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

Rutger van Haasteren, Michele Vallisneri
Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA 91109
(Dated: June 13, 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 datasets. 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 datasets assembled by the International Pulsar Timing Array collaboration, which present a serious computational challenge to existing methods.

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 local 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 (EPTA, [7,8]), the North American Nanohertz Observatory for Gravitational Waves (NANOGrav, [9,10]), and the Australian Parkes Pulsar Timing Array (PPTA, [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 datasets 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: 9,13,24], and continuous waves [for instance: 22,23]. 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 datasets such as those integrated by the International Pulsar Timing Array (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 [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. 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 likelihood of the data in the presence of stochastic signals. We also describe and test two novel Bayesian sampling schemes, inspired by Gibbs sampling [22], 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. VII we present our new-and-improved quasi-Gibbs schemes, which we test on mock data in Sec. VIII. 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 dataset 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:

- 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 \( w_{\mu} \), which are themselves Gaussian random variables with mean vector \( w_0 \) and covariance matrix \( \Sigma_{\mu \nu} \). (This is the weight-space view.)
- 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 \( m(x) = w_0 = 0 \). The duality between the two views and specifications is encapsulated by the covariance-function expansion

\[
k(x, x') = \sum_{\mu, \nu} \phi_{\mu}(x)\Sigma_{\mu \nu} \phi_{\nu}(x').
\]

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 \( \epsilon_i \) with covariance matrix \( N_{ij} \) [33]:

\[
\begin{align*}
p(y_i|w_\mu, \text{GP}) &= \frac{e^{-\frac{1}{2} \sum_{ij} (y_i - \sum_\mu \phi_{\mu}(x_i)w_\mu)(N_{ij})^{-1}(y_j - \sum_\nu \phi_{\nu}(x_j)w_\nu)}\prod_{\mu \nu \mu} w_\nu \Sigma_{\mu \nu}^{-1} w_\mu}{\sqrt{(2\pi)^n \det N}} \\
p(y_i|\text{GP}) &= \frac{e^{-\frac{1}{2} \sum_{ij} y_i (N_{ij} + K_{ij})^{-1} y_j}}{\sqrt{(2\pi)^n \det(N + K)}}, \quad \text{with} \quad K_{ij} = k(x_i, x_j) = \sum_{\mu \nu} \phi_{\mu}(x_i)\Sigma_{\mu \nu} \phi_{\nu}(x_j),
\end{align*}
\]

where \( i, j = 1, \ldots, n \) and \( \mu, \nu = 1, \ldots, m \), 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 \( w \), 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 \( \Sigma_{\mu \nu} \)] 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 \( p(y_i|w_\mu, \text{GP}) \), the marginalized likelihood \( p(y_i|\text{GP}) \) 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, \( p(y_i|\text{GP}) \) 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 \( x' = \sum_{ij} k(x', x_i)(N_{ij} + K_{ij})^{-1} y_j \) (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 re-derived and used in various forms in pulsar timing, for example for analytical marginalisation of the timing model parameters [17], and for reconstruction of DM variations [15].

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:

$$
\log p(y|w, \text{GP}) = -\frac{1}{2} \left( y - \Phi^T(x)w \right)^T N^{-1} \left( y - \Phi^T(x)w \right) - \frac{1}{2} w^T \Sigma^{-1} w - \log N_{n,N} - \log N_{m,\Sigma},
$$

$$
\log p(y|\text{GP}) = -\frac{1}{2} y^T (N + K)^{-1} y - \log N_{n,N+K},
$$

where $n$ and $m$ are the sizes of the squares matrices $N$ and $\Sigma$, and where

$$
\Phi_{\mu,i} = \phi_{\mu}(x_i), \quad K = \Phi^T \Sigma \Phi, \quad \text{and} \quad \log N_{p,X} = \frac{p}{2} \log(2\pi) + \frac{1}{2} \log \det X.
$$

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. THE 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 time of arrival (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 [34], 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 datasets 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 [37, 35], 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

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

where the vector $\epsilon$ 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 $\theta$ denotes the collection of all $\theta^{(A)}$. The crucial result from Gaussian-process theory, which enables Bayesian inference on the $\theta$, is the fact that the marginal likelihood $p(y|\theta, \text{GP})$ can be written simply as

$$
\log p(y|\theta, \text{GP}) = -\frac{1}{2} y^T (N + \sum_{(A)} K^{(A)})^{-1} y - \log N_{n,N+\sum_{(A)} K^{(A)}}, \quad \text{(6)}
$$

(In fact, this simple description requires two slight complications: first, the timing-solution errors are usually
given a special treatment, discussed in Secs. II A and IV, second, the measurement-error matrix $N$ is also parameterized 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 dataset of $n$ TOAs using Eq. (3) 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, 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 $N + K = U^TU$, which yields the determinant as the product of the squared diagonal elements, and then compute

$$y^T(N + K)^{-1}y = y^T(U^TU^{-1}y),$$

where we used the MATLAB notation $A\backslash b$ for the solution 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 \cite{38,39} 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 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. 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

$$y^{(TS)} = \sum_{a} \phi_a^{(TS)}(t)\delta\eta_a \equiv M\delta\eta,$$

where $\delta\eta_a$ is the $p$-dimensional vector of the parameter errors $\delta\eta_a = \eta^{\text{best-fit}}_a - \eta^{\text{true}}_a$, where the $\phi_a^{(TS)}(t)$ are the partial derivatives of the TOAs with respect to the $\eta_a$ evaluated at $\eta^{\text{best-fit}}$, and where $M$ is the design matrix $M_{ia} = \phi_a^{(TS)}(t_i)$. The assumption that this linear regime for the $y^{(TS)}$ 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 \cite{38,39}.

