MINIMUM REQUIREMENTS FOR DETECTING A STOCHASTIC GRAVITATIONAL
BACKGROUND USING PULSARS

J. M. Cordes1 AND R. M. Shannon1,2
1 Astronomy Department, Cornell University, Ithaca, NY 14853, USA; cordes@astro.cornell.edu, ryan.shannon@csiro.au
2 CSIRO Astronomy and Space Science, Epping, NSW 1710, Australia
Received 2011 June 14; accepted 2012 February 28; published 2012 April 18

ABSTRACT

We assess the detectability of a nanohertz gravitational wave (GW) background in a pulsar timing array (PTA) program by considering the shape and amplitude of the cross-correlation function summed over pulsar pairs. The distribution of correlation amplitudes is found to be non-Gaussian and highly skewed, which significantly influences detection and false-alarm probabilities. When only white noise combines with GWs in timing data, our detection results are consistent with those found by others. Contamination by red noise from spin variations and from any uncorrected interstellar plasma effects significantly increases the false-alarm probability. The number of arrival times (and thus the observing cadence) is important only as long as the residuals are dominated by white noise. When red noise and GWs dominate, the statistical significance of the correlation estimate can be improved only by increasing the number of pulsars. We characterize plausible detection regimes by evaluating the number of millisecond pulsars (MSPs) that must be monitored in a high-cadence, five-year timing program to detect a GW background spectrum \( h_\nu(f) = \alpha(f/f_0)^{-2/3} \) with \( f_0 = 1 \text{ yr}^{-1} \) and \( \alpha = 10^{-15} \). Our results indicate that a sample of 20 super-stable MSPs—those with rms timing residuals \( \sigma_r \lesssim 20\text{ ns}(A/10^{-15}) \) from red-noise contributions over a five-year span—will allow detection of the GW background and study of its spectrum. However, a timing program on \( \gtrsim 50-100 \) MSPs is likely needed for a complete PTA program, particularly if red noise is generally present in MSPs.

Key words: gravitational waves – pulsars: general – stars: neutron – stars: rotation

1. INTRODUCTION

There is current strong interest in exploiting the spin stability of millisecond pulsars (MSPs) to detect gravitational waves (GWs) at nanohertz frequencies (\( \sim 0.1-1 \text{ c} \text{ yr}^{-1} \)). Sources of GWs in this band include mergers of supermassive black holes (SMBHs) that collectively produce an isotropic background (Jaffe & Backer 2003; Phinney 2004; Jenet et al. 2005; Sesana et al. 2008) and in a few cases may be detected individually (GWs) at nanohertz frequencies (or \( \gtrsim 0 \) ). Timing may also detect GW backgrounds from cosmic strings (¨Olmez et al. 2010), the influence of massive gravitons (Lee et al. 2010), or solar system perturbations from primordial black holes (Seto & Cooray 2007). Detection methods have been based on finding excess variance in the timing residuals of individual sources, investigating spectral signatures in power spectra, or identifying the angular correlation expected between pulsar pairs from GWs passing through the solar system.

Astrophysical and instrumental processes limit the timing precision of any given pulsar and the overall sensitivity of a pulsar timing array (PTA). The efficacy of detection methods has received very uneven assessment with respect to contamination from different kinds of additive noise. These include both white- and red-noise processes, the latter having power strongly concentrated at lower fluctuation frequencies. We evaluate their impact on the sensitivity to a stochastic GW background, which itself comprises a red-noise process.

In this paper, we are concerned with detection of GWs as distinguished from their detailed characterization. Recent work has considered both frequentist (e.g., Jenet et al. 2005; Anholm et al. 2009; Yardley et al. 2011) and Bayesian approaches (e.g., van Haasteren et al. 2009, 2011; Finn & Lommen 2010) to the detection problem. While these methods are robust to varying degrees, our view is that detection needs to be corroborated with convincing diagnostics. We draw an analogy with the detection of a new spectral line or detection of cosmic microwave background fluctuations. Bayesian or frequentist inference can yield the probability of the existence of a source and probabilistic constraints on signal parameters, but most observers are convinced of an underlying detection and characterization with the display of a spectral line or the power versus spherical harmonic number that indicates a significant signal with respect to measurement errors. The corresponding quantity in the pulsar-GW problem is the cross-correlation between the timing residuals of pulsar pairs or some related quantity.

We derive a general expression for the signal-to-noise ratio \( \text{S/N} \) of a correlation-based detection statistic and develop a detection protocol based on the shape and amplitude of the cross-correlation function (CCF). We assess the challenges for a likely detection and estimate the minimum number of MSPs needed under different circumstances. To do so, we consider a hypothetical pulsar distribution that yields the highest possible \( \text{S/N} \) for the correlation function, all else being equal, namely, a configuration where all \( N_p \) pulsars in the PTA are in the same direction but at different distances. The GW timing perturbation comprises two terms: the “Earth” term that is correlated between different pulsars according to their angular separations and a “pulsar” term that is uncorrelated between objects for current timing data. The configuration we consider yields 100% correlation for the Earth terms but, of course, does not alter the pulsar terms. Any other configuration will yield smaller correlation and \( \text{S/N} \). We assess detectability in terms of different levels of white and red noise in timing residuals. Our main concern is with red spin noise though...
plasma propagation effects in the interstellar medium (ISM) also contribute.

Our work follows Cordes & Shannon (2010), where we assess a wide range of contributions to time of arrival (TOA) errors from the pulsar, the ISM, and from instrumental effects. White noise includes not only radiometer noise but also pulse-to-pulse phase jitter from magnetospheric activity and from an effect associated with interstellar scintillation, which are pulsar and line-of-sight (LOS) dependent, respectively. Red spin noise is common in canonical pulsars—those with periods of order 1 s and surface magnetic fields $\sim 10^{12} \text{ G}$—and is also to be expected in MSPs but at lower levels in accord with their spin parameters (Shannon & Cordes 2010). Interstellar scintillation also contributes red-noise TOA perturbations, some of which can be corrected before any analysis for GWs.

In the next section, we describe the cross-correlation analysis of a simplified timing model and develop detection criteria for assessing the presence of a GW signal. In Section 3, we describe how the prospects for detection can be maximized. We summarize our results and discuss them in broader terms in Section 4. The Appendix defines quantities used to characterize the timing residuals and describes simulations.

2. CROSS-CORRELATION-BASED GW DETECTION

We use a simplified model for the timing residuals that excludes real-world effects such as time transfer and the error in the location of the solar system barycenter (e.g., Backer & Hellings 1986). The residuals for a single pulsar are $R(t) = e(t) + p(t) + r(t)$, where $e(t)$ is the “Earth” part of the GW background and $u(t)$ includes all other processes that are uncorrelated between different pulsars. At minimum, $u(t)$ includes the GW perturbation $p(t)$ at the pulsar’s location acting on the measured signal a time $D/c$ earlier, where $D$ is the pulsar’s distance. In reality, $u$ also includes measurement errors and it is likely that red noise will also be present. We therefore write

$$u(t) = p(t) + r(t) + n(t),$$

where $r(t)$ is red noise associated with spin noise ("timing noise") in the pulsar or with multipath propagation in the ISM, and $n(t)$ is white noise that represents measurement errors of different kinds. All terms in $u$ are uncorrelated with each other as well as with $e$.

For a pair of pulsars separated by an angle $\theta$, the $e$ terms are correlated according to an angular GW correlation function $\zeta(\theta) \leq 1$ (Hellings & Downs 1983) with $\zeta(0) = 1$; this differs from other definitions in the literature that include a delta function associated with the uncorrelated $p$ terms. For our purposes, we keep the pulsar term separate.

As mentioned in Section 1, we consider a pulsar configuration that yields the maximum possible correlated signal in a PTA where all pulsars are in the same direction but at different (and unknown) distances. The $e$ terms are then identical for all LOSs. While actual PTAs can use pulsar configurations that are optimized in various respects (e.g., Burt et al. 2011; K. J. Lee 2011, private communication), they can yield no better a correlation than for the case we consider. Much more likely, the correlation will be less than what we calculate because correlation functions always maximize at zero lag.

