New advances in the Gaussian-process approach to pulsar-timing data analysis

Rutger van Haasteren[*] and Michele Vallisneri

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA

(Received 10 July 2014; published 11 November 2014)

In this work we review the application of the theory of Gaussian processes to the modeling of noise in pulsar-timing data analysis, and we derive various useful and optimized representations for the likelihood expressions that are needed in Bayesian inference on pulsar-timing-array data sets. The resulting viewpoint and formalism lead us to two improved parameter-sampling schemes inspired by Gibbs sampling. The new schemes have vastly lower chain autocorrelation lengths than the Markov-chain Monte Carlo methods currently used in pulsar-timing data analysis, potentially speeding up Bayesian inference by orders of magnitude. The new schemes can be used for a full-noise-model analysis of the large data sets currently being assembled by pulsar-timing-array collaborations, which generally present a serious computational challenge to existing methods.

DOI: 10.1103/PhysRevD.90.104012 PACS numbers: 04.30.-w, 95.30.Sf, 97.60.Gb

I. INTRODUCTION

The high-precision timing of the radio emission from pulsars has proved to be a valuable tool for probing a wide range of science. Besides great successes such as the first indirect confirmation of the emission of gravitational waves (GWs) [1] and very accurate tests of general relativity [2], pulsar timing is now used in projects that aim to *directly* detect low-frequency GWs (10^{-9}–10^{-8} Hz) from extra-Galactic sources by using a set of Galactic millisecond pulsars (MSPs) as nearly perfect Einstein clocks [3], thanks to the exceptional regularity of their pulses—once many physical effects, such as the pulsar location and motion relative to the Earth, its binary dynamics if it has a companion, the propagation of pulses through the interstellar medium, and the intrinsic evolution of pulsar spin, are modeled accurately (indeed, an accurate *timing model* must account for every rotation of the pulsar across observation epochs). The presence of GWs affects the propagation of the pulses from the pulsar to the Earth, creating detectable deviations from the strict periodicity of the pulse times of arrival (TOAs) [4–6].

In the last decade, scientists seeking to detect GWs with pulsar timing have organized in pulsar-timing-array (PTA) projects around the globe: the European Pulsar Timing Array [7,8], the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) [9,10], and the Australian Parkes Pulsar Timing Array [11,12], which have now joined into a global collaboration, the International Pulsar Timing Array (IPTA) [13,14]. Each PTA has now collected regular observations of tens of MSPs across several years, creating data sets of ever-increasing sensitivity to low-frequency GWs. As a result, a significant amount of effort has already been placed into the development of sophisticated data-analysis methods to extract GWs from pulsar TOAs, both for stochastic GW-background signals (among others [9,15–21]) and continuous waves (for instance, Refs. [22–29]). Many such methods, and especially those based on Bayesian principles, are very computationally intensive and therefore slow. Although work is ongoing on their acceleration, large modern data sets such as those integrated by the IPTA are still very challenging to analyze.

Much of the sophistication required in PTA data analysis is concerned with the description of noise. GWs must be extracted from *timing residuals* (the differences between the observed TOAs and the best timing-model fits), which include measurement errors but also other types of noise, such as "red" spin noise (or "timing noise," the long-term drifts in the rotational frequency of the pulsar), the time- and frequency-dependent delays due to pulse propagation through the interstellar medium, and effects that are correlated across pulsars, such as low-frequency drifts of atomic clocks or inaccuracies in the Solar System ephemerides. For a recent discussion of all of these, see Refs. [30,31]. Each of these noise sources must be distinguished from true GWs. The GWs themselves can have a stochastic character (as for the background from the superposition of signals from many supermassive black-hole binaries), in which case they can be extracted thanks to their correlations among pulsars.

Modern data-analysis methods model the statistics of the noise components of timing residuals as *time-correlated stochastic signals*, described by a power spectral density or a correlation function. This paper focuses on (and reviews) the description of stochastic signals as *Gaussian processes*, the generalization of random variables to functions. This description was implicit in earlier contributions (e.g., Ref. [17]), and we now make it fully explicit. Thus, we give a formal treatment of the Gaussian-process approach to pulsar-timing data analysis, and we derive (or rederive) various expressions, optimized in different ways, for the

[*] vhaasteren@gmail.com

likelihood of the data in the presence of stochastic signals. We also describe and test two novel Bayesian sampling schemes, inspired by Gibbs sampling [32], which outperform the standard Markov-chain Monte Carlo samplers used in pulsar-timing data analysis by greatly reducing the autocorrelation lengths of the chains.

The outline of this paper is as follows. We introduce Gaussian processes in Sec. II and their application to pulsar-timing data in Sec. III. In Sec. IV we discuss the analytical marginalization of likelihoods, and in Sec. V we describe low-rank approximations of covariance matrices. Both techniques are crucial to high-performance analysis methods. In Sec. VI we present our new-and-improved *quasi-Gibbs* schemes, which we test on mock data in Sec. VII. We end with our conclusions in Sec. VIII.

II. GAUSSIAN PROCESSES

Gaussian processes [33] generalize the notion of Gaussian random variables to the case of an infinite number of degrees of freedom. They provide a modern treatment for *process noise*, as defined in optimal filtering—a source of uncertainty distinct from measurement error, which represents unmodeled stochastic or systematic effects in the system under study. More formally [33], a Gaussian process is a (possible infinite) "collection of random variables, any finite number of which have a joint Gaussian distribution." This very property, which corresponds mathematically to the (always surprising) cancellations of chained exponential integrals, makes Gaussian processes especially suited to describing systems that have underlying continuous dynamics yet are necessarily measured at a finite set of *points* (which could be times, locations, or events). Thanks to this property, the likelihood of a measured data set as a function of the Gaussian-process parameters depends only on the behavior of the system at the points for which we have measurements; furthermore, it is especially convenient to interpolate or extrapolate inferences to points for which measurements were not made, or are not available.

A Gaussian process can be specified fully in one of two equivalent ways:

(i) As the sum $\sum_\mu \phi_\mu(x) w_\mu = \phi^T(x) w$ of a finite or infinite set $\{\phi_\mu(x)\}$ of deterministic basis functions, multiplied by the weights w_μ, which are themselves Gaussian random variables with mean vector w^0_μ and covariance matrix $\Upsilon_{\mu\nu}$. (This is the *weight-space* view.)

(ii) As a continuous function $f(x)$, for which we prescribe the *ensemble* mean $m(x) = \mathbb{E}[f(x)]$ and the covariance function $k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$. (This is the *function-space* view.)

In the following we will adopt the simplifying but inessential assumption that $m(x) = w^0_\mu = 0$. The duality between the two views and specifications is encapsulated by the covariance-function expansion

$$k(x, x') = \sum_{\mu,\nu} \phi_\mu(x) \Upsilon_{\mu\nu} \phi_\nu(x'). \tag{1}$$

Indeed, Mercer's theorem [33] ensures that a (possibly infinite) basis-function expansion exists for every positive-definite covariance $k(x, x')$. The power of switching between the dual views is manifest in the two equivalent expressions for the likelihood of a vector y_i of observations of the Gaussian process, taken at the set of points $\{x_i\}$, and subject to Gaussian measurement noise ϵ_i with covariance matrix N_{ij} [33],

$$\begin{aligned} p(y_i|w_\mu, \mathrm{GP}) &= \frac{\mathrm{e}^{-\frac{1}{2}\sum_{i,j}(y_i - \sum_\mu \phi_\mu(x_i) w_\mu)(N_{ij})^{-1}(y_j - \sum_\mu \phi_\mu(x_j) w_\nu)}}{\sqrt{(2\pi)^n \det N}} \\ &\quad \times \frac{e^{-\frac{1}{2}\sum_{\mu\nu} w_\mu (\Upsilon_{\mu\nu})^{-1} w_\nu}}{\sqrt{(2\pi)^m \det \Upsilon}} \\ p(y_i|\mathrm{GP}) &= \frac{\mathrm{e}^{-\frac{1}{2}\sum_{i,j} y_i (N_{ij} + K_{ij})^{-1} y_j}}{\sqrt{(2\pi)^n \det(N + K)}}, \\ \text{with } K_{ij} &= k(x_i, x_j) = \sum_{\mu\nu} \phi_\mu(x_i) \Upsilon_{\mu\nu} \phi_\nu(x_j), \end{aligned} \tag{2}$$

where i, $j = 1, \ldots, n$ and μ, $\nu = 1, \ldots, m$, and K_{ij} is the Gaussian-process *covariance matrix* (i.e., the covariance function evaluated at the measured points). The first expression in Eq. (2) shows the explicit dependence of the likelihood on the basis-function weights; the second, which is obtained by integrating over the w^μ, is in effect the *marginal likelihood* of the data given the Gaussian-process hypothesis, in a compact form that is especially useful if $k(x, x')$ [or equivalently the $\phi_\mu(x)$ and $\Upsilon_{\mu\nu}$] are taken to be functions of a vector of *hyperparameters*, such as the spectral amplitude and slope for power-law noise.

We take a moment to restate this important result: compared to the full likelihood $p(y_i|w_\mu, \mathrm{GP})$, the marginalized likelihood $p(y_i|\mathrm{GP})$ has been *integrated with respect to all possible values of the Gaussian process at the measured points and everywhere else*, subject to the probabilistic constraints given by the noisy measurements. In a Bayesian framework, $p(y_i|\mathrm{GP})$ leads directly to the posterior probability for the hyperparameters. If, conversely, we are interested in the inferred values of the Gaussian process *given the observations*, it can be shown [33] that at any points x' and x'' (whether observed or not) the process is normally distributed with mean $\bar{x}' = \sum_{i,j} k(x', x_i)(N_{ij} + K_{ij})^{-1} y_j$ (and likewise for x'') and covariance $C(x', x'') = k(x', x'') - \sum_{ij} k(x', x_i)(N_{ij} + K_{ij})^{-1} k(x_j, x'')$. This equality has been rederived and used in various forms in pulsar timing: for example, for the analytical marginalization of timing-model parameters [17] and for the reconstruction of dispersion-measure variations [34].

For later reference, we rewrite Eq. (2) using a looser notation where we omit vector indices, replace summations by vector–matrix multiplications, work with log likelihoods, and adopt a special notation for normal-distribution normalization constants,

$$\begin{aligned}\log p(y|w, \mathrm{GP}) &= -\frac{1}{2}(y - \Phi^T(x)w)^T N^{-1}(y - \Phi^T(x)w)\\ &\quad - \frac{1}{2}w^T \Upsilon^{-1} w - \log \mathcal{N}_{n,N} - \log \mathcal{N}_{m,\Upsilon},\\ \log p(y|\mathrm{GP}) &= -\frac{1}{2}y^T(N+K)^{-1}y - \log \mathcal{N}_{n,N+K}, \qquad (3)\end{aligned}$$

where n and m are the sizes of the squares matrices N and Υ and where

$$\Phi_{\mu,i} = \phi_\mu(x_i), \qquad K = \Phi^T \Upsilon \Phi, \quad \text{and}$$
$$\log \mathcal{N}_{p,X} = \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det X. \qquad (4)$$

Gaussian processes have been studied for a long time in statistics; in the last 20 years they have received renewed attention in the fields of *machine learning* and statistical inference [33].

III. GAUSSIAN-PROCESS APPROACH TO PULSAR-TIMING NOISE

In pulsar timing, the properties of the emitting system are inferred from the repeated timing of the its pulses. For millisecond pulsars, a large number of pulses is collected during each *epoch* of observation. Each single TOA is determined by *folding* the pulses with respect to fiducial period and by cross-correlating the folded *profile* to an independently determined *template*; this process produces also an estimated measurement uncertainty (known as *radiometer noise*) for each TOA [35], which can be understood qualitatively as the width of the cross-correlation pattern around its maximum. The TOAs can be predicted deterministically using models that include the astrometric and physical parameters of the source (such as its sky position and proper motion) and the intrinsic evolution of the pulsar spin frequency, as well as binary-orbit parameters for pulsars with a companion. Fitting a deterministic TOA model to a set of observed TOAs results in a *timing solution*. The differences between observed and modeled TOAs are known as *residuals*, and the best-fit model is usually chosen as the one that minimizes the root-mean-square uncertainty-weighted residual.

The gist of the Gaussian-process approach to inferring the noise properties of timing data sets and to searching for GW imprints in the TOAs is this: *the set of best-fit residuals for one or more pulsars is modeled as a sum of Gaussian processes*, which may include:

(1) effects due to the necessarily imperfect determination of the timing solution;
(2) additional observational errors not included in the cross-correlation estimate of TOA uncertainties;
(3) sources of time-correlated or uncorrelated noise intrinsic to the pulsar;
(4) effects due to the propagation of the pulses through the interstellar medium;
(5) *common-mode* effects that are correlated among multiple pulsars, such as those due to the presence of stochastic GWs or reference clock errors.

In this approach, we specify the covariance function or matrix for each Gaussian-process component (except for timing-solution errors, which are easiest to specify using basis functions) as functions of a set of hyperparameters, and we deploy the machinery of Bayesian inference to derive posterior distributions for the hyperparameters (and to characterize or marginalize over the timing-solution errors). The approach was first formulated by van Haasteren and Levin [17,36], without drawing an explicit link to the theory of Gaussian processes, and in effect rederiving basic results such as Eq. (3) as probability manipulations in Bayesian inference.

Mathematically, we write the residuals y as the sum

$$y(\theta) = \sum_{(A)} y^{(A)}(\theta^{(A)}) + \epsilon, \qquad (5)$$

where the vector ϵ denotes measurement errors (which are taken to be Gaussian with covariance matrix N); where the set $\{y^{(A)}(\theta^{(A)})\}$ includes one or more of the Gaussian processes discussed above, with $\theta^{(A)}$ the hyperparameters appropriate for each; and where θ denotes the collection of all $\theta^{(A)}$. The crucial result from Gaussian-process theory, which enables Bayesian inference on the θ, is the fact that the marginal likelihood $p(y|\theta, \mathrm{GP})$ can be written simply as

$$\begin{aligned}\log p(y|\theta, \mathrm{GP}) &= -\frac{1}{2}y^T\left(N + \sum_{(A)} K^{(A)}\right)^{-1} y\\ &\quad - \log \mathcal{N}_{n,N+\sum_{(A)}K^{(A)}}. \qquad (6)\end{aligned}$$

(In fact, this simple description requires two slight complications: first, the timing-solution errors are usually given a special treatment, discussed in Secs. III A and IV; second, the measurement-error matrix N is also parametrized by one or more hyperparameters, as described in Sec. III B below.)

