Evaluating the prevalence of spurious correlations in pulsar timing array datasets

Andrew Zic,1,2 ★ George Hobbs,2 R. M. Shannon,3,4 Daniel Reardon,3,4 Boris Goncharov,5,6 N. D. Ramesh Bhat,7 Andrew Cameron,3,4 Shi Dai,8,9,2 J. R. Dawson,1,2 Matthew Kerr,10 R. N. Manchester,2 Rami Mandow,1,2 Tommy Marshman,1,2 Christopher J. Russell,11 Nithyanandan Thyagarajan,12 X. -J. Zhu13

1 School of Mathematical and Physical Sciences, and Research Centre in Astronomy, Astrophysics and Astrophotonics, Macquarie University, NSW 2109, Australia
2 Australia Telescope National Facility, CSIRO, Space and Astronomy, PO Box 76, Epping, NSW 1710, Australia
3 Centre for Astrophysics and Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia
4 OzGrav: The ARC Centre of Excellence for Gravitational Wave Discovery, Australia
5 Gran Sasso Science Institute (GSSI), I-67100 L’Aquila, Italy
6 INFN, Laboratori Nazionali del Gran Sasso, I-67100 Assergi, Italy
7 International Centre for Radio Astronomy Research (ICRAR), Curtin University, Bentley, WA 6102, Australia
8 School of Science, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
9 National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
10 Space Science Division, Naval Research Laboratory, Washington, DC 20375-5352, USA
11 CSIRO Scientific Computing, Australian Technology Park, Locked Bag 9013, Alexandria, NSW 1435, Australia
12 Australia Telescope National Facility, CSIRO, Space and Astronomy, PO Box 1130, Bentley, WA 6102, Australia
13 Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai 519087, China

Accepted 2022 July 19. Received 2022 July 4; in original form 2022 May 4

ABSTRACT

Pulsar timing array collaborations have recently reported evidence for a noise process with a common spectrum among the millisecond pulsars in the arrays. The spectral properties of this common-noise process are consistent with expectations for an isotropic gravitational-wave background (GWB) from inspiralling supermassive black-hole binaries. However, recent simulation analyses based on Parkes Pulsar Timing Array data indicate that such a detection may arise spuriously. In this paper, we use simulated pulsar timing array datasets to further test the robustness of the inference methods for spectral and spatial correlations from a GWB. Expanding on our previous results, we find strong support (Bayes factors exceeding $10^5$) for the presence of a common-spectrum noise process in datasets where no common process is present, under a wide range of timing noise prescriptions per pulsar. We show that these results are highly sensitive to the choice of Bayesian priors on timing noise parameters, with priors that more closely match the injected distributions of timing noise parameters resulting in diminished support for a common-spectrum noise process. These results emphasize shortcomings in current methods for inferring the presence of a common-spectrum process, and imply that the detection of a common process is not a reliable precursor to detection of the GWB. Future searches for the nanohertz GWB should remain focussed on detecting spatial correlations, and make use of more tailored specifications for a common-spectrum noise process.

Key words: stars: neutron – pulsars: general – gravitational waves – methods: data analysis

1 INTRODUCTION

Pulsar timing arrays (PTAs) consist of sets of millisecond pulsars (MSPs) exhibiting high timing stability (Foster & Backer 1990). Among myriad scientific goals (Manchester et al. 2013), the primary aim of PTA experiments is the detection and characterization of the isotropic stochastic gravitational-wave background (GWB; e.g. Jenet et al. 2005). Current PTA experiments include the European PTA (EPTA; Kramer & Champion 2013), the North American Nanohertz Observatory for Gravitational waves (NANOGrav; McLaughlin 2013), the Parkes PTA (PPTA; Manchester et al. 2013), the Indian PTA (Joshi et al. 2018, InPTA;), which all comprise the International PTA (IPTA; Hobbs et al. 2010a). Other nascent PTA collaborations, such as the Chinese PTA (Lee 2016), and experiments with the MeerKAT telescope in South Africa (e.g. Bailes et al. 2020) may join efforts with the IPTA in coming years.

Some PTA collaborations have recently detected a noise process with spectral properties that appear common among all pulsars, possibly representing the emergence of the GWB signal in their datasets (e.g. Arzoumanian et al. 2020). However, because these datasets

★ E-mail: andrew.zic@mq.edu.au

© 2022 The Authors
are currently in a sub-threshold and highly model-dependent regime for GWB detection (Hazboun et al. 2020; Goncharov et al. 2021a; Romano et al. 2021; Pol et al. 2021), these findings require robust validation checks in order to understand their significance. This forms the underlying motivation behind this paper.