We do not usually deal explicitly with the covariance matrix that ensues from the basis functions $\phi_a^{(TS)}$, because it is awkward to attribute a prior covariance $\Sigma^{TS}$ to the $\delta\eta_a$. 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 $\Sigma^{TS} = \lambda I_p$.

### B. Measurement errors (EFAC, EQUAD, and jitter-like 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. In fact, the physics of the measurement suggests that separate EFACs should be used for every receiver or backend represented in the dataset. 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. Again, different EQUADs may be assigned to multiple receivers and backends.

Last, the circumstance that certain datasets (notably those collected by the NANOGrav collaboration \cite{23}) 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 caused by variable emission within pulsar magnetospheres.

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

\[ K^{(MN)} = E^2 n_i \delta_{ij} + Q^2 \delta_{ij} + J^2 \delta_{(e(i)e(j))}, \]  

(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 \( \delta \) are Kronecker deltas; and where \( e(i) \) 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 ones. 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 noiseful measurement noise \( \epsilon \) (in which case \( N = K^{(MN)} \)), or to designate EQUAD noise and jitter-like noise as separate Gaussian processes (in which case \( K^{(MN)} = N + K^{(Q)} + K^{(J)} \)). For the case of multiple receivers and backends, separate EFAC, EQUAD, and jitter-like 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/backend, 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) 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 power spectral density:

\[ P^{(PL)}(f) = A^2 (f/yr^{-1})^{-\gamma} yr^3, \]

(9)

where \( f \) is the frequency, \( A \) is a dimensionless amplitude and \( \gamma \) 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, Eq. (9) results in the correlation matrix

\[ K_{ij}^{(PL)} = k^{(PL)}(t_i, t_j) = A^2 (f_L/yr^{-1})^{-1-\gamma} \left\{ \Gamma(1-\gamma) \sin \left( \frac{\pi \gamma}{2} \right) (f_L t_{ij})^{\gamma-1} - \sum_{n=0}^{\infty} \frac{(-1)^n (f_L t_{ij})^{2n}}{(2n)!} (2n+1-\gamma) \right\}, \]

(10)

where \( \Gamma(.) \) 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 \( f_{PL} = (1/2 - \gamma/2), (1/2, 3/2 - \gamma/2), -(f_L t_{ij})^2/4)/(\gamma-1), \) where \( f_{PL} \) is the generalized hypergeometric function given by HypergeometricPFQ in Mathematica and by hyp1f2 in scipy.special.

Blandford and colleagues and later other authors 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-spindowns term of the timing model, at least for \( \gamma \) up to 7 (up to 5 using the linear term alone). For \( \gamma = 1 \), the total variance \( \int P(f)df \) becomes infinite even with the low-frequency cutoff. Thus, the spectral index \( \gamma \) is usually taken in the interval \([1, 7], \)

although imposing a high-frequency cutoff 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. Furthermore, \( K_{ij}^{(PL)} \) is a dense, full-rank matrix, so its use in computing residual likelihoods incurs the full \( O(n^3) \) cost of matrix inversion. For \( \gamma > 6 \) the matrix \( K_{ij}^{(PL)} \) 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

---

5 For a stationary process for which \( k(x', x'') = C(x' - x'') = C(2\Delta x) \), the Wiener–Khinchin theorem relates the power spectral density \( P(f) \) to the correlation function \( C(\Delta t) \) by way of \( C(\Delta t) = \int_0^\infty \cos(2\pi f \Delta t)P(f)df \). 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 \( \gamma \), so special-case expressions are required. An alternative, more benign low-frequency regularization is to redefine \( P(f) = A^2((f/yr)^2 + (f_L yr)^2)^{-\gamma/2} \), which leads to a \( C(\tau) \) expressed in terms of modified Bessel functions of the second kind.
of Fourier modes \[^{19}\]:

\[
y^{(\text{FM})}(t) = \sum_{k=1}^{q} a_k \cos(2\pi k x) + b_k \sin(2\pi k x),
\]

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 \(\phi_{t,\nu}\) instead of the covariance function, and solving for the weights \(w_{t,\nu}\) (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 \(\Sigma^{(\text{FM})}_{\mu \nu}\) as function of a set of hyperparameters. For instance, a diagonal \(\Sigma^{(\text{FM})}_{\mu \nu}\) specifying a set of variances \(\rho_{\mu}\), each shared by the cos and sin modes of the same frequency \(f_{\mu}\), 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 \(\rho_{\mu}\) is still required).