For one choice of hyperparameters, and under the assumption that K is a dense matrix, the task of evaluating a likelihood for a data set of n TOAs using Eq. (6) involves the $O(n^2)$ computation of the total covariance matrix $N + K = N + \sum_{(A)} K^{(A)}$, the $O(n^3)$ computation of its determinant and inverse,[1] and the $O(n^2 + n)$ multiplication of the inverse covariance into the y. The cost of the inverse usually dominates the accounting. Instead of computing $(N + K)^{-1}$ explicitly, one may obtain the upper-triangular decomposition $N + K = U^* U$, which yields the determinant as the product of the squared diagonal elements, and then compute $y^T (N + K)^{-1} y$ as $y^T (U \backslash (U^T \backslash y))$, where we used the MATLAB notation $A \backslash b$ for the solution x of $Ax = b$. The decomposition is again $O(n^3)$, but with a smaller numerical constant, while the linear-system solutions are $O(n^2)$.

Although the individual covariance matrices in the sum $\sum_{(A)} K^{(A)}$ are positive definite by the very definition of covariance, they may have very high condition numbers [37], and thus they may be difficult to invert (or decompose) numerically. Nevertheless, the inversion of $N + K$ is usually *regularized* by the measurement-error matrix N, which is typically diagonal, with elements that are large compared to the $K^{(A)}$. In the course of Bayesian inference, one may yet encounter corners of (hyper)parameter space where $N + K$ becomes numerically singular; it has been our practice to assign a likelihood of 0 to those locations.

We now examine the individual Gaussian-process components of pulsar-timing models and discuss the forms of covariance matrices appropriate for each.

A. Timing-solution errors

The best-fit timing solution (TS) for a set of TOAs is typically derived under the assumption that the template–profile alignment uncertainties due to radiometer noise are the only source of noise.[2] Even in that case, the resulting timing-model parameters would be slightly wrong because the minimum-residual solution always overfits the noise; in reality, the best-fit parameters will be systematically biased by the other unmodeled sources of noise.

If, however, the best-fit solution is sufficiently close to the truth and the various noise components are not too overwhelming, the component of the residuals due to timing-solution errors may be expressed as

$$y^{(\mathrm{TS})} = \sum_a \phi_a^{(\mathrm{TS})}(t) \delta\eta_a \equiv M \delta\eta, \tag{7}$$

where $\delta\eta_a$ is the p-dimensional vector of the parameter errors $\delta\eta_a \equiv \eta_a^{\text{best-fit}} - \eta_a^{\text{true}}$, where the $\phi_a^{(\mathrm{TS})}(t)$ are the partial derivatives of the TOAs with respect to the η_a, evaluated at $\eta^{\text{best-fit}}$, and where M is the *design matrix*[3] $M_{ia} = \phi_a^{(\mathrm{TS})}(t_i)$. The assumption that this linear regime for the $y^{(\mathrm{TS})}$ is actually realized in the course of Bayesian inference can be checked by carrying along the full nonlinear timing model and exploring timing-model parameter space alongside with the Gaussian-process hyperparameters [39,40].

We do not usually deal explicitly with the covariance matrix that ensues from the basis functions $\phi_a^{(\mathrm{TS})}$, because it is awkward to attribute a prior covariance Υ^{TS} to the $\delta\eta_a$, since physically motivated priors are not usually Gaussian in form. Instead, we shall see in Sec. IV how we can marginalize the likelihood with respect to an *improper* prior for $\delta\eta$, which is equivalent to taking the limit $\lambda \to \infty$ for a prior of the form $\Upsilon^{\mathrm{TS}} = \lambda I_p$. Physically motivated (non-Gaussian) priors, such as a non-negative parallax, can be incorporated numerically as discussed in Refs. [39,40].

B. Measurement errors (EFAC, EQUAD, and jitterlike noise)

We know empirically that the cross-correlation estimate of radiometer noise is not always correct; a common fix has been the inclusion in the model of a variable noise multiplier, known as EFAC (for Error Factor). In fact, the physics of the measurement suggests that separate EFACs should be used for every receiver or back end represented in the data set. We know also that there are potential sources of measurement errors that are unrelated to radiometer noise; these have been represented as a white-noise component that adds to radiometer noise in quadrature, with an amplitude parameter known as EQUAD (for Error added in Quadrature).[4] Again, different EQUADs may be assigned to multiple receivers and back ends.

Last, the circumstance that certain data sets (notably those collected by the NANOGrav collaboration [9]) include TOAs measured at the very same time and for the very same set of folded pulses, but in neighboring frequency bands, creates the possibility of noise that is largely or entirely correlated among TOAs measured simultaneously, but entirely uncorrelated among TOAs taken at different times. Some, but perhaps not all, of this noise may be understood as *pulse phase jitter* [30] caused by variable emission within pulsar magnetospheres.

In the case of a single receiver/back end, the total covariance matrix for these three noise components can be written as

$$K^{(\mathrm{MN})} = E^2 n_i \delta_{ij} + Q^2 \delta_{ij} + J^2 \delta_{e(i)e(j)}, \tag{8}$$

where the indices i and j range over the TOAs, where the n_i are the cross-correlation estimates of radiometer noise for each, where the δ are Kronecker deltas, and where $e(i)$

[1] The best known algorithms have slightly lower exponents, but they are not always available in practical computational setups.

[2] It is, however, becoming increasingly common to adopt more sophisticated noise models in timing work, following Refs. [36,38–40].

[3] The design matrix yields the least-squares timing solution as the end point of the iteration $M(\eta^{[i]})\Delta\eta^{[i+1]} = \mathrm{TOA}^{\mathrm{obs}} - \mathrm{TOA}^{\mathrm{model}}(\eta^{[i]})$, $\eta^{[i+1]} = \eta^{[i]} + \Delta\eta^{[i+1]}$.

[4] In the conventions of some timing packages, such as TEMPO2 [41], the EFAC parameter appears also in front of the EQUAD amplitude. We prefer to keep the two separate, since we believe that these hyperparameters should be uncorrelated.

indexes the epochs (i.e., reference measurement times) of each TOA. If the TOAs are sorted by epoch, the matrix $\delta_{e(i)e(j)}$ is block diagonal, with each block consisting entirely of 1's. Such a matrix has low rank corresponding to the number of epochs, which allows useful computational optimizations, discussed below in Sec. V.

It is largely a matter of taste (and sometimes, as we will see below, computational convenience) whether to include all three components in the notional measurement noise ϵ (in which case $N = K^{(\mathrm{MN})}$) or to designate EQUAD noise and jitterlike noise as separate Gaussian processes (in which case $K^{(\mathrm{MN})} = N + K^{(\mathrm{Q})} + K^{(\mathrm{J})}$). For the case of multiple receivers and back ends, separate EFAC, EQUAD, and jitterlike terms for each would appear in Eq. (8), with each set of terms applying to a disjoint subset of TOAs. If the TOAs are sorted by receiver/back end, the total covariance matrix is block diagonal, and each block has the form of Eq. (8) with different E, Q, and J.

C. Correlated pulsar noise

Millisecond pulsars are excellent clocks, but they are not perfect. Slight but measurable irregularities in their rotation (which may be due, for instance, to random angular-momentum exchanges between the normal and superfluid components of the pulsar [30]) create a time-correlated stochastic component in the TOAs that is referred to as timing noise or "red spin noise." This component of timing residuals is typically modeled as a Gaussian, stationary random process, with power-law spectral density,

$$P^{(\mathrm{PL})}(f) = A^2 (f/\mathrm{yr}^{-1})^{-\gamma}\ \mathrm{yr}^3, \tag{9}$$

where f is the frequency, A is a dimensionless amplitude, and γ is the spectral index of the power law (the alternative parametrization $\alpha = 3/2 - \gamma/2$ is also in use). By way of the Wiener–Khinchin theorem,[5] Eq. (9) results in the correlation matrix

$$\begin{aligned} K_{ij}^{(\mathrm{PL})} &= k^{(\mathrm{PL})}(t_i, t_j) \\ &= A^2 (f_L/\mathrm{yr}^{-1})^{1-\gamma} \left\{ \Gamma(1-\gamma) \sin\left(\frac{\pi\gamma}{2}\right) (f_L \tau_{ij})^{\gamma-1} \right. \\ &\quad \left. - \sum_{n=0}^{\infty} \frac{(-1)^n (f_L \tau_{ij})^{2n}}{(2n)!(2n+1-\gamma)} \right\}, \end{aligned} \tag{10}$$

where $\Gamma(\cdot)$ denotes the Euler gamma function, $\tau_{ij} = 2\pi|t_i - t_j|$ is the absolute difference of TOAs, and f_L is a low-frequency cutoff that regularizes the Wiener–Khinchin integral. The series in Eq. (10) sums up to ${}_1F_2(\{1/2-\gamma/2\}, \{1/2, 3/2-\gamma/2\}, -(f_L\tau_{ij})^2/4)/(\gamma-1)$, where ${}_1F_2$ is the generalized hypergeometric function given by HypergeometricPFQ in Mathematica and by hyp1f2 in scipy.special.

Blandford and colleagues [42] and later other authors [17,36,43] showed that the exact value of f_L is irrelevant in pulsar applications, since it is absorbed in the fitting of the linear- and quadratic-spin-downs term of the timing model, at least for γ up to 7 (up to 5 using the linear term alone). For $\gamma = 1$, the total variance $\int P(f)\mathrm{d}f$ becomes infinite even with the low-frequency cutoff. Thus, the spectral index γ is usually taken in the interval $[1, 7]$, although imposing a high-frequency cutoff[6] makes it possible to reach $\gamma = 0$, which corresponds to band-limited white noise.

The evaluation of Eq. (10) is numerically delicate, so special care and tricks are needed.[7] Furthermore, $K_{ij}^{(\mathrm{PL})}$ is a dense, full-rank matrix, so its use in computing residual likelihoods incurs the full $O(n^3)$ cost of matrix inversion. For $\gamma \gtrsim 6$ the matrix $K_{ij}^{(\mathrm{PL})}$ gains a very large condition number [on the order of $(f_L \min\tau)^{-\gamma}$] so the inversion can also be numerically unstable, although it may be regularized by the fact that we invert $N + K$ rather than K, where N is diagonal and has relatively large elements.

Both problems are solved by an alternative approach that models correlated timing noise as a sum over a set of Fourier modes (FM) [19],

$$y^{(\mathrm{FM})}(t) = \sum_{k=1}^{q} a_k \cos(2\pi k x) + b_k \sin(2\pi k x), \tag{11}$$

where $x = (t - t_0)/T$, with t_0 and T the beginning and end of the observation span, respectively. From a Gaussian-process perspective, this amounts simply to specifying the basis functions ϕ_μ instead of the covariance function and solving for the weights w_μ (here we subsume the cosines and sines, and their coefficients, into a single vector of bases of dimension $2q$). This approach offers the additional freedom of specifying the prior weight covariance $\Upsilon_{\mu\nu}^{(\mathrm{FM})}$ as a function of a set of hyperparameters. For instance, a diagonal $\Upsilon_{\mu\nu}^{(\mathrm{FM})}$ specifying a set of variances ρ_μ, each shared by the cos and sin modes of the same frequency f_μ, can be used for a form of spectral estimation [19] (which is not quite "model independent," as it is called in Ref. [19], since a prior for the ρ_μ is still required).

[5] For a stationary process for which $k(x', x'') = C(x' - x'') = C(\Delta x)$, the Wiener–Khinchin theorem relates the power spectral density $P(f)$ to the correlation function $C(\Delta t)$ by way of $C(\Delta t) = \int_0^\infty \cos(2\pi f \Delta t) P(f) \mathrm{d}f$. The total variance of the process is then $C(0)$.

[6] This can be achieved by taking the difference of two expressions of the form (10) with different f_L.

[7] Equation (10) becomes singular for some values of γ, so special-case expressions are required. An alternative, more benign low-frequency regularization is to redefine $P(f) = A^2((f\mathrm{yr})^2 + (f_L\mathrm{yr})^2)^{-\gamma/2}$, which leads to a $C(\tau)$ expressed in terms of modified Bessel functions of the second kind.

The Fourier-sum approach can be seen also as a suboptimal spectral approximation of the time-domain power-law covariance, by way of the fundamental Gaussian-process duality relation:

$$K_{ij}^{(\mathrm{PL})} = \sum_{\mu\nu} \phi_\mu^{(\mathrm{FM})}(t_i) \Upsilon_{\mu\nu}^{(\mathrm{FM})} \phi_\nu^{(\mathrm{FM})}(t_i) \tag{12}$$

with

$$\Upsilon_{\mu\nu}^{(\mathrm{FM})} = P^{(\mathrm{PL})}(f_\mu)\Delta f \delta_{\mu\nu} = P^{(\mathrm{PL})}(f_\mu)\delta_{\mu\nu}/T. \tag{13}$$

The approximation is suboptimal both because we usually sum over a small number of modes (so it is a *low-rank* approximation of a full-rank matrix) and because the modes are not the true eigenfunctions of $K_{ij}^{(\mathrm{PL})}$. However, in practice Eq. (12) can be very accurate (especially if additional, logarithmically spaced modes are added at low frequencies [44]). It can also offer very significant computational savings because the inverse of a matrix expression involving low-rank addends can be computed very efficiently. We discuss this optimization extensively in Sec. V below.