The largest contribution to the GWB is expected to come from a cosmological population of supermassive black-hole binaries (SMBHBs; Rosado et al. 2015; Sesana 2013; Wyithe & Loeb 2003), but other more exotic processes such as cosmological phase transitions (Xue et al. 2021; Arzoumanian et al. 2021; Kobakhidze et al. 2017), vibration of cosmic strings (Ölmez et al. 2010), and quantum fluctuations in the early universe (Lasky et al. 2016; Linde 1982; Starobinsky 1980; Grishchuk 1976) are also expected to contribute.

For a cosmological population of SMBHBs in circular orbits, with energy loss dominated by gravitational-wave (GW) emission within pulsar term, while the cross-correlations are sensitive primarily to fluctuations in the pulsar times of arrival (ToAs) on timescales of years (Lasky et al. 2016; Romano et al. 2021; Po et al. 2021), these findings require robust validation checks in order to understand their significance. This forms the underlying motivation behind this paper.

In the literature, this common-spectrum process is usually termed the “common red noise” (CRN), which we will also use for the remainder of this paper.

Currently, none of the PTA collaborations has reported significant evidence for HD-correlated signals in their datasets. Recently, however, NANOGrav, the PPTA, the EPTA, and the IPTA have reported strong evidence for the presence of a CRN process in recent data releases (Arzoumanian et al. 2020; Goncharov et al. 2021b; Chen et al. 2021; Antoniadis et al. 2022). Though there is some variance among the best estimates, the reported CRN properties are consistent within uncertainties. However, the inferred amplitudes for the CRN at a fixed spectral index of $\gamma = 3/3$, ranging from $\sim 2.0 \times 10^{-15}$ to $3.0 \times 10^{-15}$ at a reference frequency of 1 yr$^{-1}$, are in tension with previously set 95 per cent credible interval upper limits from NANOGrav ($A_{\text{GWB}} < 1.45 \times 10^{-15}$; Arzoumanian et al. 2018) and the PPTA ($A_{\text{GWB}} < 1.0 \times 10^{-15}$; Shannon et al. 2015).

These discrepancies have been a point of concern among PTA collaborations (e.g. Arzoumanian et al. 2020), but recent work by Johnson et al. (2022) suggests that upper limits are more likely to be under-estimated when they are formed using a subset of pulsars from a PTA, as in Shannon et al. (2015). While this offers a possible explanation for the discrepancies between CRN properties and previous upper limits, investigations by Goncharov et al. (2021b) have found that a CRN can be “detected” in simulated datasets only containing individual pulsar noise terms with disparate characteristics. This raises concerns that a CRN signal can be strongly influenced by, or arise entirely from, independent pulsar noise processes that have no relationship with the GWB.

One of the most important noise processes present in individual pulsars is “timing noise” (Groth 1975; Lyne 1999; Hobbs et al. 2010b; Cordes 2013; Parthasarathy et al. 2019), also known as “spin noise” – stochastic, time-correlated variations in the pulsar ToAs thought to be driven by rotational irregularities and other pulsar-intrinsic fluctuations. Concerns about inferences on the CRN process have arisen on the basis that some MSPs may exhibit similar timing noise characteristics (Shannon & Cordes 2010; Goncharov et al. 2021a), which may result in a false-alarm detection of a CRN. Indeed, Meyers et al. (2021) suggest that pulsar timing noise induced by spin irregularities has a spectral index of 4, close to the 13/3 value expected for a GWB. Incorrect or incomplete models for pulsar-intrinsic noise terms can bias searches for, or prevent detection of, the GWB (Goncharov et al. 2021b; Hazboun et al. 2020; Lasky et al. 2015; Cordes 2013).

It is possible that the detection of a CRN process among PTA collaborations is truly the first emerging evidence of the GWB (Romano et al. 2021; Pol et al. 2021). However, the tension with previous upper limits (Johnson et al. 2022), and false detections in simulations presented by Goncharov et al. (2021b), and in this work, warrant further investigation into present methodologies and biases involved in detecting a CRN. There have been recent efforts to improve inference methods for the CRN (e.g. Goncharov et al. 2022). As we will show in this work, developments such as these are necessary to consolidate
recent detections of a CRN as milestones toward the detection of the GWB via spatial correlations.