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:

\[
K^{(\text{PL})}_{ij} = \sum_{\mu \nu} \phi_{\nu}(\mathbf{t}_i) \Sigma^{(\text{FM})}_{\mu \nu} \phi_{\mu}(\mathbf{t}_i)
\]

(12)

with

\[
\Sigma^{(\text{FM})}_{\mu \nu} = \Sigma^{(\text{PL})}(f_{\mu}) \Delta f_{\mu \nu} = \Sigma^{(\text{PL})}(f_{\mu}) \delta_{\mu \nu}/T.
\]

(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 \(K^{(\text{PL})}\). However, in practice Eq. (12) can be very accurate (especially if additional, logarithmically spaced modes are added at low frequencies \[^{43}\]). 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 \[^{30, 44}\], given by

\[
y^{(\text{DM})} = \left(4.15 \times 10^{-3}\right) \left(\frac{\text{DM}}{\text{pc cm}^{-3}}\right) \left(\frac{\nu}{\text{GHz}}\right)^{-2}
\]

(14)

for the delay of a pulse measured at frequency \(\nu\) 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 \[^{15}\] for a discussion of previous work to characterize DM variations and their impact on pulsar-timing GW searches. In the analysis of pulsar-timing datasets 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 \[^{16}\] function. Alternatively, one can try to solve for DM variations from the multifrequency observations at each epoch, effectively generating a reduced infinite-frequency dataset \[^{45, 47}\].

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 \[^{38, 45}\]. For DM variations characterized by the power-law power spectral density

\[
P^{(\text{DM})}(f) = A_{\gamma}^{2}(f/\gamma)^{-\gamma_{\text{DM}}},
\]

(15)

the timing-residual covariance function is

\[
K^{(\text{DM})}_{ij}(t_i, t_j) = \left(4.15 \times 10^{-3}\right)^2 \times \left(\frac{\nu_{ij}}{\text{GHz}}\right)^{-\gamma_{\text{DM}}} \text{A} \rightarrow \text{A}_{\gamma} \rightarrow \gamma_{\text{DM}}
\]

(16)

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

The caveats given above for \(K^{(\text{PL})}_{ij}\) apply also to the evaluation of \(K^{(\text{DM})}_{ij}\). It is also possible to model \(y^{(\text{DM})}\) as a sum over basis functions, in analogy to Eq. (14), 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 dataset), the very low-frequency behavior of the Gaussian process is not modeled well \[^{38}\]; this can be remedied by enhancing the timing-model design matrix with a term similar to quadratic spindown, but with \(\nu^{-2}\) frequency dependence \[^{38}\], or by adding more modes at low non-Fourier frequencies \[^{43}\].

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., \[^{49}\] 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:

\[ P^{(GW)}(f) = \frac{A_{GW}^2}{12\pi^2} (f/\text{yr}^{-1})^{-\gamma_{GW}} \text{yr}^4, \]

where the \(12\pi^2\) factor follows from defining \(A_{GW}\) as the dimensionless characteristic strain \(h_c\) at \(f = 1/\text{yr}\) \cite{10}.

\[ h_c(f) = A_{GW}(f/\text{yr}^{-1})^{\alpha_{GW}} \quad \text{where} \quad \gamma_{GW} = 3 - 2\alpha_{GW}. \]

The spectral index \(\gamma_{GW}\) is 13/3 (but possibly less at low frequencies) for the background from the sum of unresolved black-hole binaries \cite{50,51}: 16/3 for a background from cosmic superstrings \cite{52,53}: 5 for a background of inflationary relics \cite{54}. Nonstrictly power-law spectra are also possible, as in the case of the QCD phase transition \cite{55}.

Thus, a GW background can be modeled as a stochastic process with a time-domain covariance matrix analog to Eq. (12), or with a Fourier-sum covariance analog to Eq. (10). However, the very concept of pulsar timing arrays robust detectors for gravitational waves. 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 \cite{56}. However, GWs are not the only types of signals that can induce correlations. Slow drifts of atomic clocks can introduce an slight error in terrestrial time standards, which would manifest themselves as a common low-frequency signal in the signals of all pulsars \cite{57}. Such a correlated signal would be a source of noise when detecting a GW background, and it must be modelled 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 \(\tau = 1\) 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 regularising 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 noise one could imagine are inaccuracies in the solar system ephemeris. Although these are unlikely to be a significant source of noise at the level of stochastic searches, if necessary they are easily modelled as a correlated stochastic signal. In that case, one should replace Eq. (20) with \(\zeta(\gamma_{ab}) = \cos(\gamma_{ab}).\)

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

\[ \log p(y|\hat{\theta}^{(\text{non-TS})}, \delta\eta_a) = -\frac{1}{2}(y - M\delta\eta)^T (N + K^{(\text{non-TS})})^{-1}(y - M\delta\eta) - \log N_{\eta} - K^{(\text{non-TS})}, \]

where \(\hat{\theta}^{(\text{non-TS})}\) denote all the model parameters other than the timing-model errors \(\delta\eta_a\), or we can treat them nonlinearily, as in Refs. \cite{58,59}, so that the residuals \(y\)
are recomputed from the full timing model for each value of the $\eta$ that we sample. This latter approach is desirable if we think that the functional dependence of the residuals on the $\eta$ 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 
\[
p(y|\theta^{(\text{non-TS})}) = \int p(y|\theta^{(\text{non-TS})}, \delta\eta_a) d(\delta\eta_a).
\]
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 \cite{vanhaasteren2010}, who showed that

\[
\int p(y|\theta^{(\text{non-TS})}, \delta\eta_a) d(\delta\eta_a) = \exp\left\{-\frac{1}{2} y^T (C - M\delta\eta) y - \frac{1}{2} \log |C| + \frac{n}{2} \log 2\pi\right\}
\]
\[
= \exp\left\{-\frac{1}{2} y^T C^{-1} y \right\}
\]
\[
= \exp\left\{-\frac{1}{2} y^T C' y \right\}
\]
\[
\text{where } C = N + K^{(\text{non-TS})}.\]

