Towards robust gravitational wave detection with pulsar timing arrays

Neil J. Cornish and Laura Sampson

Phys. Rev. D **93**, 104047 — Published 25 May 2016

DOI: 10.1103/PhysRevD.93.104047

Towards Robust Gravitational Wave Detection with Pulsar Timing Arrays

Neil J. Cornish[1] and Laura Sampson[1]

[1]*Department of Physics, Montana State University, Bozeman, MT 59717, USA.*

Precision timing of highly stable milli-second pulsars is a promising technique for the detection of very low frequency sources of gravitational waves. In any single pulsar, a stochastic gravitational wave signal appears as an additional source of timing noise that can be absorbed by the noise model, and so it is only by considering the coherent response across a network of pulsars that the signal can be distinguished from other sources of noise. In the limit where there are many gravitational wave sources in the sky, or many pulsars in the array, the signals produce a unique tensor correlation pattern that depends only on the angular separation between each pulsar pair. It is this distinct fingerprint that is used to search for gravitational waves using pulsar timing arrays. Here we consider how the prospects for detection are diminished when the statistical isotropy of the timing array or the gravitational wave signal is broken by having a finite number of pulsars and a finite number of sources. We find the standard tensor-correlation analysis to be remarkably robust, with a mild impact on detectability compared to the isotropic limit. Only when there are very few sources and very few pulsars does the standard analysis begin to fail. Having established that the tensor correlations are a robust signature for detection, we study the use of "sky-scrambles" to break the correlations as a way to increase confidence in a detection. This approach is analogous to the use of "time-slides" in the analysis of data from ground based interferometric detectors.

PACS numbers: 04.30.-w, 04.30.Tv, 97.60.Lf

I. INTRODUCTION

With the steady addition of new pulsars to the arrays and improvements in the timing sensitivity and analyses, pulsar timing is advancing rapidly as a technique for the detection of gravitational waves. Impressive new upper limits on the amplitude of power-law stochastic backgrounds are starting to challenge simple astrophysical models that attribute the background to the gravitational wave driven evolution of a population of supermassive black hole binaries on quasi-circular orbits [1–3]. These limits are dominated by the timing residuals from one or two very low noise pulsars that have been observed for many years. A detection, on the other hand, will come from combining the data of a very large number of moderately sensitive pulsars [4].

The key to making a detection, as opposed to setting an upper limit, is the unique correlation pattern that results when a gravitational wave signal passes through an array of pulsars. In the limit where there are an infinite number of isotropically distributed sources [5] (and at least two pulsars) or an infinite number of pulsars (and at least one source) [6], the correlation in the timing residuals of two pulsars a, b with an angular separation α_{ab} has the form

$$H_{ab} = \frac{3\,c_{ab}}{2}\,\ln c_{ab} - \frac{c_{ab}}{4} + \frac{1}{2}\,(1+\delta(\alpha_{ab}))\ , \qquad (1)$$

where $c_{ab} = (1-\cos\alpha_{ab})/2$. In any actual experiment, neither condition required to arrive at (1) is met. There will only be a finite number of sources contributing to the signal, and a finite number of pulsars contributing to the timing array. This means that the standard correlation analysis will be sub-optimal [6]. Here we study the impact that this has on the detectability of a stochastic background. We do this by comparing the detectability of the highly anisotropic signal formed from a finite number of black hole binaries to that of an idealized isotropic signal with the same average power level. We investigate this as a function of the number of black holes and the number of pulsars. We find that the standard correlation analysis is remarkably robust, and only results in a small loss in detection efficiency for realistic signals and array sizes. It is only in the limit of very few pulsars and very few sources that a significant loss of effectiveness occurs.

Having established the detection of a tensor correlation pattern as a robust signature of gravitational waves, we investigate the use of "sky-scrambles" to purposely break the signal correlations in the data as a test of the analysis pipelines. If the evidence for a gravitational wave signal is largest for the true pulsar sky locations, and much smaller for any of the scrambled sky locations, then we gain confidence in our models and analysis techniques. Each sky scramble produces a distinct correlation pattern, C'_{ij}, and we can define a measure of closeness based on the similarity of the correlation patterns. As hoped, we find that the evidence for a gravitational wave signal is highest for the true pulsar locations, and lowest for scrambles that are most dissimilar to the expected correlation pattern. Based on our simple measure of closeness, there are a limited number of independent sky scrambles, which limits the statistical power of the test. But we argue against trying to use the test in a frequentist framework. Rather, sky-scrambles help to validate the noise and signal models used to compute the Bayesian evidence of a signal that, like the cosmic microwave background, we only get to see once.

Our work builds on several earlier studies [7–11], where various statistics are developed to detect gravitational wave signals in pulsar timing data. Of particular rele-

vance is the 'optimal statistic' [9, 12] for the detection of the stochastic background, which is essentially a measure of how important the cross-correlations between pulsars are for describing the signal. Both our method and these frequentist analyses investigate the detection of the tensor correlation pattern between pulsars, which we again emphasize is necessary for the unambiguous detection of gravitational waves.

II. THE SIMULATED ASTROPHYSICAL SIGNAL

Electromagnetic observations of massive galaxies and galaxy mergers across cosmic history, combined with population synthesis models, suggest that the dominant source of gravitational waves in the pulsar timing band (10^{-9} Hz $\rightarrow 10^{-6}$ Hz) will be slowly evolving supermassive black hole binaries with masses in the range $10^8 M_\odot \rightarrow 10^9 M_\odot$ [13–15]. It was initially assumed that the superposition of the signals from many thousands of such systems would produce a background that is effectively stochastic and statistically isotropic, and thus amenable to detection using the cross-correlation technique developed by Hellings and Downs [5]. Recent studies of the signals produced by simulated black hole populations, though, have shown that relatively nearby and massive outliers play an important role, and can lead to significant departures from stochasticity and isotropy [16, 17].

To illustrate the importance of outliers in these populations, we simulate the gravitational wave signals from a population model provided by A. Sesana that assumes quasi-circular, gravitational wave driven orbital evolution. The gravitational waves from each binary are co-added and used to compute the signal power as a function of sky position, $h^2_{\rm ss}(\theta,\phi) = h^2_+(\theta,\phi) + h^2_\times(\theta,\phi)$, summed over frequency. Figure 1 shows the distribution of the gravitational wave power across the sky for one realization of the black hole population, as well as for a statistically isotropic stochastic background with the same average power spectrum. The difference is striking. The intensity variations for the black hole population are over one hundred times larger than for the isotropic model. The lower two panels of Figure 1 show the pulsar response to the signals as a function of pulsar sky location. (Only the Earth-term contribution is shown here. Including the pulsar terms simply adds "noise" to the maps.) Note that the pulsar response has a similar angular power distribution for both the isotropic and black hole skies, even though the underlying signals are vastly different. The explanation can be found in Figure 2, which shows the detected power in pulsars at different sky locations for a single black hole binary. The broad antenna response of the pulsars effectively blurs the underlying power distribution. Note, however, that this does not imply that pulsar timing arrays are unable to resolve small scale features - the information to reconstruct the spatial distribution resides in the cross-spectra, and this information can be used to accurately map the background [18–21].