While understanding the subtleties involved in CRN inference is important in the context of recent results, the key to unambiguously detecting the GWB lies in the spatial correlations. Therefore, understanding the robustness of spatial correlation inference techniques under different contexts is critical (Tiburzi et al. 2016; Taylor et al. 2017). Pulsar timing noise is expected to be one of the main obstacles to GWB detection (Cordes 2013; Lasky et al. 2015; Taylor et al. 2017), understanding its influence on spatial correlation inferences is particularly pertinent.

In this paper, we use simulated pulsar timing array datasets containing timing noise to explore the biases in current techniques for inferring the presence of common-spectrum and spatially correlated signals in pulsar timing array datasets. In Section 2, we describe the simulation process and present the properties of the simulated datasets. In Section 3, we present our analysis of these simulations, and discuss implications for recent pulsar timing array results and methodologies. Section 4 contains concluding discussion and remarks for this work.

2 SIMULATIONS

We constructed pulsar timing array datasets using PTASIMULATE\(^3\), a package for simulating pulsar ephemerides and ToAs in a format suitable for TEMPO2 (Edwards et al. 2006). The package can be used to inject various stochastic and deterministic signals into the ToAs, which can then be used for studying observing strategies, telescope and PTA sensitivities, GWB analyses, and many other topics relevant to PTA datasets.

While PTASIMULATE can simulate realistic datasets with a wide range of pulsar timing phenomena, in this work we chose to simulate ToAs recorded at a regular cadence and at a single frequency band, with uniform ToA uncertainties per pulsar, chosen between 90 to 500 ns based on similarities to PPTA datasets. We made these choices primarily to reduce computational costs while exploring a wide parameter space, but also to minimize dataset complexity that could obfuscate interpretation of our analysis.

Following Goncharov et al. (2021b), we simulated timing residuals for the 26 pulsars in the PPTA second data release (DR2: Kerr et al., 2020), with a regular cadence of 40 days, and over a time-span of 20 years. For each pulsar, we injected timing noise as a red noise signal with a power-law PSD \( P_{TN} \) parametrized as

\[
P_{TN}(f|P_0, \gamma, f_c) = \frac{P_0}{\left(1 + \left(\frac{f}{f_c}\right)^{\gamma}\right)^{\gamma/2}} \quad [\text{s}^2],
\]

where \( f, f_c \) are the fluctuation and corner frequencies respectively in units of \( \text{s}^{-1} \), \( P_0 \) is the PSD amplitude, and \( \gamma \) is the spectral index. This parametrization explicitly encodes a low-frequency “corner” in the PSD, below which the PSD plateaus at a constant value. In the limit \( f \gg f_c \), Eq. 5 simplifies to a standard power-law parametrization as in Eq. 4, with \( P_0 = A^2 f_c^\gamma / (12\pi^2) \).

For each pulsar, the PSD amplitude was drawn from a log-uniform distribution with a median value \( P_{0,m} = 10^{-23} \) (corresponding to \( \log_{10} A_m = -14.46 \)) and a width \( \Delta \log_{10} P_{m} \), and the spectral index was drawn from a uniform distribution with a median value \( \gamma_m = 4 \) and a width \( \Delta \gamma m \). The choice of these median values was made to approximately match the characteristics of the CRN recently detected in the PPTA DR2 (e.g. Goncharov et al. 2021b), so that we could investigate the robustness of CRN detections in similar datasets. In Goncharov et al. (2021b), we simulated datasets where the full width of the input uniform distribution for the PSD amplitude (hereafter termed \( \Delta \log_{10} P_0 \)) and spectral index (hereafter termed \( \Delta \gamma \)) was increased simultaneously (i.e., we only explored a one-dimensional path in the \((\Delta \log_{10} P_0, \Delta \gamma)\) parameter space). In this work, we extended this analysis by exploring the \((\Delta \log_{10} P_0, \Delta \gamma)\) parameter space in a regularly-spaced 11 \( \times \) 11 grid, where \( \Delta \log_{10} P_{0,m} \) varied from 0 to 14, and \( \Delta \gamma \) varied from 0 to 8, around the central values \( \log_{10} P_{0,m} = -23 (\log_{10} A_m = -14.46) \) and \( \gamma_m = 4.0 \). In this description, \((\Delta \log_{10} P_0, \Delta \gamma) = (0, 0)\) corresponds to a true common-spectrum red noise process, and increases of \( \Delta \log_{10} P_0 \) and \( \Delta \gamma \) correspond to increasingly disparate pulsar timing noise properties.