D. Propagation through the interstellar medium

Pulsar radio signals travel across the electromagnetically dispersive interstellar medium, incurring a frequency-dependent, stochastic phase delay known as dispersion-measure (DM) noise [30,45], given by

$$y^{(\mathrm{DM})} = (4.15 \times 10^{-3}\ \mathrm{s}) \left(\frac{\mathrm{DM}}{\mathrm{pc\,cm}^{-3}}\right) \left(\frac{\nu}{\mathrm{GHz}}\right)^{-2} \tag{14}$$

for the delay of a pulse measured at frequency ν with respect to a (hypothetical) pulse at infinite frequency. The time-dependent quantity DM is the column density of free electrons along the (time-changing) line of sight from the pulsar to the radiotelescope. See Lee and colleagues [34] for a discussion of previous work to characterize DM variations and their impact on pulsar-timing GW searches. In the analysis of pulsar-timing data sets that comprise observations at multiple frequencies, DM variations have been modeled with timing-model parameters that describe DM(t) as a piecewise constant [9] or linear [46] function. Alternatively, one can try to solve for DM variations from the multifrequency observations at each epoch, effectively generating a reduced infinite-frequency data set [34,47].

In the context of the Gaussian-process approach, DM noise can be modeled as a correlated Gaussian process, with an additional dependence on the frequency at which each TOA was determined [34,39]. For DM variations characterized by the power-law power spectral density

$$P^{(\mathrm{DM})}(f) = A_{\mathrm{DM}}^2 (f/\mathrm{yr}^{-1})^{-\gamma_{\mathrm{DM}}}\ \mathrm{yr}^3, \tag{15}$$

the timing-residual covariance function is

$$K_{ij}^{(\mathrm{DM})} = k^{(\mathrm{DM})}(t_i, t_j) = (4.15 \times 10^{-3}\ \mathrm{s})^2 \times \left(\frac{\nu_i \nu_j}{\mathrm{GHz}}\right)^{-2} |k^{(\mathrm{PL})}(t_i, t_j)|_{A \to A_{\mathrm{DM}}, \gamma \to \gamma_{\mathrm{DM}}}, \tag{16}$$

where the last term is given by the red-noise power-law covariance Eq. (10) after replacing A and γ with their DM counterparts. For a Kolmogorov DM spectrum resulting from plasma turbulence, $\gamma_{\mathrm{DM}} = 11/3$ [46,48].

The caveats given above for $K_{ij}^{(\mathrm{PL})}$ apply also to the evaluation of $K_{ij}^{(\mathrm{DM})}$. It is also possible to model $y^{(\mathrm{DM})}$ as a sum over basis functions, in analogy to Eq. (11), using either a "spectral-estimation" or power-law prior. If the basis functions are Fourier modes at multiples of the fundamental frequency $1/T$ (with T the duration of the data set), the very low-frequency behavior of the Gaussian process is not modeled well [39]; this can be remedied by enhancing the timing-model design matrix with a term similar to quadratic spin-down, but with ν^{-2} frequency dependence [39], or by adding more modes at low non-Fourier frequencies [44].

E. Gravitational waves and clock errors

Pulsar TOAs carry an imprint of the space-time perturbations (i.e., GWs) that they traverse as they travel from their neutron-star source to the Earth [6]. For an individual source of plane GWs, the frequency-shifting *Doppler response* of the pulsar-to-radiotelescope baseline includes an *Earth term* proportional (times geometric factors) to the GW strain at the event (time and place) of pulse reception and a *pulsar term* proportional to the GW strain at the event of pulse emission [4]. Integrating both terms yields the TOA response, modulo a constant time offset that is degenerate with the initial-phase parameter of the timing model (see, e.g., Ref. [49] for the case of GWs from a black-hole binary). In the Gaussian-process approach to pulsar-timing analysis, such a deterministic signal would not be modeled as a stochastic process, but rather it would be subtracted from the residuals before evaluating their likelihood.

By contrast, a *stochastic* background of GWs can be modeled as a Gaussian process and included in Eq. (5). Various commonly considered backgrounds have a power-law power spectral density,

$$P^{(\mathrm{GW})}(f) = \frac{A_{\mathrm{GW}}^2}{12\pi^2} (f/\mathrm{yr}^{-1})^{-\gamma_{\mathrm{GW}}}\ \mathrm{yr}^3, \tag{17}$$

where the $12\pi^2$ factor follows from defining A_{GW} as the dimensionless characteristic strain h_c at $f = 1/\mathrm{yr}$ [16],

$$h_c(f) = A_{\mathrm{GW}} (f/\mathrm{yr}^{-1})^{\alpha_{\mathrm{GW}}} \quad \text{where } \gamma_{\mathrm{GW}} = 3 - 2\alpha_{\mathrm{GW}}. \tag{18}$$

The spectral index γ_{GW} is 13/3 (but possibly less at low frequencies) for the background from the sum of unresolved black-hole binaries [50–52], 16/3 for a background from cosmic superstrings [53,54], and 5 for a background of inflationary relics [55]. Nonstrictly power-law spectra are also possible, as in the case of the QCD phase transition [56].

Thus, a GW background can be modeled as a stochastic process with a time-domain covariance matrix analog to Eq. (10), or with a Fourier-sum covariance analog to Eq. (12). However, the very concept of pulsar-timing array depends on the fact that the TOA imprints of stochastic GWs are *correlated* among different pulsars. For an isotropic background, the correlation between the GW-induced residuals $y_{ia}^{(\mathrm{GW})}$ (for pulsar a) and $y_{jb}^{(\mathrm{GW})}$ (for pulsar b) is given by

$$K_{iajb}^{(\mathrm{GW})} = \zeta(\gamma_{ab}) k^{(\mathrm{GW})}(t_{ia}, t_{jb}), \tag{19}$$

where $\zeta(\gamma_{ab})$ is the Hellings–Downs coefficient [57,58] for the angle γ_{ab} between the pulsars:

$$\zeta(\gamma_{ab}) = \frac{3}{2}\sin^2\left(\frac{\gamma}{2}\right)\log\sin^2\left(\frac{\gamma}{2}\right) - \frac{1}{4}\sin^2\left(\frac{\gamma}{2}\right) + \frac{1}{2}, \qquad \zeta(0) = 1. \tag{20}$$

The correlations have a more complicated structure if the GW polarizations are not the two quadrupolar modes predicted by general relativity [58–60] or if the background is not isotropic [21,61].

It follows that the full GW-background covariance matrix for a pulsar-timing-array data set can be very large ($N \times N$, where $N = \sum_a n_a$ is the sum of the TOA counts for the individual pulsars); it is also dense, so its inversion is a computationally expensive proposition. In Sec. V A we will see that modeling the GW background as a Fourier sum (and matching Fourier frequencies among pulsars) offers a useful shortcut.

The fact that correlations between a multitude of pulsars are used as a detection mechanism makes pulsar-timing arrays robust detectors for GWs. Noise can generally be expected to be uncorrelated between pulsars, and even without doing proper parameter estimation and noise analysis, a stochastic GW background can still be detected when enough pulsars are observed (Jenet *et al.* [62], Siemens *et al.* [63]). However, GWs are not the only types of signals that can induce correlations. Slow drifts of atomic clocks can introduce a slight error in terrestrial time standards, which would manifest themselves as a common low-frequency signal in the signals of all pulsars (Hobbs *et al.* [64]). Such a correlated signal would be a source of noise when detecting a GW background, and it must be modeled appropriately. The clock signal consists of a time-correlated stochastic signal that is common to all pulsars. As such, we use the same models as for a GW background, except that we take $\zeta = 1$ instead of Eq. (20). Although the covariance matrix component of the clock signal is actually singular, in all realistic scenarios this is always regularized by the other constituents of the covariance matrix. If no regularizing signal is present in the model as could be the case with mock data, it is trivial to replace it with a rank-reduced expansion similar to what we did in the previous sections.

Besides clock errors, another possible source of timing noise are inaccuracies in the Solar-System ephemeris. Although these are unlikely to be a significant source of noise for stochastic-GW searches, they are easily modeled as a correlated stochastic signal, for which one should replace Eq. (20) with $\zeta(\gamma_{ab}) = \cos(\gamma_{ab})$ [65].

IV. MARGINALIZING OVER TIMING-SOLUTION ERRORS

As mentioned above, the timing-solution parameter errors $\delta\eta_a$ are usually given a special treatment: we can include them among the inferred parameters in a Bayesian analysis (i.e., among the parameters that would be sampled explicitly in a Markov-chain Monte Carlo run) and use a likelihood in the form

$$\begin{aligned}\log p(y|\theta^{(\text{non-TS})}, \delta\eta_a) &= -\frac{1}{2}(y - M\delta\eta)^T (N + K^{(\text{non-TS})})^{-1}(y - M\delta\eta) \\ &\quad - \log \mathcal{N}_{n,N+K^{(\text{non-TS})}},\end{aligned} \tag{21}$$

where the $\theta^{(\text{non-TS})}$ denote all the model parameters other than the Timing-Solution errors $\delta\eta_a$, or we can treat them nonlinearly, as in Refs. [39,40], so that the residuals y are recomputed from the full timing model for each value of the η that we sample. This latter approach is desirable if we think that the functional dependence of the residuals on the η may be significantly nonlinear within the relevant parameter ranges.

Otherwise, Eq. (21) can be marginalized analytically over the $\delta\eta_a$ by computing the integral $\int p(y|\theta^{(\text{non-TS})}, \delta\eta_a)\mathrm{d}(\delta\eta_a)$. When doing so, we are in effect assuming an improper (infinitely vague) prior for the $\delta\eta_a$, which is acceptable from a Bayesian perspective as long as the observed data is informative with respect to those parameters. The first authors to propose this marginalization were van Haasteren and Levin [17], who showed that

$$\begin{aligned}p(y|\theta^{(\text{non-TS})}) &= \int \frac{\exp\{-\frac{1}{2}(y - M\delta\eta)^T C^{-1}(y - M\delta\eta)\}}{\sqrt{(2\pi)^n |C|}} d(\delta\eta_a) \\ &= \frac{\exp\{-\frac{1}{2}y^T(C^{-1} - C^{-1}M(M^T C^{-1} M)^{-1} M^T C^{-1})y\}}{\sqrt{(2\pi)^{n-m}|C||M^T C^{-1} M|}} \\ &\equiv \frac{\exp\{-\frac{1}{2}y^T C' y\}}{\sqrt{(2\pi)^{n-m}|C||M^T C^{-1} M|}},\end{aligned} \tag{22}$$

where $C = N + K^{(\text{non-TS})}$.[8]

[8] To perform this integral, we remember the field-theoretical version of Gaussian integrals, $\int \mathrm{e}^{-\frac{1}{2}x^T A x + J^T x} dx = \sqrt{(2\pi)^{(\mathrm{size}\,A)}|A^{-1}|}\mathrm{e}^{\frac{1}{2}J^T A^{-1} J}$, and identify $A = M^T C^{-1} M$ and $J^T = y^T C^{-1} M$.

A derivation of the van Haasteren–Levin result can also be given that remains closer in spirit to the logic of Gaussian processes. For that, we remember that the $\delta\eta_a$ can be seen as the weights of the basis functions $\phi^{(\mathrm{TS})}(t)$ (the columns of the design matrix M). We can then use Eq. (3) with $K^{(\mathrm{TS})} = M\Upsilon^{(\mathrm{TS})}M^T$ and $\Upsilon^{(\mathrm{TS})} = \lambda I_p$ and take the limit $\lambda \to \infty$ corresponding to an infinitely vague prior for the TS weights:

$$\begin{aligned}
\lim_{\lambda\to\infty} \log p(y|\theta^{(\text{non-TS})}) &= \lim_{\lambda\to\infty}\left\{-\frac{1}{2}y^T(C+M\lambda M^T)^{-1}y - \frac{1}{2}\log|C+M\lambda M^T| - \frac{n}{2}\log 2\pi\right\}\\
&= \lim_{\lambda\to\infty}\left\{-\frac{1}{2}y^T C^{-1}y + \frac{1}{2}y^T C^{-1}M(\lambda^{-1}I_p + M^T C^{-1}M)^{-1}M^T C^{-1}y\right.\\
&\quad\left. -\frac{1}{2}\log|C| - \frac{1}{2}\log|\lambda^{-1}I_p + M^T C^{-1}M| - \frac{p}{2}\log\lambda - \frac{n}{2}\log 2\pi\right\}\\
&= -\frac{1}{2}y^T C' y - \frac{1}{2}\log|C| - \frac{1}{2}\log|M^T C^{-1}M| - \frac{n-m}{2}\log 2\pi(+\text{infinite constant}).
\end{aligned}\tag{23}$$

In the second row of Eq. (23) we used the Woodbury formula and the matrix determinant lemma [66],

$$\begin{gathered}
(A+UWV^T)^{-1} = A^{-1} - A^{-1}U(W^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1},\\
\det(A+UWV^T) = \det(W^{-1}+V^TA^{-1}U)\det W \det A;
\end{gathered}\tag{24}$$

we will have occasion to use these formulas repeatedly in the rest of this paper, and we will discuss their computational significance in Sec. V.

Van Haasteren and Levin [36] later derived an alternative form for the $\delta\eta_a$-marginalized likelihood, which exploits the singular-value decomposition (SVD) $M = U\Sigma V^*$ [67]. If M is an $n\times p$ matrix, then U and V are orthogonal matrices of sizes $n\times n$ and $p\times p$, respectively, while Σ is an $n\times p$ diagonal matrix. If we partition U as $[FG]$, where F comprises the first p columns, we see that F spans range(M), while G spans the subspace orthogonal to range(M). Heuristically, we may reason that the projection FF^Ty of the residuals involves components that can be reabsorbed by a change in the $\delta\eta_a$, so these components are in effect *unobserved* from the Gaussian-process perspective; a likelihood can then be written directly for the $(n-p)$-dimensional observable data vector G^Ty (or more precisely, for the coefficients of the y over the partial orthonormal basis given by the G columns):

$$p(y|\theta^{(\text{non-TS})}) = \frac{\exp\{-\frac{1}{2}y^TG(G^TCG)^{-1}G^Ty\}}{\sqrt{(2\pi)^{n-p}|G^TCG|}}. \tag{25}$$

In Appendix A we demonstrate that Eqs. (22) and (25) are indeed equivalent up to a multiplicative constant that does not affect Bayesian calculations.