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 $\phi^{(T)}(t)$ (the columns of the design matrix $M$). We can then use Eq. \[(3)\] with $K^{(T)} = M\Sigma^{(T)} M^T$ and $\Sigma^{(T)} = \lambda I_p$, and take the limit $\lambda \to \infty$ corresponding to an infinitely vague prior for the TS weights:

\[
\lim_{\lambda \to \infty} \log p(y|\theta^{(\text{non-TS})}) = \lim_{\lambda \to \infty} \left\{-\frac{1}{2} y^T C^{-1} y - \frac{1}{2} \log |C| + \frac{n}{2} \log 2\pi\right\}
\]
\[
= \lim_{\lambda \to \infty} \left\{-\frac{1}{2} y^T C^{-1} y + \frac{1}{2} y^T C^{-1} M(\lambda^{-1} I_p + M^T C^{-1} M)^{-1} M^T C^{-1} y - \frac{1}{2} \log |C| - \frac{1}{2} \log |\lambda^{-1} I_p + M^T C^{-1} M| - \frac{p}{2} \log \lambda - \frac{n-m}{2} \log 2\pi\right\}
\]
\[
= -\frac{1}{2} y^T C' y - \frac{1}{2} \log |C| - \frac{1}{2} \log |M^T C^{-1} M| - \frac{n-m}{2} \log 2\pi (+ \text{infinite constant})
\]

In the second row of Eq. \[(23)\] we used the Woodbury formula and the matrix determinant lemma \cite{woodbury1950}:

\[
(A + UWV^T)^{-1} = A^{-1} - A^{-1}U(W^{-1} + V^T A^{-1} U)^{-1} V^T A^{-1},
\]
\[
\det(A + UWV^T) = \det(W^{-1} + V^T A^{-1} U) \det W \det A;
\]

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 \cite{vanhaasteren2010} later derived an alternative form for the $\delta\eta_a$-marginalized likelihood, which exploits the singular-value decomposition (SVD) $M = U\Sigma V^*$ \cite{hastie2009}. If $M$ is an $n \times n$ matrix, then $U$ and $V$ are orthogonal matrices of sizes $n \times n$ and $p \times p$ respectively, while $\Sigma$ 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 $FF^T y$ of the residuals involves components that can be reabsorbed by a change in the $\delta\eta_a$, so these components are in effect \textit{unobserved} from the Gaussian-process perspective; a likelihood can then be written directly for the $(n-p)$-dimensional observable data vector $GG^T y$ (or more precisely, for the coefficients of the $y$ over the partial orthonormal basis given by the $G$ columns):

\[
p(y|\theta^{(\text{non-TS})}) = \exp\left\{-\frac{1}{2} y^T G[G^T CG]^{-1} G^T y \right\}
\]
\[
\sqrt{(2\pi)^{n-p}[G^T CG]}.
\]
In App. A we demonstrate that Eq. (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-TS 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 $G^T CG$, and by $O((n - p)n^2)$ matrix multiplications (some of these can be avoided by storing the matrices $G^T K^{(A)}(\theta^{(A)}) G$ for varying values of $\theta^{(A)}$, and interpolating); 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 certain of 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 $l \ll n$.

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:

\[
(D + USU^T)^{-1} = D^{-1} - D^{-1}U^{-1}(S^{-1} + UTD^{-1}U)^{-1}UTD^{-1}.
\] (26)

We see that the matrix inversions in this reworked expression are those of $D$ (an $O(n)$ operation), $S$ (an $O(l^2)$ operation), and $(S^{-1} + UTDU)$ (again $O(l^2)$), gaining us an impressive speedup. The corresponding lemma for the determinant is $|D + USU^T| = |D||S||S^{-1} + UTD^{-1}U|$, 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 $M^T C^{-1} M$ is still necessary (in addition, computing $M^T C^{-1} M$ 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

\[
y^T G(D + USU^T) G^{-1} y = y^TW y - y^T W U (S^{-1} + UT W) U^T W y,
\] (27)

with

\[
W = G(D^T DG)^{-1} G^T.
\] (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 $\theta$ is 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

\[
G(D^T DG)^{-1} G^T = D^{-1} - D^{-1} F(D^{-1} F)^{-1} F D^{-1} - 1
\] (29)

(see Eq. (A6) in App. 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-backend/receiver EFAC and EQUAD noise). We can then diagonalize the two simultaneously with a non-orthogonal basis transformation:

\[
G(D^T DG) = aG^T AG + bG^T BG = L V (aI + bQ)V^T L^T,
\] (30)

with

\[
L L^T = G^T AG \quad \text{and} \quad V Q V^T = L^{-1} G^T B G L^{-1},
\] (31)

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

\[
y^T G(D^T DG)^{-1} G^T y = y^T G (V^T L)^{-1} (aI + bQ)^{-1} (L)^{-1} G^T y,
\] (32)

\[
\det(G^T DG) = \det(G^T AG) \det(aI + bQ).
\] (33)