For each choice of \( \Delta \log_{10} P_0 \) and \( \Delta \gamma \), we drew a PSD amplitude \( P_0 \) and spectral index \( \gamma \) for each pulsar, and simulated 100 realizations of timing noise according to these chosen parameters. Altogether with 121 choices of timing noise parameters, with 100 realizations each, this resulted in 12,100 simulated datasets. The large number of realizations enabled us to explore detection statistics and perform false alarm analyses across the \((\Delta \log_{10} P_0, \Delta \gamma)\) parameter space. We note that the values of \( \log_{10} P_0 \) and \( \gamma \) for each pulsar were drawn independently from the uniform distribution with widths \( \Delta \log_{10} P_0 \), \( \Delta \gamma \), with no regard given to the properties of any particular pulsar in real datasets.

In Figure 1 we show timing residuals in one realization of a simulation with \( \Delta \log_{10} P_0 = 2.8, \Delta \gamma = 4.0 \), and in Figure 2, we show the PSDs for the 26 pulsars, with increasing degrees of variation in the injected timing noise parameters.

We do not inject any GWB or other spatially-correlated signals into our simulations. While the GWB signal likely exists within real datasets, even if it is low in amplitude (Bonetti et al. 2018; Dvorkin & Barausse 2017; Taylor et al. 2016; Shannon et al., 2015), our analysis on datasets containing only independent pulsar timing noise terms allows us to test current methodologies in the “worst-case” scenario of PTA datasets dominated by pulsar noise terms. By doing this, we aim to investigate the extent that recent detections of a CRN (Arzoumanian et al. 2020; Goncharov et al. 2021b; Chen et al. 2021; Antoniadis et al. 2022) could be influenced by timing noise. Furthermore, there have already been several detailed GWB injection-recovery analyses presented elsewhere, so we do not repeat those analyses here (e.g., Taylor et al. 2022; Hazboun et al. 2020; Vigeland et al. 2018; Tiburzi et al. 2016).

3 COMMON-SPECTRUM AND SPATIAL CORRELATION ANALYSIS

We used an established Bayesian inference procedure (see Antoniadis et al. 2022; Chen et al. 2021; Goncharov et al. 2021b; Arzoumanian et al. 2020) for our analysis. To summarize, we use the multivariate Gaussian likelihood to model the data (Taylor et al. 2017; Arzoumanian et al. 2016). Next, we construct the so-called design matrix with LIBTEMPO (Vallisneri 2020) and TEMPO2 (Edwards et al. 2006) to marginalize the likelihood over the terms in the deterministic pulsar timing model. We used ENTERPRISE (Ellis et al. 2019) to

---

3 https://bitbucket.org/psrsoft/ptasimulate

---

4 The subscript \( m \) indicates the median value of the distribution of injected timing noise parameters.
Figure 1. Timing residuals for a simulated dataset with $\Delta \log_{10} P_0 = 2.8$, $\Delta \gamma = 4.0$. We show a 1 $\mu$s scale bar in the left of each sub-figure for reference. Note that the simulated residuals are not necessarily reflective of the residuals for each pulsar in real datasets.

Figure 2. Power spectral densities of timing residuals from four simulated datasets, each with timing noise sampled from distributions of different widths (rows 1 to 4, coloured blue) and for the PPTA DR2 dataset (row 5, coloured black). The spectra have been calculated using a generalized least-squares technique (Coles et al. 2011). Distribution widths for the simulated input timing noise parameters, $\Delta \log_{10} P_0$ – $\Delta \gamma$, are $0.0 – 0.0$, $1.4 – 1.6$, $2.8 – 3.2$, and $2.8 – 4.0$, from top toward bottom panels. We also show a reference spectrum (orange) in each panel, with $A = 2.2 \times 10^{-15}$, $\gamma = 13/3$, representing the CP2 model found in PPTA DR2 (Goncharov et al. 2021b).

We perform Bayesian model selection and parameter estimation using PTMCMCSAMPLER (Ellis & van Haasteren 2019). We used the hybrid Bayesian-frequentist optimal statistic (Anholm et al. 2009; Demorest et al. 2013; Chamberlin et al. 2015, see Section 3.3.1) to evaluate evidence for spatial correlations in the simulated datasets.

We employed three models to evaluate the simulated datasets in our analysis:
Figure 3. realization-averaged Bayes factors for model CP1 (top) and CP2 (bottom) over TN, as a function of $\Delta \log_{10} P_0$ and $\Delta \gamma$. Both CP1 and CP2 are heavily favoured across a wide range of variations in input timing noise parameters. White stars indicate the samples of the $(\Delta \log_{10} P_0, \Delta \gamma)$ parameter space plotted in Figure 2.