The computation of the "M-matrix" marginal likelihood [Eq. (22)] is again dominated by the $O(n^3)$ inversion and determinant of the non-Timing-Solution covariance C; it involves also $O(pn^2)$ and $O(p^2n)$ matrix–matrix multiplications, and the $O(p^3)$ inversion and determinant of $M^TC^{-1}M$, as well as negligible quadratic-order matrix–vector multiplications. The computation of the "G-matrix" marginal likelihood [Eq. (25)] is dominated by the $O((n-p)^3)$ inversion and determinant of the projected covariance G^TCG, and by $O((n-p)n^2)$ matrix multiplications (some of these can be avoided by storing the matrices $G^TK^{(A)}(\theta^{(A)})G$ for varying values of $\theta^{(A)}$ and interpolating [36]); it requires also the $O(np^2)$ SVD decomposition of M, which can be performed once and for all when we set up Bayesian inference.

V. LOW-RANK FORMULATIONS FOR CORRELATED NOISE

As we have seen so far, the bottleneck in the evaluation of Gaussian-process marginal likelihoods is the $O(n^3)$ computation of the inverse and determinant of the total covariance matrix $N + \sum_{(A)}K^{(A)}$. It is possible to improve on this situation by exploiting the specific structure of the individual covariance matrices. For instance, the measurement-noise covariance matrix N and some among the $K^{(A)}$ are diagonal, with trivial $O(n)$ inverses. By contrast, other $K^{(A)}$ represent *correlated noise* and therefore *a small number of effective degrees of freedom*; these matrices are usually *severely rank deficient* (at least numerically, which is why they are so hard to invert), and they can be represented accurately by a *truncated eigenvector expansion* USU^T, where U is $n\times l$ with $l \ll n$ [37].

Thus, we are left with the task of computing the inverse of the sum of a diagonal matrix D with a low-rank matrix USU^T. This is where the Woodbury lemma (24) comes to the rescue. Indeed, its principal application in the literature is the *low-rank update of an inverse*, which is just what we need:

$$(D+USU^T)^{-1}=D^{-1}-D^{-1}U^{-1}(S^{-1}+U^TD^{-1}U)^{-1}U^TD^{-1}. \tag{26}$$

We see that the matrix inversions in this reworked expression are those of D [an $O(n)$ operation], S [an $O(l^3)$ operation], and $(S^{-1}+U^TDU)$ [again $O(l^3)$], gaining us an impressive speedup. The corresponding lemma for the determinant is $|D+USU^T|=|D||S||S^{-1}+U^TD^{-1}U|$, which reduces the original $O(n^3)$ computation to $O(n)$ and $O(l^2)$ operations.

The correlated-noise expansion is compatible with the M-matrix formulation of Sec. IV, although the $O(p^3)$ inversion of $M^TC^{-1}M$ is still necessary [in addition, computing $M^TC^{-1}M$ itself is $O(pn^2)$]. An alternative way to include the M-matrix marginalization is to replace U in Eq. (26) with the concatenation $U'=[MU]$, adopting an infinitely vague prior for the timing-model parameters, as we did in Eq. (23).

With a little more work, the correlated-noise expansion is also compatible with the G-matrix formulation, where it leads to

$$\begin{aligned} &y^TG(G^T(D+USU^T)G)^{-1}G^Ty \\ &\quad = y^TWy - y^TWU(S^{-1}+U^TWU)^{-1}U^TWy, \end{aligned} \tag{27}$$

with

$$W=G(G^TDG)^{-1}G^T. \tag{28}$$

Now, the computation of the "weight" matrix W involves an $O(n^3)$ inverse, which can be computed once and for all at the beginning of inference if its only dependence on the θ is a multiplicative constant such as an EFAC. If the dependence of W is more complicated, we can still avoid the $O(n^3)$ scaling by rewriting

$$G(G^TDG)^{-1}G^T = D^{-1}-D^{-1}F(F^TD^{-1}F)^{-1}F^TD^{-1} \tag{29}$$

[see Eq. (A6) in Appendix A], where F is the $n\times p$ orthogonal complement of G (see Sec. IV), and the required matrix inversions are therefore $O(n)$ and $O(p^3)$.

A computationally efficient expression for W is also available when D is the sum $aA+bB$ of two constant components, each multiplied by its own multiplicative hyperparameter (as in the case of single receiver/back end EFAC and EQUAD noise). We can then diagonalize the two simultaneously with a nonorthogonal basis transformation,

$$G^TDG = aG^TAG + bG^TBG = LV(aI+bQ)V^TL^T, \tag{30}$$

with

$$LL^T=G^TAG \quad \text{and} \quad VQV^T=L^{-1}G^TBGL^{T-1}, \tag{31}$$

where I is the identity matrix, L is a lower-diagonal Cholesky decomposition [67], and VQV^T is an eigendecomposition, with Q a diagonal matrix. The quantities required in Eq. (27) are now trivial to calculate:

$$\begin{aligned} &yG(G^TDG)^{-1}G^Ty \\ &\quad = y^TG(V^TL^T)^{-1}(aI+bQ)^{-1}(LV)^{-1}G^Ty, \end{aligned} \tag{32}$$

$$\det(G^TDG)=\det(G^TAG)\det(aI+bQ). \tag{33}$$

Because we need to calculate LVG^Ty (or any other combination like LVG^TU) only once, the computational burden of the inverse is $O(n)$, and evaluating Eq. (27) is $O(nl)$ and $O(l^3)$.

We note that, besides the low-rank expansions we outline in this section, another similar computational trick has been explored in Ref. [68]. Instead of using the Woodbury lemma to expand the low-rank representation of the covariance matrix, the data were compressed to a similar low-rank basis. The low-rank basis was not based on a frequency representation of the signal as we do in the next few sections but was a high-fidelity basis derived from a Fisher-information matrix approximation of the likelihood. Linear interpolation of the compressed covariance matrices was subsequently used to obtain the covariance function for various model parameters. In Appendix B we discuss linear data compression in the context of the more versatile frequency representation of signals, but in the rest of the paper we focus on the uncompressed data.

In the rest of this section we discuss the applications of low-rank expansions: in Sec. V A for correlated timing noise, in Sec. V B for jitterlike noise in multifrequency data sets, and in Sec. V C to define a notion of coarse-grained residuals per epoch.

A. Low-rank expansions by Fourier sums

The Fourier-sum approach discussed in Secs. III C–III E for correlated timing noise, DM variations, and GWs leads directly to a low-rank approximation for the covariance matrix, which is obtained by setting, in the language of Eqs. (12) and (26), $U_{i\mu}=\phi^{(\mathrm{FM})}(t_i)$ and $D_{\mu\nu}=\Upsilon^{(\mathrm{FM})}_{\mu\nu}$. As a reminder, μ ranges from 1 to $2q$ and indexes the Fourier basis functions $\cos(2\pi f_\mu t_i)$ and $\sin(2\pi f_\mu t_i)$ with f_μ a multiple of $1/T$, the inverse duration of the observation; the matrix $\Upsilon^{(\mathrm{FM})}_{\mu\nu}$ is diagonal, with equal elements for each set of two bases of the same frequency. In the case of DM variations, the basis functions would be $(\nu_i/GHz)^{-2}\cos(2\pi f_\mu t_i)$ and $(\nu_i/GHz)^{-2}\sin(2\pi f_\mu t_i)$, following Eq. (16).

If we are modeling correlated noise, DM variations, and GWs all together by way of low-rank expansions, we need to include a separate set of basis functions (and diagonal priors) for each. The resulting global F matrix is obtained by stacking the individual F's horizontally, and the global Υ is the block-diagonal matrix of the individual Υ's. However, since the bases for correlated noise and GWs are the same, except possibly for a different choice of q, the corresponding F matrix needs to be included only once, and the two diagonal prior matrices can be summed. This means that the correlated-noise and GW hyperparameters will be correlated (partially or entirely, depending on the structure of the priors).

In the case of multipulsar analysis, the Fourier-sum modeling of GWs poses a challenge to the derivation of low-rank expressions. Each pulsar gets its own Fourier basis, but each such basis represents a Gaussian process that is correlated with the GW processes of the other pulsar. Let us label the residuals as y_{ai}, where a indexes the pulsar and i ranges over the residuals of each (which can be different numbers). If we use the same set $\{f_\mu\}$ of Fourier frequencies for all pulsars (based, e.g., on the duration of the longest data set), the resulting multipulsar GW covariance is given by

$$\begin{aligned} K^{(\mathrm{GW})}_{aibj} &= \sum_{ab\mu\nu} \Phi_{a\mu}(t_{ai}) \Upsilon^{(\mathrm{GW})}_{a\mu b\nu} \Phi_{b\nu}(t_{bj}) \\ &= \sum_{ab\mu\nu} \Phi_{a\mu}(t_{ai}) (\Upsilon^{(\mathrm{GW})}_{\mu\nu} \zeta_{ab}) \Phi_{b\nu}(t_{bj}) \end{aligned} \tag{34}$$

[see Eq. (19)]. Now, $\Upsilon^{(\mathrm{GW})}_{\mu\nu}$ is diagonal, but ζ_{ab} is dense, so its inverse is potentially expensive. If we order the Fourier coefficients in blocks corresponding to the N pulsars, the matrix $\Upsilon_{a\mu b\nu}$ appears to be made up of $N \times N$ blocks, each of which is a diagonal matrix. By contrast, if we order the coefficients in blocks corresponding to each f_μ, then the matrix $\Upsilon_{a\mu b\nu}$ is block diagonal, with each block a dense matrix given by $\rho_\mu \gamma_{ab}$; thus, its inverse is just $\rho_\mu^{-1} \gamma_{ab}^{-1}$, which incurs an acceptable computational cost $O(qN^3)$.

In principle γ_{ab}^{-1} could be saved and reused; however, if we are also modeling correlated noise with Fourier sums that share the same basis functions as the GWs, the resulting prior would be

$$\Upsilon_{a\mu b\nu} = \Upsilon^{(\mathrm{GW})}_{\mu\nu} \zeta_{ab} + \Upsilon^{(\mathrm{red})}_{a,\mu\nu} \delta_{ab}; \tag{35}$$

in this case each block is given by $\rho^{(\mathrm{GW})}_\mu \gamma_{ab} + \rho^{(\mathrm{red})}_{a,\mu} \delta_{ab}$, and it must be inverted for each choice of $\rho^{(\mathrm{red})}_{a,\mu}$.

B. Low-rank expansions for jitterlike noise

The covariance matrix corresponding to jitterlike noise, as described in Sec. III B, can be expressed *exactly* as the low-rank expression

$$C_{\mathrm{J}} = UEU^T, \tag{36}$$

where E is a diagonal matrix with entries J_e^2 corresponding to squared amplitude of jitterlike noise at each epoch (usually the same for all epochs corresponding to measurements with the same receiver/back end) and where $U_{ie} = 1$ if measurement i belongs to epoch e, 0 otherwise. If the residuals are sorted by epoch, the structure of the expansion is graphically obvious:

$$\begin{pmatrix} 1 & & & \\ 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & 1 \\ & & & \vdots \\ & & & 1 \end{pmatrix} \begin{pmatrix} J_1^2 & & & \\ & J_2^2 & & \\ & & \ddots & \\ & & & J_{n_e}^2 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 & & & & & & & & & \\ & & & & 1 & 1 & \cdots & 1 & & & & & \\ & & & & & & & & \ddots & & & & \\ & & & & & & & & & 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{37}$$

This representation can be used together with the Fourier sums of Sec. V A (for correlated noise, DM variation, and GWs) by stacking the F and U matrices as well as the priors. Otherwise, jitterlike noise can be kept in the matrix D of Eqs. (26) and (27). D is then block diagonal (for sorted residuals) rather than diagonal, but its inverse can be computed very efficiently. Each block D_e has the form of Eq. (8),

$$D_e^{-1} = (N_e + J_e^2 u_e u_e^T)^{-1}, \tag{38}$$

where $(N_e)_{ij} = (E_e^2 n_i + Q_e^2)\delta_{ij}$, with indices ranging over the epoch only, and $u_e^T = (1, 1, \ldots, 1)^T$. By Woodbury's lemma,

$$D_e^{-1} = N_e^{-1} - \frac{N_e^{-1} u_e u_e^T N_e^{-1}}{\alpha_e}, \quad \text{with}$$
$$\alpha_e = J_e^{-2} + u_e^T N_e^{-1} u_e, \tag{39}$$

which is $O(b^2)$, with b the dimension of the block. Altogether D^{-1} can be computed in $O(n\bar{b})$ time, $\bar{b}$ the average number of residuals in an epoch.

C. Low-rank expansion of coarse-grained residuals

A low-rank expansion can also be used to define a statistically principled notion of *coarse-grained residual* per epoch for multifrequency data sets such as NANOGrav's [69]. The idea is to write the total Gaussian-process covariance matrix as $N + C = N + U\tilde{C}U^T$, where the $n \times n$ matrix N includes the measurement noise components (such as EFAC and EQUAD noise) that are independent for each residual, while the $n_e \times n_e$ matrix $\tilde{C}$ (with n_e the number of epochs) describes components such as jitterlike noise, correlated noise, and GWs[9] that depend only on the observation time of each epoch and are therefore entirely correlated among residuals in the same epoch; thus, the "exploder" matrix U has the same structure as in Eq. (37).

