# Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. I. The optimal statistic

Michele Vallisneri,[1,2,*] Patrick M. Meyers,[2,†] Katerina Chatziioannou,[3,2,‡] and Alvin J. K. Chua[4,5,§]

[1]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA*
[2]*Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*
[3]*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[4]*Department of Physics, National University of Singapore, Singapore 117551*
[5]*Department of Mathematics, National University of Singapore, Singapore 119076*

(Received 11 July 2023; accepted 4 October 2023; published 6 December 2023)

A gravitational-wave background can be detected in pulsar-timing-array data as Hellings-Downs correlations among the timing residuals measured for different pulsars. The *optimal statistic* implements this concept as a classical null-hypothesis statistical test: a null model with no correlations can be rejected if the observed value of the statistic is very unlikely under that model. To address the dependence of the statistic on the uncertain pulsar noise parameters, the pulsar-timing-array community has adopted a hybrid classical-Bayesian scheme [S. J. Vigeland *et al.*, Phys. Rev. D **98**, 044003 (2018).] in which the posterior distribution of the noise parameters induces a posterior distribution for the statistic. In this article we propose a rigorous interpretation of the hybrid scheme as an instance of posterior predictive checking, and we introduce a new summary statistic (the Bayesian signal-to-noise ratio) that should be used to accurately quantify the statistical significance of an observation instead of the mean posterior signal-to-noise ratio, which does not support such a direct interpretation. In addition to falsifying the no-correlation hypothesis, the Bayesian signal-to-noise ratio can also provide evidence supporting the presence of Hellings-Downs correlations. We demonstrate our proposal with simulated datasets based on NANOGrav's 12.5-yr data release. We also establish a relation between the posterior distribution of the statistic and the Bayes factor in favor of correlations, thus calibrating the Bayes factor in terms of hypothesis-testing significance.

DOI: 10.1103/PhysRevD.108.123007

## I. INTRODUCTION AND SUMMARY

In June 2023, four international pulsar-timing-array (PTA) efforts (NANOGrav, EPTA, PPTA, and CPTA) reported compelling evidence [1–4] for the existence of a low-frequency gravitational-wave background (GWB), as expected from the binaries of supermassive black holes at galactic centers [5–8], although more exotic sources are also possible [6,9]. These results had been prefigured by findings of excess timing noise [10–13] with spectra that were consistent across pulsars and plausible with respect to astronomical expectations [14]. These findings were suggestive but by no means conclusive, since the excess noise may have arisen from sources other than the GWB, such as intrinsic pulsar spin noise [15,16].

Indeed, in Refs. [1–4] evidence for the GWB was established using a different criterion: the presence of specific correlations between the time series of residuals of different pulsars. For an isotropic GWB, these correlations follow a geometric relation (a function of the angle $\zeta$ between the pulsars) first derived by Hellings and Downs [17]. While other systematic effects (such as clock and ephemeris errors [18,19]) may generate different angular patterns, it is difficult to imagine an explanation other than the GWB for a manifest Hellings-Downs pattern in the data.

The problem of GWB detection with PTAs data then turns into the probabilistic characterization of observed inter-pulsar correlations. Bayesian approaches have been the tool of choice, since they allow for a principled treatment of all unknown variables needed to fully describe the data, such as the geometric and kinematic parameters of the pulsars, the levels and spectral shapes of radiometric noise, intrinsic spin noise, and dispersion-measure noise from the interstellar medium, and of course the GWB parameters (see Refs. [20,21] and references therein). In the dominant Bayesian approach, the statistical evidence in favor of Hellings-Downs correlations is quantified as the Bayes factor between two data models

*michele.vallisneri@jpl.nasa.gov
†pmeyers@caltech.edu
‡kchatziioannou@caltech.edu
§alvincjk@nus.edu.sg

 

that include all the elements listed above, and that are identical *except* for the inter-pulsar correlation coefficients of the shared common-spectrum process, which are set to zero for the null model (common-spectrum uncorrelated red noise, or CURN) and to the Hellings-Downs pattern for the isotropic-GWB model (HD). Thus, this model comparison begins with the well-established finding of excess timing-residual power, and attempts to attribute its origin to either independent processes in each pulsar, or to the phase-coherent delays induced by the GWB. See Refs. [10,22] for examples of this analysis. A complementary Bayesian strategy based on posterior predictive model checking [23] is discussed in a companion to this paper [24].

By contrast, classical frequentist statistics offers an attractive formulation of GWB detection based on *null hypothesis statistical testing* [25], as used in LIGO's historical black-hole binary detection [26]. For GWB searches with PTAs, null hypothesis testing is implemented with the *optimal statistic* [27–29], which was devised as a direct measure of Hellings-Downs correlations. The basic idea is that if we observe a value of the optimal statistic much larger than expected under the null hypothesis of no inter-pulsar correlations, we can reject that scenario and conclude that correlations are present. The $p$-value (the probability of obtaining the observed value of the optimal statistic, or larger, under the null hypothesis) quantifies the statistical significance of this conclusion.

Unfortunately, this simple test cannot be implemented in practice because the optimal statistic depends parametrically on the unknown pulsar noise parameters. Vigeland *et al.* [29] introduced a hybrid scheme in which the posterior distribution of pulsar noise parameters (obtained from Bayesian inference) induces a distribution of the observed optimal statistic, and proposed that the posterior mean of the signal-to-noise ratio (SNR)—i.e., of the optimal statistic divided by its standard deviation across noise realizations—could be used as a measure of statistical significance. However, the mean posterior SNR does not correspond to a $p$-value for the optimal statistic, so its hypothesis-testing interpretation is questionable. Partly because of this, and because of the perception that Bayesian model comparison accounts more fully for the uncertainties in the data model, the optimal statistic has been relegated to a secondary role in PTA searches for the GWB.

In this article we submit that the optimal statistic should remain a tool of first choice in these searches, on par with Bayesian model comparison. Specifically, in Sec. II we show that the hybrid optimal statistic can be interpreted rigorously in the framework of posterior predictive model checking [23], leading to a self-consistent Bayesian generalization of statistical testing that can falsify the no-correlation hypothesis while accounting for the uncertainty in pulsar noise parameters. We introduce a new detection statistic (the *Bayesian SNR*, henceforth BSNR) that maps to a single well-defined $p$-value and provides a direct measure of statistical significance. In Sec. III we demonstrate this scheme by examining the distribution of hybrid SNRs and BSNR in two sets of simulated datasets (with and without a loud GWB injection). In Sec. IV we discuss the role that the BSNR can play in discriminating a true Hellings-Downs-correlated signal from spurious processes with other spatial correlations. In Sec. V we exploit the fact that the (squared) optimal statistic approximates the log ratio of the HD and CURN likelihoods to formulate detection statements based on the distribution of the posterior log likelihood ratio (PLLR, [30,31]). We show that the PLLR is related to the HD-vs.-CURN Bayes factor, providing a frequentist calibration for its value, and confirming that the commonly used $\exp(\mathrm{SNR}^2/2)$ heuristic for the Bayes factor can overpredict its value. In Sec. VI we present our conclusions.

## II. NULL HYPOTHESIS TESTS AND THE BAYESIAN $p$-VALUE

Classical null hypothesis testing [25] proceeds by assigning a function $D$ of the data $y$ to serve as a *test statistic*, and then rejecting the null hypothesis $H_0$ when the observed value $D_{\mathrm{obs}} = D(y_{\mathrm{obs}})$ is in the extreme tail of its background distribution

$$p(D|H_0) = \int p(D|y)p(y|H_0)\mathrm{d}y; \tag{1}$$

that is, when it is very unlikely that $H_0$ could produce data that result in $D_{\mathrm{obs}}$. The tail area $P(D > D_{\mathrm{obs}}|H_0)$ (or $P(D < D_{\mathrm{obs}}|H_0)$ as appropriate) is known as the one-sided $p$-value. Just how small a $p$-value should justify the rejection of the null hypothesis depends on extra-statistical considerations, and has been the subject of considerable debate. Crucially, the $p$-value quantifies the probability that $H_0$ would generate the observed data, and *not* the probability that $H_0$ is true given the data, which depends on the *base rate* of $H_0$ in "similar" experiments.

In the PTA context, the optimal statistic is used as follows to implement null hypothesis testing:

(i) we construct the optimal statistic (which we will again denote as $D$) specifically to quantify the strength of Hellings-Downs correlations in the data; by design, the optimal statistic embodies a fiducial GWB spectral shape (but not its overall amplitude), as well as a fiducial noise model for each pulsar;

(ii) we compute the background distribution $p(D|H_0)$ under the null hypothesis $H_0$ (≡CURN) that the GWB is not present (i.e., a CURN signal appears with the same spectrum across the array,

but its realizations in different pulsars are spatially uncorrelated);

(iii) we obtain the observed optimal statistic $D_{\rm obs} = D(y_{\rm obs})$, and reject $H_0$ if the $p$-value $\int P(D > D_{\rm obs}|H_0)\mathrm{d}D$ is sufficiently small.