Figure 1 (left panel) shows simulated time series for a 20-MSP PTA. Each row corresponds to one of the 20 pulsars. The rms of each term ($e$, $p$, $r$) is 20 ns. For the $e$ and $p$ terms, we have used the same rms value as appropriate for a stochastic GW background. The time series show the $e$ terms to be identical for all 20 pulsars, as assumed, while the $p$ and $r$ terms are different. The sum of all terms and the residuals from

![Figure 1](image-url)
The ratio of rms values ($\sigma_w/\sigma_p$) varies by more than an order of magnitude before fitting but covers a somewhat smaller range after fitting for red processes with $\alpha = 13/3$, the value expected from a GW background produced by merging SMBHs (Jaffe & Backer 2003). Histograms are also shown for pure white noise ($\alpha = 0$). In the Appendix, we define $\psi$ as the ratio of the variance of an individual realization to the ensemble-average variance. The ratio of rms values ($\sqrt{\psi}$) varies by more than an order of magnitude before fitting but covers a somewhat smaller range after fitting for red processes with $\alpha \geq 2$ compared to a very narrow range for white noise.

The model for $R$ is a sum of Gaussian distributed processes because all terms originate from conditions that usually will satisfy the central limit theorem (CLT). Conceivable exceptions include events from focusing in the ISM (Coles et al. 2010) or events intrinsic to the pulsar (e.g., glitches) that might induce non-Gaussian statistics for some objects. Assuming Gaussian statistics, the CCF estimate is a “sufficient” statistic (e.g., Box & Tiao 1973, pp. 60–63) for the true function. That is, it contains all the relevant information in the data needed to characterize the background. The estimator for the CCF between the $i$th and $j$th pulsars is an integral over the interval $[0, T]$

$$\hat{C}_{ij}(\tau) = T^{-1} \int_0^{T-|\tau|} dt \, R_i(t) \overline{R_j(t + \tau)}.$$  

This pairwise correlation, when averaged over all pairs into bins for the angular separation for each pair, yields the two-dimensional CCF $\hat{C}(\theta, \tau)$, where $\theta$ is the angular separation, as defined explicitly in the Appendix.

The temporal aspect of the CCF is an important discriminator between GW perturbations and those from additive noise. If GWs were to dominate all other contributions to the residuals, the temporal CCF would maximize at $\tau = 0$ in the limit of a large number of pulsar pairs. For a small number, the pulsar term ($p$) alone can shift the CCF away from the origin and red and white noise will exacerbate the shift. The left-hand panel of Figure 2 demonstrates this feature where CCFs are shown for 20 realizations of a 20-pulsar PTA. The left column of traces (one for each realization) shows cases where the rms GW, red-noise, and white-noise contributions are equal over the data span $T$, $\sigma_w = \sigma_p = \sigma_e = \sigma_n$. For all 20 cases, the CCF maximizes at or very close to $\tau = 0$. The right-hand set of traces shows CCFs where the GW contribution has been set to zero. In this case, the maximum correlation is generally at a lag $\tau \neq 0$ and the zero-lag value can be negative or positive, as expected because the CCF estimate is a sum over uncorrelated red- and white-noise signals. The ensemble mean of the CCF for the no-GW case is of course identically zero. These results suggest that a convincing

\footnote{The estimator for actual data would involve a sum over discrete samples; for our analysis it is more convenient to use continuous notation. It is straightforward to translate our results to those that would apply for discrete data, as described in the Appendix.}
GW detection must show a temporal correlation that maximizes near $\tau = 0$; otherwise, contamination by uncorrelated red noise is indicated. We discuss this criterion along with others in Section 2.2.

In the Appendix, we calculate the mean and variance of the correlation function and their variations over an ensemble. Much of our focus is on the maximum correlation $\hat{C}_{00} \equiv \hat{C}(\theta = 0, \tau = 0)$ from which we define the $S/N$ or significance as $S = \frac{(\hat{C}_{00})}{\sigma_{C}}$. For the pulsar configuration we consider, all pulsar pairs contribute to the $\theta = 0$ value of the correlation function. Figure 2 (right-hand panel) shows histograms of $\hat{C}_{00}$ obtained from simulations with and without a GW contribution and both before and after removing a parabolic fit to account for pulsar spin-down. Under relevant circumstances, the distribution is asymmetric, with a long tail for positive values while the mode and median are less than the mean. This counterintuitive result, discussed in the Appendix, occurs when the correlated quantity includes a red process with a steep power spectrum. The time series for a red process is effectively dominated by a small quantity includes a red process with a steep power spectrum. The result, discussed in the Appendix, occurs when the correlated quantity includes a red process with a steep power spectrum. The long tail in the distribution of $\hat{C}_{00}$ influences the detection probability and the false-alarm rate significantly. One way to get a more symmetric distribution is to average multiple estimates of the CCF. Multiple estimates can be obtained by sub-dividing timing residuals into $M$ blocks, each of length $T/M$, as mentioned in Shannon & Cordes (2010). The CFT will apply to the average so the distribution should tend to a Gaussian form. We discuss this further in the next section where we find that $M \approx 3–6$ is optimal under realistic circumstances.

2.1. Statistical Significance of the Zero-lag Cross-correlation

The zero-lag correlation function—when averaged over many pulsar pairs or an ensemble of noise terms—receives contributions only from the $e$ part of the timing perturbation so that $\hat{C}_{00} \rightarrow e^2$. We relate $e^2$, which is the variance of a single realization of $e(t)$, to the ensemble variance $\sigma_{gw}^2$ using $\sigma_{gw}^2 = \psi \sigma_{gw}^2$. From our previous discussion related to Figure 1 (right-hand panel), we recall that $\psi$ can vary considerably from an average value of unity. The statistical significance (i.e., the $S/N$) of the zero-lag correlation function is then

$$S = \frac{(\hat{C}_{00})}{\sigma_{C}} = \frac{\psi \sigma_{gw}^2}{\sigma_{C}},$$

where the rms variation $\sigma_{gw}^2$ includes contributions from the GWs themselves along with uncorrelated red and white noise. For arbitrary combinations of red and white noise and GWs, the full expression in Equation (A19) is lengthy because there are 12 non-vanishing terms out of the total of 42 for the second moment calculation. Also we have included smoothing of the data and analysis of the full time series of length $T$ in blocks of length $M$. All relevant quantities are defined in the Appendix.

Here, we discuss scaling laws for several limiting cases and their interpretation, leaving the detailed derivation of the full expression for $S$ in the Appendix. For cases discussed here we assume a large number of MSPs, $N_p \gg 1$, though the full expression is for an arbitrary number.

The GW-dominated regime includes only the perfectly correlated $e$ term (for our assumed configuration of pulsars) and the uncorrelated $p$ term, yielding

$$S \approx \frac{1}{2} \frac{\psi N_p M}{\omega_{gw}},$$

where $\omega_{gw} \sim 1$ is also a dimensionless timescale. Now $S$ grows as $\sqrt{N_p}$ and benefits from blocking in $M$ blocks, becoming arbitrarily large with $\sqrt{M}$ subject to the requirement that the continuum approximation holds for arbitrarily small $T/M$ (i.e., a sufficient number of samples in an interval $T/M$). When white noise contributes, however, there is a distinct maximum in $S$ versus $M$.

The opposite, noise-limited regime is where the GW variance is negligible compared to the total variance $\sigma_{gw}^2$ in the uncorrelated noise. The significance for this case is

$$S \approx \frac{N_p}{\sigma_{gw}} \left(\frac{\psi \sigma_{gw}}{\sigma_{gw}^2}\right)^2,$$

where $\omega_{gw}$ is a dimensionless timescale defined in the Appendix. In this regime, no matter whether the dominant uncorrelated term is red or white noise, the significance scales linearly with the number of pulsars but with a coefficient that is much smaller than the number of pulsars but with a coefficient that is much smaller than 1 (by definition) $\sigma_{gw}^2/\sigma_{gw}^2 \ll 1$. For the specific case where $u$ comprises only white noise, the scaling is

$$S \approx \frac{N_p \sqrt{N_t N_s}}{\sqrt{2}} \left(\frac{\psi \sigma_{gw}}{\sigma_{gw}^2}\right).$$

The significance clearly improves with a larger number of TOAs, $N_t$, and from smoothing by $N_s$ samples. Any reduction in the rms white noise ($\sigma_{gw}$) through greater telescope sensitivity and other measures also improves the significance. Blocking of the data in this regime actually degrades the significance because the GW variance is evaluated as $\sigma_{gw}(T/M) \leq \sigma_{gw}(T)$.

Some general conclusions follow from the above extreme cases and inspection of the full expression for $S$ in Equation (A19), which includes terms that scale as $C(1), 1/N_p, 1/N_s$, and $1/(N_p N_s)$. The number of TOAs, $N_t$, is important only as long as the white-noise part of the residuals is sizable. If less than other terms, the number of TOAs—and thus any cadence in acquiring them—becomes unimportant. A larger number of pulsars can reduce the effects of both the red and white noise from non-GW contributions in addition to reducing the uncorrelated pulsar part of the GWs. For very large $N_t$ and $N_p$, $S$ reduces to that for the GW-only case.