Because we need to calculate $L V G^T y$ (or any other combination like $L V G^T U$) 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 [64]. Instead of using the Woodbury lemma to expand the low-rank representation of the covariance matrix, the data was 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 on 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. [VA] for correlated timing noise; in Sec. [VB] for jitter-like noise in multifrequency datasets; in Sec. [VC] to define a notion of epoch-averaged residuals. An alternative approach, not based on low-rank expansions: in Sec. [VA] for correlated timing noise, as described in Sec. [VB]. 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], \( U_{\mu t} = \phi^{(FM)}(t_i) \) and \( D_{\mu \nu} = \Sigma^{(FM)}_{\mu \nu} \). As a reminder, \( \mu \) ranges from 1 to 2\( q \), and indexes the Fourier basis functions \( \cos(2\pi f_{\mu} t_i) \) and \( \sin(2\pi f_{\mu} t_i) \) with \( f_{\mu} \) a multiple of \( 1/T \), the inverse duration of the observation; the matrix \( \Sigma^{(FM)}_{\mu \nu} \) 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 \( (\nu_i/\text{GHz})^{-2} \cos(2\pi f_{\mu} t_i) \) and \( (\nu_i/\text{GHz})^{-2} \sin(2\pi f_{\mu} t_i) \), 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 \( \Sigma \) is the block-diagonal matrix of the individual \( \Sigma \)'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 multi-pulsar 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 dataset), the resulting multipulsar GW covariance is given by

\[
K^{\text{(GW)}}_{a ib j} = \sum_{ab\mu\nu} \Phi_{a\mu}(t_{ai})\Sigma^{\text{(GW)}}_{a\mu b \nu} \Phi_{b \nu}(t_{bj}) = \sum_{ab\mu\nu} \Phi_{a\mu}(t_{ai})\Sigma^{\text{(GW)}}_{\mu\nu} \Phi_{b \nu}(t_{bj})
\]

[34]. Now, \( \Sigma^{\text{(GW)}}_{\mu\nu} \) is diagonal, but \( \Sigma^{\text{(red)}}_{a \mu b \nu} \) is dense, so its inverse is potentially expensive. If we order the Fourier coefficients in blocks corresponding to the \( N \) pulsars, the matrix \( \Sigma^{\text{(GW)}}_{a \mu b \nu} \) 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 \( f_{\mu} \), then the matrix \( \Sigma^{\text{(GW)}}_{a \mu b \nu} \) is block-diagonal, with each block a dense matrix given by \( \rho_{\mu \gamma} \); thus, its inverse is just \( \rho_{\mu \gamma}^{-1} \), which incurs an acceptable computational cost \( O(qN^3) \).

In principle \( \gamma_{ab}^{-1} \) 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

\[
\Sigma^{\text{(red)}}_{a \mu b \nu} = \Sigma_{a \mu b \nu}^{\text{(GW)}} + \Sigma_{a \mu}^{\text{(red)}} \delta_{ab};
\]

in this case each block is given by \( \rho_{\mu \gamma}^{\text{(GW)}} + \rho_{\mu \gamma}^{\text{(red)}} \delta_{ab} \), and it must be inverted for each choice of \( \rho_{\mu \gamma}^{\text{(red)}} \).

### B. Low-rank expansions for jitter-like noise

The covariance matrix corresponding to jitter-like noise, as described in Sec. [III B], can be expressed exactly as the low-rank expression

\[
C_J = U E U^T
\]

where \( E \) is a diagonal matrix with entries \( J_{e}^2 \) corresponding to squared amplitude of jitter-like noise at each epoch (usually the same for all epochs corresponding to measurements with the same receiver/backend), 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:
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, jitter-like 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 $(N_e)_{ij} = (E_e^2 u_i + Q_e^2) \delta_{ij}$, with indices ranging over the epoch only, and $u_e^T = (1, 1, \ldots, 1)^T$. By Woodbury’s lemma,

$$D_e^{-1} = N_e^{-1} - \frac{N_e^{-1} u_e u_e^T N_e^{-1}}{\alpha_e}, \tag{39}$$

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

### C. Low-rank expansion of epoch-averaged residuals

A low-rank expansion can also be used to define a statistically principled notion of epoch-averaged residual for multifrequency datasets such as NANOGrav’s [Michele, unpublished NANOGrav memo]. The idea is to write the total Gaussian-process covariance matrix as $N + C = N + U U^T$, 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 $C$ (with $n_e$ the number of epochs) describes components such as jitter-like noise, correlated noise, and GWs 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

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

where we have neglected logarithms of $2\pi$. In Eq. (40) the $n$-dimensional vector of residuals $y$ is replaced by the $n_e$-dimensional vector of epoch-averaged residuals

\[\tilde{y} = U^T \chi y, \quad \tilde{\chi}^2 = \tilde{y}^T W \tilde{y}, \quad \text{and} \quad X = U^T W U, \quad \text{with} \quad W = G(G^T D G)^{-1} G^T.\]

\[\text{DM fluctuations require a slightly more complicated description where } U \text{ gains } n \text{ rows, with the same structure as its those of Eq. (37), but each multiplied by } u_i^2.\]
Unfortunately epoch averaging 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 backend/receiver combinations), so the full set of residuals must be carried along throughout the analysis to recompute $\tilde{y}$, $\tilde{x}^2$, and $X$ as the hyperparameters change. If a single EFAC and EQUAD describe the entire dataset, then epoch-averaged residuals can be used by way of the two-component expansion of Eq. (30).