As anticipated above, moving beyond the rejection of CURN and claiming GWB detection would require additional lines of evidence to conclude that HD is indeed the best explanation for the data.

The "classic" optimal statistic is formulated as

$$D(y) = \frac{\sum_{i\neq j} y_i^T C_i^{-1}\tilde{\Gamma}_{ij}C_j^{-1}y_j}{\sum_{i\neq j}\mathrm{tr}[C_i^{-1}\tilde{\Gamma}_{ij}C_j^{-1}\tilde{\Gamma}_{ji}]}, \tag{2}$$

where the sum is over all pairs of pulsars in the array; $y_k$ is the vector of residuals for pulsar $k$; $C_k$ is the covariance matrix for those residuals, including measurement noise, intrinsic noise, timing-model errors, and common red noise; and $\tilde{\Gamma}_{ij} = \Sigma(t_i, t_j) \times f_{\rm HD}(\zeta_{ij})$ is the correlation matrix for GWB-induced residuals in pulsars $i$ and $j$. Here $\Sigma(t_i, t_j)$ is set by the spectral content of the GWB and it is normalized so that in ensemble average $\langle y_{i,\rm gw} y_{j,\rm gw}^T\rangle = A_{\rm gw}^2\tilde{\Gamma}_{ij}$ for a GWB of amplitude $A_{\rm gw}$; while $f_{\rm HD}(\zeta_{ij})$ is the Hellings-Downs function of the pulsar-pair angle $\zeta_{ij}$ [17]. See Ref. [32] for details about the Gaussian-process formulation of the PTA likelihood.

Under $H_0$, $D(y)$ follows a generalized $\chi^2$ distribution [33] with an expectation value of zero. In the presence of the GWB, $D(y)$ follows a noncentral generalized $\chi^2$ distribution [34,35], and its expectation value is $\langle D(y)\rangle = A_{\rm gw}^2$. Thus, $D(y)$ may serve as an estimator of $A_{\rm gw}$. Both distributions have been approximated as normal in the optimal-statistic literature:

$$p(D(y)|H_0) \simeq \mathcal{N}(0, \sigma_0^2),$$
$$\sigma_0^2 = \left(\sum_{i\neq j}\mathrm{tr}[C_i^{-1}\tilde{\Gamma}_{ij}C_j^{-1}\tilde{\Gamma}_{jk}]\right)^{-1}, \tag{3}$$

and

$$p(D(y)|A_{\rm gw}) \simeq \mathcal{N}(A_{\rm gw}^2, \sigma_0^2). \tag{4}$$

For simplicity, we will use these approximations throughout this article, but we caution that their exact distributions [33–35] should be used to interpret optimal-statistic results with real data: this is especially important because $p(D(y)|H_0)$ has more substantial tails than the normal distribution, and $p(D(y)|A_{\rm gw})$ has variance larger than $\sigma_0^2$ because of GWB-induced correlations between the summands of Eq. (2).

The $p$-value for $D_{\rm obs} \equiv D(y_{\rm obs})$ is $1 - \mathrm{CDF}(D_{\rm obs}|H_0) \equiv \int P(D > D_{\rm obs}|H_0)\mathrm{d}D$. If the normal approximation is taken at face value, it follows from Eq. (3) that the combination $\mathrm{SNR} \equiv D(y)/\sigma_0$ can be interpreted as a *signal-to-noise ratio* for null hypothesis testing, so the $p$-value is $\mathrm{erfc}(\mathrm{SNR}/\sqrt{2})/2$, where erfc is the complementary error function (e.g., $p = 0.02, 1.3\times10^{-3}, 3.2\times10^{-5}, 2.9\times10^{-7}$ for SNR $= 2, 3, 4, 5$ respectively). While this definition of SNR is used broadly in the optimal-statistic literature, it is important to remember that because of the normal approximation an observed SNR of X does not actually imply "X$\sigma$" significance.

This optimal-statistic $p$-value refers to the *assumed* population $p(y|\theta_0, H_0)$ of datasets that are generated under the null hypothesis $H_0$ with the assumed pulsar noise parameters $\theta_0$, which enter $D(y)$ through the $C_k$. We will write $D(y;\theta_0)$ to emphasize this dependence. When analyzing real data, we face the problem that the noise parameters $\theta$ must themselves be estimated from the data. The simplest approach is setting $\theta$ to their maximum-likelihood or maximum *a posteriori* values $\hat{\theta}(y_{\rm obs})$. Unfortunately, the optimal statistic is very sensitive to pulsar noise assumptions, so fixing them in this way can distort hypothesis-testing conclusions. (It can also lead to biased $A_{\rm gw}$ estimates, but that is less of a concern for this article.) To address this problem, Vigeland *et al.* [29] suggested that we consider the distributions of $D(y_{\rm obs};\theta)$ and SNR induced by the Bayesian posterior $p(\theta|y_{\rm obs}, H_0)$. In this approach the SNR gains a notion of Bayesian uncertainty. hypothesis-testing "significance" is quoted as the mean posterior SNR:

$$\overline{\mathrm{SNR}} = \int \frac{D(y_{\rm obs};\theta)}{\sigma_0(\theta)} p(\theta|y_{\rm obs}, H_0)\mathrm{d}\theta. \tag{5}$$

Because the $p$-value is a nonlinear function of the SNR, this average cannot be mapped to the $p$-value of the optimal statistic with respect to any background population. In other words, in performing this marginalization we lose relevant information about the distribution of the optimal statistic, so $\overline{\mathrm{SNR}}$ is not a direct measure of hypothesis-testing significance.

There is however a straightforward way to build a statistically meaningful statistic from the posterior distribution of $D(y_{\rm obs};\theta)$: instead of marginalizing the SNR, we marginalize the $p$-value itself. In addition to making sense intuitively, this procedure admits a principled null-hypothesis-testing interpretation in terms of Bayesian model checking [36,37]. In this framework, the *Bayesian p-value* of the observed data is computed over the population of conditional *data replications* generated under $H_0$ with noise parameters $\theta \sim p(\theta|y_{\rm obs}, H_0)$:

$$p_{\rm B}(y_{\rm obs}) \equiv \int P\left[D\left(y_{\rm rep}^{(\theta)};\theta\right) > D(y_{\rm obs};\theta)\right] \times p(\theta|y_{\rm obs}, H_0)\mathrm{d}\theta. \tag{6}$$

where $y_{\rm rep}^{(\theta)} \sim p(y_{\rm rep}^{(\theta)}|\theta, H_0)$. If we assume that $D(y_{\rm obs};\theta)$ is normally distributed, as in Eq. (3), we obtain

$$p_{\rm B}(y_{\rm obs}) \simeq \int \frac{1}{2}\mathrm{erfc}(\mathrm{SNR}(y_{\rm obs};\theta)/\sqrt{2})\, p(\theta|y_{\rm obs}, H_0)\mathrm{d}\theta. \tag{7}$$

The Bayesian $p$-value characterizes how often the null model, *conditioned on the data*, would result in the values of the statistic that we observe—concretely, how often intrinsic pulsar noise with parameter distributions inferred from $y_{\rm obs}$ would yield the spatial correlations that we measure. We propose that $p_B(y_{\rm obs})$ should be used to establish the presence of spatial correlations in PTA data by rejecting the null model in hypothesis testing. We may also map $p_B(y_{\rm obs})$ to an effective *Bayesian SNR*,

$$\mathrm{BSNR} \equiv \sqrt{2}\ \mathrm{erfc}^{-1}(2p_{\rm Bayes}), \tag{8}$$

thus associating a Gaussian $\sigma$ level with the rejection of $H_0$. Note that the erfc appears by *definition* in Eq. (8), but is only justified under the normal approximation in Eq. (7).

The BSNR is skewed toward the lower tail of the posterior $D(y_{\rm obs};\theta)$ distribution, because smaller SNRs yield much greater $p$-values, which dominate Eq. (6) and therefore Eq. (8). Qualitatively, we are averaging a risk (that CURN should yield extreme data), so we pay the most attention to the riskiest pulsar-noise configurations (those that minimize observed correlations). By contrast, $\overline{\mathrm{SNR}}$ is a direct average of SNR, so it overemphasizes the highest SNRs, which however provide little probability mass to the $p$-value. Thus, BSNR should be quoted instead of $\overline{\mathrm{SNR}}$ as the measure of null-hypothesis-testing significance.