We illustrate these trends in Figure 3. In the left panel, the significance $S$ is plotted against the number of blocks for cases that include red and white noise added to the GW perturbation. The plotted values are based on $S = 1$ and on a total of $N_t = 10^7$ TOAs over 5 years for 20 pulsars. After a second-order fit to a $T = 5$ yr data span, the rms of the $e$ or $p$ term is $\sigma_{gw}(T) \approx 20$ ns, as is consistent with a GW spectrum having $A = 10^{-15} \text{ yr}^{-1/3}$, (e.g., Shannon & Cordes 2010). The labeled values for the rms red noise also refer to a five-year data span. The rms GW term...
scales as $T^{x_r}$ with $x_r = 5/3$, while the rms red-noise term scales with an exponent $x_r = 2$ based on the work of, e.g., Shannon & Cordes (2010).

The curve for GWs only (no red or white noise) increases monotonically with $M$, but for curves with non-zero noise there is a distinct maximum in $S$ that separates the GW-dominated and noise-dominated regimes. The optimal number of blocks is $M \approx 3$–6 for the cases shown. In the noise-limited regime, $S$ scales as $\psi N_p \sqrt{N_t N_r} M^{-2x_r}$, so smoothing improves $S$ but blocking degrades it. When not white noise limited, $S$ no longer depends on $N_t$, so smoothing has no effect. This may be seen in Figure 3 (left panel) for the curves labeled 20, 50, and 100 ns, which converge at large $S$ for both smoothing values shown, $N_r = 1$ and 30. In the GW-dominated case, there is no dependence on $N_t, N_r$, or on the scaling exponent, $x_r$.

Figure 3 (right panel) shows $S$ plotted against the number of pulsars for several different values of red and white noise. The curves in the figure were calculated for $M = 3$ blocks.

If red noise is absent, $S \propto \sqrt{N_t N_p}$ is linear in the number of pulsars when $S \ll 1$ but has a shallower dependence $S \propto \sqrt{N_p}$ when $S$ is large enough to provide a confident detection. If we let $S_{pe}$ be the value for the GW-only case of Equation (4), it can be shown that $S \ll S_{pe}$ in the white-noise-limited case. Thus, for $S$ large enough to correspond to a plausible detection, it is not likely to scale with $N_p$ as it does in the noise-limited regime but instead will scale more slowly with $N_p$.

### 2.2. Detection Criteria

Over an ensemble, the CCF vanishes unless there is a significant correlated term from the $e(t)$ term. However, deviations from ensemble-average statistics in data sets that are finite in both the data span $T$ and the number of pulsars $N_p$ will produce false positive and false negative detections from the uncorrelated terms in the timing residuals, like those shown in Figure 1.

A detection protocol for the GW background can exploit the following aspects of timing residuals and their correlations in the presence of GWs:

1. The timing residuals must include contributions from one or more of the predicted isotropic GW backgrounds that typically will appear as a stochastic process with a red power spectrum (Jenet et al. 2005). If GWs from any discrete source are significant, there should be a corresponding departure from white-noise statistics described by a spectrum that depends on the nature of the single-source emission (Lommen & Backer 2001).

2. The GW signal should appear consistently in the residuals for all pulsars in a PTA but with an expected large rms variation between objects because of the realization-to-realization volatility expected for a steep GW spectrum. While the $e$ terms represent the same realization, the pulsar term is different for each pulsar.

3. The maximum of the CCF should be at or near zero time lag, $\tau \approx 0$, depending on how strong the GWs are relative to red-noise contributions. Uncorrelated contributions produce estimation errors in $\tilde{C}_{00}$ that peak at arbitrary time lags in estimates using a finite number of pulsars and thus can induce false non-detections and false positive detections.

4. The zero-lag amplitude of the CCF must be significantly larger than expected when only uncorrelated terms contribute to the time series.

5. The correlation versus angular separation between pulsars should be consistent with that expected (e.g., Equation (5) of Hellings & Downs 1983) for an isotropic background or the equivalent angular correlation for a discrete source.

---

Footnote 4: We note that errors in the location of the solar system barycenter or from instrumentation that affect the timing residuals from all pulsars would also induce a non-zero correlation.
6. Any correlation established using one set of pulsars can be checked using a completely independent set of pulsars.

When white- and red-noise processes are significant, the estimated CCF has a high probability of peaking at \( \tau \neq 0 \). If white noise dominates the timing residuals of all pulsars, the CCF itself will vary rapidly with time lag and its formal maximum could be at any lag. Red noise by definition has a long correlation time so the CCF will appear quite smooth and most likely peak far away from the origin. After a second-order polynomial fit, the red-noise residuals will typically have two zero crossings so the CCF maximum is likely to be more centered than the pre-fit CCF, an effect we see in simulations.

### 2.3. Detection Metrics

The detection significance \( S \) is the ratio of the ensemble-average mean to the rms of the zero-lag cross-correlation. It is easily related to the spectrum for GWs that also characterizes the ensemble, but it cannot be estimated from limited data obtained on a finite number of pulsars except perhaps when the situation is one of the extremes we considered earlier (white noise limited or GW dominated). Here, we characterize the estimated correlation function using empirical metrics and then define a detection protocol. We present these metrics as useful examples. Through simulations we then relate the implied detection probability and false-alarm probability to the significance \( S \).

We define two metrics that characterize the shape of the CCF,

\[
m_1 = \frac{\tilde{C}_{00}}{\tilde{C}_{\text{min}}}, \quad m_2 = \frac{\tau_{C,\text{max}}}{\tau_{\text{max}}};
\]

\( m_1 \) is the ratio of the CCF at zero lag to the most negative value; \( m_2 \) is the offset of the CCF maximum relative to the maximum calculated lag, which in our simulation is \( \tau_{\text{max}} = T/2 \).

The two metrics characterize the shape of the CCF to provide direct empirical measures for assessing the presence of a GW signal. The metric \( m_1 \) is an amplitude measure similar to \( S \) but is based on a single realization of the correlation estimate and is normalized by the minimum of the particular CCF, not the rms value over an ensemble. Figure 4 shows a simulation-based scatter plot of \( m_1 \) and \( m_2 \) that displays a peak in \( m_1 \) near \( \tau = 0 \). The fraction of points in the peak depends on the strength of the GWs compared to other contributions and on the number of pulsars. For residuals dominated by uncorrelated red noise, there is a sizable fraction (\( \sim 15\% \)) of cases with peak values that can mimic a GW detection, while the remainder are spread over \( |m_2| \leq 1 \). We define a joint detection criterion that comprises a lower bound \( m_1 > m_{1}\text{min} \) and an upper bound \( |m_2| \leq m_{2}\text{max} \).

We also define the detection fraction \( f_d \) as the fraction of PTA realizations in a simulation that satisfy the detection criteria. The corresponding false-alarm fraction is defined as the detection fraction in the absence of any GW signal.

Inspection of Figure 4 suggests a threshold \( m_{1}\text{min} = 1 \), which is plausible since the CCF of a noise-only signal is likely to have approximately equal positive- and negative-going excursions. Also consistent with the figure is a threshold \( m_{2}\text{max} = 0.1 \), which enforces the zero time-lag nature of the “Earth” part of the GW signal but rejects cases where noise processes steer the CCF maximum away from zero lag. Figure 2 shows cases where the CCFs satisfy the detection criteria (large amplitude and peak at \( \tau = 0 \)) and others that do not. Conversely, when there is no GW contribution, red noise can cause false positive cases that satisfy the criteria. We quantify these features in the discussion that follows.

![Figure 4](image_url)  
**Figure 4.** Plot of metrics \( m_1 \) vs. \( m_2 \) that characterize the shape of the CCF of the post-fit timing residuals, as defined in the text; points are shown for 100 realizations of each PTA, defined as the number of pulsars, the rms red noise, and the rms white noise, as labeled. The dashed lines denote the minimum threshold for the vertical axis and the maximum departure of the CCF peak from zero lag in order to provide a plausible detection. The open circles are for a case with no-GW contribution, while the open squares and filled circles have a post-fit GW rms of 20 ns in each time series.