A Woodbury expansion yields the likelihood in the form

$$\begin{aligned} &-\frac{1}{2} y^T (N + U\tilde{C}U^T)^{-1} y - \frac{1}{2} \log |N + U\tilde{C}U^T| \\ &\quad = -\frac{1}{2} y^T N^{-1} y - \frac{1}{2} \log |N| + \frac{1}{2} y^T N^{-1} U (\tilde{C}^{-1} + U^T N^{-1} U)^{-1} U^T N^{-1} y - \frac{1}{2} \log |\tilde{C}| |\tilde{C}^{-1} + U^T N^{-1} U| \\ &\quad = -\frac{1}{2} \tilde{\chi}^2 - \frac{1}{2} \log |N| + \frac{1}{2} \tilde{y}^T (\tilde{C}^{-1} + X)^{-1} \tilde{y} - \frac{1}{2} \log |\tilde{C}| |\tilde{C}^{-1} + X|, \end{aligned} \tag{40}$$

where we have neglected logarithms of 2π. In Eq. (40) the n-dimensional vector of residuals y is replaced by the n_e-dimensional vector of coarse-grained residuals $\tilde{y} = U^T N^{-1} y$. Thus, in principle a full multifrequency data set can be condensed into $\tilde{y}$, plus the white-noise $\tilde{\chi}^2$ of the observation and the $n_e \times n_e$ matrix $X = U^T N^{-1} U$ of averaged measurement noise. The marginalization over the timing-model parameters can also by accommodated, in the G-matrix formulation of Eq. (25), by redefining $\tilde{y} = U^T W y$, $\tilde{\chi}^2 = \tilde{y}^T W \tilde{y}$, and $X = U^T W U$, with $W = G(G^T D G)^{-1} G^T$.

Unfortunately, coarse graining per epoch is not useful in practice because the measurement-noise matrix N is usually a function of several hyperparameters (the EFACs and EQUADs for the various receiver/back end combinations), so the full set of residuals must be carried along throughout the analysis to recompute $\tilde{y}$, $\tilde{\chi}^2$, and X as the hyperparameters change. If a single EFAC and EQUAD describe the entire data set, then coarse-grained residuals can be used by way of the two-component expansion of Eq. (30).

VI. QUASI-GIBBS SCHEMES FOR BAYESIAN INFERENCES ON PULSAR-TIMING DATA SETS

Performing Bayesian inference for model parameters and hyperparameters requires the exploration of a high-dimensional parameter space to build a representation of the posterior parameter distributions. Reducing the number of search parameters by marginalizing over some of them analytically, as we discussed in Sec. IV, can be part of the solution, but it is not the entire story. The reason is that stochastic methods such as Markov-chain Monte Carlo (MCMC) are typically used to explore the space of the remaining parameters, so the efficiency of an inference scheme depends crucially on the number of likelihood evaluations required to sample the posteriors broadly and accurately enough as well as the computational cost of an individual likelihood evaluation. The need to choose wisely is especially pointed for large data sets such as the upcoming IPTA data releases, which may contain many tens of thousands of TOAs, requiring (in principle) the inversion of matrices with billions of elements.

MCMC methods explore parameter posteriors by using (in effect) a guided random walk: they generate a sequence of *samples* the distribution of which converges asymptotically to the posterior. The rate of convergence, regardless of the dimension of parameter space, can be characterized as $1/\sqrt{N}$, where N is the number of samples [strictly speaking, it is the fractional error of integrated quantities such as $\int \phi(x) p(x) \mathrm{d}x$ that scales as $\langle \phi \rangle / \sqrt{N}$, with $\langle \phi \rangle$ the variance of the function $\phi(x)$]. The N in this scaling, however, is really the number of *statistically independent* samples, which is related to the length of the chain by a multiplicative constant that depends on the dimension of parameter

[9] DM fluctuations require a slightly more complicated description where U gains n rows, with the same structure as those of Eq. (37) but each multiplied by ν_i^{-2}.

space, on the structure of the posterior, and on the particular scheme used to generate the chain. For an actual chain, the multiplicative constant is characterized well by the *sample autocorrelation function* (ACF), defined as

$$\mathrm{ACF}_t(x) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-t}(x_i - \bar{x})(x_{i+t} - \bar{x})}{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{41}$$

where x and N are the vector and number of samples, $\bar{\cdot}$ indicates the sample mean, and t is the sample lag. The lag at which the ACF drops by a factor e is known as the exponential autocorrelation length [70]; sampling schemes that yield lower autocorrelation lengths for all parameters require correspondingly fewer samples to achieve the same accuracy.

In this section we present two schemes, inspired by Gibbs sampling (see Sec. VI A below), that result in much lower autocorrelation lengths than the state-of-the-art methods currently in use. In Sec. VI B we introduce a scheme optimized for spectral estimation (i.e., for correlated-noise and GW models with free Fourier-sum coefficients); in Sec. VI C we describe a modified scheme that is useful for model spectra (e.g., power-law correlated noise and GWs).

A. Gibbs sampling for pulsar-timing analysis

The simplest MCMC schemes are based on the *Metropolis–Hastings rule*: each new sample in the sequence $\{\theta^{(n)}\}$ is generated by first *proposing* a new parameter vector $\theta^{(n+1)}$ from a proposal distribution $q(\theta^{(n+1)}|\theta^{(n)})$ (which often describes a local perturbation) then *accepting* it with probability given by the Metropolis–Hastings ratio

$$\frac{p(\theta^{(n+1)}|\mathrm{data})}{p(\theta^{(n)}|\mathrm{data})} \times \frac{q(\theta^{(n)}|\theta^{(n+1)})}{q(\theta^{(n+1)}|\theta^{(n)})}. \tag{42}$$

The resulting *detailed balance* (the fact that the flow of samples between two locations in parameter space is proportional to the ratio of the posteriors) guarantees the existence of an equilibrium distribution. If the proposal is such that the chain is *ergodic* (it can reach any corner of parameter space), convergence is assured in the limit of infinite samples. Choosing the proposal distribution smartly (see, e.g., Ref. [71]) is paramount to achieving good chain *mixing* (low autocorrelation lengths).

In a *Gibbs scheme* [32], by contrast, at each step one modifies only a subset of parameters (often just one) and does so by drawing the new value directly from the *conditional* probability distribution of the modified parameters given the unmodified ones. If the blocks of parameters that are modified together are chosen to minimize correlations between blocks, the resulting chain mixing is very good, because all parameters are in effect drawn from the global posterior, except for the effects of residual interblock correlations.

For pulsar-timing analysis, the opportunity of using Gibbs sampling is motivated by a crucial observation on the full unmarginalized likelihood for the case of Fourier-sum correlated noise and GWs (see Sec. V A):

$$p(y|\theta) = \frac{\exp\{-\frac{1}{2}(y - M\delta\eta - Fa)^T N^{-1}(y - M\delta\eta - Fa)\}}{\sqrt{(2\pi)^n|N|}} \times \frac{\exp\{-\frac{1}{2}a\Upsilon^{-1}a\}}{\sqrt{(2\pi)^p|\Upsilon|}}. \tag{43}$$

(Here the a and F are the weights and basis matrix of the Fourier-sum Gaussian processes, and Υ encodes their priors; the $\delta\eta$ and M are the timing-model parameter errors and design matrix, and the infinitely vague prior is implicit.) Equation (43) can be interpreted as a conditional probability for the a and $\delta\eta$ given the hyperparameters that define N and Υ. Furthermore, the conditional probability is Gaussian, which makes it easy to sample from it, as we see below. If the hyperparameters that define Υ are given in the "spectral-estimation" form $\Upsilon_{\mu\nu} = \rho_\mu \delta_{\mu\nu}$ [see Eq. (12)], then the ρ_μ can also be drawn easily from their conditional posteriors, given the weights and the other hyperparameters. Last, Eq. (43) requires the inversion of diagonal matrices only [an $O(n)$ operation], so it can be evaluated very efficiently.

The reason why the full Fourier-sum likelihood has not been used so far in pulsar-timing analysis is that the resulting increase in computational efficiency is outweighed by the increased autocorrelation lengths in MCMC schemes that evolve the hyperparameters together with the weights in perturbative fashion. A quasi-Gibbs, blocked sampling scheme overcomes this problem (the scheme is not quite Gibbs because we still need Metropolis–Hastings updates for the hyperparameters that appear nontrivially in the likelihood). We describe it in the next section.

B. Quasi-Gibbs, blocked sampling scheme for spectral estimation

In this sampling scheme we successively modify the values of blocks of parameters, holding all the others fixed (hence, the scheme is *blocked*). The choice of blocks aims at two goals: the covariance between parameters in separate groups should be minimized to improve the ACF, and it should be possible to sample directly from the conditional probabilities for each block, or at least to evaluate them cheaply. Thus, we choose the following groups:

(1) *Quadratic parameters*, consisting of the timing-model parameter errors $\delta\eta$ and the Fourier coefficients a for both correlated noise, GWs, and DM

variations. (For a single pulsar, GWs would be degenerate with correlated noise.)

(2) Hyperparameters describing white noise, and optionally jitterlike noise, using Eqs. (38) and (39); jitterlike noise could also be modeled with quadratic parameters, per Eq. (36).

(3) Hyperparameters describing priors for correlated-noise and GW Fourier coefficients.

(4) Hyperparameters describing priors for DM-variation Fourier coefficients.

We cycle through these four steps, resampling the parameters in each block while holding the others fixed to their most recent value; at the end of each cycle we obtain a full Markov-chain sample. We now discuss each step in detail.

(1) Sampling the quadratic parameters: As mentioned before, the quadratic parameters are the weights of the basis functions $\phi_\mu^{(\mathrm{TS})}(t_i) = M_{i\mu}$ and $\phi_\mu^{(\mathrm{FM})}(t_i) = F_{i\mu}$ of the timing model and the correlated noise, respectively. We denote them collectively as $w^T = (\delta\eta^T, a^T)$ and with $\Phi = (M; F)$. Fixing all the hyperparameters θ, the log-posterior probability of the w can be rewritten as

$$\begin{aligned}\log P(w|y,\theta,\mathrm{GP})\\ &= -\frac{1}{2}(w - Q^{-1}\Phi N^{-1}y)^T Q(w - Q^{-1}\Phi N^{-1}y)\\ &\quad -\frac{1}{2}\log\det Q + \mathrm{const},\end{aligned}\tag{44}$$

with

$$Q = \Phi N^{-1}\Phi^T + \Upsilon^{-1},\tag{45}$$

where N, Υ, and the additive constant are functions of the hyperparameters and of the residuals and where we interpret Υ^{-1} in the broad sense explained in Sec. IV: by assuming an infinitely vague prior for the timing-model parameters, we set Υ^{-1} to zero in their subspace. Equation (44) states that the quadratic parameters w are distributed according to a multivariate normal distribution with mean $\bar{w} = Q^{-1}\Phi N^{-1}y$ and covariance Q^{-1}. We can draw from this distribution by computing $w_{\mathrm{new}} = \bar{w} + L\epsilon$, with ϵ a vector of zero-mean, unit-norm, uncorrelated normal deviates (see, e.g., Ref. [72]) and L a square root of Q^{-1} (i.e., $LL^T = Q^{-1}$). For numerical stability, we first evaluate Q^{-1} with a QR decomposition (the product of an orthogonal matrix and an upper-triangular matrix [67]), then use an SVD decomposition [67] to compute the square root.

In multipulsar data sets, the effects of GWs on the timing residuals of different pulsars are correlated [see Eq. (34)]; thus, so are the posterior distributions of the GW Fourier coefficients for each pulsar [by way of Eq. (35)]. If we were to use the procedure that we have just outlined to draw new GW quadratic parameters, we would have to do so for all the pulsars at once, which can be very computationally expensive. Instead, the step can be performed separately for each pulsar a by conditioning the corresponding w_a on the most recent w_b for all $b \neq a$. Expanding Eq. (43) for a prior matrix Υ that includes cross terms between pulsars and collecting all the terms that involve the w_a results in the conditional probability

$$\begin{aligned}\log P(w_a|w_{b\neq a}, y, \theta, \mathrm{GP})\\ &= -\frac{1}{2}(w_a - Q_a^{-1}z_a)^T Q_a(w_a - Q_a^{-1}z_a)\\ &\quad -\frac{1}{2}\log\det Q_a + \mathrm{const}(\theta, w_{b\neq a}),\end{aligned}\tag{46}$$

where

$$\begin{aligned}z_a &= \Phi_a N_a^{-1}y_a + \sum_{b\neq a}(\Upsilon^{-1})_{ab}w_b,\\ Q_a &= \Phi_a N_a^{-1}\Phi_a^T + (\Upsilon^{-1})_{aa},\end{aligned}\tag{47}$$

from which w_a can be drawn directly with the covariance-square-root procedure. Because of the structure of the multipulsar prior [Eq. (35)], computing the submatrices $(\Upsilon^{-1})_{aa}$ and $(\Upsilon^{-1})_{ab}$ does not require the inversion of the full Υ, but only of each pulsar block, which is much cheaper.

(2) Sampling the white-noise hyperparameters: The conditional probability for the white-noise hyperparameters θ_w that determine N in Eq. (43) is very simple,

$$\begin{aligned}\log P(\theta_w|y_{\mathrm{red}}, w, \theta_p, \mathrm{GP})\\ &= -\frac{1}{2}y_{\mathrm{red}}^T N^{-1}y_{\mathrm{red}} - \frac{1}{2}\log\det N - \log p(\theta_w),\end{aligned}\tag{48}$$

where the reduced residuals $y_{\mathrm{red}} = y - \Phi w$ are obtained by subtracting the most recent realization of the quadratic-parameter Gaussian processes from the residuals and where $p(\theta_w)$ is the prior for the θ_w. We cannot draw directly from this distribution, but we can approximate such a draw by performing a sequence of perturbative Metropolis–Hastings updates (in effect, a small MCMC run) for the θ_w. Because of the form of Eq. (48), this is not costly. (In our tests, we performed the small MCMC run with an adaptive Metropolis sampler, allowing for significant burn in on the first iteration, and then using the adaptively tuned proposal covariance in subsequent iterations. Each small MCMC is run for longer than a full autocorrelation length, as estimated in the first iteration.)

(3) and (4) Sampling the Fourier-sum hyperparameters: The following description applies to the Fourier-sum hyperparameters for correlated timing noise and GWs, and for DM variations. We denote either set as θ_p. The conditional probability for the θ_p, fixing everything else, is given by

$$\log P(\theta_p|y,\theta_w,w,\mathrm{GP}) = -\frac{1}{2}w^T\Upsilon^{-1}w - \frac{1}{2}\log\det\Upsilon - \log p(\theta_p), \quad (49)$$

where, again, Υ^{-1} is identically zero in the subspace of the timing-model parameters, which do not appear in this equation. (This is not an inherent restriction of our scheme, but it is our choice.)