If we do want to use $\overline{\mathrm{SNR}}$, we need to turn to a different scheme. As for any statistic, the significance of an observed $\overline{\mathrm{SNR}}$ can be obtained *empirically* by sampling its distribution over a relevant population, such as simulations or "bootstrapped" data models in which inter-pulsar correlations are masked by randomizing the phases of red-noise Fourier components (*phase shifts*, [38]) or by randomly assigning pulsar positions when computing the Hellings-Downs function (*sky scrambles*, [38,39]).

These options are correct both formally and substantially, but they answer different questions about the data (because they reference different background populations), so they provide information complementary, but not equivalent, to the BSNR. Specifically, with simulations we ask how often we would observe a given $\overline{\mathrm{SNR}}$ (or a larger value) over the simulated $H_0$ population, which could have $\theta$ fixed to $\mathrm{argmax}_\theta p(\theta|y_{\rm obs}, H_0)$, or distributed with that posterior. This option is very costly because a new posterior $p(\theta|y_{\rm sim}, H_0)$ must be obtained for every simulation. For sky scrambles [38,39], we randomize pulsar sky positions and ask how often the observed pulsar-pair correlations would happen to conform to the resulting Hellings-Downs patterns (i.e., we explore whether the observed $\overline{\mathrm{SNR}}$ is the product of a lucky sky configuration). For phase shifts [38], we ask how often intrinsic red noise with the observed spectral amplitudes, but with random phases, would produce the observed $\overline{\mathrm{SNR}}$.

## III. SIMULATIONS: SNR DISTRIBUTIONS AND BAYESIAN P-VALUES

In this section we obtain the posterior SNR distribution and compute the BSNR for a set of simulated datasets created to resemble the 12.5-yr NANOGrav data release [40], to get a sense of what we should expect for undetectable and detectable GWBs, and to understand how the BSNR summarizes the distributions. All simulations comprise the 47 pulsars in the release, "observed" at the same TOAs, but with residuals drawn randomly according to *maximum-a-posteriori* noise hyperparameters $\theta_{\rm sim}$ determined from the real dataset. See the Appendix for technical details about our simulations and Bayesian inference.

Figure 1 shows the posterior distribution of $\mathrm{SNR}(y_{\rm obs};\theta)$ induced by $p(\theta|y_{\rm obs}, \mathrm{CURN})$ in two representative simulations: the first (top) with no injection of a GWB or any other common noise, and the second (bottom) with a loud power-law GWB with amplitude $A_{\rm gw} = 10^{-14}$ and spectral slope $\gamma = 13/3$ (quoted according to the conventions of [10]). The optimal statistic is also built under the assumption of a $\gamma = 13/3$ GWB spectrum. The vertical bars show $\overline{\mathrm{SNR}}$, BSNR, the maximum-likelihood $\mathrm{SNR}(y_{\rm obs};\hat{\theta})$, and the *true* $\mathrm{SNR}(y_{\rm obs};\theta_{\rm sim})$, which of course is not accessible for real data.

The no-injection dataset (top panel) produces a posterior SNR distribution consistent with the null hypothesis, as expected. The four SNR statistics cluster closely. By contrast, the loud-injection dataset (bottom panel) produces a posterior SNR distribution that is inconsistent with the null hypothesis, with mean and mode approximately $3.5\sigma$ from zero, close to the true SNR. The maximum-likelihood SNR is significantly higher at 4.6, and the Bayesian SNR is lower at 2.9. In other words, given the uncertainty in the determination of the noise parameters, we only reach a $p$-value $\simeq 10^{-3}$ ($\simeq 3\sigma$) rather than the more significant $p$-values incorrectly implied by $\overline{\mathrm{SNR}}$ and $\mathrm{SNR}(y_{\rm obs};\hat{\theta})$.

Figure 2 shows posterior optimal-statistic SNR distributions for 66 12.5-yr-like datasets with no GWB or common noise (left, blue), and 69 12.5-yr-like datasets with loud $A_{\rm gw} = 10^{-14}$ GWB injections (right, orange). (The peculiar numbers of datasets resulted from a fraction of simulations failing on our computing cluster because of out-of-memory errors.) The dots mark BSNR values, which are used to sort the simulations. Figure 3 shows histograms of $\overline{\mathrm{SNR}}$ and BSNR across the simulations.

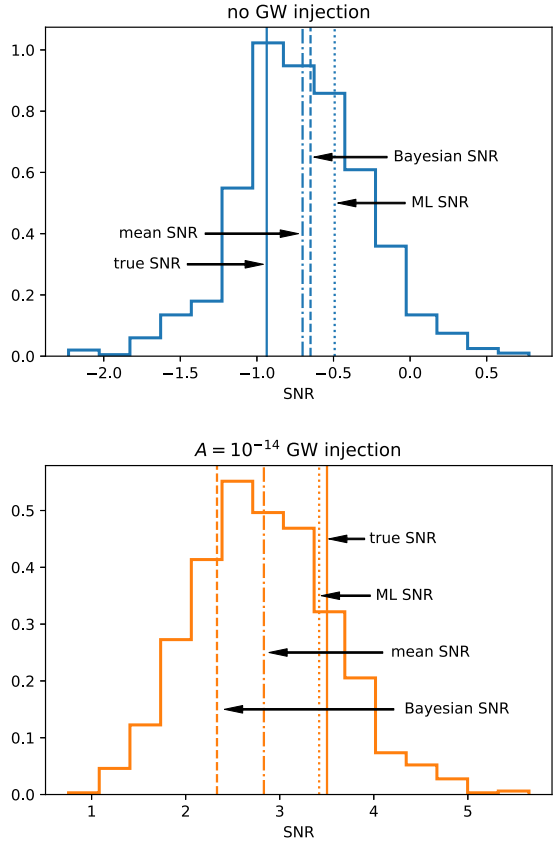


FIG. 1. Optimal-statistic SNR posterior distributions for two simulated datasets with no common-process injection (top) and with an $A_{\rm gw} = 10^{-14}$ GWB injection (bottom). The no-injection SNR distribution is consistent with the null hypothesis, and the four SNR statistics are close in value. The loud-injection SNR distribution is inconsistent with the null hypothesis, with BSNR $\simeq 3$ and therefore average $p$-value $\simeq 10^{-3}$, significantly lower than implied by the $\overline{\rm SNR}$ and (especially so) by ${\rm SNR}(y_{\rm obs}; \hat{\theta})$.

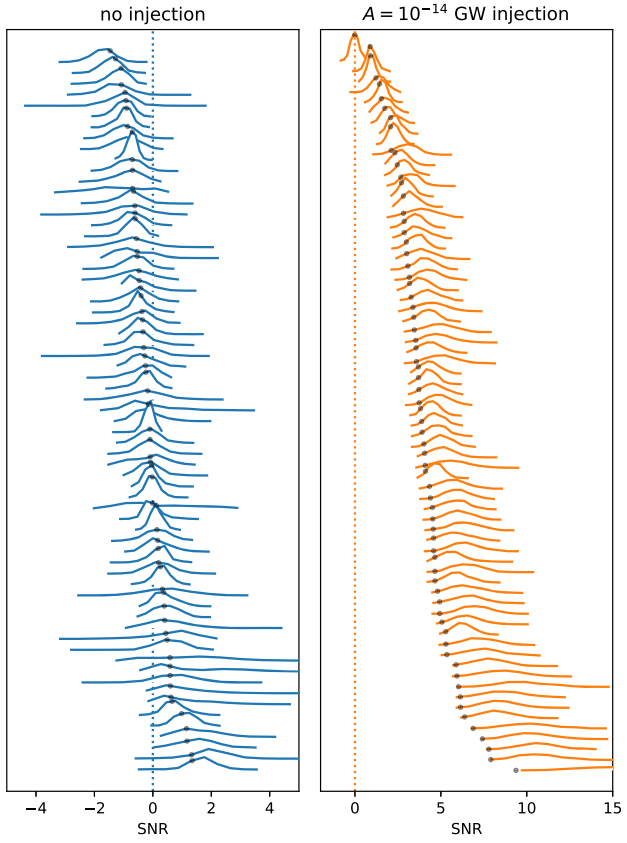


FIG. 2. Posterior distributions of optimal-statistic SNR for 66 simulated datasets with no GWB (left, blue) and 69 datasets with loud GWB injections (right, orange). The dots show Bayesian SNRs [Eqs. (6) and (8)]. Simulations are sorted according to optimal-statistic SNR.