- TN: Independent timing noise for each pulsar, parametrized by a red power-law spectrum as in Eq. 4 (this is generally the correct description, except in simulations with $\gamma_{CP} = 0$, which can be described as a true spatially-uncorrelated common-spectrum process).
- CP1: Independent red-spectrum timing noise (as with model TN) for each pulsar, but with the addition of a common-spectrum noise process with varying spectral index $\gamma_{CP}$ and amplitude $A_{CP}$.
- CP2: Same as model CP1, but with the CRN spectral index $\gamma$ held fixed to the fiducial value of 13/3 expected for a classical GWB.

The red noise terms were evaluated using a Fourier series with a fundamental frequency corresponding to the inverse of the dataset observing time-span $1/T_{\text{obs}}$. We held white noise hyper-parameters (known in the pulsar timing community as EFAC and EQUAD) fixed at 1 and 0 respectively, as our simulations did not incorporate any deviations of the white noise characteristics from the injected values.

In our standard Bayesian posterior sampling and model selection runs, we used uniform priors on the timing noise and CRN spectral indices ($\gamma \in U[0,10]$), and log-uniform priors on the timing noise and CRN amplitudes ($\log A \in U[-20,-6]$).

3.1 Model selection analysis

We performed model selection for models CP1 and CP2 over TN using the product-space method in a “hypermodel” framework (Carlin & Chib 1995; Hee et al. 2016; Taylor et al. 2020). To provide adequate dynamic range for measuring very large or very small Bayes factors, we sampled with $1 \times 10^6$ iterations. Because of computational costs, we only analysed ten noise realizations for each cell across the $(\Delta \log_{10} P_0, \Delta \gamma)$ parameter space, meaning that we only processed 1210 out of 12100 simulated datasets for the model selection analysis. After performing the model selection procedure on each noise realization, we computed the realization-averaged Bayes factor in log-space.

5 When referring to either model CP1 and CP2 generally, we use CP.
The results from this search are shown in Figures 3 and 4. On average, we find strong support for model CP1 and CP2 over TN, for a region of parameter space spanning up to $\Delta \gamma \sim 4$ and $\Delta \log_{10} P_0 \sim 4$ ($\log_{10} B_{TN}^{CP} \geq 5.8$), and more moderate support for values of $\Delta \log_{10} P_0$ up to 6, and $\Delta \gamma$ up to 7. That is, CP models remain the preferred model on average over TN under $\sim$ six orders-of-magnitude variations in timing noise amplitude, and variations in the timing noise spectral index by a range of $\sim 7$. In Figure 4, we show the maximum Bayes factors across all realizations. We find that the Bayes factors are limited by the number of posterior samples over the entire region in Figure 3 where the CP models are not strongly favored. This indicates that even if there is moderate support on average for a CRN process for a given choice of $(\Delta \log_{10} P_0, \Delta \gamma)$, strong support for CP models is found in at least one out of ten realizations.

Our simulations demonstrate that model comparison of CP1 or CP2 against TN alone is insufficient to claim detection of CRN. If the inference of a CRN was to be used as preliminary evidence for a GWB, then the CRN models (CP1 and CP2) should only be favored in our simulations when $\Delta \log_{10} P_0$ and $\Delta \gamma$ are close to 0. Furthermore, the support should quickly decline as $\Delta \log_{10} P_0$ and $\Delta \gamma$ increase. Instead, we see a more gradual decline on average as the span of timing noise parameters increases.

### 3.1.1 The distribution of Bayes factors

When comparing appropriately-specified cosmological models, Bayes factors can be expected to exhibit scatter of about an order of magnitude due to cosmic variance (e.g. Joachimi et al. 2021). In principle, this scatter may be used to set appropriate (and more conservative) decision thresholds in Bayesian model comparison. Similar boot-strapping approaches have already been developed for frequentist detection statistics for spatial correlations in pulsar timing array analysis (Taylor et al. 2017). We investigated the underlying Bayes factor distributions to determine whether it is possible to calibrate Bayes factors for the CRN under current procedures. To improve our sample statistics, we group cells in the $\Delta \log_{10} P_0, \Delta \gamma$ parameter space by average Bayes factor values.