Specifying a detection criterion and corresponding significance requires consideration of the tradeoff between the detection fraction, \( f_d \), and the false-alarm fraction, \( f_{fa} \). Figure 5 (left panel) shows “receiver operating characteristic” (ROC) curves calculated by varying \( m_{1}\text{min} \) and \( m_{2}\text{max} \) to alter the detection and false-alarm fractions. Each curve corresponds to a particular PTA (number of pulsars and levels of red and white noise). Ideally, one would like to have 100% detection fraction with no false alarms. The cases \( (N_p, \sigma_r, \sigma_n) = (50, 10 \text{ ns}, 20 \text{ ns}) \) and \( (N_p, \sigma_r, \sigma_n) = (20, 0 \text{ ns}, 100 \text{ ns}) \) come closest to this ideal. All of the other cases, which have larger noise levels or smaller numbers of pulsars in the PTA, depart significantly from the ideal. With 20 pulsars having 20 ns of red noise and negligible white noise, a 90% detection fraction comes at the expense of an 11% false-alarm fraction. A larger number of pulsars (such as 50 pulsars with 20 ns red noise or 100 pulsars with 50 ns of red noise) decrease the false-alarm fraction to 8%.

Our method uses the CCF as a test statistic that we characterize using the \( m_1-m_2 \) metrics that are required to satisfy certain conditions. This protocol needs to be related to the significance \( S \) that in turn can be related to theoretical GW spectra (i.e., ensemble averages) and sources of noise. The mapping of significance \( S \) to detection fraction is shown in the right-hand panel of Figure 5. Most of the curves shown are for equal levels of red and white noise and different pulsar numbers (solid curves). Detection fractions \( f_d \geq 0.8 \) require \( S \geq 1.5 \) and \( f_d \geq 0.95 \) requires \( S \geq 2 \). A case with 100 ns white noise (dashed curve) shows that \( N_p = 20 \) pulsars yields \( S \approx 3 \) and a detection fraction \( >80\% \) and minimal false-alarm fraction as shown in the left-hand panel. Our results are therefore broadly consistent with those of Jenet et al. (2005), who consider only white-noise
TOA errors. The primary conclusion from Figure 5 is that red noise drastically alters the detection and false-alarm fractions and therefore also any assessment of GW detectability.

Figure 6 shows how the required significance places joint constraints on the amount of red and white noise. Contours are shown for $S = 2$ and $3$ vs. the ratios $\sigma_n/\sigma_{gw}$ and $\sigma_r/\sigma_{gw}$ for PTAs with $N_p = 20$ pulsars (dashed lines) and $N_p = 50$ pulsars (solid lines). The left-hand panel shows results for no blocking ($M = 1$), and the right-hand panel is for three blocks, $M = 3$. The results apply for $N_t = 10^3$ arrival times obtained over $T = 5$ yr.

Figure 5. Left: ROC curves showing detection fraction vs. false-alarm fraction for different detection criteria obtained by varying the thresholds for $m_1$ and $m_2$. The detection fractions were calculated for the same GW strength (20 ns over 5 yr) but different numbers of pulsars and for different levels of red and white noise, as labeled. The false-alarm fraction was obtained by turning off the GW signal and keeping the red- and white-noise levels the same as for the GW “on” case. Right: detection fraction plotted against the expected significance $S$ of the CCF. The five points for each curve correspond to PTAs with $N_p = 4, 8, 20, 50$, and 100 pulsars. Solid lines: the rms red noise and white noise are equal as labeled for each curve near the point corresponding to $N_p = 4$. Dashed line: a case with white noise with 100 ns rms (no red noise) added to the GW signal.

Figure 6. Contours of fixed significance $S = 2$ and 3 vs. the ratios $\sigma_n/\sigma_{gw}$ and $\sigma_r/\sigma_{gw}$ for PTAs with $N_p = 20$ pulsars (dashed lines) and $N_p = 50$ pulsars (solid lines). The left-hand panel shows results for no blocking ($M = 1$), and the right-hand panel is for three blocks, $M = 3$. The results apply for $N_t = 10^3$ arrival times obtained over $T = 5$ yr.

2.4. Comparison with Other Detection Approaches

Our expression for $S$ is similar to Equation (12) of Jenet et al. (2005), who define their detection significance as the $S/N S_f$ of a weighted correlation quantity, as discussed in the Appendix. However, their expression does not explicitly account for red noise.
and white noise individually; instead, a quantity $\chi$ is defined that measures the degree of whiteness of the timing residuals. Any non-white components are assumed to arise solely from the GW background and not from any additional spin-noise or ISM contribution. As a consequence, we do not expect the GW background and not from any additional spin-noise or ISM contribution. As a consequence, we do not expect the

This value is sufficient to yield a

In the regime where detections can be made, the

For the same cases, we find

Increasing the number of pulsars helps in any regime, though it has greater impact in the noise-limited case (red or white) where the detection significance $S$ is linear in $N_p$. Detection of GWs almost certainly will occur in the regime where there is a shallower dependence of $S$ on $N_p$. In the extreme case where the residuals are GW dominated, $S$ increases only as the square root of the number of pulsars.

In the white-noise-limited regime, increasing $Z_{wn}$ through a combination of more pulsars, greater timing throughput, smoothing, and a decrease in timing error per TOA will increase $S$.

Because red noise is spectrally similar to the GW contribution and does not average out significantly in the CCF because of its long correlation time, the primary means for increasing $S$ is to sum over many pulsars and to use blocking.

In the GW-dominated regime, increasing the blocking $M$ and the number of pulsars $N_p$ are the only options. Figure 3 (left-hand panel) shows that $M$ cannot be increased arbitrarily because eventually the rms noise in the interval $T/M$ will overwhelm the GW signal, which decreases with smaller $T$.

The sensitivity of a PTA to GWs of course improves with total observing span. In the white-noise-dominated case, these improvements are the greatest, with $S \propto T^{2+\epsilon/1+2} \propto T^{23/6}$ for the SMBH-produced GW spectrum, where we assume that the TOAs are obtained with fixed cadence, i.e., $N_t \propto T$. If data are smoothed over a time proportional to $T$, the significance increases even faster with data span, $S \propto T^{13/5}$. However, a detection is not likely in this regime because $S$ is small by definition. In the regime where detections can be made, the longer observing span only enables a larger amount of sub-blocking so that $S \propto \sqrt{T}$.

The most difficult hindrance to overcome is the presence of red noise from spin variations and from any residual ISM effects that cannot be corrected. The range of red timing noise levels in MSPs is not known definitively, but our recent assessment (Shannon & Cordes 2010) suggests that it is larger in objects with larger spin–frequency derivatives. Latent red noise may emerge in many MSPs when more sensitive (i.e., lower white noise) and longer time-span observations are obtained. The same can be said for the correlated GW signal. Greater timing throughput is needed in order to make it possible to detect the GWs. This would allow more MSPs to be timed but also increase the observing time per pulsar so that white noise from interstellar refraction and diffraction can be mitigated by using higher frequencies and by appropriate fitting across wide bandwidths. Larger telescopes and bandwidths can minimize radiometer noise but will have no effect on achromatic pulse phase jitter. Longer integration times are the only recourse for reducing jitter, and they also will minimize radiometer noise.

The most difficult hindrance to overcome is the presence of red noise from spin variations and from any residual ISM effects that cannot be corrected. The range of red timing noise levels in MSPs is not known definitively, but our recent assessment (Shannon & Cordes 2010) suggests that it is larger in objects with larger spin–frequency derivatives. Latent red noise may emerge in many MSPs when more sensitive (i.e., lower white noise) and longer time-span observations are obtained. The same can be said for the correlated GW signal. Greater timing throughput is needed in order to make it possible to detect the GWs. This would allow more MSPs to be timed but also increase the observing time per pulsar so that white noise from interstellar refraction and diffraction can be mitigated by using higher frequencies and by appropriate fitting across wide bandwidths. Larger telescopes and bandwidths can minimize radiometer noise but will have no effect on achromatic pulse phase jitter. Longer integration times are the only recourse for reducing jitter, and they also will minimize radiometer noise.

The most difficult hindrance to overcome is the presence of red noise from spin variations and from any residual ISM effects that cannot be corrected. The range of red timing noise levels in MSPs is not known definitively, but our recent assessment (Shannon & Cordes 2010) suggests that it is larger in objects with larger spin–frequency derivatives. Latent red noise may emerge in many MSPs when more sensitive (i.e., lower white noise) and longer time-span observations are obtained. The same can be said for the correlated GW signal. Greater timing throughput is needed in order to make it possible to detect the GWs. This would allow more MSPs to be timed but also increase the observing time per pulsar so that white noise from interstellar refraction and diffraction can be mitigated by using higher frequencies and by appropriate fitting across wide bandwidths. Larger telescopes and bandwidths can minimize radiometer noise but will have no effect on achromatic pulse phase jitter. Longer integration times are the only recourse for reducing jitter, and they also will minimize radiometer noise.