In the no-injection datasets, SNRs cluster around zero as expected, with a few tails extending to SNRs $\simeq 4$. BSNRs are usually close to $\overline{\rm SNR}$s. The blue histograms in Fig. 3 show the ensemble distributions of $\overline{\rm SNR}$ and BSNR across simulations, which are concentrated around zero. There is no expectation that BSNRs (or $\overline{\rm SNR}$s) would be distributed normally [41].

In the loud-injection datasets, SNRs are broadly distributed between 0 and 15, with many very convincing detections but also a few false dismissals straddling zero. (These may occur when intrinsic red noise and the GWB pulsar terms, which contribute half of the GWB variance for each pulsar, conspire to obscure the GWB correlations.) As anticipated in the discussion of Eq. (6), the BSNR sits on the lower tail of the SNR distributions (except for the false negatives, in which the tails do not represent extreme $p$-values).

Across simulations (Fig. 3, orange histograms on the right), BSNRs cluster around 4, while $\overline{\rm SNR}$s cluster around 5 with similar tails. Thus, our admittedly small sample suggests that in approximately half of the realizations of a 12.5-yr-like dataset, the null CURN hypothesis could be falsified with 4-$\sigma$ Bayesian $p$-value significance, but that the test would be inconclusive ($< 3$-$\sigma$) in a quarter of

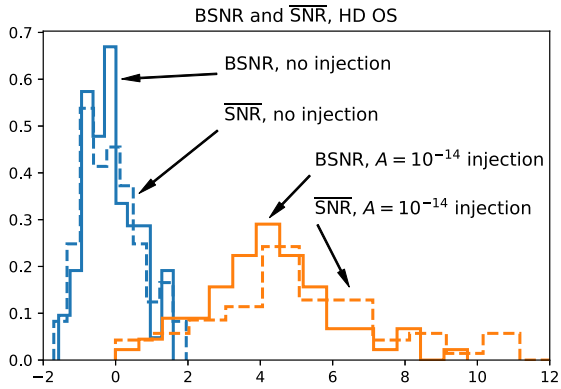


FIG. 3. Distributions of optimal-statistic BSNR (solid) and $\overline{\rm SNR}$ (dashed) over 66 simulated datasets with no GWB (left, blue) and 69 datasets with loud GWB injections (right, orange).

realizations. Incorrectly taking $\overline{\text{SNR}}$s at face value would have overstated significance obviously, if not dramatically, since $\overline{\text{SNR}} > 4$ in 70% of realizations, and $< 3$ in 17% of realizations.

Comparing these significance levels with those measured on "bootstrapped" background populations is computationally difficult for many of our loud-injection simulations, which have very high significance levels and therefore would require very large backgrounds. Taking as example one simulation that results in moderate BSNR of 1.9 and $\overline{\text{SNR}}$ of 2.13, we perform 10,000 sky scrambles [38,39] and 10,000 phase shifts [38], and find a sky-scramble (phase-shift) significance of 1.9-$\sigma$ (2.4-$\sigma$) for $\overline{\text{SNR}}$. Thus in this case BSNR and sky-scrambled $\overline{\text{SNR}}$ agree. Phase-shifted backgrounds embody less variation, because they effectively fix the amplitudes of red-noise Fourier components and vary only their relative phases, it is reasonable that the phase-shift significance would skew higher.

## IV. INTERPRETING THE OPTIMAL STATISTIC FOR ALTERNATIVE CORRELATION FUNCTIONS

GWB searches in PTA data must account for systematic error sources such as long-term oscillations in the time standard [42] and errors in the Solar System ephemerides [19], both of which are used to refer telescope observations to an inertial frame at rest with respect to the Solar System barycenter. These errors create timing residuals that are correlated across pulsars, albeit with non-Hellings–Downs geometry: $\langle \delta y_i \delta y_j^T \rangle \propto \tilde{\Gamma}_{ij}^{\text{clk}} = 1 \times \Sigma(t_i, t_j)$ for clock errors and $\propto \tilde{\Gamma}_{ij}^{\text{ephem}} = \cos\zeta_{ij} \times \Sigma(t_i, t_j)$ for ephemeris errors [cf. the definition of $\zeta_{ij}$ below Eq. (2)]. PTA analysts refer to these systematic errors as *monopole* and *dipole*, respectively, with reference to the angular dependence of the interpulsar correlations that they embody. By replacing $\tilde{\Gamma}_{ij}$ with $\tilde{\Gamma}_{ij}^{\text{clk}}$ or $\tilde{\Gamma}_{ij}^{\text{ephem}}$ in Eqs. (2) and (3), we obtain optimal-statistic variants $D^{\text{clk}}$ and $D^{\text{ephem}}$ that target monopolar and dipolar correlations, and that have been used to diagnose the presence of correlated systematics in PTA datasets, e.g., [10]. In this section we focus on the monopole optimal statistic, examine its distribution in our simulated datasets, and discuss how Hellings-Downs and monopole signals could be distinguished using the optimal statistic.

Figure 4 shows posterior SNR distributions (curves) and Bayesian SNRs (dots) for the monopole optimal statistic in the same no-injection and loud-GWB-injection simulations as Sec. III. Datasets are still sorted according to the Hellings-Downs BSNR, and Hellings-Downs SNR distributions are overdrawn more faintly for comparison. Like the standard optimal statistic, the monopole optimal statistic is built under the assumption of a $\gamma = 13/3$ power law.

In the no-injection datasets, monopole SNRs cluster around zero in a manner similar to the Hellings-Downs

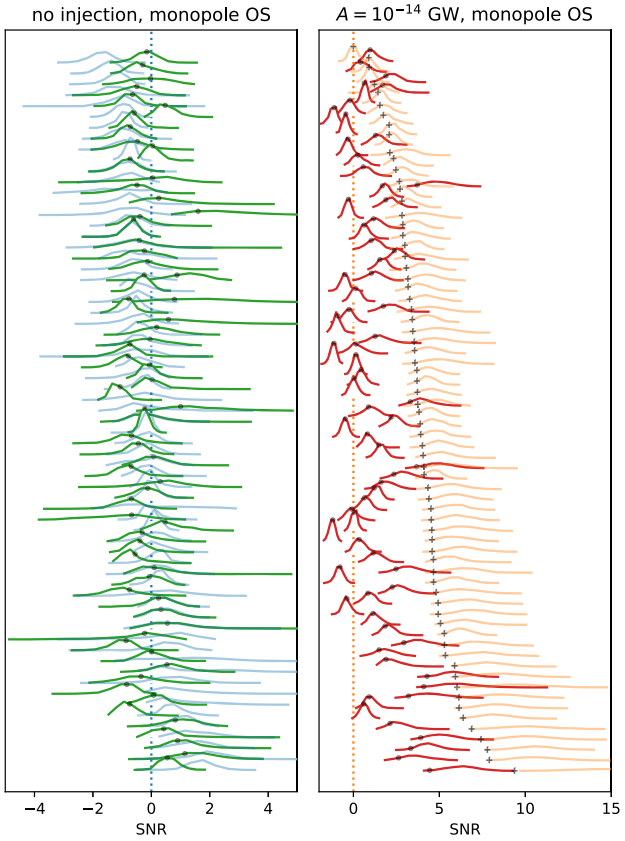


FIG. 4. Posterior distributions of *monopole* optimal-statistic SNR for 66 simulated datasets with no GWB (left, green) and 69 datasets with loud GWB injections (right, red). Black dots show Bayesian SNRs. The fainter blue and orange curves show the corresponding Hellings-Downs SNR distributions, identical to Fig. 2. Simulations are sorted according to the Hellings-Downs optimal-statistic SNR, with plus-shaped markers showing the Hellings-Downs optimal-statistic BSNR.

optimal statistic; their distributions across simulations are shown in Fig. 5. The monopole and Hellings-Downs SNRs are largely uncorrelated, with Pearson $r$ coefficients $\simeq 0.25$ for both $\overline{\text{SNR}}$s and BSNRs. Overall, the monopole optimal statistic fails (correctly) to reject the null hypothesis.