The spectral-estimation model discussed in Sec. III C includes an independent variance parameter ρ_μ on the diagonal of Υ for each modeled frequency. Each ρ_μ applies to a cosine and a sine mode; we will denote their weights as a_μ and b_μ. If we adopt $1/\rho_\mu$ Jeffreys priors for each ρ_μ [73], Eq. (49) becomes fully separable, and we can write

$$P(\rho_\mu|a_\mu,b_\mu,\theta_w,\mathrm{GP}) = \frac{(a_\mu^2+b_\mu^2)\exp\left(-\frac{1}{2}\frac{a_\mu^2+b_\mu^2}{\rho_\mu}\right)}{\rho_\mu^2}. \quad (50)$$

We can draw samples from this distribution analytically, even if we adopt a proper Jeffreys prior with compact support $\rho_{\mu,\mathrm{min}} < \rho_\mu < \rho_{\mu,\mathrm{max}}$. To do so, we pick η uniformly in the interval $[0, 1-\exp(\tau/\rho_{\mu,\mathrm{max}} - \tau/\rho_{\mu,\mathrm{min}})]$, with $\tau = (a_\mu^2+b_\mu^2)/2$, and we compute

$$\rho_{\mu,\mathrm{new}} = \frac{\tau}{\tau/\rho_{\mu,\mathrm{max}} - \log(1-\eta)}. \quad (51)$$

With a more general prior $p(\theta_p)$, we can still use the small-MCMC strategy discussed above for θ_w.

This scheme is analog to augmented/missing-data methods used in machine learning [70]: if we think of the Fourier coefficients as *unobserved data* rather than model parameters, then at the beginning of each cycle we are in effect *imputing* their values (according to their conditional probability with the current hyperparameters) to "complete" the data set and evaluate model-parameter likelihoods with greater convenience.

In actual use, this scheme turns out to be very efficient, with extremely low autocorrelation lengths (see Sec. VII A). This is because nearly all the parameters in the different blocks turn out to be nearly uncorrelated; the only significant correlations are between the quadratic-spin-down timing-model parameter and the lowest Fourier coefficients, which do not increase the overall autocorrelation length significantly. In addition, the Fourier coefficients are also effectively uncorrelated among themselves, because the corresponding modes are approximately orthogonal (they would be exactly orthogonal if the TOAs were sampled regularly). This does not matter to their update step, since we are drawing from the joint posterior; however, this non-correlation helps chain mixing, because it means that each pair of (a_μ, b_μ) interacts (and correlates) with a single ρ_μ that is updated in a different block.

However, if we apply the quasi-Gibbs scheme to a model of correlated noise where the Fourier and timing-model coefficients are correlated more strongly through the hyperparameters (as in a model with power-law spectral densities), the autocorrelation lengths increase sharply. To illustrate this problem, in Fig. 1 we show the correlation profile of the correlated-noise power-law parameters (amplitude and spectral slope), as estimated in a standard marginalized-poster MCMC, together with the much smaller conditional-correlation profiles (white curves) that is "seen" in one of the hyperparameter block updates of the quasi-Gibbs scheme, where all the Fourier coefficients are fixed to specific values. The limited extension of the effective correlation profiles greatly increases the autocorrelation times of the hyperparameters in the quasi-Gibbs chain.

C. Collapsed quasi-Gibbs sampling scheme for modeled spectra

To improve this behavior, we need to sample the Fourier coefficients and their hyperparameters simultaneously. This is what we do in the modified scheme described here, which trades some computational efficiency for shorter autocorrelation lengths. This scheme has the same four

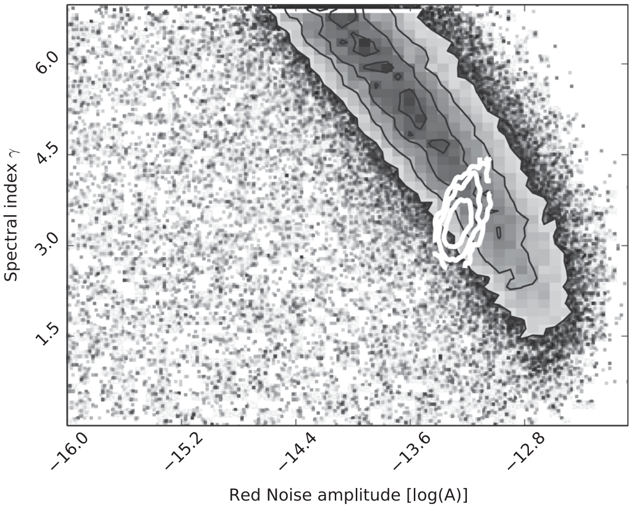

FIG. 1. Comparison of the correlation profiles for the correlated-noise amplitude and spectral slope parameters in a full MCMC run (larger density profile) and as seen in a correlated-noise hyperparameter block update, where the Fourier coefficients are fixed to specific "imputed" values. The runs were performed on NANOGrav's 5-year J1910 + 1256 data set, which includes DM corrections as part of the timing model [9].

steps as the quasi-Gibbs scheme of the last section, but we modify the step 3/4 where we update the hyperparameters of the modeled spectra. For these we adopt the following procedure: a) we first draw new hyperparameters θ_p from a perturbative proposal; b) we then generate new quadratic parameters w_p directly from their conditional posterior given the new θ_p; finally c) we accept the new (θ_p, w_p) according to the Metropolis–Hastings rule.

The Metropolis–Hastings ratio for the entire step is then

$$\frac{p(\theta_p^{(n+1)}, w_p^{(n+1)}|y, \ldots)}{p(\theta_p^{(n)}, w_p^{(n)}|y, \ldots)} \times \frac{q(w_p^{(n)}|w_p^{(n+1)})q(\theta_p^{(n)}|\theta_p^{(n+1)})}{q(w_p^{(n+1)}|w_p^{(n)})q(\theta_p^{(n+1)}|\theta_p^{(n)})}, \tag{52}$$

where we do not indicate the dependence of the probabilities on all the hyperparameters and coefficients that are not updated in this step. However, since the proposal for w_p is just its conditional given the new hyperparameters,

$$q(w_p^{(n+1)}|w_p^{(n)}) = p(w_p^{(n+1)}|\theta_p^{(n+1)}, y, \ldots), \tag{53}$$

and since the overall posterior probability can be factorized as

$$p(\theta_p, w_p|y, \ldots) = p(\theta_p|y, \ldots)p(w_p|\theta_p, y, \ldots), \tag{54}$$

where $p(\theta_p|y, \ldots)$ is marginalized over the w_p, the Metropolis–Hastings ratio simplifies (*collapses*) to

$$\frac{p(\theta_p^{(n+1)}|y, \ldots)}{p(\theta_p^{(n)}|y, \ldots)} \times \frac{q(\theta_p^{(n)}|\theta_p^{(n+1)})}{q(\theta_p^{(n+1)}|\theta_p^{(n)})}, \tag{55}$$

with

$$\begin{aligned} \log p(\theta_p|y, \ldots) \\ &= -\frac{1}{2} y^T (N^{-1} - N^{-1}\Phi(\Phi^T N^{-1}\Phi + \Upsilon^{-1})^{-1}\Phi^T N^{-1})y \\ &\quad -\frac{1}{2}\log|\Phi^T N^{-1}\Phi + \Upsilon^{-1}| + \text{const.} \end{aligned} \tag{56}$$

Thus, we are just taking a Metropolis–Hastings step over the θ_p using the marginalized posterior, and we can wait to draw new $w_p^{(n+1)}$ from the conditional probability given the $\theta_p^{(n+1)}$ only if the step is accepted (conveniently, we already have the appropriate Q^{-1} covariance to do so).

In addition to reworking steps 3/4, we need also to adjust the parameter blocks, by including among the quadratic parameters that are updated also the timing-model parameters that are significantly covariant with them (i.e., the quadratic spin-down in the correlated-noise block and the DM parameter[10] in the DM-variation block). From a computational-cost standpoint, this scheme is comparable to an MCMC based on the fully marginalized posterior. However, the resulting autocorrelation lengths are much improved by the blocked updates of uncorrelated parameter subsets (see Sec. VII B).

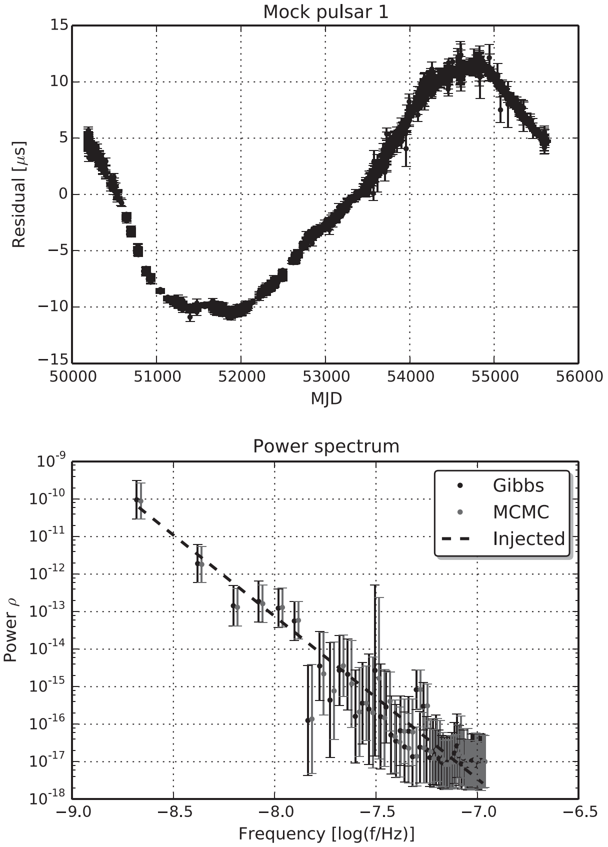

FIG. 2. Mock data and spectral estimation in the test of the quasi-Gibbs scheme. Top: Mock residuals for pulsar J0437-4715 used in the test of Sec. VII A. We generated 1,500 TOAs over a time span of 6,000 days, injecting power-law timing noise [Eq. (9)] with parameters $A = 3 \times 10^{-14}$, and $\gamma = 4.33$. Bottom: Recovered Fourier-mode variances for the Metropolis and quasi-Gibbs samplers. The error bars show 1-σ standard deviations, and the spectrum of injected noise is shown as the dashed line.

VII. TESTS OF THE QUASI-GIBBS SCHEMES USING MOCK DATA

In this section we test the performance (and basic correctness) of our quasi-Gibbs sampling schemes using simulated timing residuals, which we obtain using the LIBSTEMPO interface [74] to the TEMPO2 timing package [41]. In Sec. VII A we compare the spectral-estimation quasi-Gibbs method of Sec. VI B with a standard MCMC method, applying both to a single-pulsar data set. In Sec. VII B we compare the more general quasi-Gibbs method of Sec. VI C with again a standard MCMC method, apply both to a multipulsar data set that contains a GW background.

[10] And when not accurately modeling the lowest DM variation frequencies, also the first and second time derivatives of the DM.

A. Test of spectral-estimation quasi-Gibbs scheme

The single-pulsar mock data set for this test is based on the timing model of pulsar J0437-4715 in the pulsar catalog of the Australia Telescope National Facility [75]. This is one of the IPTA pulsars with the lowest TOA uncertainty, and it has been observed regularly. We generate timing residuals with real-world characteristics: the TOAs are sampled unevenly (in the modified Julian date interval 50,000–56,000), they reflect strong timing noise, and their TOA uncertainties are varying. As typical for actual data collected with ever-evolving observation systems, we partition the data set in 15 blocks, all corresponding to different hardware, each with a different EFAC and EQUAD. The residuals for this mock data set are shown in Fig. 2.

In our test we determine the power spectral density of injected noise, in the style of Eq. (11) and Ref. [19], using two sampling schemes: a "vanilla" adaptive Metropolis MCMC sampling method (as described in the appendix of Ref. [69]) and the quasi-Gibbs method of Sec. VI B. With both methods we adopt the same noise model: white noise with $15 + 15$ EFAC and EQUAD hyperparameters, plus correlated noise described by 50 Fourier modes at frequency multiples of $1/T$, with 50 independent variance parameters describing the spectral density. For the adaptive Metropolis sampler, the posterior is marginalized analytically over all quadratic parameters, so the total dimension of parameter space is 80. For the quasi-Gibbs sampler, the unmarginalized posterior is a function of 214 parameters: 30 white-noise hyperparameters, 50 spectral-density prior variances, 100 Fourier coefficients, and 34 timing-model parameters, of which 12 model the unknown phase offsets between different observing systems.

The Metropolis sampler was run for 4 million steps and the quasi-Gibbs scheme for 30,000. The resulting estimates of power spectral density, shown in Fig. 2 (bottom panel), agree very well. The autocorrelation functions, shown in Fig. 3, differ greatly, with much shorter autocorrelation lengths in the quasi-Gibbs scheme—that is why we needed only 30,000 steps for it. The observant reader will note that the autocorrelation length of the Fourier-mode variances is 1 in the quasi-Gibbs scheme. This is the lowest possible, indicating that our samples are virtually independent draws from the posterior; no sampler can do better. In addition, we note that since the Fourier-mode variances are inherently uncorrelated their autocorrelation length does not depend on the number of frequencies included in the model, in sharp contrast to Metropolis samplers.