The most difficult hindrance to overcome is the presence of red noise from spin variations and from any residual ISM effects that cannot be corrected. The range of red timing noise levels in MSPs is not known definitively, but our recent assessment (Shannon & Cordes 2010) suggests that it is larger in objects with larger spin–frequency derivatives. Latent red noise may emerge in many MSPs when more sensitive (i.e., lower white noise) and longer time-span observations are obtained. The same can be said for the correlated GW signal. Greater timing throughput is needed in order to make it possible to detect the GWs. This would allow more MSPs to be timed but also increase the observing time per pulsar so that white noise from interstellar refraction and diffraction can be mitigated by using higher frequencies and by appropriate fitting across wide bandwidths. Larger telescopes and bandwidths can minimize radiometer noise but will have no effect on achromatic pulse phase jitter. Longer integration times are the only recourse for reducing jitter, and they also will minimize radiometer noise.

The most difficult hindrance to overcome is the presence of red noise from spin variations and from any residual ISM effects that cannot be corrected. The range of red timing noise levels in MSPs is not known definitively, but our recent assessment (Shannon & Cordes 2010) suggests that it is larger in objects with larger spin–frequency derivatives. Latent red noise may emerge in many MSPs when more sensitive (i.e., lower white noise) and longer time-span observations are obtained. The same can be said for the correlated GW signal. Greater timing throughput is needed in order to make it possible to detect the GWs. This would allow more MSPs to be timed but also increase the observing time per pulsar so that white noise from interstellar refraction and diffraction can be mitigated by using higher frequencies and by appropriate fitting across wide bandwidths. Larger telescopes and bandwidths can minimize radiometer noise but will have no effect on achromatic pulse phase jitter. Longer integration times are the only recourse for reducing jitter, and they also will minimize radiometer noise.
For $\psi = 1$, $M = 1$, and $w_{pp} = 0.4$ a minimum $S$ of 2 requires $N_p = 6$ pulsars, as seen in Figure 3. With no timing noise or white noise, $M$ can be made arbitrarily large (subject to sampling rates). However, even optimistic values of white noise and red noise (e.g., 20 ns for each) indicate that $S$ is maximized for $M \approx 2–3$, so $N_p \gtrsim 7/\psi$ is needed for $M = 3$. The actual variance of the “Earth” part of the GW background could have $\psi$ much larger or smaller than unity, so detection of GWs with a small number of pulsars may be marginal. A larger threshold $S_{\text{min}} = 5$ will require $N_p \gtrsim 25/\psi$ for $M = 3$. Our results suggest that an increase in the cadence of timing measurements to increase the total number of TOAs for each object ($N$) may be a necessary but insufficient course to take for GW detection.

To obtain our results, we have assumed that the Earth term $e(t)$ is the same for all pulsars in a hypothetical spatial configuration. For realistic distributions of pulsars on the sky, the correlation amplitude is reduced by $\overline{\xi}^2 \approx 0.6$, thus increasing the number of required pulsars by about a factor of $1/(\overline{\xi}^2)^2 \approx 3$. We obtain $\xi^2 \approx 0.6$ by summing the square of the Hellings and Downs curve over pulsar sky distributions obtained by Monte Carlo. For PTAs with a range of red- and white-noise levels, weights for each pulsar can be introduced in the double sum in Equation (A2). For a nominal level of white noise, for example, with some objects having smaller and others larger rms values, a weighted sum will yield a larger $S/N$ than the case where all objects have the same rms white noise. If we take the nominal value as the minimum in the sample, however, the optimally weighted $C$ will have lower $S/N$.

Currently, only two MSPs (J1713+0747, J1909–3744) are known to have rms timing residuals less than 40 ns over a five-year interval (Demorest et al. 2012) and one other less than 100 ns, J0437–4715 (Manchester 2010). An aggressive campaign is needed to find more MSPs with timing noise substantially less than 100 ns in a five-year span. The timing noise scaling law of Shannon & Cordes (2010) suggests that such MSPs will have small spin-down rates and may be less luminous if the radio beam luminosity correlates with the energy loss rate. This implies that MSP surveys may need to be more sensitive than at present.

The best strategy is to identify $\sim 20$ “super”-stable MSPs with rms timing noise of 20 ns or less over time spans of 5 years or more if a detection threshold $S_{\text{min}} = 2$ is considered sufficient for detection. A larger $S_{\text{min}} = 5$ requires $\sim 50$ objects. However, if no such super-stable objects exist and MSPs more typically have 20 ns rms timing noise or larger over 5 years, many more MSPs will need to be timed, perhaps exceeding 100 MSPs. Even if the super-stable regime applies, once a detection is made, possibly using existing telescopes to time known stable objects along with any new discoveries in the near term, a more detailed analysis of the GW spectrum will be desired, and that certainly will require a much larger set of MSPs and overall greater throughput of the timing program.

Each MSP needs a careful error budget analysis. This would include a detailed characterization of the red- and white-noise levels, including a breakdown of each from different physical causes. The two kinds of noise can be distinguished through appropriate use of the structure function of timing residuals (e.g., Cordes & Downs 1985). Departures from white noise need to be characterized according to amplitude and spectrum and classified as contributions from red noise of any kind and from changes in instrumentation, which can cause jumps in pulse phase between epochs. A change-point analysis (e.g., O’Ruanaidh & Fitzgerald 1996, chapter 5) on timing residuals can identify the amplitudes of such jumps whether or not their occurrence epochs are known. MSPs with significant red noise that is demonstrated to be from non-GW causes should be rejected because they do not contribute to the sensitivity of a PTA to a stochastic background of GWs.

4. DISCUSSION

The main results of our paper are as follows.

The CCF is the primary statistic that we consider for a hypothetical pulsar distribution that yields the maximum possible $S/N$ for the stochastic GW background, all else being equal. For a realistic distribution, the equivalent quantity would be a weighted sum similar to the quantity $\rho$ defined by Jenet et al. (2005), but with a time-lag argument included. We have shown that the correlated GW signal contributes variance that can differ markedly from the ensemble value if the stochastic GW signal has a steep power-law spectrum, like that expected from mergers of SMBHs. This stochasticity of the sample variance can either greatly enhance or diminish the chances of detecting the signal. The CCF has amplitudes over an ensemble that have a positive skewed distribution that influences the detection and false-alarm fractions. We have taken this into account in our analysis, and we also suggest that sub-dividing the entire span of timing residuals into $M \approx 3$ sub-spans for each pulsar will reduce the skewness and increase the statistical significance. Such blocking is equivalent to high-pass filtering the data. It requires a large enough cadence for TOA measurements that there are an ample number of samples within each interval of length $T/M$.

The number of pulsars needed for a likely detection of a stochastic background of nanohertz GWs depends strongly on the levels of white noise and especially the red noise in the data. In related papers (Cordes & Shannon 2010), we have shown that both kinds of noise are likely to be present due to torque noise in the pulsar, magnetospheric motions of emission regions, and interstellar plasma phenomena. Red- and white-noise timing residuals dramatically alter the achievable $S/N$ of the CCF. When residuals are white noise dominated, improvements can be made by increasing the net integration time per pulsar and by smoothing individual measurements over a time shorter than the smallest GW period that is likely to be identified. Marginal gains can also be made from blocking of the data.

If, however, the detection statistic is dominated by red noise with a power spectrum similar to that of the stochastic GW power spectrum, smoothing or other increases in net integration time per TOA will not help. The best recourse is to increase the number of pulsars in the PTA.

A minimum of seven MSPs is needed for a plausible detection under optimistic red- and white-noise levels, as described in the previous section. A pulsar sample that is distributed on the sky will increase this number by a factor of $\sim 3$, and verification with a completely independent sample will require another doubling. The program going forward therefore requires an aggressive search campaign to discover more MSPs and to identify the most spin-stable objects. In parallel, it is important to characterize the timing error budget for each MSP. It is possible that the most stable objects are also those with smaller radio luminosities. The scaling law for red noise identified by Shannon & Cordes (2010) implies that objects with larger spin-down rates have larger...
noise levels. While not known for certain, the radio luminosity likely also is larger for these objects. Further study of spin noise in MSPs is needed to ascertain whether it can be mitigated, as suggested by Lyne et al. (2010). We consider it likely that 50–100 spin-stable MSPs will be needed for a comprehensive pulsar timing program that fully characterizes the stochastic background of GWs at nanohertz frequencies by detailed study of the GW spectrum after a secure detection has been made.