In the loud-injection datasets (right) monopole SNRs are distributed broadly, although not as much as the Hellings-Downs SNRs, and they extend from $-2$ to 10. Across simulations, monopole BSNRs average to 1.2 and monopole $\overline{\text{SNR}}$s to 1.5; only 4% of the former and 13% of the latter are greater than 4 (see Fig. 5). Monopole and Hellings-Downs SNRs are significantly correlated, with Pearson $r$ coefficients $\simeq 0.5$ for both $\overline{\text{SNR}}$s and Bayesian SNRs. Overall, the monopole optimal statistic is much less effective than the Hellings-Downs optimal statistic at rejecting the null hypothesis when a Hellings-Downs-correlated signal is present. Clearly that must be because the statistic encodes the wrong correlation pattern. Even so, the Hellings-Downs signal *can* excite the monopole

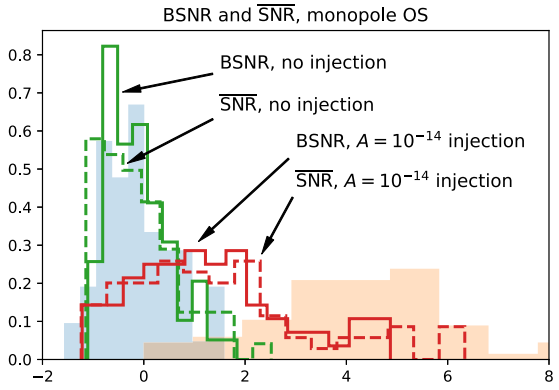


FIG. 5. Distributions of *monopole* optimal-statistic BSNR (solid) and $\overline{\text{SNR}}$ (dashed) over 66 simulated datasets with no GWB (leftmost, green) and 69 datasets with loud GWB injections (rightmost, red). The fainter blue and orange areas show the corresponding Hellings-Downs BSNR distributions from Fig. 3.

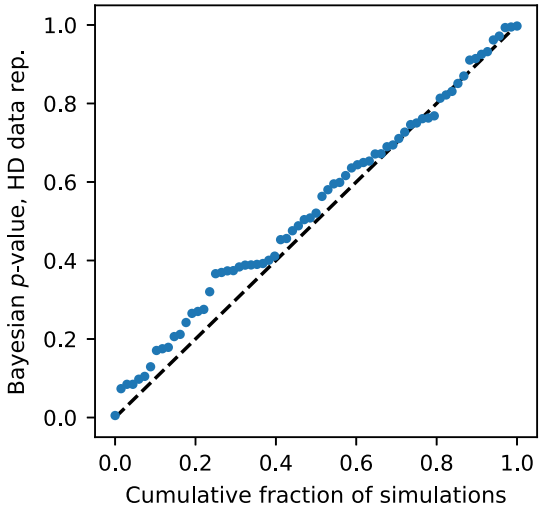


FIG. 6. Cumulative distribution of Bayesian $p$-values for $D(y_{\rm obs};\theta)$ under the HD hypothesis across 69 loud-injection simulations. As expected, the distribution is approximately uniform. The sampling error of each $p$-value is $\simeq 0.01$.

statistic, and in some cases (four simulations, or about 6%) it can even produce $\overline{\text{SNR}}$ and BSNR greater than their Hellings-Downs counterparts. However, none of these four simulations lead to convincing CURN rejections.

In other words, a Hellings-Downs signal can still be picked up by the monopole optimal statistic, typically (but not always) with suboptimal SNR. The converse is also true. Thus, a large monopole SNR does not by itself indicate that systematic residuals with clock-like correlations are present in the data. More generally, while it is tempting to compare the Hellings-Downs and monopole SNRs to determine which alternative hypothesis is favored by the data, that is not something we can do within null hypothesis testing or its Bayesian extension, in which $p$-values are always computed for the null hypothesis, and therefore carry no quantitative information about the alternatives.

In fact, that is the very restriction that we need to address in order to test alternative hypotheses within the optimal-statistic framework. For instance, having falsified CURN because we observed a high Hellings-Downs BSNR, we may now check the HD model by computing the Bayesian $p$-value of $D(y_{\rm obs};\theta)$ *under the HD hypothesis*. That is, we consider the distribution of the optimal statistic over HD data replications, modifying Eq. (6) by replacing $p(\theta|y_{\rm obs},\text{CURN})$ with $p(\theta|y_{\rm obs},\text{HD})$ and having $y_{\rm rep}^{(\theta)} \sim p(y_{\rm rep}^{(\theta)}|\theta,\text{HD})$. The resulting $D(y_{\rm rep};\theta)$ is centered on $A_{\rm gw}^2$, so we should see that $D(y)$ tracks the estimated GWB amplitude across the posterior $p(A_{\rm gw}|y_{\rm obs},\text{HD})$. If HD is the correct hypothesis, we expect to find an unremarkable $p$-value, neither too small nor too close to 1. A very low $p$-value would instead point to mismodeling.

In Fig. 6 we perform this check on our loud-injection simulations and find $p$-values that are distributed approximately uniformly between 0 and 1, as expected. A perfect uniform distribution would only obtain in the limit of exact noise-parameter determination, because we would be evaluating the $p$-value of each simulation $y_{\rm sim}^{(k)}$ against the true distribution $D(y_{\rm sim};\theta_{\rm sim}^{(k)})$. Because the approximation of Eq. (4) is inadequate for such strong GWB signals, to build Fig. 6 we evaluate Eq. (6) empirically as

$$\frac{1}{NM}\sum_{k}^{N}\sum_{l}^{M}\Theta[D(y_{\rm obs};\theta^k) - D(y^{k,l};\theta^k)], \tag{9}$$

where $y^{k,l} \sim p(y|\theta^k,\text{HD})$ and $\theta^k \sim p(\theta|y_{\rm obs},\text{HD})$.

To check whether we can exclude that the optimal statistic is excited by clock error, we would instead compute the Bayesian $p$-value of $D(y_{\rm obs};\theta)$ *under a monopole hypothesis*, using $p(\theta|y_{\rm obs},\text{CLK})$ and $p(y_{\rm rep}^{(\theta)}|\theta,\text{CLK})$ for Eq. (6). A small $p$-value would falsify the clock-noise hypothesis, while an unremarkable $p$ would suggest that the model is viable. Performing this check on our loud-injection simulations yields $p$-values too small too be measured using Eq. (9), but all indeed $\lesssim 0.01$.

Sardesai and Vigeland [43] propose an extension of the optimal-statistic framework (the multicomponent optimal statistic), in which the pulsar-pair correlations

$$\rho_{ij} = \frac{y_i^T C_i^{-1}\tilde{\Gamma}_{ij}C_j^{-1}y_j}{\text{tr}(C_i^{-1}\tilde{\Gamma}_{ij}C_j^{-1}\tilde{\Gamma}_{ji})}, \tag{10}$$

are fit to a linear model with components corresponding to different correlation patterns, with errors $\delta\rho_{ij}$ derived under the null hypothesis and assumed to be Gaussian and independent. (In fact, if spatially correlated processes with variance $A_\alpha^2$ are present in the data, they will induce correlations of order $O(A_\alpha^2)$ and $O(A_\alpha^4)$

among the $\rho_{ij}$, which will bias regression unless taken into account iteratively.)

SNRs for each component are defined as the $z$-score of the corresponding linear coefficient. For a single component, this reproduces the formal SNR of the standard optimal statistic [cf. the discussion below Eq. (4)]. In the general case, the multicomponent optimal statistic attempts to disentangle the cross-sensitivities of the individual optimal-statistic variants, but it can only do so in the context of regression rather than detection. That is, if we *assume* that a certain set of spatially correlated processes are present in the data, the multicomponent optimal statistic will produce estimates or their relative amplitudes. When augmented with procedures such as the Akaike information criterion [44], the multicomponent optimal statistic will also select a best-fitting model among multiple options [43]. Neither result can be interpreted easily as rejecting the null hypothesis, or providing quantifiable evidence of an alternative.

## V. THE OPTIMAL STATISTIC AS AN APPROXIMATE LIKELIHOOD RATIO

So far we have focused on computing optimal-statistic $p$-values for null hypothesis testing and its Bayesian extension. The $p$-values can falsify the null CURN hypothesis, thus confirming the presence of correlations, and they can also verify that the data are consistent with HD, by failing to falsify that hypothesis. Moving beyond this strictly falsificationist viewpoint, the optimal statistic is also related (at least approximately) to the ratio of the CURN and HD likelihoods [45]. In this section we describe how the likelihood ratio can be used directly to discriminate between the two models, and how it provides a link between the posterior distribution of the SNR and the CURN-vs.-HD Bayes factor.

To see how the optimal statistic is related to the likelihood ratio, we begin with the PTA likelihood in its marginalized form [see, e.g., [32]]:

$$\log p(y|\theta) = -\frac{1}{2}\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|2\pi\mathsf{K}| \quad \text{with} \quad \mathsf{K} = \mathsf{B} + A^2\tilde{\Gamma}; \tag{11}$$

here $\mathbf{y}$ is the concatenation of residuals for all array pulsars; $\mathsf{B}(\theta)$ is a block-diagonal covariance matrix in which each block $B_i$ describes the noise processes that are individual to a pulsar, including measurement noise, intrinsic spin noise, and timing-model errors, but not common red noise; and $A^2\tilde{\Gamma}(\theta)$ represents the covariance of the common red-noise process. For CURN, $\tilde{\Gamma}(\theta)$ is block-diagonal, with blocks given by $\Sigma_{ii}$; for HD, $\tilde{\Gamma}(\theta)$ has the same diagonal blocks, plus off-diagonal blocks given by $\Sigma_{ij}f_{\mathrm{HD}}(\zeta_{ij})$.