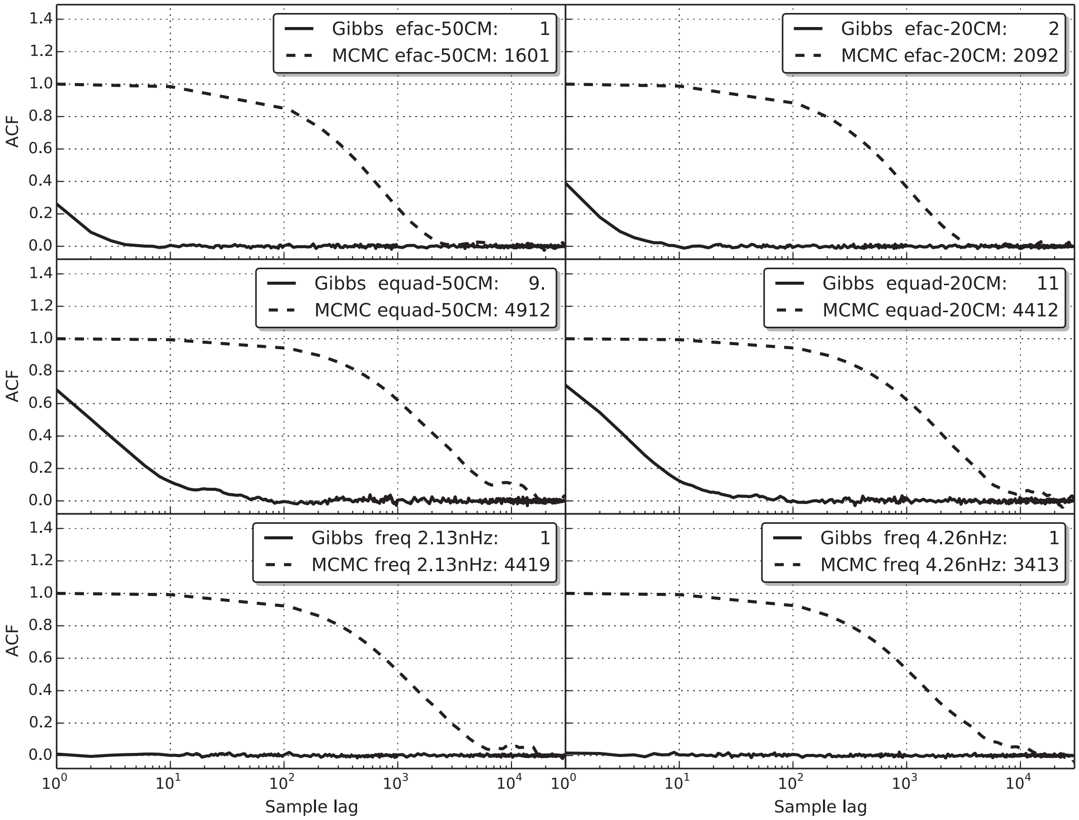

FIG. 3. Autocorrelation as a function of sample lag for various model parameters (EFAC and EQUAD for 50- and 20-cm receivers, and Fourier-mode variances at 2.13 and 4.26 nHz), as measured in the adaptive-MCMC chain (dashed) and the quasi-Gibbs chain (solid), both run on the mock J0437-4715 data set. The legends show the autocorrelation lengths, which were typically 400–1,000 times shorter with the quasi-Gibbs sampler.

B. Test of collapsed quasi-Gibbs scheme

For this test we use mock data for an entire PTA: specifically, the second "open" data set in the IPTA Mock Data Challenge [76], which consists of white radiometer (EFAC) noise in 36 pulsars (with different EFACs) plus a coherently injected GW background with $h_c(1\ \mathrm{yr}^{-1}) = 5 \times 10^{-14}$ and $\gamma = 4.33$. For this set we need the collapsed Gibbs sampler both because the GW background is parametrized as a frequency-domain power law and because the data set includes multiple pulsars.

In our test we determine the EFACs and the level and shape of the GW background using two sampling schemes: again the vanilla adaptive Metropolis MCMC of Ref. [69] and the collapsed quasi-Gibbs sampler of Sec. VI C. We assume that the GW-background covariance matrix is characterized well by a low-rank expansion that includes 30 frequency components. The MCMC scheme, which uses a fully marginalized posterior, explores a 38-dimensional parameter space (36 EFACs plus the GW-background A and γ), while the quasi-Gibbs scheme must deal with a multitude of extra parameters: $2 \times 36 \times 30 = 2,160$ frequency modes and $36 \times (\text{an average of } 12) = 441$ timing-model parameters, for a whopping total of 2,639.

The autocorrelation functions of the two MCMC chains are shown in Fig. 4. As it was the case in the first test, the quasi-Gibbs scheme vastly outperforms the adaptive Metropolis MCMC, although it must contend with a much larger parameter space. The smaller autocorrelation lengths result from the fact that the Metropolis–Hastings updates are never performed on all the parameters at once. In fact, the GW-background steps are two dimensional, and the noise steps are one dimensional.

In a more realistic analysis we would have to model also correlated spin noise for every pulsar. The corresponding hyperparameters are highly covariant with those of the GW background, and together they would create a 74-dimensional covariant block, a very significant increase. However, that is as bad as it gets; all the other parameters (such as white-noise, jitterlike-noise, and DM-variation hyperparameters) would not further increase autocorrelation lengths. Since 74 covariant dimensions are manageable with modern computing systems, our scheme makes a full-IPTA-sized, full-parameter-set analysis feasible.

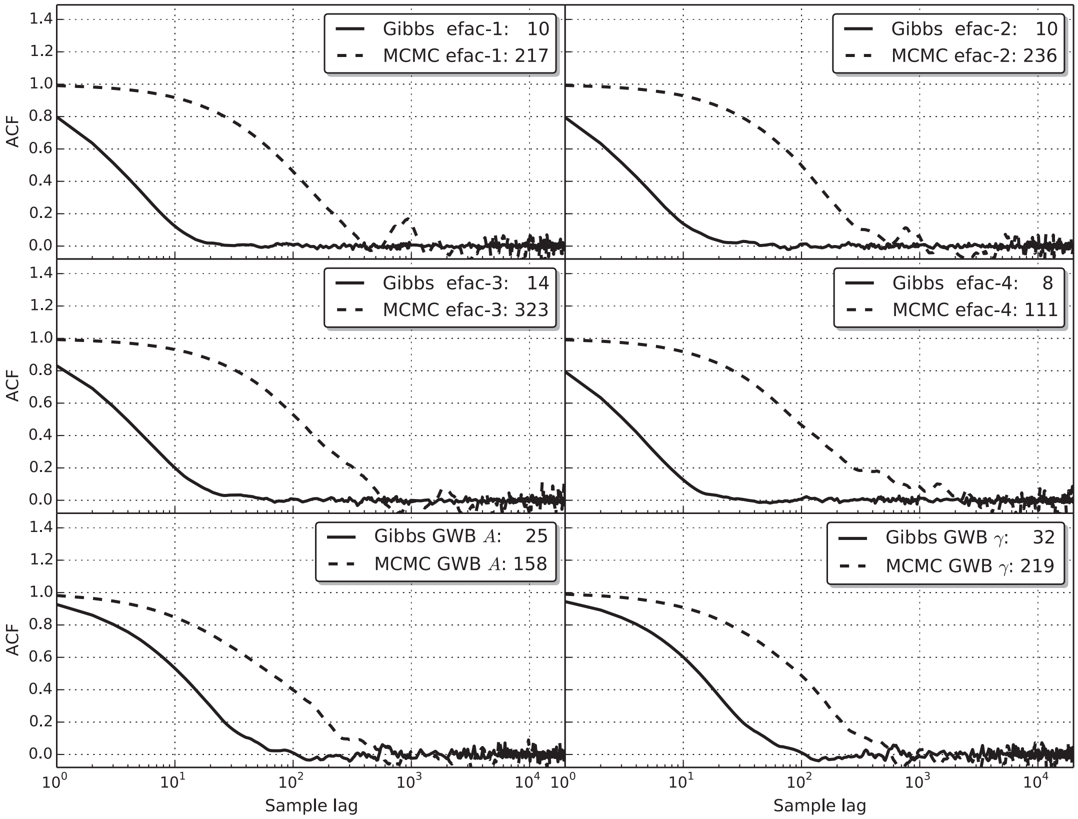

FIG. 4. Similar to Fig. 3, for the adaptive-MCMC chain (dashed) and collapsed quasi-Gibbs chain (solid), both run on the multipulsar data set form the first IPTA Mock Data Challenge. We plot the autocorrelation functions for four EFAC parameters, and for the two GW-background parameters; the legends show the autocorrelation lengths, which are always shorter for the Gibbs sampler, although not as dramatically as in Fig. 3.

VIII. CONCLUSIONS

In this paper we have reviewed the description of stochastic signals in pulsar-timing data analysis, which we have recast in the language of Gaussian processes. In this formal context we have rederived and optimized various expressions that are used in Bayesian inference. For some, the Gaussian-process description offers a more insightful interpretation; for others, it allows computationally more efficient implementations.

The Bayesian-inference schemes in current use have trouble scaling up to large modern data sets such as those assembled by current PTA collaborations. Their analysis should include full pulsar noise models as in Ref. [69], resulting in a very large parameter space to explore. Even with the optimized likelihood expressions that we reviewed in this manuscript, the ensuing MCMC autocorrelation lengths are so large that practical analysis becomes computationally challenging. In this paper we have addressed this problem by constructing two sampling schemes inspired by Gibbs sampling.

The first scheme is very well suited to power-spectral-density estimation in single-pulsar data sets, where we parametrize the power spectrum by independent variance parameters at frequencies multiples of $1/T$, with T the length of the data set. Currently this is done in practice for few frequencies (up to ~20 [19]). However, an extended analysis should include many more Fourier modes, possibly all the way up to the Nyquist frequency. In our scheme we partition parameter space in several blocks; for some of them we can draw samples directly from the conditional block posterior; others allow very rapid conditional-posterior evaluations. The parameters in different blocks are almost uncorrelated, resulting in greatly reduced chain autocorrelation lengths. With tests on mock data we demonstrated that the autocorrelation lengths obtained with our Gibbs-inspired sampler are nearly optimal for all Fourier-sum variances, which makes extended spectral analysis practical for single pulsars.

The second scheme, which we named a collapsed quasi-Gibbs sampler, is well suited for the Bayesian analysis of very large multipulsar data sets. Unlike the first scheme, this sampler does rely on perturbative Metropolis–Hastings updates, so autocorrelation lengths cannot be minimal. However, by combining blocked updates with the direct sampling of quadratic parameters from their conditional posteriors, we were still able to reduce autocorrelation lengths significantly compared to more conventional MCMC methods. Furthermore, in our Gibbs-like scheme the autocorrelation lengths are much less dependent on the number of noise parameters. This makes full noise modeling in Bayesian methods practical in very large data sets: we look forward to actually tackling them in their full glory.

ACKNOWLEDGMENTS

We are grateful to many NANOGrav and European Pulsar Timing Array colleagues for helpful discussions, to Scott Ransom for use of the NRAO Nimrod cluster, where our tests were run, and to Stephen Taylor and an anonymous referee for useful comments. R. v. H. is supported by NASA Einstein Fellowship Grant No. PF3-140116. M. V. is supported by the Jet Propulsion Laboratory RTD program. The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship is acknowledged.

APPENDIX A: EQUIVALENCE OF THE M- AND G-MATRIX FORMULATIONS

The equivalence of Eqs. (22) and (25) (i.e., Eq. (18) of Ref. [17] and Eq. (15) of Ref. [36]) is established by the following derivation, which is implied but not shown in Ref. [36]. Consider the full SVD decomposition $M = U\Sigma V^*$ [with dimensions $(n\times n)\times(n\times p)\times(p\times p)$], which is equivalent to the reduced decomposition $F\hat{\Sigma}V^*$ [with dimensions $(n\times p)\times(p\times p)\times(p\times p)$], where $U=[FG]$. In particular, the p columns of F span the range of M, while the $n-p$ columns of G form the orthonormal completion of F to a full n-dimensional basis.

We first concentrate on the determinants that appear at the denominator of Eq. (22), obtaining

$$\begin{aligned}|M^TC^{-1}M| &= |V\hat{\Sigma}F^TC^{-1}F\hat{\Sigma}V^*| = |\hat{\Sigma}F^TC^{-1}F\hat{\Sigma}|\\ &= |\hat{\Sigma}|^2|F^TC^{-1}F| \end{aligned}\tag{A1}$$

(since orthogonal transformations leave determinants invariant and the determinant of the product of square matrices is the product of their determinants); and

$$\begin{aligned}|C| &= |U^TCU| = |G^TCG||(F^TC^{-1}F)^{-1}|\\ &= |G^TCG|/|F^TC^{-1}F|,\end{aligned}\tag{A2}$$

where the second equality can be read off from the block matrix identity

$$\begin{aligned}U^TCU &= \begin{pmatrix} G^TCG & G^TCF \\ F^TCG & F^TCF \end{pmatrix}\\ &= \begin{pmatrix} G^TCG & 0 \\ F^TCG & I \end{pmatrix}\\ &\quad\times\begin{pmatrix} I & (G^TCG)^{-1}G^TCF \\ 0 & F^TCF - F^TCG(G^TCG)^{-1}G^TCF \end{pmatrix}\\ &= \begin{pmatrix} G^TCG & 0 \\ F^TCG & I \end{pmatrix}\begin{pmatrix} I & (G^TCG)^{-1}G^TCF \\ 0 & (F^TC^{-1}F)^{-1} \end{pmatrix}.\end{aligned}\tag{A3}$$

Thus, the normalization factor of Eq. (22) is given by $\sqrt{(2\pi)^n|\hat{\Sigma}|^2|G^TCG|}$; we may drop $|\hat{\Sigma}|$, which is essentially arbitrary (it is the Jacobian of the coordinate transformation $\eta' = (F^TF)^{-1}F^TM\eta$, while we take infinitely vague priors for these parameters), and adjust the 2π exponent to match the $n - p$ dimension of G^TCG.