Our procedure for testing the tensor correlation analysis on realistic black hole populations and pulsar timing arrays is to compare the evidence for detections between simulated black hole populations and statistically isotropic signals with the same average power level. To do this we first simulate the response to a particular realization of the black hole population model for an array of pulsars, and from this compute the average power spectrum (averaged over the pulsars), $S_h(f)$. We then simulate a statistically isotropic, unpolarized Gaussian stochastic background with this same power spectrum. For reference, we also consider the standard power law spectrum for quasi-circular binaries whose evolution is gravitationally wave driven, which have characteristic strain $h_c(f) = A(f/f_y)^{-2/3}$ and $S_h(f) = A^2(f/f_y)^{-4/3}/(12\pi^2 f^3)$ where $f_y = 1/\text{year}$. Figure 3 shows examples of the average power spectrum for the three models for a 20 pulsar array. The power law model is generated with $A = 10^{-15}$. The slight differences in the black hole spectrum and the equivalent isotropic spectrum are due to the fact that this is a single realization of the stochastic power spectrum, averaged over the array.

The reference astrophysical model we are using - based on a quasi-circular, gravitational wave driven merger model - likely *overestimates* the degree of isotropy we can expect in reality. Environmental effects, such as stellar scattering and gas driven mergers, along with orbital eccentricity will act to reduce the number of binary systems in each frequency bin [22, 23]. As a crude proxy for these effects, we also produce simulated backgrounds that randomly down-sample the full black hole population by factors of 10 and 100, and, as an extreme case, produce backgrounds that include only the 10 brightest sources from the full simulation (in contrast, the full population model includes over 22,000 sources in the observation band). In addition to considering different numbers of sources in the signal, we also investigate the impact of including different numbers of pulsars in the array. We investigate simulated arrays with 5 equally sensitive pulsars, 20 equally sensitive pulsars, and the 36 pulsars of varying sensitivity taken from the first International Pulsar Timing Array (IPTA) mock data challenge. The simulated data sets include white noise at a range of levels, but no simulated red noise or residuals from the quadratic spin-down model. The effects of red noise and timing errors are, however, included in the analysis.

III. ANALYSIS

We apply Bayesian inference and model selection to analyze the simulated data sets. The analysis procedure is very similar to that described in Ref. [24]. The likelihood of observing data d for a given model set of model

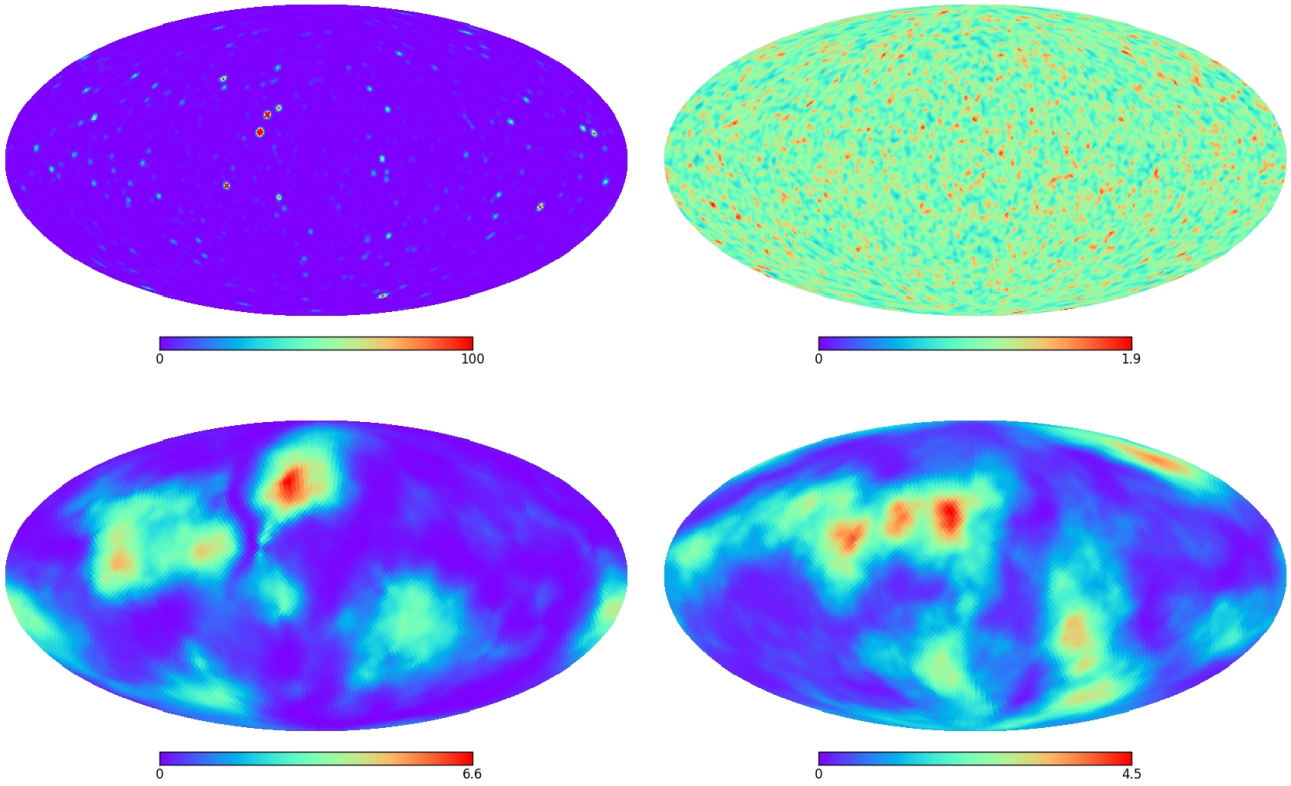

FIG. 1: The upper two maps show the distribution of the gravitational wave power across the sky, scaled by the all-sky average for two signal model. The map on the upper left is for a simulated black hole population, smoothed with a two degree Gaussian blur and clipped at a contrast of 100 to enhance weaker features. The peak intensity in the un-clipped map exceeds 800. The map on the upper right is for a statistically isotropic signal with the same power spectrum as the black hole simulation. The map has been smoothed with a two degree Gaussian blur. Clearly, the power distribution from a realistic black hole population is far from isotropic. The lower two panels show the detected power in the Earth-term for pulsars at different sky locations, in other words, the raw signals convolved with the antenna patterns, summed and squared. Despite the large differences in the underlying power distribution, the response to the anisotropic BH background (lower left) is qualitatively identical to the response to the statistically isotropic signal (lower right).