Expanding $\log p(y|\theta,\mathrm{HD})$ to linear order with respect to the $f_{\mathrm{HD}}(\zeta_{ij})$ coefficients yields [45]

$$\log p(y|\theta,\mathrm{HD}) \simeq -\frac{1}{2}\sum_i y_i^T C_i^{-1} y_i - \frac{1}{2}\sum_i \log|2\pi C_i| + \frac{1}{2}A^2\sum_{i\neq j} y_i^T C_i^{-1}\tilde{\Gamma}_{ij}C_j^{-1}y_j, \tag{12}$$

where $C_i = B_i + A^2\Sigma_{ii}$. Given that the first two terms in the sum add up to $\log p(y|\theta,\mathrm{CURN})$ and that the third term is proportional to the unnormalized optimal statistic [cf. Eq. (2)], it follows that

$$\frac{p(y|\theta,\mathrm{HD})}{p(y|\theta,\mathrm{CURN})} \simeq \exp\left\{\frac{1}{2}A^2 D(y;\theta)/\sigma_0^2(\theta)\right\}, \tag{13}$$

where of course $A$ is itself a component of $\theta$.

Likelihood ratios are broadly used as detection statistics to discriminate between pairs of models. Indeed, under broad conditions they are Neyman–Pearson-optimal [46], yielding the lowest rate of false dismissals for a chosen rate of false alarms. In our context, however, for any given dataset $y$ there is no single value of $D(y;\theta)$ and $\sigma(\theta)$, but rather posterior distributions induced by $p(\theta|y)$. Likewise, $p(\theta|y)$ induces a posterior distribution of the log likelihood ratio (PLLR). Dempster [30,31] has suggested that the PLLR can be used directly to make detection statements such as "under $X$% of the CURN posterior, HD is $Y$ times more likely than CURN." For instance, one may require $Y = 100$ and $X = 95\%$ to claim a detection.

Figure 7 shows PLLR distributions for the loud-injection simulations discussed above. Our example criterion is satisfied in 48% of the simulations. PLLR statements can be seen as posterior-predictive extensions of classical detection theory. Furthermore, they provide an interesting complement to Bayes factors: instead of "integrating, then

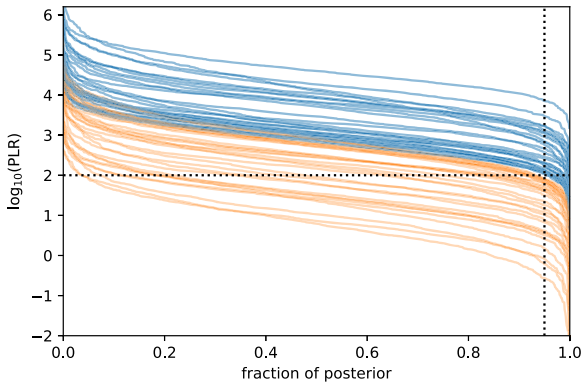


FIG. 7. Distribution of the posterior log likelihood ratio (PLLR), as approximated with the optimal statistic using Eq. (13), in the 69 loud-injection simulations. If we set our detection criterion as PLLR > 100 over 95% of the posterior, we conclude that HD is "detected" over CURN in 48% of our simulations: those that cross the 95%-percentile vertical line with values above the PLLR = 100 horizontal line, shown in blue here, while those that do not are shown in orange.

comparing," as we do with the Bayesian evidence, with PLLRs we "compare, then integrate" [31].

Finally, we demonstrate how the PLLR links the posterior distribution of the optimal-statistic SNR to the HD-vs.-CURN Bayes factor. This relationship was examined empirically by Pol and colleagues [49, Fig. 1], who find that, for small SNR, $\log \mathrm{BF}^{\mathrm{HD}}_{\mathrm{CURN}} \simeq \mathrm{SNR}^2/2$. Because the HD and CURN models share the same parameters and priors, the marginalized PLLR yields

$$\begin{aligned}
&\int \log\frac{p(y|\theta,\mathrm{HD})}{p(y|\theta,\mathrm{CURN})}\,p(\theta|y,\mathrm{CURN})\mathrm{d}\theta \\
&= \int \log\left(\frac{p(y|\mathrm{HD})}{p(y|\mathrm{CURN})}\frac{p(\theta|y,\mathrm{HD})}{p(\theta|y,\mathrm{CURN})}\right)p(\theta|y,\mathrm{CURN})\mathrm{d}\theta \\
&= \log\frac{p(y|\mathrm{HD})}{p(y|\mathrm{CURN})}\int p(\theta|y,\mathrm{HD})\mathrm{d}\theta \\
&\quad + \int \log\frac{p(\theta|y,\mathrm{HD})}{p(\theta|y,\mathrm{CURN})}\,p(\theta|y,\mathrm{CURN})\mathrm{d}\theta \\
&= \log \mathrm{BF}^{\mathrm{HD}}_{\mathrm{CURN}} - D_{\mathrm{KL}}(\mathrm{CURN}||\mathrm{HD}). \qquad (14)
\end{aligned}$$

In the second line we have used the Bayes theorem plus the fact that $p(\theta|\mathrm{HD}) = p(\theta|\mathrm{CURN})$ to rewrite the likelihood ratio as the Bayes factor times the ratio of posteriors. The $D_{\mathrm{KL}}(\mathrm{CURN}||\mathrm{HD})$ at the tail end of Eq. (14) is the Kullback–Leibler divergence [47] from HD to CURN, a non-negative measure of the discrepancy between the two distributions.

Now, if we approximate the PLLR using the optimal statistic [Eq. (13)] and replace the amplitude parameter $A^2$ with its optimal-statistic estimator $D(y;\theta)$, we obtain a heuristic relation between the HD-vs.-CURN Bayes factor and the optimal-statistic SNR,

$$\log \mathrm{BF}^{\mathrm{HD}}_{\mathrm{CURN}} - D_{\mathrm{KL}}(\mathrm{CURN}||\mathrm{HD}) = \overline{\mathrm{SNR}^2}/2, \qquad (15)$$

which *calibrates* the Bayes factor by linking it to a frequentist scheme. For instance, we may say that 3-$\sigma$ optimal statistic corresponds very roughly to $\mathrm{BF} \simeq 90$, while 4-$\sigma$ optimal statistic maps to $\mathrm{BF} \simeq 3,000$. Figure 8 shows that Eq. (15) is approximately realized in our loud-injection simulations, in agreement with Ref. [48]. However, $\overline{\mathrm{SNR}^2}/2$ values consistently overestimate Bayes factors, especially for larger SNRs. Kullback–Leibler corrections are small. For this figure, Bayes factors and divergences were computed by reweighting a moderate number of posterior samples [49]; thus, they are somewhat noisy, but not nearly enough to explain the spread observed in Fig. 8, which must instead originate in the approximations made to obtain Eq. (15).

Thinking back to the optimal statistic as a detection statistic, the integral over the noise-parameter posterior that defines the Bayesian $p$-value is dominated by the parameters that yield lower SNRs and larger $p$-values, and therefore higher risk; by contrast, Eq. (15) contains $\overline{\mathrm{SNR}^2}/2$, emphasizing higher SNRs, and therefore greater confidence. After all, even if the hypothesis-testing and model-comparison approaches can be related, they answer fundamentally different questions about detection.

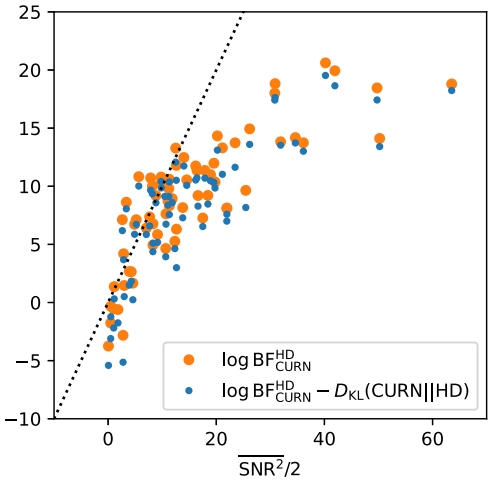


FIG. 8. Posterior mean of $\mathrm{SNR}^2/2$ vs log HD-vs.-CURN Bayes factors, with and without the correction of Eq. (15), in our loud-injection simulations. Bayes factors and Kullback–Leibler divergences are computed using the reweighting scheme of Hourihane *et al.* [49]; errors are ∼1, and not shown for clarity. Note the large range of $\overline{\mathrm{SNR}^2}$ and Bayes factors, obtained even if all simulations have the same pulsar-noise and GWB parameters. Equation (15) is realized approximately, with a large vertical spread. Bayes factors appear to saturate toward the high end of the $\overline{\mathrm{SNR}^2}/2$ distribution.