We thank members of the NANOGrav collaboration for useful comments on this work. We also thank an anonymous referee for many suggestions for making the presentation clearer. Our work was supported by the NSF through a subaward to Cornell University from Partnerships for International Research and Education award 0968296 to West Virginia University.

APPENDIX

PROPERTIES OF A CORRELATION-BASED DETECTION STATISTIC

We use the signal model $R(t) = e(t) + u(t)$ mentioned in Section 2 of the main text, where $e$ is the “Earth” part of GW timing perturbation that is correlated between different LOSs, and $u = p + r + n$ includes uncorrelated perturbations comprising the “pulsar” GW perturbation ($p$), red noise ($r$) from spin variations and the ISM, and white noise ($n$) from radiometer noise, pulse jitter, and diffractive interstellar scintillations. While the $e$ and $p$ terms from the stochastic background produced by SMBH mergers have red power spectra, here and in the main text we reserve the term “red noise” for the $r$ term.

A.1. Correlation Function Definitions

The CCF between the $i$th and $j$th pulsars is estimated by an integral over the interval $[0, T]$

$$\hat{C}_{ij}(\tau) = T^{-1} \int_0^{T-|\tau|} dt \ R_i(t) R_j(t + \tau).$$  \hspace{1cm} (A1)

The pairwise correlation can be summed over all pairs into angular bins for the angular separation

$$\hat{C}(\theta, \tau) = \frac{1}{N_\theta(\theta)} \sum_{i < j} \sum_{\theta_i, \theta_j - \theta} \frac{1}{T} \int_0^{T-|\tau|} dt \ x_i(t) x_j(t + \tau),$$  \hspace{1cm} (A2)

where the sums are for pulsar pairs with angular separations in an angular bin centered on $\theta$ and $N_\theta(\theta)$ is the number of pairs within the bin. In practice, there is a choice of normalizing coefficient between $1/T$ as shown or $1/(T - |\tau|)$. This issue is secondary to our discussion, which focuses on the $\tau = 0$ case and on the S/N of the CCF where leading constants cancel out.

To calculate the ensemble average of the mean and second moment of $\hat{C}$, we define the temporal correlation function for $e$ as

$$\langle e(t)e(t') \rangle \equiv \sigma_e^2(T) \rho_e(t, t'),$$

where $\sigma_e^2(T) \equiv \frac{1}{T} \int_0^T dt \ \langle e^2(t) \rangle$.  \hspace{1cm} (A3)

With these definitions, $\rho_e(t, t')$ is a dimensionless function having unit area over the interval $[0, T]$. Thus, while $\langle e(t)e(t') \rangle$ is independent of the particular interval, the factorization is in terms of quantities that depend on the length of the interval, as is convenient for calculations. This form makes explicit the non-stationary aspect of the GW perturbation that, for realistic values of $T$, has a variance that depends on the length of the time series$^5$ and a normalized correlation function that depends on both $t$ and $t'$. If the process has stationary statistics, $\sigma_e^2$ is independent of $T$ and $\rho_e$ depends only on the time difference $t - t'$. Similar definitions for $\sigma_u^2(T)$ and $\rho_u(t, t')$ have the same properties.

Later we will also need the dimensionless cross-correlation time defined as

$$w_{uu} = T^{-2} \int_0^T dt dt' \ \rho_u^2(t, t').$$  \hspace{1cm} (A4)

and the dimensionless cross-correlation time for $e$ and $u$

$$w_{eu} = T^{-2} \int_0^T dt dt' \ \rho_e(t, t') \rho_u(t, t').$$  \hspace{1cm} (A5)

For a pair of pulsars the cross-correlation becomes

$$\langle e_i(t) e_j(t') \rangle = \sigma_e^2(T) \rho_e(t, t') \xi_{ij},$$  \hspace{1cm} (A6)

where $\xi_{ij} = \xi(\theta_{ij})$ is the Hellings–Downs angular correlation with the normalization $\xi(0) = 1$. A similar definition is used for the cross-correlation of $u$ except that $\xi_{ij}$ is replaced with a Kronecker delta,

$$\langle u_i(t) u_j(t') \rangle = \sigma_u^2(T) \rho_u(t, t') \delta_{ij}.$$  \hspace{1cm} (A7)

The ensemble mean of the correlation estimate is (for narrow bins in $\theta$)

$$\langle \hat{C}(\theta, \tau) \rangle = \sigma_e^2(T) \xi(\theta) T^{-1} \int_0^T dt \ \rho_e(t, t + \tau).$$  \hspace{1cm} (A8)

A.2. Estimation Error for $\hat{C}_{00}$

We want to know the statistical significance of $\hat{C}$. Over an ensemble, its maximum is at zero time lag and zero angular lag,

$$\langle \hat{C}_{00} \rangle = \langle \hat{C}(0, 0) \rangle = \sigma_e^2(T),$$  \hspace{1cm} (A9)

so it suffices to consider $\hat{C}_{00}$ when addressing the minimal requirements for detection. To calculate the estimation error for $C_{00}$, we need the mean square

$$\langle \hat{C}_{00}^2 \rangle = N_X^{-2} \sum_{i < j} \sum_{i' < j'} \sigma_e^4(T) \int_0^T \int_0^T dt dt' \ 
\times \langle R_i(t) R_i(t') R_j(t') R_j(t) \rangle,$$  \hspace{2cm} (A10)

where now $N_X = N_p(N_p - 1)/2$ is the number of unique pairs of pulsars for the extreme (and best) case we consider, namely, where all pulsars are in the same direction. Formally, there are $4^4$ terms in the fourfold summation, but most terms are cross products that vanish in the ensemble average. To evaluate the second moment, we take into account the following:

1. Assume that $R(t)$ is a Gaussian random process so that the fourth moment can be expanded into products of second moments.

$^5$ This assumes that the spectrum for the GW perturbation is very broad and increases rapidly with decreasing frequency. If its lower cutoff frequency $f_{\text{min}} \ll 1/T$, then as $T$ gets longer, the variance in the time series will grow as some power of $T$. 

The Astrophysical Journal, 750:89 (13pp), 2012 May 10

Cordes & Shannon
2. For the term involving only the Earth perturbation, \( \langle e^2(t)e^2(t') \rangle \) formally expands into three terms. However, to model the observational situation, we consider \( e(t) \) to represent a single realization of the random process, so we do not want to include all the variance in the three terms. Instead, we use only the term \( \langle e^2(t)e^2(t') \rangle \rightarrow \langle e^2(t) \rangle \langle e^2(t') \rangle \) because it will yield the square of the mean correlation estimate, \( \langle \xi_0 \rangle^2 \). The other two terms quantify the ensemble variance of \( e \) that we wish to exclude for the reasons stated.

3. The expectation of the fourth product of \( u, \langle u_i(t)u_j(t')u_j(t')u_j(t') \rangle \), has only two non-zero terms because the third involves second moments that require \( i = j \) and \( i' = j' \), which are excluded since the sums only include terms with \( i < j \) and \( i' < j' \). Therefore,

\[
\langle u_i(t)u_j(t')u_j(t')u_j(t') \rangle = \langle u_i(t)u_j(t') \rangle \langle u_j(t')u_j(t') \rangle = \sigma_u^4(T)\rho_u(T, t')\delta_{ij}\delta_{ij'} + \delta_{ij'}\delta_{ij}'. \tag{A11}
\]

For simplicity of our bottom-line expression, we have assumed that the variance is the same for all LOSs. Relaxing this assumption allows a less useful expression for the goals of this paper, which are to discuss how GW detectability depends on the mixture of red and white noise. The fourfold sum of the delta functions over the four indices yields \( N_p \).

4. Similarly, the cross terms involving the second moments of \( e \) and \( u \):

\[
\langle e(t)e(t') \rangle \left[ \langle u_i(t)u_j(t') \rangle + \langle u_i(t)u_j(t') \rangle + \langle u_i(t)u_j(t') \rangle \right] + \langle u_j(t)u_j(t') \rangle = \sigma_u^2(T)\sigma_u^2(T)\rho_u(t, t')\rho_u(t', t') \times (\delta_{ij} + \delta_{ij'} + \delta_{ij'} + \delta_{ij}). \tag{A12}
\]

The summation of the delta functions over all indices yields \( 2N_p(N_p - 1) \).