parameters $\vec{\lambda}$ is

$$p(d|\vec{\lambda}) = \frac{\exp\left(-\frac{1}{2}\sum_{ab}\sum_{ij} r_{ai} C^{-1}_{(ai)(bj)} r_{bj}\right)}{\sqrt{(2\pi)^M \det C}}, \qquad (2)$$

where C is the covariance matrix, which depends on both the noise in the individual pulsars and on the GW background, and $r = d - t$ denotes the timing residuals after the subtraction of the (deterministic) timing model t from the data d. The indices a and b label individual pulsars, and run from 1 to the number of pulsars, N_p. The indices i and j label the data samples, i.e. individual frequency bins. Since our simulated data is stationary, the correlation matrix is diagonal in i, j and $C_{(ai)(bj)} \to C_{ab}(f_i)\delta_{ij}$. The simulated data set consists of $N = 512$ samples per pulsar, evenly spaced in time at weekly intervals, giving a total data set of size $M = N\,N_p$ spanning just under $T = 10$ years. The analysis is carried out in the Fourier domain, where the quadratic timing model for pulsar a has the form

$$t_a(f_k) = \frac{\alpha_a}{f_k^2} + \frac{i\beta_a}{f_k}, \qquad (3)$$

where $f_k = k/T$ for integer k (see [24], Sec. III for details.) The covariance matrix is given by

$$C_{ab}(f) = S_h(f) H_{ab} + \delta_{ab}\left\{S_{n_a} + S_{r_a}(f/f_y)^{r_a}\right\}, \qquad (4)$$

where $S_h(f)$ is the PSD of the GW background, S_{n_a} is the PSD of the white noise, S_{r_a} is the amplitude of the PSD of the red noise, and r_a is the spectral slope of the red noise (which should not be confused with the r_a from Eq. (2), which represents the residuals in pulsar a!). In the sky-scramble analysis the tensor correlation matrix, H_{ab}, is replaced by a scrambled version that is derived by randomly choosing false sky locations for each pulsar. We consider two models for $S_h(f)$, a simple power law $S_h(f) = S_g(f/f_y)^\gamma$, and a

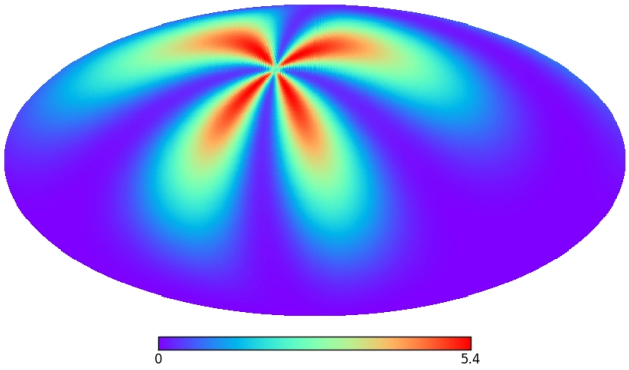

FIG. 2: The detected power for pulsars at different sky locations for a single BH located at the center of the "petals". The orientation of the petals rotates depending on the polarization of the signal. The broad spread in the power distribution for PTAs explains why the response to isotropic and anisotropic signals is so similar.

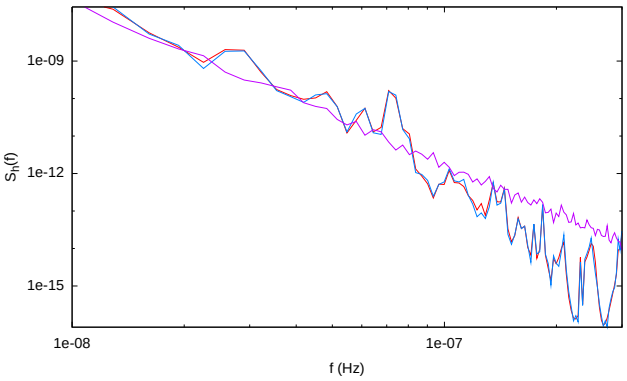

FIG. 3: The average power spectrum in a 20 pulsar array for a black hole population (in red), a stochastic, isotropic model with the same average power level as the black hole population (in blue), and a power law, stochastic, isotropic model with amplitude $A = 10^{-15}$.

more general bin-by-bin model $S_h(f_k) = P_k$, where P_k is the power spectral density in the $k^{\rm th}$ frequency bin. There are $N/2$ of these P_k parameters, given an observation time of T and a Nyquist frequency of $0.5/dt$, so $N_{\rm bin} = 0.5 * T/dt = N/2$. The full parameter vector $\vec{\lambda}$ for the timing plus noise model has the $5N_p$ parameters $\vec{\lambda} \rightarrow \{\alpha_a, \beta_a, S_{n_a}, S_{r_a}, r_a\}$, while the power-law gravitational wave model has the $2 + 5N_p$ parameters $\vec{\lambda} \rightarrow \{S_g, \gamma, \alpha_a, \beta_a, S_{n_a}, S_{r_a}, r_a\}$, and the bin-by-bin gravitational wave model has the $N/2 + 5N_p$ parameters $\vec{\lambda} \rightarrow \{P_k, \alpha_a, \beta_a, S_{n_a}, S_{r_a}, r_a\}$. The priors $p(\vec{\lambda})$ on the power spectral density parameters $\{S_g, P_k, S_{n_a}, S_{r_a}\}$ are taken to be uniform in the logarithm across the range $[10^{-35}{\rm Hz}^{-1}, 10^{-4}{\rm Hz}^{-1}]$. The priors on the spectral slope parameters $\{\gamma, r_a\}$ are taken to be uniform in the range $[-2, -6]$, and the priors on the timing model parameters $\{\alpha_a T^2, \beta_a T\}$ are taken to be uniform in the range $[-0.8, 0.8]$.

In the analyses that follow we are less interested in the posterior distributions for the model parameters, $p(\vec{\lambda}|d)$, than we are in the model evidence $p(d) = \int p(d|\vec{\lambda})p(\vec{\lambda})d\vec{\lambda}$ for the various models. In particular, we compute the Bayes factor for a detection as the evidence ratio between the signal model and the noise model. The evidence is computed using the thermodynamic integration technique [25], which returns the evidence as a natural by-product of the parallel tempered Markov Chain Monte Carlo scheme [26] that we use to map the posterior distributions. The implementation of the MCMC algorithm is as described in Ref. [24], with two additional features: an additional move that proposes to transfer power between the signal and red noise models, and an adaptive scheme for the temperature ladder.

The adaptive scheme is as follows: we begin with 50 chains equally spaced between in log temperature between 1 and 10^6, and perform an MCMC run with this spacing while keeping track of the acceptance rate for parallel tempering moves between all adjacent pairs of chains. If one of these acceptance rates falls below 1%, we stop the run and insert 3 chains with temperatures evenly spaced between the two chains that lost contact. We continue this process until the MCMC runs for 10^6 iterations without adding any chains. We then use this final temperature ladder (usually including $\sim 90 - 100$ chains) to perform the evidence calculation, using an MCMC run of 1.5 million iterations.

As indicated, the new move we have implemented shifts power between the GW signal and the independent red noise in each pulsar. To do this, we calculate the average level of red noise in all of the pulsars, $\bar{S}_r = \sum_i S_r^i/NP$, and propose that the GW amplitude takes this value by proposing from a Gaussian centered at this value with a width of $\sigma_1 = 0.5$. We simultaneously propose that the red noise level in each pulsar be drawn from a Gaussian centered at the current red noise level, with a width of $\sigma_2 = \sigma_1/\sqrt{NP}$. We find that this proposal greatly aids the mixing between models. A nearly identical proposal can be used to move power between the red noise in the individual pulsars and the common red noise level, if such a term is present in the model.

IV. DETECTING ANISOTROPIC BACKGROUNDS

Several methods have been proposed for detecting and mapping anisotropic gravitational wave backgrounds with pulsar timing arrays [6, 18, 19, 21, 27], but it remains to be seen if these methods are more effective at making a first detection than the standard tensor correlation analysis. In the case of the cosmic microwave background radiation, the uniform glow was detected long before the first anisotropies were seen, but in that case the anisotropies are tiny compared to the overall power.