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

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. [VI] 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 datasets such as the upcoming IPTA data releases, which may contain many tens of thousands 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 whose distribution 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 $\int \phi(x)p(x)\,dx$ that scales as $(\phi)/\sqrt{N}$, with $(\phi)$ is 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 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 well characterized by the sample autocorrelation function (ACF), defined as

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

where $x$ and $N$ are the vector and number of samples, $\bar{x}$ 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 [65]; 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. [VIA] 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

$$\frac{p(\theta^{(n+1)}|\text{data})}{p(\theta^{(n)}|\text{data})} \times \frac{q(\theta^{(n)}|\theta^{(n+1)})}{q(\theta^{(n+1)}|\theta^{(n)})}.$$  (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. Choosing the proposal distribution smartly [see, e.g. 66] 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 inter-block 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. [VIA]):

$$p(y|\theta) \propto \exp\left\{-\frac{1}{2}(y - M\delta\eta - Fa)^T N^{-1}(y - M\delta\eta - Fa)\right\} \times \exp\left\{-\frac{1}{2} a\Sigma^{-1} a\right\}$$

$$\frac{1}{\sqrt{(2\pi)^n|N|}} \times \frac{1}{\sqrt{(2\pi)^p|\Sigma|}}.$$  (43)
are given in the “spectral–estimation” form \( \Sigma \) it, as we see below. If the hyperparameters that define \( \Sigma \) are given in the “spectral–estimation” form \( \Sigma_{\mu\nu} = \rho_{\mu}\delta_{\mu\nu} \) (see Eq. (12)), then the \( \rho_{\mu} \) 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 jitter-like noise, using Eqs. (38) and (39); jitter-like 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 \( \phi_{\mu}^{(TS)}(t_i) = M_{i\mu} \) and \( \phi_{\mu}^{(FM)}(t_i) = F_{i\mu} \) of the timing model and the correlated noise respectively. We denote them collectively as \( w = (\delta \eta^T, a^T) \) and with \( \Phi = (M; F) \). Fixing all the hyperparameters \( \theta \), the log-posterior probability of the \( w \) can be rewritten as

\[
\log P(w|y, \theta, \text{GP}) = \frac{1}{2} (w - Q^{-1} \Phi N^{-1} y)^T Q (w - Q^{-1} \Phi N^{-1} y) - \frac{1}{2} \log \det Q + \text{const}
\]

with

\[
Q = \Phi N^{-1} \Phi^T + \Sigma^{-1}
\]

where \( N, \Sigma, \) and the additive constant are functions of the hyperparameters and of the residuals, and where we interpret \( \Sigma^{-1} \) in the broad sense explained in Sec. [IV] by assuming an infinitely vague prior for the timing-model parameters, we set \( \Sigma^{-1} \) to zero in their subspace. Equation (44) states that the quadratic parameters \( w \) are distributed according to a multivariate normal distribution with mean \( \hat{w} = Q^{-1} \Phi N^{-1} y \) and covariance \( Q^{-1} \). We can draw from this distribution by computing \( w_{\text{new}} = \hat{w} + L \epsilon \), with \( \epsilon \) a vector of zero-mean, unit-norm, uncorrelated normal deviates (see, e.g., [67]), 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 [63], then use an SVD decomposition [63] to compute the square root.

In multipulsar datasets, 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 \( \Sigma \) that includes cross terms between pulsars and collecting all the terms that involve the \( w_a \) results in the conditional probability

\[
\log P(w_a|w_b \neq a, y, \theta, \text{GP}) = \frac{1}{2} (w_a - Q_a^{-1} z_a)^T Q_a (w_a - Q_a^{-1} z_a) - \frac{1}{2} \log \det Q_a + \text{const}(\theta, w_b \neq a)
\]

where

\[
z_a = \Phi_a N_a^{-1} y_a + \sum_{b \neq a} (\Sigma_{ab}^{-1})_{ab} w_b,
\]

\[
Q_a = \Phi_a N_a^{-1} \Phi_a^T + (\Sigma_{aa}^{-1})_{aa},
\]
from which $w_a$ can be drawn directly with the covariance-square-root procedure. Because of the structure of the multipulsar prior [Eq. (33)], computing the submatrices $(\Sigma^{-1})_{aa}$ and $(\Sigma^{-1})_{ab}$ does not require the inversion of the full $\Sigma$, but only of each pulsar block, which is much cheaper.

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

$$\log P(\theta_w|y_{\text{red}}, w, \theta_p, \text{GP}) =$$

$$-\frac{1}{2}y_{\text{red}}^T N^{-1} y_{\text{red}} - \frac{1}{2} \log \det N - \log p(\theta_w),$$

(48)

where the reduced residuals $y_{\text{red}} = \Phi w$ 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 $\theta_w$. 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 $\theta_w$. Because of the form of Eq. (48), this is not costly.