We show Bayes factor distributions in Figure 5, for $\Delta \log_{10} P_0, \Delta \gamma$ cells where $B_{TN}^{CP1} < 10^{-2}$ (top), $10^{-2} < B_{TN}^{CP1} < 10^{-2}$ (middle), $B_{TN}^{CP1} > 10^{2}$ (bottom). These groupings represent cells where CP is, on average, strongly disfavored, weakly disfavored/favored, and strongly favored (respectively). While a minority of Bayes Factors have values representing moderate to strong evidence against CP1, most Bayes Factors in the sample are peaked at the boundary values close to $10^{46}$, which are set by the number of our posterior samples in the hypermodel framework. Furthermore, there are very few Bayes factors within intermediate values, with only a small tail weighted toward $B_{TN}^{CP1} < 1$. These features suggest that many more posterior samples are required to resolve the true underlying Bayes factor distributions. More importantly, these distributions highlight the improper performance of CRN inference under current models, priors, and model selection procedures. While further investigations of these underlying Bayes factor distributions may be a topic of interest for future work, we suggest that improvements in the underlying inference procedure (e.g. Goncharov et al. 2022) is a more appropriate route toward robust CRN inference in future.

![Figure 5. Relative distributions of Bayes factors for CP1 over TN, grouped by realisations with $B_{TN}^{CP1} < 10^{-2}$ (top), $10^{-2} < B_{TN}^{CP1} < 10^{2}$ (middle), $B_{TN}^{CP1} > 10^{2}$ (bottom).](image)

### 3.1.2 The effect of prior volumes

We consider the possibility that the spurious support for CP models arises from the choice of priors on timing noise parameters. This is motivated by the fact that when a CRN is well-constrained in our simulations, the estimated timing noise amplitudes for most pulsars tend to drop to very small values, and the spectral indices become unconstrained. In our standard analyses, and the analysis presented in Goncharov et al. (2021b), we use wide priors on timing noise parameters. This is reflective of our lack of a-priori knowledge of the timing noise properties of the pulsars, particularly since the spectral parameters for timing noise and the GWB are highly covariant.

If the priors on timing noise parameters are chosen such that their range more closely reflects the true range of injected values, the support for CP models may diminish. To test this, we selected simulated datasets with $(\Delta \log_{10} P_0, \Delta \gamma) = (1.4, 0.8)$ and $(2.8, 1.6)$. As before, we performed our model selection analysis on 10 realizations, but this time with a gradually decreasing prior width on $\log A$ and $\gamma$, until the priors approached the delta function at the median timing noise parameters.

The results are shown in Figure 6, where we plot realization-averaged Bayes factors for CP2 over TN, as a function of $\Phi = (\Delta \log_{10} A \Delta \gamma)_{\text{prior}} / (\Delta \log_{10} A \Delta \gamma)_{\text{prior}} - \gamma$ the ratio between the volumes of simulated and prior timing noise parameter distributions. Heuristically, when the ratio $\Phi$ is small, the priors span a wider range than the range of injected timing noise parameters, and vice versa for...
large values of $\Phi$, Figure 6 shows that as the ratio approaches unity, the support for the CRN quickly diminishes.

For simulations with a wider range of timing noise prescriptions, such as $(\Delta \log_{10} A, \Delta \gamma) = (2.8, 1.6)$ shown in Figure 6, Bayes factors begin to increase again at large values of $\Phi$. This is likely because the timing noise priors are too restrictive given the variance of timing noise prescriptions in the simulated datasets, causing the CP model to be favoured over TN once again (even though neither model represents a good description of the data with the choice of priors). Overall, these effects highlight the sensitivity of CRN inference to the choice of priors on timing noise parameters.

It is not possible to accurately bound the priors on timing noise parameters to match the true distributions over all pulsars in real datasets – we do not know the true underlying distribution of pulsar timing noise parameters a-priori, particularly for low-amplitude timing noise. As mentioned above, red noise in pulsar timing residuals could be ascribed to either timing noise, or a GWB signal, or a combination of these. The results presented above are simply an exercise in demonstrating the sensitivity of current methodologies to the choice of priors.

### 3.2 Common-spectrum process and timing noise characteristics

To better understand the origin of spurious CRN detections, we now consider the relationship between the CRN inferred in our standard analysis and the injected timing noise. We performed Bayesian parameter estimation for model CP1 across the timing noise parameter space. In Figure 7 we show the difference between the posterior median CRN parameters and the central values of the injected timing noise parameters, $\log_{10} A_m = -14.46$, and $\gamma_m = 4$. In the region of parameter space where CP1 is preferred, the recovered CRN is close to the central timing noise parameters, indicating that the inferred CRN is consistent with the ensemble average of independent pulsar timing noise terms.