The much larger anisotropy of the nanoHertz gravitational wave sky may mean that a model that allows for anisotropy will improve the prospects for detection. But it is not obvious that this will be true, because while the data might be better fit by an anisotropic model, such models are necessarily more complicated than the isotropic model, and this added complexity comes at a price. We defer the comparison of the efficacy of isotropic and anisotropic models for future study, and instead consider the simpler question of how effective the standard isotropic analysis is when applied to realistic anisotropic signals using realistic numbers of pulsars in the array. We accomplish this by comparing the detectability of anisotropic signals from a population of black holes to the detectability of a statistically isotropic background with an identical power spectrum.

We consider two measures of detectability, the first being the Bayes factor, or evidence ratio, between the signal and noise models. Recall that the noise model includes individual red and white noise components for each pulsar, which are uncorrelated between pulsars, and the signal model includes a common stochastic component with a red power spectrum, with the characteristic Hellings-Downs correlation pattern between pulsars. Unfortunately, as we show in section §V, this signal-to-noise model Bayes factor will imply the detection of signals that do not have the correct tensor correlation pattern. (Though with lower significance than similarly bright signals that do.) Because the tensor correlation pattern is key to any claim of a gravitational wave detection with pulsar timing, we go on to consider a second measure of detectability - this measure compares the evidence for the tensor correlation model to a model with a diagonal correlation matrix. This diagonal model corresponds to a common level of red noise present in all pulsars.

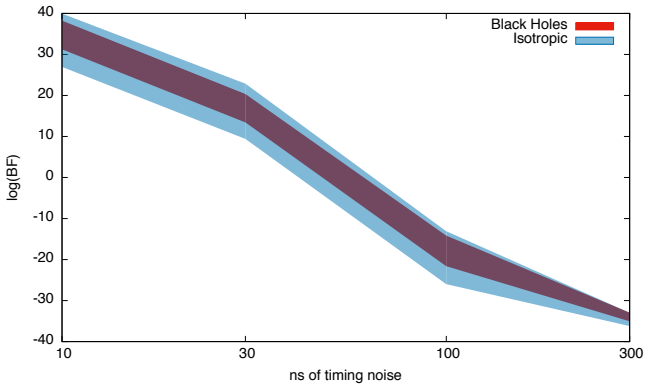

FIG. 4: The detectability of the full simulated black hole population (red) and a statistically isotropic stochastic signal with the same power spectrum (blue) as a function of the white timing noise level in a simulated pulsar timing array with 20 equally sensitive pulsars, as measured by the signal-model to noise-model Bayes factor. The spread in the Bayes factors is computed by considering 10 realizations of each signal type. The shaded bands covers a one standard deviation spread about the mean.

Even within one class of signals there is considerable variation in detectability from realization to realization. To account for this we consider multiple realizations for each signal type and aggregate the results. Figure 4 compares the detectability of the signal from a full black hole population to that of a statistically isotropic stochastic signal with the same power spectrum, as a function of the white timing noise level in a simulated pulsar timing array with 20 equally sensitive pulsars. (In this figure we define detectability in terms of the evidence ratio between the signal and noise models). Remarkably, we see that there is no discernible difference between the detectability of the two types of signal, even though the black hole signal is far from isotropic and the array is comprised of relatively few pulsars.

The signal model used here, and in all other figures, is that of a pure power law. The bin-by-bin model has many more parameters than the power-law model, and is therefore less effective at detecting a power-law signal, but it has added flexibility that can better capture the non–power-law spectra that arise for the very sparse black hole population models. We calculated the evidence for both types of signal model for many of the simulations discussed in this section, and consistently found that the pure power-law model is preferred. In this analysis, the added complexity of the bin-by-bin model overwhelmed the benefit of a better fit to the spectral shape. Figure 3 illustrates why this is the case - the difference between the simulated spectrum and a pure power law is quite small. The simulated non-power-law signals are better recovered with a power-law model than with the highly flexible bin-by-bin model. There will be spectra that are so poorly described by a power law that the bin-by-bin is preferred, but we did not encounter any examples of this type in our analysis.

Figure 5 shows similar results as Figure 4, but for smaller black hole populations (i.e., more anisotropic skies). Once again the anisotropic signals are almost as detectable as their isotropic equivalents. Differences only become apparent for exceedingly sparse black hole populations with only a handful of sources. One may be concerned that the simulations that use only the ten brightest black holes to generate the background may not be well described by a power law. Examining the power spectra for these cases, though, shows that they can indeed be fit by power laws, although typically with slopes that are steeper than for the full population. We additionally investigate how the size of the array affects the results by considering a smaller array made up of 5 equally sensitive pulsars. The upper panel in Figure 6 shows the detectability of the full black hole population and its isotropic equivalent, while the lower panel in Figure 6 shows a more extreme case, with just 10 black hole binaries and 5 pulsars in the array. In this case the correlation analysis does perform worse on the black hole population than on the isotropic equivalent, but the difference is still within the uncertainty from realization to realization.

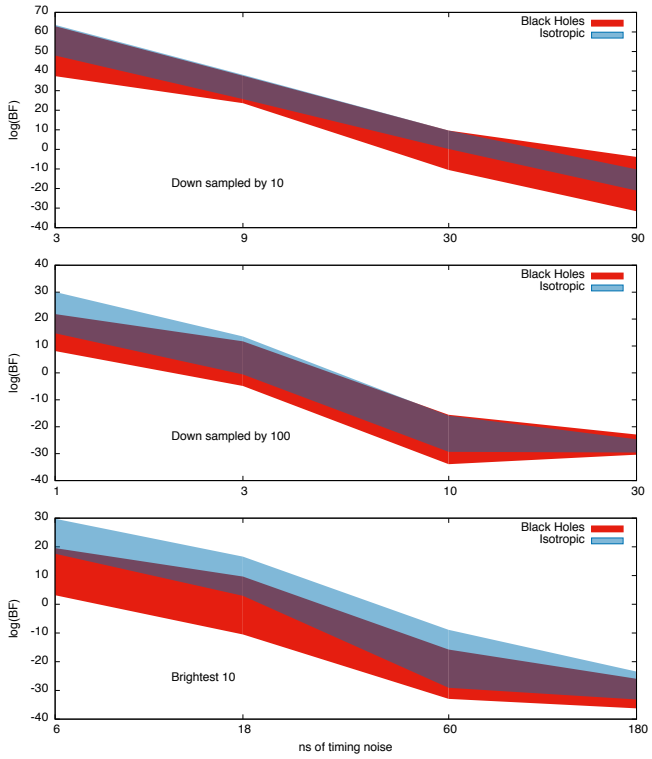

FIG. 5: The detectability of down sampled black hole populations (red) and a statistically isotropic stochastic signal with the same power spectrum (blue) as a function of the white timing noise level in a simulated pulsar timing array with 20 equally sensitive pulsars, as measured by the signal-model to noise-model Bayes factor. The upper panel is for the full population down sampled by a factor of ten, the middle panel is down sampled by one-hundred, and the lower panel is for the ten brightest binaries. The spread in the Bayes factors is computed by considering 10 realizations of each signal type. The shaded bands covers a one standard deviation spread about the mean.