## VI. CONCLUSION

In this article we examine the role of the optimal statistic [27–29], and especially of its hybrid variant [29] in establishing the presence of inter-pulsar correlations in pulsar-timing-array data. The logic is that of null-hypothesis statistical testing: by observing an extreme value of the optimal statistic, we are able to reject a null model that contains a common-spectrum signal but no inter-pulsar correlations. The strength of this conclusion is encapsulated by the $p$-value—the probability that the null model could have produced data that results in an equally extreme statistic.

The fact that we must simultaneously fit for the unknown pulsar noise parameters leads to the hybrid frequentist–Bayesian approach, in which we obtain a posterior distribution for the optimal statistic from the noise-parameter posteriors. We show that the hybrid optimal statistic can be understood in the framework of Bayesian model checking [23], in which the $p$-value is evaluated with respect to data replications generated from the null model by drawing model parameters from their posteriors. This Bayesian $p$-value maps to a new summary statistic, the Bayesian SNR, which should be used to characterize the statistical

significance of the observed correlations. Computed against different posteriors, the Bayesian SNR can also provide principled evidence *for* Hellings-Downs correlations, by failing to reject the HD model, and by rejecting models with alternative correlation patterns such as CLK.

By contrast, the posterior SNR of the hybrid SNR (i.e., the $\overline{\text{SNR}}$) *cannot* be mapped to a $p$-value for the optimal statistic. Instead, a $p$-value *for the* $\overline{\text{SNR}}$ can be established empirically with respect to a population of simulations or of "bootstrapped" datasets [38,39] for which certain model details are altered to effectively erase interpulsar correlations. These are different tests that answer different detection questions with narrower definitions of the null hypothesis, so they should be used in complement to the BSNR.

We also consider the optimal statistic as an approximation to the HD–CURN log likelihood ratio. We suggest that the posterior SNR can be used to formulate detection statements based on Dempster's posterior distribution of the likelihood ratio [30]. However these statements may be biased by the approximation of the log likelihood as linear in the Hellings-Downs coefficient, especially so for loud correlated signals at the edge of detection.

Last, we show that the mean HD–CURN log likelihood ratio is related to the HD-vs.-CURN Bayes factor by way of the Kullback–Leibler divergence between the two posteriors. Since the ratio is approximately $\overline{\text{SNR}^2}/2$, this relation provides a qualitative calibration of Bayes factors in terms of the hypothesis-testing SNRs that may be expected for similar datasets. The relation also justifies the commonly used heuristic $\log \text{BF} \simeq \text{SNR}^2/2$, but our experiments (as displayed in Fig. 8) suggest that the heuristic is realized only very approximately.

## ACKNOWLEDGMENTS

The authors would like to thank Bence Bécsy for identifying a problem in original version of Figure 8. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF award No. ACI-1548562, and specifically the Bridges-2 system at the Pittsburgh Supercomputing Center, supported by NSF award No. ACI-1928147. M. V., P. M. M., and K. C., acknowledge support through the NSF Physics Frontiers Center awards No. 1430284 and No. 2020265. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

## APPENDIX: INFERENCE AND SIMULATION

Throughout this article, we computed CURN, HD, and CLK posteriors by evaluating the marginalized PTA likelihood [Eq. (11)] with ENTERPRISE [50], and sampling it with PTMCMCSAMPLER [51], following all prescriptions adopted in the NANOGrav 12.5-yr GWB analysis [10]. Pulsar noise parameters were set to the maximum *a posteriori* values obtained in single-pulsar "noise runs" on NANOGrav 12.5-yr data [52], except for intrinsic–red-noise (log) amplitudes and spectral indices, which were MCMC-sampled alongside $\log_{10} A_{\text{curn}}$ (or $\log_{10} A_{\text{hd}}$ or $\log_{10} A_{\text{clk}}$, as appropriate). We used uniform priors of $[-18, -11]$ for the $\log_{10}$ amplitude quantities, and of $[0, 7]$ for red-noise spectral indices. The spectral index of the common process was fixed to $13/3$. The optimal statistic was evaluated with ENTERPRISE using the matrix components of these likelihoods, drawing intrinsic–red-noise parameters from the appropriate posterior chains, and keeping the other pulsar-noise parameters fixed.

Simulated datasets were obtained under the HD model by fixing *all* noise parameters to 12.5-yr single-pulsar maximum *a posteriori* values, augmented by full-array maximum *a posteriori* values for the intrinsic–red-noise amplitudes and spectral indices. To draw random realizations of all noise processes, we decomposed the matrices $B_i$ that appear in Eq. (11) as

$$B_i = N_i + F_i \Phi_i F_i^\dagger, \tag{A1}$$

where $N_i$ is an $n_i^{\text{obs}} \times n_i^{\text{obs}}$ diagonal matrix (with $n_i^{\text{obs}}$ the number of measured residuals for pulsar $i$), $\Phi$ is a $n_i^{\text{gp}} \times n_i^{\text{gp}}$ diagonal matrix (with $n_i^{\text{gp}}$ the total number of Gaussian-process basis components for matrix $i$), and $F_i$ is a $n_i^{\text{obs}} \times n_i^{\text{gp}}$ the rectangular matrix of Gaussian-process basis vectors. For each simulated dataset and each pulsar we then obtained

$$y_i^{\text{noise}} = \sqrt{N_i}\epsilon + F^\dagger \sqrt{\Phi_i}\zeta, \tag{A2}$$

with $\epsilon$ ($\zeta$) an $n_i^{\text{obs}}$-dimensional ($n_i^{\text{gp}}$-dimensional) vector of independent unit Gaussian deviates. The components of $\Phi_i$ were set to the appropriate power laws for noise processes, and to $10^{-14} \times n_{\text{obs}}$ s$^2$ for timing-model errors, so that each timing-model parameter could contribute the same variance of $10^{-14}$ s$^2$ (on average) to each residual. For each simulated dataset the Hellings-Downs process was sampled jointly for all pulsars as

$$\mathbf{y}^{\text{hd}} = A\mathsf{G}\mathsf{L}\boldsymbol{\xi}, \tag{A3}$$

where $\mathsf{G}$ is a block-diagonal matrix in which each $n_i^{\text{obs}} \times n^{\text{hd}}$-dimensional block $G_i$ encodes the Hellings-Downs basis vectors for pulsar $i$; $\mathsf{L}\mathsf{L}^\dagger$ is the Cholesky decomposition of the Hellings-Downs covariance matrix $\tilde{\Gamma}$; and $\boldsymbol{\xi}$ is an $(n^{\text{psr}} \times n^{\text{hd}})$-dimensional vector of independent unit Gaussian deviates. We set $\log_{10} A_{\text{hd}}$ to $-18$ for the no-injection datasets and to $-14$ for the loud-injection datasets. Although we ran 100 simulations for each case, only 66 and 69 respectively were completed due to memory limitations on the Bridges-2 computing cluster.

See Refs. [10,22,32] for details on the Gaussian-process formulation of the PTA likelihood.

[1] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Evidence for a gravitational-wave background, Astrophys. J. Lett. **951**, L8 (2023).

[2] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European Pulsar Timing Array III. Search for gravitational wave signals, Astron. Astrophys. **678**, A50 (2023).

[3] D. J. Reardon *et al.*, Search for an isotropic gravitational-wave background with the Parkes Pulsar Timing Array, Astrophys. J. Lett. **951**, L6 (2023).

[4] H. Xu *et al.*, Searching for the Nano-hertz stochastic gravitational wave background with the Chinese Pulsar Timing Array data release I, Res. Astron. Astrophys. **23**, 075024 (2023).

[5] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Constraints on supermassive black hole binaries from the gravitational-wave background, Astrophys. J. Lett. **952**, L37 (2023).

[6] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European Pulsar Timing Array: V. Implications for massive black holes, dark matter and the early Universe, arXiv:2306.16227.

[7] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European Pulsar Timing Array IV. Search for continuous gravitational wave signals, arXiv:2306.16226.