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 $\theta_p$. The conditional probability for the $\theta_p$, fixing everything else, is given by

$$\log P(\theta_p|y, \theta_w, w, \text{GP}) =$$

$$-\frac{1}{2}w^T \Sigma^{-1} w - \frac{1}{2} \log \det \Sigma - \log p(\theta_p),$$

(49)

where, again, $\Sigma^{-1}$ 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 $\rho_\mu$ on the diagonal of $\Sigma$ for each modeled frequency. Each $\rho_\mu$ applies to a cosine and a sine mode; we will denote their weights as $a_\mu$ and $b_\mu$. If we adopt $1/\rho_\mu$ Jeffreys priors for each $\rho_\mu$, Eq. (49) becomes fully separable, and we can write

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

(50)

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

$$\rho_\mu,\text{new} = \frac{\tau}{\rho_\mu,\text{max} - \log (1 - \eta)}. $$

(51)

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

This scheme is analog to augmented/missing-data methods used in machine learning [65]: 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 dataset, 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-spindown 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 noncorrelation helps chain mixing, because it means that each pair of $(a_\mu, b_\mu)$ interacts (and correlates) with a single $\rho_\mu$ that is updated in a different block.

However, if we apply 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 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 run, 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 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 $\theta_p$ from a perturbative proposal; b) we then generate new quadratic parameters $w_p$ directly from their conditional posterior given the new $\theta_p$; finally c)
And when not accurately modeling the lowest DM variation frequencies, also the first and second time derivatives of the DM.

Thus we are just taking a Metropolis–Hastings step over the \( \theta_p \) 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 step 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 spindown in the correlated-noise block, and the DM parameter 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).

\[
\log p(\theta_p|y, \ldots) = -\frac{1}{2}y^T \left( N^{-1} - N^{-1} \Phi (\Phi^T N^{-1} \Phi + \Sigma^{-1})^{-1} \Phi^T N^{-1} \right) y - \frac{1}{2} \log |\Phi^T N^{-1} \Phi + \Sigma^{-1}| + \text{const.}
\]

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 \texttt{libstempo} interface to the \texttt{Tempo2} timing package. In Sec. VII A we compare the spectral-estimation quasi-Gibbs method of Sec. VII B with a standard MCMC method, applying both to a single-pulsar dataset. In Sec. VII B we compare the more general quasi-Gibbs method of Sec. VII C with again a standard MCMC method, applying both to a multipulsar dataset that contains a GW background.

A. Test of spectral-estimation quasi-Gibbs scheme

The single-pulsar mock dataset for this test is based on the timing model of pulsar J0437-4715 in the ATNF pulsar catalog. 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
MJD 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 dataset in 15 blocks, corresponding to different hardware, each with a different EFAC and EQUAD. The residuals for this mock dataset 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 [71]), 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 un marginalized 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.

### B. Test of collapsed quasi-Gibbs scheme

For this test we use mock data for an entire PTA: specifically, the second “open” dataset in the IPTA Mock Data Challenge [72], which consists of white radiometer (EFAC) noise in 36 pulsars (with different EFACs) plus a coherently injected GW background with $h_c(1\text{ yr}^{-1}) = 5 \times 10^{-14}$ and $\gamma = 4.33$.

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. [71], 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 $\gamma$), while the quasi-Gibbs scheme must deal with a multitude of extra parameters: $2 \times 36 \times 30 = 2,160$ frequency modes, and $36 \times \text{(average of 12)} = 441$ timing-model parameters, for a whopping total of 2,639.

The autocorrelation functions of the two MCMC chains are shown in Fig. 1. 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.

![Mock pulsar 1](mock_pulsar.png)

![Power spectrum](power_spectrum.png)

**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.
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 dataset. The legends show the autocorrelation lengths, which were typically 400–1,000 shorter with the quasi-Gibbs sampler.

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, jitter-likenoise, 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.

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 datasets such as those assembled by the International Pulsar Timing Array. Their analysis should include full pulsar noise models as in Arzoumanian et al. [71], 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 datasets, where we parametrize the power spectrum by independent variance parameters at frequencies multiples of \( 1/T \), with \( T \) the length of the dataset. Currently this is done in practice for few frequencies (up to \( \sim 20 \) [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 param-
FIG. 4. Similar to Fig. 3 for the adaptive-MCMC chain (dashed) and collapsed quasi-Gibbs chain (solid), both run on the multipulsar dataset from the first IPTA Mock Data Challenge. We plot the autocorrelation functions for four EFAC parameters, and for the two GWB parameters; the legends show the autocorrelation lengths, which are always shorter for the Gibbs sampler, although not as dramatically as in Fig. 3.

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 full IPTA-sized multipulsar datasets. 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 IPTA-sized datasets: we look forward to actually tackling them in their full glory.

Acknowledgments. We are grateful to many NANOGrav and EPTA colleagues for helpful discussions, and to Scott Ransom for use of the NRAO Nimrod cluster, where our tests were run. RvH is supported by NASA Einstein Fellowship grant PF3-140116. MV was 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. Copyright 2014 California Institute of Technology. Government sponsorship 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. [35]) is established by the following derivation, which is implied but not shown in Ref. [35]. 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 $F \Sigma V^*$ [with dimensions $(n \times p) \times (p \times p) \times (p \times p)$], where $U = [F \ G]$. 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

\[ |M^T C^{-1} M| = |V \hat{\Sigma} F^T C^{-1} F \hat{\Sigma}^*| = |\hat{\Sigma} F^T C^{-1} F \hat{\Sigma}| = |\hat{\Sigma}|^2 |F^T C^{-1} F| \]  

(A1)

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

\[ |C| = |U^T C U| = |G^T C G|(F^T C^{-1} F)^{-1} = |G^T C G|/[F^T C^{-1} F], \]  

(A2)