Finally, we explore the effect of anisotropy on the detection of a stochastic GW background using the pulsars from the IPTA. Instead of injecting a particular noise level into all of the pulsars in the array, we scale the actual noise level for each of the pulsars by the same factor. One might expect that this array would behave much like the 20 pulsar array explored previously in this section. It turns out, though, that there are a few pulsars in the array that are much better-timed than the others, and so dominate detection. Figure 7 shows the results from this study, presented in the same format as previous results in this section. The results are much more like the 5 pulsar array than the 20 pulsar case.

While our analysis indicates that the standard correlation analysis is almost as effective at detecting anisotropic signals as it is at detecting the isotropic signals it was designed for, it is unclear if the signal model is picking up the full tensor correlation pattern (1), or merely a common red component in the timing residuals - i.e.

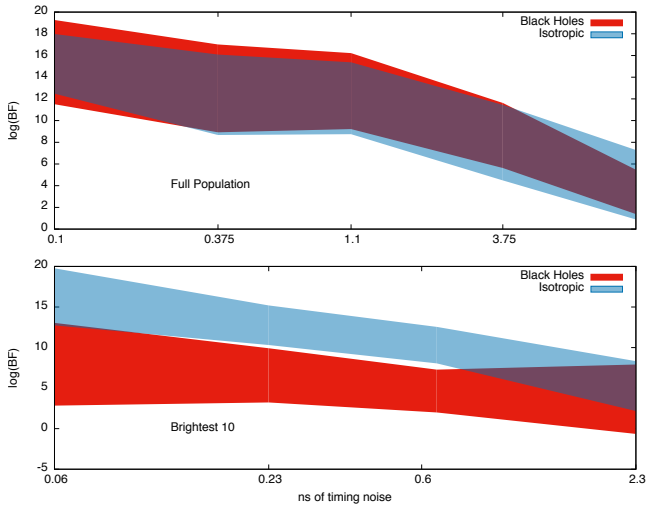

FIG. 6: The detectability of signals from a population of black holes (red) and a statistically isotropic stochastic signal with the same power spectrum (blue) for a small pulsar timing array made up of 5 equally sensitive pulsars, as measured by the signal-model to noise-model Bayes factor. The upper panel is for the full black hole population, while the lower panel is for the brightest 10 black holes. The spread in the Bayes factors is computed by considering 10 realizations of each signal type. The shaded bands covers a one standard deviation spread about the mean.

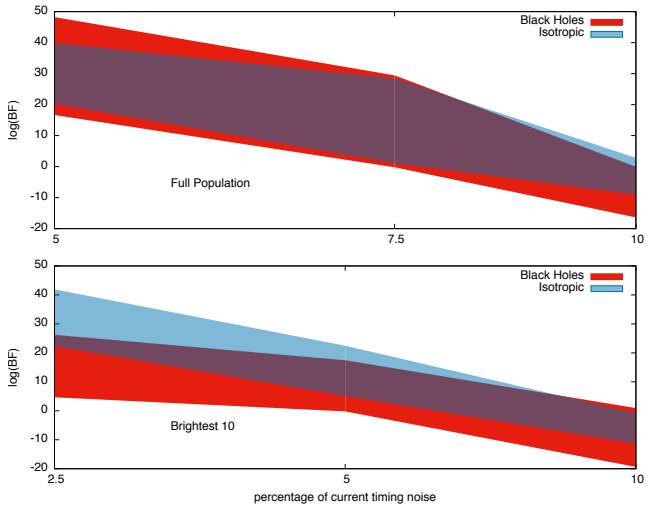

FIG. 7: The detectability of signals from a population of black holes (red) and a statistically isotropic stochastic signal with the same power spectrum (blue) for the IPTA pulsar array, again for the full spectrum in the upper panel and the brightest 10 black holes in the lower panel. The horizontal axis shows the percentage of the actual timing noise in each pulsar that is included in the injected data. The results are more similar to the 5 pulsar array than the 20 pulsar case, despite the fact that there are 36 pulsars in the IPTA. This is because a few of the pulsars are much better timed than the others.

just the diagonal terms in (1). While the gravitational wave signal model includes a diagonal component, there could conceivably be correlated power on the diagonal due to some common red noise process. For example, there may be some physical process that is shared by all neutron stars that produces a characteristic spectrum of red timing noise [28]. To be sure that it is a gravitational wave signal that has been detected we need clear evidence that the off-diagonal, cross-correlation terms follow the Hellings-Downs curve. One approach is to try and infer the correlation pattern directly from the data, using techniques such as a cubic spline fit to the correlation pattern [29], or by directly inferring the correlation of each pulsar pair [30] and comparing this to the Hellings-Downs curve. Another approach is to apply Bayesian model selection between a model with the full tensor correlation curve H_{ab} and a model with a common red noise term, given by the diagonal correlation model $H'_{ab} = \delta_{ab}$. We opt to follow the later approach, which was first described in Ref. [31]. The common red noise term can either be considered in addition to the gravitational wave model as an extra term in the noise model, or by comparing "signal" models with correlation matrices given by H_{ab} and H'_{ab}. We settle on the latter approach, as including an extra common red noise model made it very difficult to compute reliable estimates for the evidence. This is because the large correlations between the common red noise model, the per-pulsar red noise model, and the signal model impeded mixing of the Markov chains, and led to a series of steep transitions in the thermodynamic integration integrand. Moreover, even with reliable evidence estimates produced by using vast numbers of steps in the temperature ladder, the results are highly dependent on the choice of the priors on the various parameters in each model, which is always an issue when comparing models of different dimension. Comparing the full and diagonal correlation models is much easier, and the choice of priors has much less of an effect, as the two models share the same parameters, effectively canceling the prior dependence in the evidence ratios. In the language of Ref. [31], we are comparing the evidence of models $M_{\rm gw}$ and $M_{\rm corr}$, whereas our earlier results compared the gravitational wave model $M_{\rm gw}$ to the noise model $M_{\rm null}$. We demand that the evidence for $M_{\rm gw}$ exceeds the evidence for both $M_{\rm null}$ and $M_{\rm corr}$ to claim a detection. Figure 8 shows the Bayes factors between the full tensor correlation model $M_{\rm gw}$ and the noise model $M_{\rm null}$, and the Bayes factors between the diagonal correlation model $M_{\rm corr}$ and the noise model $M_{\rm null}$ for a simulated pulsar timing array with 20 equally sensitive pulsars. Each panel shows the results for an increasingly anisotropic signal from a population of black holes, with the spread in Bayes factors computed by considering multiple (10) realizations of each population to account for cosmic variance. (We chose 10 realizations because of the computational cost involved in running each simulation. The resulting spread in Bayes factors is perhaps a conservative estimate, but the general trends should be robust.)