This is perhaps not surprising, considering the heuristic description of the CRN as a spectral process present among all pulsars, and suggests that timing noise can bias estimates for a CRN when it is lower in amplitude than typical timing noise terms. Indeed, in Goncharov et al. (2021b), we showed that a CRN signal need not be present in all pulsars to be inferred – some pulsars have a substantially lower red noise level than the majority of pulsars in both real and simulated PTA datasets, but this has little impact on the detection of an apparent CRN signal.

Some trends in the characteristics of the inferred CRN are apparent. In Figure 7 bottom, a systematic trend of steeper spectral indices, and lower amplitudes for the inferred common spectrum process is evident when $\Delta \gamma$ is large, $\Delta \log_{10} P_0$ is small. Similarly, when $\Delta \log_{10} P_0$ is large and $\Delta \gamma$ is small, the inferred CRN spectrum tends to be slightly shallower, and slightly lower in amplitude.

### 3.3 Optimal statistic analysis for spatial correlations

#### 3.3.1 The optimal statistic

It is important to consider how spurious detections of a CRN may also result in spurious detections of spatial correlations, and if so, how often. To do this, we employed the optimal statistic, $\hat{A}^2$ (Anholm et al. 2009; Demorest et al. 2013; Chamberlin et al. 2015; Vigeland...
et al. 2018), which is a frequentist estimator of the amplitude of spatially-correlated noise processes. It is constructed as the weighted sum of inter-pulsar spatial correlations accounting for pulsar-specific and inter-pulsar noise covariances, and is given by (Chamberlin et al. 2015)

\[ A^2 = \frac{\sum_{ab} \delta t_a^T C_a^{-1} \tilde{S}_{ab} C_b^{-1} \delta b}{\sum_{ab} \text{tr} \left( C_a^{-1} \tilde{S}_{ab} C_b^{-1} S_{ba} \right)}, \]  

(6)

where \( \delta t_a \) is the vector of timing residuals for pulsar \( a \), \( C_a = \langle \delta t_a \delta t_a^T \rangle \) is the autocovariance matrix, \( \tilde{S}_{ab} = S_{ab} A_{\text{GWB}}^{-2} \) is the GWB amplitude-normalized cross correlation matrix, with \( S_{ab} = \langle \delta t_a \delta t_b^T \rangle_{ab} \). The optimal statistic signal-to-noise (S/N) ratio gives a measure of the significance for \( A_{\text{GWB}} \neq 0 \), and is given by

\[ \rho = \frac{\sum_{ab} \delta t_a^T C_a^{-1} \tilde{S}_{ab} C_b^{-1} \delta b}{\left[ \sum_{ab} \text{tr} \left( C_a^{-1} \tilde{S}_{ab} C_b^{-1} S_{ba} \right) \right]^{1/2}}. \]  

(7)

In the standard approach, the pulsar-intrinsic red noise terms are first jointly sampled with the CRN terms using Bayesian parameter estimation, and held fixed at the maximum-likelihood values when computing the optimal cross-correlation statistic (Chamberlin et al. 2015). However, this method does not fully account for the degeneracy between pulsar-intrinsic red noise and the red noise induced by the GWB, resulting in biased estimates of the GWB amplitude (Vigeland et al. 2018). To address this, Vigeland et al. (2018) developed the noise-marginalized optimal statistic, which estimates the cross-correlation amplitude using posterior samples from the pulsar-intrinsic noise terms from the joint pulsar-intrinsic and CRN parameter estimations (e.g., parameter estimation of model CP2). The optimal statistic has been deployed as a complement to fully Bayesian characterisation of inter-pulsar spatial correlations in PTA datasets (Arzoumanian et al. 2018, 2020; Antoniadis et al. 2022). The results from optimal statistic analysis of spatial correlations are broadly consistent with the fully Bayesian measurements of spatial correlations, but there are some key differences; namely, the presence of a monopole-correlated signal is marginally supported in optimal statistic analyses of recent PTA datasets (Antoniadis et al. 2022; Arzoumanian et al. 2020), but is not supported by Bayesian analyses.

3.3.2 Optimal statistic false detection analysis

To assess the robustness of spatial correlation inference in the presence of a wide range of timing noise characteristics, we computed the optimal statistic for HD, dipole, and monopolar correlations on our simulated datasets containing only pulsar timing noise. We used posterior samples from Bayesian parameter estimation runs of model CP2 over our simulated datasets to calculate both the standard (maximum-likelihood) and noise-marginalized optimal statistics. In both cases, we also calculated the optimal statistic S/N \( \rho \) and the inter-pulsar covariance measured in angular bins to investigate the significance of any false detections.