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

\[
U^T C U = \begin{pmatrix} G^T C G & G^T C F \\ F^T C G & F^T C F \end{pmatrix} = \begin{pmatrix} G^T C G & 0 \\ F^T C G & I \end{pmatrix} \begin{pmatrix} I & (G^T C G)^{-1} G^T C F \\ 0 & F^T C F - F^T C G (G^T C G)^{-1} G^T C F \end{pmatrix} = \begin{pmatrix} G^T C G & 0 \\ F^T C G & I \end{pmatrix} \begin{pmatrix} I & (G^T C G)^{-1} G^T C F \\ 0 & (F^T C^{-1} F)^{-1} \end{pmatrix}. \]

(A3)

Thus the normalization factor of Eq. (22) is given by \( \sqrt{(2\pi)^n |\Sigma|^2 |G^T C G|} \); we may drop \( |\hat{\Sigma}| \), which is essentially arbitrary (it is the Jacobian of the coordinate transformation \( \eta' = (F^T F)^{-1} F^T M \eta \), while we take infinitely vague priors for these parameters), and adjust the \( 2\pi \) exponent to match the \( n - p \) dimension of \( G^T C G \).

Moving on to the main quadratic expression in Eq. (22), we rewrite \( C' = C^{-1} - C^{-1} M (M^T C^{-1} M)^{-1} M^T C^{-1} \) as

\[
C' = C^{-1} - C^{-1} (F \hat{\Sigma} V^*) (V \hat{\Sigma} F^T C^{-1} F \hat{\Sigma} V^*)^{-1} (V \hat{\Sigma} F^T) C^{-1} = C^{-1} - C^{-1} F (F^T C^{-1} F)^{-1} F^T C^{-1} \]

(A4)

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

\[
C' (F^T y) = (C^{-1} F - C^{-1} F (F^T C^{-1} F)^{-1} F^T C^{-1} F) (F^T y) = (C^{-1} F - C^{-1} F) (F^T y) = 0. \]

(A5)

We are then left with

\[
y^T C' y = (G^T y)^T G^T C' G (G^T y) = (G^T y)^T (G^T C^{-1} G - G^T C^{-1} F (F^T C^{-1} F)^{-1} F^T C^{-1} G) (G^T y) = (G^T y)^T (G^T C G)^{-1} (G^T y), \]

(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 [64], 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 \( \hat{y} = H y \), 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 the \( \phi_p \) 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. \( y^T C^{-1} y \) 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
\[ y^T C^{-1} y = y^T (D + USU^T)^{-1} \approx y^T H (H^T DH + H^T USU^TH)^{-1} H^T y + y^T H_c (H_c^T D H_c)^{-1} H_c y. \]  

(B1)

Here \( H \) and \( H_c \) are matrices with the properties \( H^T H = 0 \), and \( HH^T + H_c H_c^T = I \), 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:

\[
H_c^T U = 0, \quad HH^T U = U, \quad H_c^T D H = 0 \quad (B2)
\]

The authors 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 \( W \Sigma V^* = G^T U U^T G \). Here \( H \) consists of the first \( l \) columns of \( W \), with \( l \) the number of non-singular values in \( \Sigma \).

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 \( O(t^3) \) inversion of \( H^T CH \), or the \( O(t^2) \) operation of the multiplication \( H^T C \).

Our presentation of data compression differs from the “ABC method” originally presented by van Haasteren \[64\], which did not include both terms of the separated likelihood function. By using only the data \( H^T y \) and \( H^T y \), 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 \( H^T D H \neq 0 \). We do note that, even when our likelihood function is not fully separable, Eq. (B1) represents a 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 gone 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.

\[ 11 \text{ This is typically the number of columns in } U, \text{ but it can be smaller (numerically) when the timing basis is sufficiently close to the basis in } U. \]
[63] L. Trefethen and D. Bau, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997).

[64] R. van Haasteren, *MNRAS* **429**, 55 (2013), arXiv:1210.0584 [astro-ph.IM].

[65] J. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, 2001).

[66] G. Roberts, A. Gelman, and W. Gilks, *Ann. of Appl. Prob.* **7**, 110 (1997).

[67] 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, 1992).

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

[69] M. Vallisneri, “*libstempo, a python wrapper for TEMPO2*,” (2012), https://github.com/vallis/libstempo.

[70] R. N. Manchester et al., “ATNF Pulsar Catalog,” http://www.atnf.csiro.au/research/pulsar/psrcat.

[71] Z. Arzoumanian, A. Brazier, S. Burke-Spolaor, S. J. Chamberlin, S. Chatterjee, J. M. Cordes, P. B. Demorest, X. Deng, T. Dolch, J. A. Ellis, R. D. Ferdman, N. Garver-Daniels, F. Jenet, G. Jones, V. M. Kaspi, M. Koop, M. Lam, T. J. W. Lazio, A. N. Lommen, D. R. Lorimer, J. Luo, R. S. Lynch, D. R. Madison, M. McLaughlin, S. T. McWilliams, D. J. Nice, N. Palliyaguru, T. T. Pennucci, S. M. Ransom, A. Sesana, X. Siemens, I. H. Stairs, D. R. Stinebring, K. Stovall, J. Swiggum, M. Vallisneri, R. van Haasteren, Y. Wang, and W. W. Zhu, ArXiv e-prints (2014), arXiv:1404.1267.

[72] M. J. Keith, K. J. Lee, and F. Jenet, in preparation.