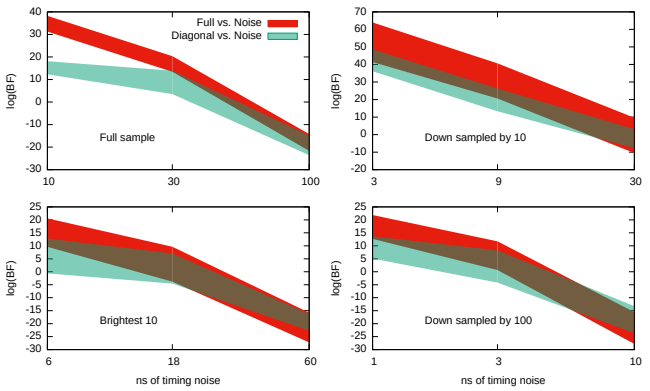

FIG. 8: The detectability of simulated black hole populations as a function of the noise level for a simulated pulsar timing array with 20 equally sensitive pulsars. The red bands show the spread in log Bayes factors for the full tensor correlation model computed from multiple realizations of a black hole population model. The green bands show the spread in log Bayes factors for the diagonal correlation model applied to the same set of simulated signals. The gravitational wave signal model is favored when the log Bayes factor for the tensor correlation model versus the noise model is positive *and* that it exceeds the log Bayes factor for the diagonal correlation model versus the noise model. From top to bottom the simulations are for the full black hole background, the populations downsampled by ten then one hundred, and finally for just the ten brightest systems.

Figure 9 shows the log Bayes factors between the full tensor correlation model $M_{\rm gw}$ and the diagonal common noise model $M_{\rm corr}$, for both anisotropic black hole populations and their isotropic equivalents, with the spread showing the cosmic variance derived by considering multiple realizations. Values greater than zero indicate that the Hellings-Downs correlation curve has been detected in the data. As expected, the evidence ratio for the full and diagonal models tends to unity as the noise level increases, because these models have the same prior volume. We see that the results for isotropic and anisotropic signals are essentially indistinguishable for the relatively non-downsampled cases, which implies that not only is a common red noise component being detected in both cases, but so is the tensor correlation pattern. For the highly downsampled cases (the lower panels of this figure), the detectability of the Hellings-Downs pattern is much more realization-dependent for the black hole populations than for the isotropic backgrounds. This indicates a greater difficulty in differentiating between a GW signal and a different correlated noise source (i.e. clock noise) for these highly anisotropic signals.

V. SKY SCRAMBLES

We have found the tensor correlation pattern to be a remarkably robust signature for detecting gravitational waves with pulsar timing arrays. As the existing pulsar

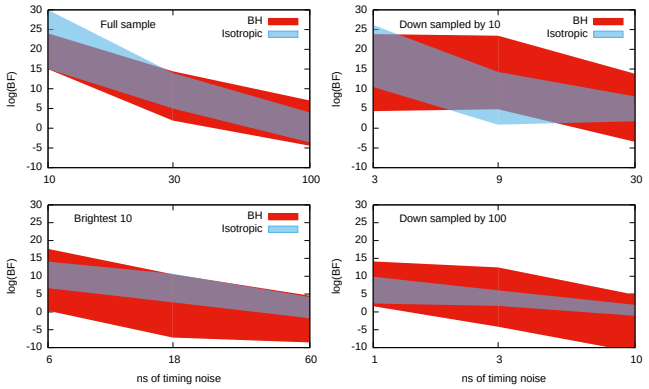

FIG. 9: The log Bayes factors between the full and diagonal correlation model. Values greater than zero indicate that the Hellings-Downs correlation pattern has been detected. The red bands are for simulated black hole populations, and the blue bands are for the isotropic equivalents. For the highly downsampled cases (lower panels), the detectability of the Hellings-Downs pattern is much more realization-dependent for the black hole populations than for the isotropic backgrounds, indicating a greater difficulty in differentiating between a GW signal and a different correlated noise source.

timing arrays continue to collect data at improved sensitivity, and as more pulsars are added to the arrays, we should start to see the first hints of correlated gravitational wave power in the timing residuals [4]. The evidence for a signal will then grow steadily with time, until eventually the evidence becomes overwhelming. However, the evidence we compute is between *our model for the signal* and *our model for the noise*, and deficiencies in either of these models could lead to false positives or false negatives. Assuming that general relativity provides a faithful description of gravity in the regime probed by pulsar timing, our model for the signals and how they perturb the timing residuals should be reliable, but the many potential sources of noise are less well understood. Ideally, we would like to be able to study the noise properties in data that is free of gravitational waves, but there is no way to shield our detector from gravitational wave signals.

The same challenge occurs in the analysis of data from the ground-based LIGO/Virgo interferometers [32, 33], where studies have shown the noise to be both non-stationary and non-Gaussian [34–36], with frequent loud transient features, or glitches [37]. While it is not possible to remove gravitational wave signals from the data, it is possible to destroy signal coherence across the detector network by introducing artificial time delays between the detectors during the analysis [38]. Because the noise in each detector is assumed to be uncorrelated to begin with, the 'time slides' preserve the statistical properties of the noise. The groups that analyze the data from ground based detectors approach the detection problem in a frequentist framework, using some detection statistic to identify candidate events. The distribution of triggers from the time slides are used to compute false alarm rates, and any triggers in the zero-lag data that exceed a pre-ordained false alarm rate (such as one per millennium) are deemed detection candidates. To establish a false alarm threshold of one per millennium for a year long data set requires thousands of independent time slides. This can be achieved for the ground-based detectors, as the correlation time of typical signals in the ground based interferometer band are generally less that a second.

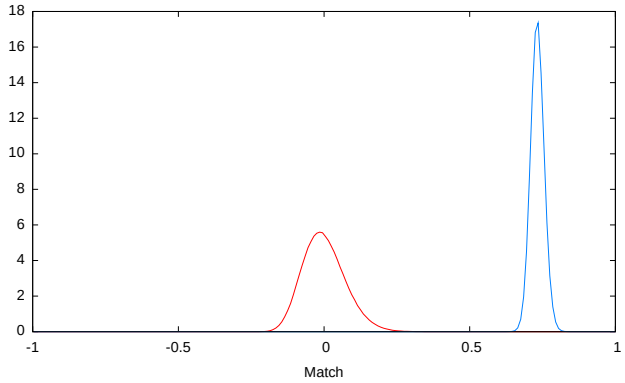

FIG. 10: Histograms of the full match M (in blue), and the off-diagonal match $\overline{M}$ for randomly drawn sky-scrambles of an equal-sensitivity 20 pulsar timing array

In the case of pulsar timing, though, the correlation time of the gravitational wave signals is expected to be comparable to the duration of the available data sets, making it impossible to generate independent time-slides. Instead, we can break the *spatial* correlations by artificially scrambling the pulsar positions used in the gravitational wave analyses (the true positions still have to be used in the timing model). We can define the match between two sets of correlation matrices H_{ab} and H'_{ab} as

$$M = \frac{\sum_{a,b} H_{ab} H'_{ab}}{\left(\sum_{a,b} H_{ab} H_{ab} \sum_{a,b} H'_{ab} H'_{ab}\right)^{1/2}} \qquad (5)$$

Because the correlation matrices are dominated by their diagonal terms, the match M will always be greater than zero. Since what we are really interested in are the cross-correlation terms, we can define a modified match that excludes the diagonal contributions:

$$\overline{M} = \frac{\sum_{a\neq b} H_{ab} H'_{ab}}{\left(\sum_{a\neq b} H_{ab} H_{ab} \sum_{a\neq b} H'_{ab} H'_{ab}\right)^{1/2}} \qquad (6)$$