Moving on to the main quadratic expression in Eq. (22), we rewrite $C' = C^{-1} - C^{-1}M(M^TC^{-1}M)^{-1}M^TC^{-1}$ as

$$\begin{aligned} C' &= C^{-1} - C^{-1}(F\hat{\Sigma}V^*)(V\hat{\Sigma}F^TC^{-1}F\hat{\Sigma}V^*)^{-1}(V\hat{\Sigma}F^T)C^{-1} \\ &= C^{-1} - C^{-1}F(F^TC^{-1}F)^{-1}F^TC^{-1}, \end{aligned} \tag{A4}$$

where we have used the fact that for unitary V and invertible X, $(VXV^*)^{-1} = VX^{-1}V^*$ and that for diagonal $\hat{\Sigma}$ and invertible Y, $(\hat{\Sigma}Y\hat{\Sigma})^{-1} = \hat{\Sigma}^{-1}Y^{-1}\hat{\Sigma}^{-1}$. Now, if we apply C' to the data y rewritten as $(FF^T + GG^T)y$, we see that the terms that involve F^Ty (either on the right or the left) vanish trivially. For instance,

$$\begin{aligned} C'(FF^Ty) &= (C^{-1}F - C^{-1}F(F^TC^{-1}F)^{-1}F^TC^{-1}F)(F^Ty) \\ &= (C^{-1}F - C^{-1}F)(F^Ty) = 0. \end{aligned} \tag{A5}$$

We are then left with

$$\begin{aligned} y^TC'y &= (G^Ty)^TG^TC'G(G^Ty) \\ &= (G^Ty)^T(G^TC^{-1}G - G^TC^{-1}F(F^TC^{-1}F)^{-1}F^TC^{-1}G) \\ &\quad\times(G^Ty) \\ &= (G^Ty)^T(G^TCG)^{-1}(G^Ty), \end{aligned} \tag{A6}$$

where the last equality can be proved by direct matrix multiplication. We thus recover Eq. (25).

APPENDIX B: DATA COMPRESSION

As we have seen in Sec. V, we have focused our efforts on overcoming the bottleneck in evaluating the likelihood on low-rank expansions of the covariance matrix. We observed that the covariance matrix is the sum of a diagonal matrix and a rank-reduced matrix, and we applied the Woodbury lemma in various ways, thereby accurately approximating the likelihood function.

Another approach that utilizes the rank deficiency of various components in the covariance matrix was formulated by van Haasteren [68], who observed that one is usually not interested in all the parameters $\theta^{(\text{non-TS})}$ in the likelihood function, which allows for the likelihood function to be modified in a way that retains sensitivity only to the parameters of interest. This was presented in the form of linear data compression $\hat{y} = Hy$, with the compression matrix H constructed in a way to maximize sensitivity to some subset of $\theta^{\text{non-TS}}$ with its number of columns as low as possible. In the language of this paper, it means that the information about our signal of interest is encoded in a small subset of ϕ_μ functions of the Gaussian process. By using a data vector of reduced size, the transformed covariance matrix is reduced in size as well, which in turn reduces the computational burden. Here we present these ideas in a slightly altered way to conform to the formalism presented in this work.

Essentially, to evaluate the likelihood, we want to approximate two quantities, $y^TC^{-1}y$ and $\det C$ [or when including the timing model, these same quantities with the G-matrix inserted as in Eq. (25)]. For some combinations of D and U, it is possible to use the approximation

$$\begin{aligned} y^TC^{-1}y &= y^T(D + USU^T)^{-1} \\ &\approx y^TH(H^TDH + H^TUSU^TH)^{-1}H^Ty \\ &\quad + y^TH_c(H_c^TDH_c)^{-1}H_cy. \end{aligned} \tag{B1}$$

Here H and H_c are matrices with the properties $H_c^TH = 0$, and $HH^T + H_cH_c^T = I$, and they must be constructed for a specific problem. It is only possible to find suitable H and H_c when the following requirements can be satisfied:

$$H_c^TU = 0, \qquad HH^TU = U, \qquad H_c^TDH = 0. \tag{B2}$$

We found that these criteria are sufficiently satisfied only in limited cases, mainly when the columns of U consist of a basis of Fourier modes as described in Sec. V A. The matrix H can be constructed from U analogous to how the G matrix was constructed from M in Sec. IV with an SVD. Including the marginalization over the timing model with the G-matrix formalism, we end up with $(H, H_c) = W$, with W from the SVD $W\Sigma V^* = G^TUU^TG$. Here H consists of the first l columns of W, with l the number of nonsingular values in Σ.[11]

With Eq. (B1) we have made the likelihood function separable, with one piece greatly rank reduced and the other part large but with a diagonal covariance matrix. The bottleneck will be the $O(l^3)$ inversion of H^TCH or the $O(ln^2)$ operation of the multiplication H^TC.

Our presentation of data compression differs from the "ABC method" originally presented by van Haasteren [68], which did not include both terms of the separated likelihood function. By only using the data H^Ty, and not H_c^Ty, the ABC method loses sensitivity to some model parameters, and the actual value of the likelihood is changed. Bayesian model selection is not possible in that case, or when $H_c^TDH \neq 0$. We do note that, even when our likelihood function is not fully separable, Eq. (B1) represents a

[11] This is typically the number of columns in U, but it can be smaller (numerically) when the timing basis is sufficiently close to the basis in U.

fully valid way to do analyze the observations. It is equivalent to partitioning the data in two separate components and analyzing the components simultaneously. Some correlation information may have gotten lost, but the result is still internally consistent for *any* H. This does not mean that the parameter estimates are the same for any H. Since the data is changed, the actual estimates can vary.

[1] J. H. Taylor and J. M. Weisberg, Astrophys. J. **253**, 908 (1982).

[2] M. Kramer *et al.*, Science **314**, 97 (2006).

[3] R. S. Foster and D. C. Backer, Astrophys. J. **361**, 300 (1990).

[4] F. B. Estabrook and H. D. Wahlquist, Gen. Relativ. Gravit. **6**, 439 (1975).

[5] M. V. Sazhin, Sov. Astron. **22**, 36 (1978).

[6] S. Detweiler, Astrophys. J. **234**, 1100 (1979).

[7] R. van Haasteren *et al.*, Mon. Not. R. Astron. Soc. **414**, 3117 (2011).

[8] M. Kramer and D. J. Champion, Classical Quantum Gravity **30**, 224009 (2013).

[9] P. B. Demorest *et al.*, Astrophys. J. **762**, 94 (2013).

[10] M. A. McLaughlin, Classical Quantum Gravity **30**, 224008 (2013).

[11] R. N. Manchester *et al.*, Publ. Astron. Soc. Aust. **30**, e017 (2013).

[12] G. Hobbs, Classical Quantum Gravity **30**, 224007 (2013).

[13] G. Hobbs *et al.*, Classical Quantum Gravity **27**, 084013 (2010).

[14] R. N. Manchester, Classical Quantum Gravity **30**, 224010 (2013).

[15] F. A. Jenet, G. B. Hobbs, K. J. Lee, and R. N. Manchester, Astrophys. J. **625**, L123 (2005).

[16] F. A. Jenet, G. B. Hobbs, W. van Straten, R. N. Manchester, M. Bailes, J. P. W. Verbiest, R. T. Edwards, A. W. Hotan, J. M. Sarkissian, and S. M. Ord, Astrophys. J. **653**, 1571 (2006).

[17] R. van Haasteren, Y. Levin, P. McDonald, and T. Lu, Mon. Not. R. Astron. Soc. **395**, 1005 (2009).

[18] M. Anholm, S. Ballmer, J. D. E. Creighton, L. R. Price, and X. Siemens, Phys. Rev. D **79**, 084030 (2009).

[19] L. Lentati, P. Alexander, M. P. Hobson, S. Taylor, J. Gair, S. T. Balan, and R. van Haasteren, Phys. Rev. D **87**, 104021 (2013).

[20] R. M. Shannon *et al.*, Science **342**, 334 (2013).

[21] S. R. Taylor and J. R. Gair, Phys. Rev. D **88**, 084001 (2013).

[22] D. R. B. Yardley *et al.*, Mon. Not. R. Astron. Soc. **407**, 669 (2010).

[23] V. Corbin and N. J. Cornish, arXiv:1008.1782.

[24] A. Sesana and A. Vecchio, Phys. Rev. D **81**, 104008 (2010).

[25] R. van Haasteren and Y. Levin, Mon. Not. R. Astron. Soc. **401**, 2372 (2010).

[26] K. J. Lee, N. Wex, M. Kramer, B. W. Stappers, C. G. Bassa, G. H. Janssen, R. Karuppusamy, and R. Smits, Mon. Not. R. Astron. Soc. **414**, 3251 (2011).

[27] J. A. Ellis, X. Siemens, and J. D. E. Creighton, Astrophys. J. **756**, 175 (2012).

[28] S. Babak and A. Sesana, Phys. Rev. D **85**, 044034 (2012).

[29] A. Petiteau, S. Babak, A. Sesana, and M. de Araújo, Phys. Rev. D **87**, 064036 (2013).

[30] J. M. Cordes and R. M. Shannon, arXiv:1010.3785.

[31] R. M. Shannon and J. M. Cordes, Astrophys. J. **725**, 1607 (2010).

[32] S. Geman and D. Geman, *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images* (IEEE, New York, 1984), Vol. 6, p. 721.

[33] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, Adaptative Computation and Machine Learning Series (Mit, Cambridge, MA, 2006).

[34] K. J. Lee *et al.*, Mon. Not. R. Astron. Soc. **441**, 2831 (2014).

[35] D. R. Lorimer and M. Kramer, *Handbook of Pulsar Astronomy*, edited by R. Ellis, J. Huchra, S. Kahn, G. Rieke, and P. B. Stetson (Cambridge University Press, Cambridge, England, 2004).

[36] R. van Haasteren and Y. Levin, Mon. Not. R. Astron. Soc. **428**, 1147 (2013).

[37] G. Golub and C. Van Loan, *Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences* (Johns Hopkins University Press, Baltimore, 2012).

[38] W. Coles, G. Hobbs, D. J. Champion, R. N. Manchester, and J. P. W. Verbiest, Mon. Not. R. Astron. Soc. **418**, 561 (2011).

[39] L. Lentati, P. Alexander, M. P. Hobson, F. Feroz, R. van Haasteren, K. J. Lee, and R. M. Shannon, Mon. Not. R. Astron. Soc. **437**, 3004 (2014).

[40] S. J. Vigeland and M. Vallisneri, Mon. Not. R. Astron. Soc. **440**, 1446 (2014).

[41] G. B. Hobbs and R. T. Edwards, http://www.sf.net/projects/tempo2.

[42] R. Blandford, R. W. Romani, and R. Narayan, J. Astrophys. Astron. **5**, 369 (1984).

[43] K. J. Lee, C. G. Bassa, G. H. Janssen, R. Karuppusamy, M. Kramer, R. Smits, and B. W. Stappers, Mon. Not. R. Astron. Soc. **423**, 2642 (2012).

[44] R. van Haasteren and M. Vallisneri, arXiv:1407.6710.

[45] J. W. Armstrong, Nature (London) **307**, 527 (1984).

[46] M. J. Keith *et al.*, Mon. Not. R. Astron. Soc. **429**, 2161 (2013).

[47] T. T. Pennucci, P. B. Demorest, and S. M. Ransom, Astrophys. J. **790**, 93 (2014).

[48] X. P. You *et al.*, Mon. Not. R. Astron. Soc. **378**, 493 (2007).

[49] F. A. Jenet, A. Lommen, S. L. Larson, and L. Wen, Astrophys. J. **606**, 799 (2004).

[50] A. H. Jaffe and D. C. Backer, Astrophys. J. **583**, 616 (2003).

[51] A. Sesana, A. Vecchio, and C. N. Colacino, Mon. Not. R. Astron. Soc. **390**, 192 (2008).

[52] A. Sesana, Mon. Not. R. Astron. Soc. **433**, L1 (2013).

[53] T. Damour and A. Vilenkin, Phys. Rev. D **71**, 063510 (2005).

[54] X. Siemens, V. Mandic, and J. Creighton, Phys. Rev. Lett. **98**, 111101 (2007).

[55] L. P. Grishchuk, Physics Usp. **48**, 1235 (2005).

[56] C. Caprini, R. Durrer, and X. Siemens, Phys. Rev. D **82**, 063511 (2010).

[57] R. W. Hellings and G. S. Downs, Astrophys. J. **265**, L39 (1983).

[58] K. J. Lee, F. A. Jenet, and R. H. Price, Astrophys. J. **685**, 1304 (2008).

[59] M. E. D. S. Alves and M. Tinto, Phys. Rev. D **83**, 123529 (2011).

[60] S. J. Chamberlin and X. Siemens, Phys. Rev. D **85**, 082001 (2012).

[61] C. M. F. Mingarelli, T. Sidery, I. Mandel, and A. Vecchio, Phys. Rev. D **88**, 062005 (2013).

[62] F. A. Jenet, T. Creighton, and A. Lommen, Astrophys. J. Lett. **627**, L125 (2005).

[63] X. Siemens, J. Ellis, F. Jenet, and J. D. Romano, Classical Quantum Gravity **30**, 224015 (2013).

[64] G. Hobbs *et al.*, Mon. Not. R. Astron. Soc. **427**, 2780 (2012).

[65] N. J. Cornish and R. van Haasteren (unpublished).

[66] W. Hager, SIAM Rev. **31**, 221 (1989).

[67] L. Trefethen and D. Bau, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997).

[68] R. van Haasteren, Mon. Not. R. Astron. Soc. **429**, 55 (2013).

[69] Z. Arzoumanian *et al.*, arXiv:1404.1267.

[70] J. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics (Springer, New York, 2001).

[71] G. Roberts, A. Gelman, and W. Gilks, Ann. Appl. Probab. **7**, 110 (1997).

[72] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C. The Art of Scientific Computing* (Cambridge University Press, Cambridge, England, 1992).

[73] E. T. Jaynes and G. L. Bretthorst, *Probability Theory* (Cambridge University Press, Cambridge, England, 2003).

[74] M. Vallisneri, https://github.com/vallis/libstempo.

[75] R. N. Manchester *et al.*, http://www.atnf.csiro.au/research/pulsar/psrcat.

[76] M. J. Keith, K. J. Lee, and F. Jenet (unpublished).