5. The ensemble average of Equation (A2) involves double time integrals of the product or square of normalized correlation functions. Combined with the leading \( T^{-2} \) factor, the integrals of the different terms yield the dimensionless correlation times defined above. For a stationary process, \( \omega_{uu} \) would be equal to the correlation time divided by \( T \).

For the white-noise contribution to \( u \), for example, the correlation time should be equated to the sample interval and \( \omega_{uu} \sim 1/N_i \), where \( N_i \) is the number of discrete samples. For a large number of samples, \( \omega_{uu} \ll 1 \) as expected for the large number of degrees of freedom in the white noise. For non-stationary red noise, however, \( \omega_{uu} \) is of order unity because the variance is dominated by fluctuations with characteristic times of order \( T \) and thus a small number of degrees of freedom.

6. The Earth term \( e(t) \) represents a single realization from an ensemble and therefore can have a sample variance that differs significantly from the ensemble variance. The pulsar terms in the GW perturbation (included in \( u \)) sample \( N_p \) different realizations and thus yield sums that are more representative of the ensemble variance. However, the \( p \) term for a particular pulsar may have variance that differs significantly from that of the ensemble. Assuming \( N_p \gg 1 \), we introduce a fudge factor \( \psi \) and write \( \sigma_u^2(T) = \psi \sigma_u^2(T) \), where \( \psi \) has unit mean but can be much larger or smaller than unity for any given realization, particularly so for red processes with steep power spectra (see the main text and Figure 1).

Assembling these results and (again) associating the fourth moment of \( e \) with the square of \( \langle \xi_0 \rangle \), we obtain the correlation variance

\[
\sigma_c^2 \equiv \langle \xi_0^2 \rangle - \langle \xi_0 \rangle^2 = \frac{4}{N_p} \left[ \sigma_u^2(T)\sigma_u^2(T)\omega_{uu} + \sigma_u^2(T)\omega_{uu} \right]. \tag{A13}
\]

A.3. Statistical Significance of the Correlation Estimate

We define the significance of the correlation function as the ensemble-average \( S/N \)

\[
S \equiv \langle \xi_0 \rangle = \frac{\sigma_u^2(T)}{2N_p^{1/2}} \left[ \sigma_u^2(T)\sigma_u^2(T)\omega_{uu} + \sigma_u^2(T)\omega_{uu} \right]^{1/2}. \tag{A14}
\]

Inspection of Equation (A14) indicates that the noise-limited regime is where the \( \sigma_u^4 \) in the denominator is much larger than the first term. The opposite case, where the \( \sigma_u^4 \sigma_u^2 \) term dominates the \( \sigma_u^4 \) term, defines the GW-dominated regime.

A.4. Combinations of White, Red, and GW Fluctuations

Realistically, the residuals are a combination of red and white noise along with the GW perturbations, so we need to consider \( u = p + r + n \). We can use Equations (A2)–(A13) to derive a general expression for \( \sigma_c^2 \) by expanding \( \sigma_u^2 \omega_{uu} \) and \( \sigma_u^4 \omega_{uu} \) in Equation (A13) and introducing other variances and dimensionless correlation times.

The cross-product that appears in Equation (A2) becomes \( \sigma_u^2 \sigma_u^2 \omega_{uu} \rightarrow \sigma_u^2 \sigma_u^2 \rho_u + \sigma_u^2 \rho_u + \sigma_u^2 \rho_u \), where we drop the \( T \) arguments for the variances and the \( t, t' \) arguments for correlation functions for clarity. When integrated over \( t \) and \( t' \) and normalized by \( T^{-2} \), we get \( \sigma_u^2 \sigma_u^2 \omega_{uu} \rightarrow \sigma_u^2 \sigma_u^2 \omega_{uu} + \sigma_u^2 \omega_{uu} + \sigma_u^2 \omega_{uu} \). Dimensionless correlation times involving \( n \) are determined by the correlation time for the white noise, which is much narrower than for the red noise or the GWs, so \( \omega_{uu} \rightarrow 1/N_i \).

Similarly, the product in Equation (A11) becomes \( \sigma_u^4 \rho_r^2 \rightarrow \sigma_u^4 \rho_r^2 + \sigma_u^4 \rho_r + \sigma_u^4 \rho_r \), and its normalized double time integral is \( \sigma_u^4 \omega_{uu} \rightarrow \sigma_u^4 \omega_{uu} + \sigma_u^4 \omega_{uu} + \sigma_u^4 \omega_{uu} + 2\sigma_u^2 \rho_u \omega_{uu} + 2\sigma_u^2 \rho_u \omega_{uu} + 2\sigma_u^2 \rho_u \omega_{uu} \). As above, the dimensionless timescales involving white noise become \( \omega_{uu} = \omega_{uu} = \omega_{uu} = 1/N_i \). We also let \( \omega_{wp} \rightarrow \omega_{wp} \) even though this is strictly an approximation because \( e \) is a single realization while \( p \) includes contributions from \( N_p \) LOSs. Similarly, we let \( \omega_{wr} \rightarrow \omega_{wr}, \omega_{wp} \rightarrow \omega_{wp}, \) and \( \omega_{wr} \rightarrow \omega_{wr} \).

We define variance ratios that characterize the strengths of the red noise and white noise in terms of the GW variance

\[
\xi \equiv \frac{\sigma_u^2}{\sigma_{gw}^2}, \quad \eta \equiv \frac{\sigma_u^4}{\sigma_{gw}^4}. \tag{A15}
\]

All quantities on both sides of these equations except \( \sigma_u^2 \) are functions of the data span \( T \).
We can now substitute into Equation (A13) and arrange terms to get
\[
\sigma_{\xi \epsilon N}^2 = \frac{4\psi \eta}{N_p} \left\{ w_{gg} + \xi w_{gr} + \frac{(w_{gg} + \xi^2 w_{rr} + 2\xi w_{gr})}{2\psi(N_p - 1)} \right\} + \frac{\eta}{N_t} \left[ 1 + \frac{(\eta + 2 + 2\xi)}{2\psi(N_p - 1)} \right].
\] (A16)

**Smoothing.** Smoothing the data can be beneficial for increasing the significance \(S\), at least when white noise matters, i.e., in the low-signal regime described earlier. Smoothing over \(N_t\) samples reduces the white-noise variance, \(\eta \rightarrow \eta/N_t\), but the number of independent samples also decreases, \(N_t \rightarrow N_t/N_t\), so that the ratio \(\eta/N_t\) remains constant and the correlation variance becomes
\[
\sigma_{\xi \epsilon N}^2 = \frac{4\psi \eta}{N_p} \left\{ w_{gg} + \xi w_{gr} + \frac{(w_{gg} + \xi^2 w_{rr} + 2\xi w_{gr})}{2\psi(N_p - 1)} \right\} + \frac{\eta/N_t}{N_t} \left[ 1 + \frac{(\eta/N_t + 2 + 2\xi)}{2\psi(N_p - 1)} \right].
\] (A17)

**Blocking.** The variance can be further reduced by analyzing the full data span \(T\) in blocks of size \(T/M\) and averaging the individual correlation functions over blocks. The variance will be reduced by a leading factor \(1/M\), but individual terms within Equation (A17) will also be affected. In particular, \(\eta\) and \(\xi\) depend on the data span length used to calculate the correlation function, so we replace them by \(\eta \rightarrow \eta_M\) and \(\xi \rightarrow \xi_M\). Also, the number of time samples in each block becomes \(N_t \rightarrow N_t/M\). Overall this yields our final expression for the variance of the correlation function:
\[
\sigma_{\xi \epsilon N}^2 = \frac{4\psi \eta}{M N_p} \left\{ w_{gg} + \xi_M w_{gr} + \frac{(w_{gg} + \xi_M^2 w_{rr} + 2\xi_M w_{gr})}{2\psi(N_p - 1)} \right\} + \frac{\eta_M M}{N_t} \left[ 1 + \frac{(\eta_M/N_t + 2 + 2\xi_M)}{2\psi(N_p - 1)} \right].
\] (A18)

The resulting significance using \(\langle \hat{C}_{i0} \rangle = \psi \sigma_{\epsilon N}^2\) is
\[
S = \frac{\sqrt{\psi M N_p}}{2} \left\{ w_{gg} + \xi w_{gr} + \frac{(w_{gg} + \xi^2 w_{rr} + 2\xi w_{gr})}{2\psi(N_p - 1)} \right\}^{-1/2} + \frac{\eta_M M}{N_t} \left[ 1 + \frac{(\eta_M/N_t + 2 + 2\xi_M)}{2\psi(N_p - 1)} \right].
\] (A19)