[8] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Bayesian limits on gravitational waves from individual supermassive black hole binaries, Astrophys. J. Lett. **951**, L50 (2023).

[9] A. Afzal *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Search for signals from new physics, Astrophys. J. Lett. **951**, L11 (2023).

[10] Z. Arzoumanian *et al.*, The NANOGrav 12.5 yr data set: Search for an isotropic stochastic gravitational-wave background, Astrophys. J. Lett. **905**, L34 (2020).

[11] B. Goncharov *et al.*, On the evidence for a common-spectrum process in the search for the nanohertz gravitational-wave background with the Parkes Pulsar Timing Array, Astrophys. J. Lett. **917**, L19 (2021).

[12] S. Chen *et al.*, Common-red-signal analysis with 24-yr high-precision timing of the European Pulsar Timing Array: Inferences in the stochastic gravitational-wave background search, Mon. Not. R. Astron. Soc. **508**, 4970 (2021).

[13] J. Antoniadis *et al.*, The International Pulsar Timing Array second data release: Search for an isotropic gravitational wave background, Mon. Not. R. Astron. Soc. **510**, 4873 (2022).

[14] H. Middleton, A. Sesana, S. Chen, A. Vecchio, W. Del Pozzo, and P. A. Rosado, Massive black hole binary systems and the NANOGrav 12.5 yr results, Mon. Not. R. Astron. Soc. **502**, L99 (2021).

[15] A. Zic *et al.*, Evaluating the prevalence of spurious correlations in pulsar timing array data sets, Mon. Not. R. Astron. Soc. **516**, 410 (2022).

[16] B. Goncharov, E. Thrane, R. M. Shannon, J. Harms, N. D. R. Bhat, G. Hobbs, M. Kerr, R. N. Manchester, D. J. Reardon, C. J. Russell, X.-J. Zhu, and A. Zic, Consistency of the Parkes Pulsar Timing Array signal with a nanohertz gravitational-wave background, Astrophys. J. Lett. **932**, L22 (2022).

[17] R. W. Hellings and G. S. Downs, Upper limits on the isotropic gravitational radiation background from pulsar timing analysis, Astrophys. J. Lett. **265**, L39 (1983).

[18] C. Tiburzi, G. Hobbs, M. Kerr, W. A. Coles, S. Dai, R. N. Manchester, A. Possenti, R. M. Shannon, and X. P. You, A study of spatial correlations in pulsar timing array data, Mon. Not. R. Astron. Soc. **455**, 4339 (2016).

[19] M. Vallisneri *et al.*, Modeling the uncertainties of solar system ephemerides for robust gravitational-wave searches with Pulsar-Timing Arrays, Astrophys. J. **893**, 112 (2020).

[20] R. van Haasteren, Y. Levin, P. McDonald, and T. Lu, On measuring the gravitational-wave background using Pulsar Timing Arrays, Mon. Not. R. Astron. Soc. **395**, 1005 (2009).

[21] S. R. Taylor, *Nanohertz Gravitational Wave Astronomy* (CRC Press, Boca Raton, FL, 2021).

[22] Z. Arzoumanian *et al.*, The NANOGrav 11 year data set: Pulsar-timing constraints on the stochastic gravitational-wave background, Astrophys. J. **859**, 47 (2018).

[23] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis, London, 2013).

[24] P. Meyers, M. Vallisneri, K. Chatziioannou, and A. J. K. Chua, following paper, Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. II. Posterior predictive distributions and pseudo-Bayes factors Phys. Rev. D **108**, 123008 (2023).

[25] C. Pernet, Null hypothesis significance testing: A short tutorial, F1000Research 10.12688/f1000research.6963.3 (2015), 4.

[26] B. P. Abbott *et al.*, Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).

[27] M. Anholm, S. Ballmer, J. D. E. Creighton, L. R. Price, and X. Siemens, Optimal strategies for gravitational wave stochastic background searches in pulsar timing data, Phys. Rev. D **79**, 084030 (2009).

[28] S. J. Chamberlin, J. D. E. Creighton, X. Siemens, P. Demorest, J. Ellis, L. R. Price, and J. D. Romano, Time-domain implementation of the optimal cross-correlation statistic for stochastic gravitational-wave background searches in Pulsar Timing Data, Phys. Rev. D **91**, 044048 (2015).

[29] S. J. Vigeland, K. Islo, S. R. Taylor, and J. A. Ellis, Noise-marginalized optimal statistic: A robust hybrid frequentist-Bayesian statistic for the stochastic gravitational-wave background in pulsar timing arrays, Phys. Rev. D **98**, 044003 (2018).

[30] A. P. Dempster, The direct use of likelihood for significance testing, Stat. Comput. **7**, 247 (1997).

[31] I. Smith and A. Ferrari, The posterior distribution of the likelihood ratio as a measure of evidence, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics Conference Series Vol. 1305, edited by A. Mohammad-Djafari, J.-F. Bercher, and P. Bessiére (Springer Cham, Switzerland, 2011), pp. 391–398.

[32] R. van Haasteren and M. Vallisneri, New advances in the Gaussian-process approach to pulsar-timing data analysis, Phys. Rev. D **90**, 104012 (2014).

[33] J. S. Hazboun, P. M. Meyers, J. D. Romano, X. Siemens, and A. M. Archibald, Analytic distribution of the optimal cross-correlation statistic for stochastic gravitational-wave-background searches using pulsar timing arrays, arXiv:2305.01116.

[34] J. D. Romano, J. S. Hazboun, X. Siemens, and A. M. Archibald, Common-spectrum process versus cross-correlation for gravitational-wave searches using pulsar timing arrays, Phys. Rev. D **103**, 063027 (2021).

[35] B. Allen and J. D. Romano, The Hellings and Downs correlation of an arbitrary set of pulsars, Phys. Rev. D **108**, 043026 (2023).

[36] D. B. Rubin, Bayesianly justifiable and relevant frequency calculations for the applied statistician, Ann. Stat. **12** 1151 (1984).

[37] A. Gelman, X. Meng, and H. Stern, Posterior predictive assessment of model fitness via realized discrepancies, Statistica Sinica **6**, 733 (1996), https://www.jstor.org/stable/24306036.

[38] S. R. Taylor, L. Lentati, S. Babak, P. Brem, J. R. Gair, A. Sesana, and A. Vecchio, All correlations must die: Assessing the significance of a stochastic gravitational-wave background in pulsar timing arrays, Phys. Rev. D **95**, 042002 (2017).

[39] N. J. Cornish and L. Sampson, Towards robust gravitational wave detection with pulsar timing arrays, Phys. Rev. D **93**, 104047 (2016).

[40] M. F. Alam *et al.*, The NANOGrav 12.5 yr data set: Observations and narrowband timing of 47 millisecond pulsars, Astrophys. J. Suppl. Ser. **252**, 4 (2021).

[41] A. Gelman, Two simple examples for understanding posterior p-values whose distributions are far from uniform, Electron. J. Stat. **7**, 2595 (2013).

[42] G. Hobbs *et al.*, A pulsar-based time-scale from the International Pulsar Timing Array, Mon. Not. R. Astron. Soc. **491**, 5951 (2020).

[43] S. C. Sardesai and S. J. Vigeland, Generalized optimal statistic for characterizing multiple correlated signals in pulsar timing arrays, arXiv:2303.09615.

[44] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Selected Papers of Hirotugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa (Springer, New York, New York, NY, 1998), pp. 199–213.

[45] J. A. Ellis, X. Siemens, and R. van Haasteren, An efficient approximation to the likelihood for gravitational wave stochastic background detection using Pulsar Timing Data, Astrophys. J. **769**, 63 (2013).

[46] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. R. Soc. A **231**, 289 (1933).

[47] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat. **22**, 79 (1951).

[48] N. S. Pol *et al.* (NANOGrav Collaboration), Astrophysics milestones for Pulsar Timing Array gravitational-wave detection, Astrophys. J. Lett. **911**, L34 (2021).

[49] S. Hourihane, P. Meyers, A. Johnson, K. Chatziioannou, and M. Vallisneri, Accurate characterization of the stochastic gravitational-wave background with pulsar timing arrays by likelihood reweighting, Phys. Rev. D **107**, 084045 (2023).

[50] J. A. Ellis, M. Vallisneri, S. R. Taylor, and P. T. Baker, ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust PulsaR Inference SuitE (2019), ascl:1912.015.

[51] J. Ellis and R. van Haasteren, github.com/jellis18/ptmcmcsampler: Official release (2017), 10.5281/zenodo.1037579.

[52] NANOGrav Collaboration, 12.5-year GWB analysis pulsar noise file, https://github.com/nanograv/12p5yr_stochastic_analysis (2020).