,

with $1 \leq k, l \leq m$, but the generalization to a different dimensionality is straightforward. The two-dimensional marginalized posterior is constructed by integrating over all elements of ξ, except for ξ_k and ξ_l. This integration is analogous to what we did with equations (5)–(15). We therefore construct two auxiliary matrices similar to $\mathbf{F}$ and $\mathbf{G}$ of equation (14):

$$\mathbf{L}_F = (\cdots \quad \hat{e_{k-1}} \quad \hat{e_{k+1}} \quad \cdots \quad \hat{e_{l-1}} \quad \hat{e_{l+1}} \quad \cdots)$$
$$\mathbf{L}_G = (\hat{e}_k \quad \hat{e}_l), \tag{B6}$$

where the $\hat{e}_i$ are the basis vectors of $\mathbb{R}^m$. Similar to equation (15), the two-dimensional marginalized posterior now becomes

$$P(\xi_k, \xi_l | \delta t) = \left\langle \frac{\exp\left(\frac{-1}{2} \Delta \boldsymbol{\xi}^{\mathrm{T}} \mathbf{L}_G \left(\mathbf{L}_G^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{L}_G\right)^{-1} \mathbf{L}_G^{\mathrm{T}} \Delta \boldsymbol{\xi}\right)}{\sqrt{(2\pi)^2 \det\left(\mathbf{L}_G^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{L}_G\right)}} \right\rangle, \tag{B7}$$

where

$$\Delta \boldsymbol{\xi} = \begin{pmatrix} \xi_k - \chi_k \\ \xi_l - \chi_l \end{pmatrix}. \tag{B8}$$

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