For realistic networks in which the noise varies between pulsars and with frequency, the sums in the above expression should be replaced by noise-weighted sums of

the form

$$\sum_{a,b} H_{ab} H'_{ab} \rightarrow \sum_{a,b} \int \frac{H_{ab} H'_{ab}}{S_a(f) S_b(f)} \, df \,, \tag{7}$$

where $S_a(f)$ is the sensitivity curve for pulsar a. The sensitivity curve is derived by convolving the noise spectrum with the gravitational wave response and the timing model. Figure 10 shows a histogram of M and $\overline{M}$ for randomly drawn sky locations for an equal-sensitivity 20 pulsar network. The width of the distributions scale inversely with the effective number of pulsars in the array, which we define as

$$N_{\rm eff} = \frac{\sum_{a=1}^{N_p} \int S_a(f)^{-1} df}{S_{\rm max}^{-1}} \,, \tag{8}$$

where $S_{\rm max}^{-1} = \max_{1 \leq a \leq N_p} \int S_a(f)^{-1} df$. For an equally sensitivity network $N_{\rm eff} = N_p$, while for a heterogeneous network $N_{\rm eff} < N_p$. For example, the $N_p = 36$ pulsar network used in the first IPTA mock data challenge has $N_{\rm eff} = 4.35$, which nicely explains the result from the previous section that shows that the IPTA array behaves more like an array with $N_p = 5$ than with $N_p = 20$.

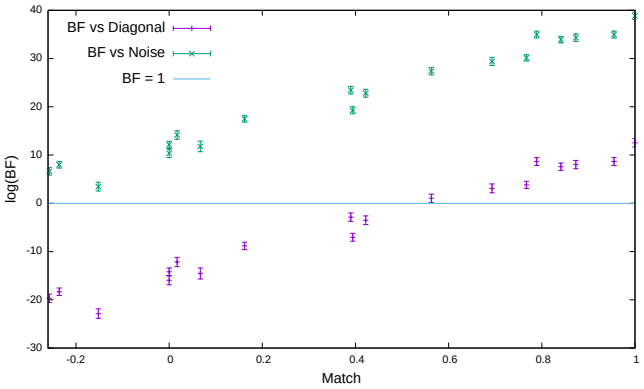

FIG. 11: Log Bayes factors in favor of a detection as a function of the match $\overline{M}$ for an equal-sensitivity 20 pulsar timing array. The green points compare the full signal model to the noise model, while the purple points compare the full signal model to a signal model with a diagonal correlation matrix.

We expect to see a correlation between the match, $\overline{M}$, and the Bayes factor between the signal and noise model. Figure 11 shows a roughly linear relationship between $\overline{M}$ and ln(BF). The simulations were for an equal-sensitivity 20 pulsar network for two noise levels using one realization of the full black hole population mode. Because random scrambles rarely produce matches above $\overline{M} = 0.2$ for a 20 pulsar network, the high match examples were found by applying small random perturbations to the true pulsar locations. We see that even scrambles with $\overline{M} < 0$ can produce Bayes factors in favor of the signal model, as the diagonal components of H'_{ab} pick up power from both the Earth-term and the Pulsar-term. Another way of saying this is that the full match, M, which is greater than zero for all scrambles, is the relevant quantity when comparing the signal model to the standard noise model described by (4). As described in the previous section, we can focus the analysis on the off-diagonal, cross-correlation pattern by adding a common red "noise" term as a diagonal component of C_{ab} to the standard noise model - we call this the diagonal model. This results in the Bayesian equivalent of the frequentist optimal statistic defined in Ref. [9, 12].

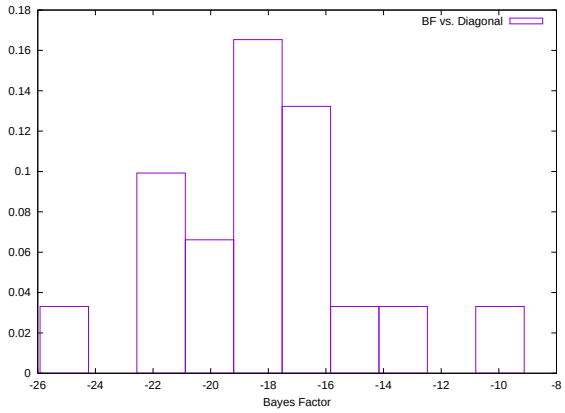

FIG. 12: A histogram of the log Bayes factors between the full correlation model and a diagonal correlation model for the 18 independent sky-scrambles for an equal-sensitivity 20 pulsar timing array. Here independent is defined as matches with the true correlation pattern and each other of $\overline{M} < 0$. For reference, the analysis with un-scrambled sky locations gave a log Bayes factor of 12.

In the LIGO/Virgo context, thousands of independent time-slides are used to establish the false alarm rate. In the pulsar timing case we are unable to generate nearly so many independent sky-scrambles, at least when using $\overline{M}$ as the sole measure of independence. For example, if we define two correlation patterns H_{ab} to be independent if their match $\overline{M} < 0$, then the number of independent scrambles is roughly equal to $N_{\rm eff}$. This number is derived from numerical experiments in which successive skies were generated randomly and added to the group of independent skies if they were independent of all the other skies in the group. If the criteria for independence is relaxed to, say, $\overline{M} < 0.1$, then the number of independent skies grows by an order of magnitude. Figure 12 shows the distribution of log Bayes factors between the full and diagonal signal models for 18 mutually independent scrambles (defined as having $\overline{M} < 0$), using the same data set as Figure 11. Notice that none of the analyses with scrambled sky locations give a Bayes factor in favor of a gravitational wave signal with a tensor correlation pattern being present, while the analysis using the correct sky locations gave overwhelming evidence for such a signal. With just 10 or 20 independent scrambles it is

not possible to make very interesting statements about "false alarm rates", which in any case are rather meaningless in the context of measuring a single realization of a signal. Rather, the sky scrambles, along with analyses of simulated signals, can be used as a way of testing the models being used in the Bayesian analysis. It may also be that using the match to define independence significantly underestimate the number of independent sky scrambles. For example, there are many scrambles that give the same match value $\overline{M}$ to the original array, but have very different collections of pulsar pairs in the positive and negative sectors of the Hellings-Downs curve.

VI. SUMMARY

The detection of a stochastic gravitational wave background with PTAs requires the detection of cross-correlations between the timing residuals in multiple pulsars. When either the gravitational wave signal or the pulsar array is spatially isotropic, the values of these cross-correlations are uniquely determined by the tensor nature of the radiation via the Hellings-Downs curve. Standard pipelines for analyzing the stochastic background in PTA data assume an isotropic background when searching for these cross-correlations. We have shown that, despite the fact that realistic gravitational wave skies can contain a large degree of anisotropy, the isotropic search is remarkably robust, and only leads to loss of detection efficiency in extreme cases. This robustness being established, we have shown that we can break the expected correlations by scrambling the location of the pulsars in the sky, and that these sky scrambles result in a lower evidence for the signal model. Thus we can build confidence in the detection of a stochastic background of gravitational waves be establishing that the evidence for such a detection shrinks as the correlation matrix for the pulsars in the array is made more and more dissimilar to the Hellings-Downs values.