For a dimensionless strain amplitude spectrum \(h_i(f) = Af^{\alpha_i}\), the spectrum of timing residuals \(\propto f^{2\alpha_i-3}\) and the rms residual scales as \(\sigma_{\epsilon N}(T) \propto T^{\alpha_i}\) with \(x_p = 1 - \alpha_G\) for \(\alpha_G < 1\). For the GW background produced by merging SMBHs (Jaffe & Backer 2003), \(x_p = -3/2\) and \(x_p = 5/3\). Similarly, red timing noise has been characterized with an exponent \(x_r \approx 2 \pm 0.2\) corresponding to a spectrum \(\propto f^{-(5-10)}\) (Shannon & Cordes 2010). The dimensionless ratios then become \(\eta_M = \eta_1 M^{2x_r}\) and \(\xi_M = \xi_1 M^{2x_r-x_2}\), where \(\eta_1\) and \(\xi_1\) are the values for the full-length time series (\(M = 1\)). Nominal values of the exponents \(x_r\) and \(x_2\), red timing noise grows more quickly with \(T\) than does the GW perturbation. This suggests that there is an optimal time span \(T\) that maximizes the GW perturbation against white noise while not letting the red noise dominate the variance. This in turn suggests that there is an optimal number of blocks to use for any given value of \(T\).

In contrast to white noise, the steep red power spectra for \(e, p, r\) yield dimensionless timescales \(\gg 1/N_t\). Using simulations like those described below, we find \(w_{erp} \approx 0.36\) for red noise created with an \(f^{-13/3}\) spectrum after removal of a second-order polynomial. For red noise consistent with timing noise in pulsars (\(\propto f^{-5}\)), we obtain \(w_{erp} \approx 0.44\). For a flat spectrum (white noise), we verify that \(w_{erp} = 1/N_t\) to within statistical errors in simulations.

**A.5. Relationship to Other Definitions of Detection Significance**

Our definition for \(S\) has some similarity to that defined by Jenet et al. (2005), but with a crucial difference. Their Equation (4) is essentially a matched filter based on the Hellings and Downs angular correlation. In our notation, the numerator of their equation is (with subscript “\(J\)” for Jenet et al.)
\[
\hat{C}_J = \frac{1}{N_x} \sum_{i<j} \langle \hat{C}_{ij}(\theta_{ij}) - \langle \hat{C}_{ij} \rangle \rangle \langle \xi_{ij} - \langle \xi \rangle \rangle.
\] (A20)
The angular separation of the \(i\)th and \(j\)th objects is \(\theta_{ij}\), and barred quantities are sample means over all pulsar pairs of \(\hat{C}_{ij}\) and \(\xi_{ij} \equiv \xi(\theta_{ij})\), respectively. The ensemble mean is
\[
\langle \hat{C}_J \rangle = \sigma_{\epsilon N}^2 \langle \xi^2 - \langle \xi \rangle^2 \rangle.
\] (A21)

For the compact pulsar configuration we consider (all pulsars in the same direction but at different, unknown distances), \(\langle \hat{C}_J \rangle\) vanishes because \(\langle \xi^2 \rangle = \langle \xi \rangle = 1\) and thus cannot be used to quantify detection in this case. If we redefine the weighted correlation as
\[
\hat{C}_J' = \frac{1}{N_x} \sum_{i<j} \langle \hat{C}_{ij}(\theta_{ij}) \rangle \xi_{ij},
\] (A23)
the Jenet et al. test statistic becomes
\[
S_J' = \frac{\hat{C}_J'}{\sigma_{C_J}},
\] (A24)
which is identical to our definition for \(S\) when the pulsar configuration is compact with \(\theta_{ij} = 0\) for all \(i, j\).

Later in the paper we will relate our results to an arbitrary configuration of pulsars by considering \(S_J\), which simply multiplies our result for \(S\) by the mean-square angular correlation over the sample, \(\bar{\xi}^2\).

**A.6. Generation of Simulated Time Series**

For the simulations reported in the main text, we generate realizations of red noise by shaping complex white noise in the frequency domain and performing an inverse discrete Fourier transform. We fill an array that includes Fourier components with periods that are four times longer than our desired time series so that low-frequency components are not underestimated.
We then select 1/4 of the time series. To suitably mimic the analysis of pulsar timing data, we subtract a straight line whose end points equal the first and last data points. This accounts for the fact that prior to doing a least-squares fit to timing data, a preliminary timing model is first removed. In this way, we compare pre- and post-fit variances that are close to representing those that would result in actual applications.

A.7. Non-Gaussianity of the Zero-lag CCF

In the main text, we describe the skewness of the distribution of $\hat{C}_{10}$ toward positive values. The skewness is generic for time series that include a red-noise component. Two effects lead to this result. First, inspection of Equation (A2) shows that the correlation function for a single pair of pulsars is an integral (or sum) of the products of two time series. The CLT will apply if each time series includes many independent fluctuations over the interval $[0, T]$. The sum over all pairs will also satisfy the conditions for the CLT, and $\hat{C}_{10}$ will have a Gaussian distribution. For red-noise processes, however, each time series is dominated by order only one fluctuation, so the CLT will not apply to the single-pair integral. The second effect is that when the CLT does not apply to the CCF for a single pair it also does not apply to the sum over all pairs, in part because a given time series contributes to $N_p - 1$ terms in the sum and the terms are not independent.

REFERENCES

Anholm, M., Ballmer, S., Creighton, J. D. E., Price, L. R., & Siemens, X. 2009, Phys. Rev. D, 79, 084030
Backer, D. C., & Hellings, R. W. 1986, ARA&A, 24, 537
Box, G., & Tiao, G. 1973, Bayesian Inference in Statistical Analysis, Addison-Wesley Series in Behavioral Science: Quantitative Methods (New York: Addison-Wesley)
Burt, B. J., Lommen, A. N., & Finn, L. S. 2011, ApJ, 730, 17
Coles, W. A., Rickett, B. J., Gao, J. J., Hobbs, G., & Verbiest, J. P. W. 2010, ApJ, 717, 1206
Cordes, J. M., & Downs, G. S. 1985, ApJS, 59, 343
Cordes, J. M., & Shannon, R. M. 2010, arXiv:1010.3785
Demorest, P. B., et al. 2012, arXiv:1201.6641
Finn, L. S., & Lommen, A. N. 2010, ApJ, 718, 1400
Hellings, R. W., & Downs, G. S. 1983, ApJ, 265, L39
Jaffe, A. H., & Backer, D. C. 2003, ApJ, 583, 616
Jenet, F. A., Hobbs, G. B., Lee, K. J., & Manchester, R. N. 2005, ApJ, 625, L123
Jenet, F. A., Lommen, A., Larson, S. L., & Wen, L. 2004, ApJ, 606, 799
Lee, K., Jenet, F. A., Price, R. H., Wex, N., & Kramer, M. 2010, ApJ, 722, 1589
Lommen, A. N., & Backer, D. C. 2001, ApJ, 562, 297
Lyne, A., Hobbs, G., Kramer, M., Stairs, I., & Stappers, B. 2010, Science, 329, 408
Manchester, R. N. 2010, Highlights Astron., 15, 233
O’Ruanaidh, J. K., & Fitzgerald, W. J. 1996, Numerical Bayesian Methods Applied to Signal Processing (New York: Springer)
Olmez, S., Mandic, V., & Siemens, X. 2010, Phys. Rev. D, 81, 104028
Phinney, S. 2004, APS Meeting Abstracts, L9002
Sesana, A., Vecchio, A., & Colacino, C. N. 2008, MNRAS, 390, 192
Seto, N., & Cooray, A. 2007, ApJ, 659, L33
Shannon, R. M., & Cordes, J. M. 2010, ApJ, 725, 1607
Smith, T. L., Kamionkowski, M., & Wandelt, B. D. 2011, Phys. Rev. D, 84, 063013
van Haasteren, R., & Levin, Y. 2010, MNRAS, 401, 2372
van Haasteren, R., Levin, Y., Janssen, G. H., et al. 2011, MNRAS, 414, 3117
van Haasteren, R., Levin, Y., McDonald, P., & Lu, T. 2009, MNRAS, 395, 1005
Yardley, D. R. B., Coles, W. A., Hobbs, G., et al. 2011, MNRAS, 414, 1777
Yardley, D. R. B., Hobbs, G., Jenet, F. A., et al. 2010, MNRAS, 407, 669