Acknowledgments

We benefited from discussion with participants in the Aspen Center for Physics summer program on "Computation, systematics, and inference for pulsar-timing arrays, and beyond", including Steven Taylor, Jonathan Gair, Stanislav Babak, Alberto Sesana, Xavier Siemens, Sean McWilliams, Justin Ellis, Rutger van Haasteren Joe Romano and Chiara Mingarelli. The Aspen Center for Physics is supported by National Science Foundation grant PHY-106629. NJC was supported by NSF Physics Frontiers Center Award PFC-1430284. LS was supported by Nicolás Yunes' NSF CAREER Award PHY-1250636.

[1] R. M. Shannon, V. Ravi, L. T. Lentati, P. D. Lasky, G. Hobbs, M. Kerr, R. N. Manchester, W. A. Coles, Y. Levin, M. Bailes, et al., Science **349**, 1522 (2015), http://www.sciencemag.org/content/349/6255/1522.full.pdf, URL `http://www.sciencemag.org/content/349/6255/1522.abstract`.

[2] L. Lentati, S. R. Taylor, C. M. F. Mingarelli, A. Sesana, S. A. Sanidas, A. Vecchio, R. N. Caballero, K. J. Lee, R. van Haasteren, S. Babak, et al., MNRAS **453**, 2576 (2015), 1504.03692.

[3] Z. Arzoumanian, A. Brazier, S. Burke-Spolaor, S. Chamberlin, S. Chatterjee, B. Christy, J. Cordes, N. Cornish, P. Demorest, X. Deng, et al., ArXiv e-prints (2015), 1508.03024.

[4] X. Siemens, J. Ellis, F. Jenet, and J. D. Romano, Class.Quant.Grav. **30**, 224015 (2013), 1305.3196.

[5] R. W. Hellings and G. S. Downs, Astrophysical Journal - Letters (1983).

[6] N. J. Cornish and A. Sesana, Class.Quant.Grav. **30**, 224005 (2013), 1305.0326.

[7] J. M. Cordes and R. M. Shannon, Astrophys. J. **750**, 89 (2012), 1106.4047.

[8] D. R. B. Yardley, W. A. Coles, G. B. Hobbs, J. P. W. Verbiest, R. N. Manchester, W. van Straten, F. A. Jenet, M. Bailes, N. D. R. Bhat, S. Burke-Spolaor, et al., MNRAS **414**, 1777 (2011), 1102.2230.

[9] M. Anholm, S. Ballmer, J. D. E. Creighton, L. R. Price, and X. Siemens, Phys. Rev. D **79**, 084030 (2009), 0809.0701.

[10] S. J. Chamberlin, J. D. E. Creighton, X. Siemens, P. Demorest, J. Ellis, L. R. Price, and J. D. Romano, Phys. Rev. D **91**, 044048 (2015), 1410.8256.

[11] L. S. Finn, S. L. Larson, and J. D. Romano, Phys. Rev. D **79**, 062003 (2009), 0811.3582.

[12] S. J. Chamberlin, J. D. E. Creighton, P. B. Demorest, J. Ellis, L. R. Price, J. D. Romano, and X. Siemens, ArXiv e-prints (2014), 1410.8256.

[13] J. S. B. Wyithe and A. Loeb, Astrophys. J. **590**, 691 (2003), astro-ph/0211556.

[14] A. H. Jaffe and D. C. Backer, Astrophys. J. **583**, 616 (2003), astro-ph/0210148.

[15] M. Rajagopal and R. W. Romani, Astrophys. J. **446**, 543 (1995), astro-ph/9412038.

[16] P. A. Rosado and A. Sesana, MNRAS **439**, 3986 (2014), 1311.0883.

[17] V. Ravi, J. S. B. Wyithe, G. Hobbs, R. M. Shannon, R. N. Manchester, D. R. B. Yardley, and M. J. Keith, Astrophys. J. **761**, 84 (2012), 1210.3854.

[18] C. M. F. Mingarelli, T. Sidery, I. Mandel, and A. Vecchio, Phys. Rev. D **88**, 062005 (2013), 1306.5394.

[19] S. R. Taylor and J. R. Gair, Phys. Rev. **D88**, 084001 (2013), 1306.5395.

[20] S. R. Taylor, C. M. F. Mingarelli, J. R. Gair, A. Sesana, G. Theureau, S. Babak, C. G. Bassa, P. Brem, M. Burgay, R. N. Caballero, et al., Physical Review Letters **115**, 041101 (2015), 1506.08817.

[21] N. J. Cornish and R. van Haasteren (2014), 1406.4511.

[22] B. Kocsis and A. Sesana, MNRAS **411**, 1467 (2011), 1002.0584.

[23] V. Ravi, J. S. B. Wyithe, R. M. Shannon, and G. Hobbs,

Mon. Not. Roy. Astron. Soc. **447**, 2772 (2015), 1406.5297.
[24] L. Sampson, N. J. Cornish, and S. T. McWilliams, Phys.Rev. **D91**, 084055 (2015), 1503.02662.
[25] P. M. Goggans and Y. Chi, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. J. Erickson and Y. Zhai (2004), vol. 707 of *American Institute of Physics Conference Series*, pp. 59–66.
[26] R. H. Swendsen and J.-S. Wang, Physical Review Letters **57**, 2607 (1986).
[27] J. Gair, J. D. Romano, S. Taylor, and C. M. Mingarelli, Phys. Rev. **D90**, 082001 (2014), 1406.4664.
[28] R. M. Shannon and J. M. Cordes, Astrophys. J. **725**, 1607 (2010), 1010.4794.
[29] S. R. Taylor, J. R. Gair, and L. Lentati, Phys. Rev. **D87**, 044035 (2013), 1210.6014.
[30] L. Lentati, P. Alexander, M. P. Hobson, S. Taylor, J. Gair, S. T. Balan, and R. van Haasteren, Phys. Rev. **D87**, 104021 (2013), 1210.3578.
[31] J. A. Ellis, X. Siemens, and R. van Haasteren, Astrophys. J. **769**, 63 (2013), 1302.1903.
[32] J. Aasi et al. (LIGO Scientific), Class.Quant.Grav. **32**, 074001 (2015), 1411.4547.
[33] T. Accadia, M. Agathos, A. Allocca, P. Astone, G. Ballardin, et al., pp. 261–270 (2015).
[34] J. Aasi et al. (VIRGO), Class.Quant.Grav. **29**, 155002 (2012), 1203.5613.
[35] N. J. Cornish and T. B. Littenberg, Class.Quant.Grav. **32**, 135012 (2015), 1410.3835.
[36] T. B. Littenberg and N. J. Cornish, Phys.Rev. **D91**, 084034 (2015), 1410.3852.
[37] L. Blackburn, L. Cadonati, S. Caride, S. Caudill, S. Chatterji, et al., Class.Quant.Grav. **25**, 184004 (2008), 0804.0800.
[38] M. Was, M.-A. Bizouard, V. Brisson, F. Cavalier, M. Davier, et al., Class.Quant.Grav. **27**, 015005 (2010), 0906.2120.