In Figures 8 and 9 we show the number of realizations (out of 100) with \( \rho > 3 \) for the maximum-likelihood and noise-marginalized optimal statistic (respectively), as a proxy for the number of false detections of spatial correlations. We also show the overall fraction of realizations with \( \rho > 3 \) over all values of \( \Delta \text{log}_{10} P_{\text{p}} \) and \( \Delta \gamma \) in Table 1. Figure 10 shows the overall distribution of \( \rho \) for the maximum likelihood and noise-marginalized optimal statistic. These results highlight that the noise-marginalized optimal statistic produces substantially fewer false detections than the maximum-likelihood method. This is consistent with the reduced bias for the noise-marginalized optimal statistic found by Vigeland et al. (2018). However, Vigeland et al. (2018) report that the maximum-likelihood optimal statistic systematically under-estimates the true GWB amplitude. Here, we find that the bias of the maximum-likelihood optimal statistic appears to work in the opposite sense, in that it results in more false detections of spatial correlations. This could be explained by the fact that the maximum-likelihood method places excessive weight on the CRN terms, while the noise-marginalized optimal statistic marginalizes out these biases. In any case, the message from these results is in agreement with previous analyses (Vigeland et al. 2018): the noise-marginalized optimal statistic is a more accurate and robust tool for measuring spatial correlations in the presence of pulsar timing noise.

The false detection rate is strongly dependent on the choice of maximum-likelihood or noise-marginalized methods, but only moderately dependent on the overlap reduction function being considered. HD correlations appear to have the lowest false detection rate, followed by dipolar and monopolar correlations having higher false detection rates. A possible reason for this effect is that the simplest ORF in terms of functional form is the monopolar ORF, and a small D.C. offset in the average inter-pulsar covariances could result in a modest signal-to-noise detection of monopolar correlations. On the other hand, dipolar and HD correlations have a more complex spatial signature, which may be more difficult to encounter by chance. We also note that the S/N distributions shown in Figure 10 are positively-skewed for all spatial correlations. In Figure 11, we show optimal statistic results from a realization with noise-marginalized optimal statistic S/N of \( \rho = 3.2 \) for HD correlations. In Figure 12, we show the associated spatial covariance between pulsar pairs, \( \hat{A}^2 \Gamma(\xi_{ab}) \), binned in angular intervals.

### Table 1
| ORF      | Max. Likelihood | Marginalized |
|----------|-----------------|--------------|
| HD       | 0.020           | 0.003        |
| Dipole   | 0.023           | 0.004        |
| Monopole | 0.024           | 0.008        |

### 4 DISCUSSION AND CONCLUSIONS

Recently, several PTA collaborations have reported detections of a CRN, following early indications that such a process may exist (Lentati et al. 2015; Arzoumanian et al. 2018). The detection of a CRN process is consistent with the presence of a GWB (Romano et al. 2021), leading to cautious optimism for future GWB detection prospects among the community. However, PTA collaborations also recognize that the detection of a CRN process is not necessarily related to the presence of the GWB in current datasets.

Along this cautionary line, Goncharov et al. (2021b) raised the possibility of timing noise “masquerading” as a CRN. In this work, we have ventured further down this avenue, and have found that a CRN process can be falsely detected under a very wide range of pulsar timing noise conditions, under currently-used hypotheses and methodologies. We have shown that these spurious detections are highly sensitive to the span of the priors on timing noise parameters: when the prior distributions more closely match the real distribution of timing noise parameters, support for a CRN diminishes. While there have been previous and ongoing efforts to account for individual pulsar noise terms (e.g. Goncharov et al. 2021a), our results are
Figure 8. Percentage of realizations with a maximum-likelihood optimal statistic S/N greater than 3, for HD (top), dipolar (middle), and monopolar (bottom) spatial correlations, as a function of the variation in input timing noise parameters, $\Delta \log_{10} P_0$ and $\Delta \gamma$. Contours are as in Figure 7.

Figure 9. Same as Figure 8, but for the noise-marginalized optimal statistic.
not surprising given the degeneracy between pulsar timing noise and autocorrelation terms of the GWB (Hazboun et al. 2020; Lasky et al. 2015; Coles et al. 2011; Shannon & Cordes 2010). The overwhelming false support for the presence of a CRN under disparate timing noise conditions, under standard assumptions, choices of priors, and models, should lower confidence in the interpretation of recent CRN detections as “pre-cursor” detections of the GWB.

On the other hand, while spatial correlations remain undetected, our analysis shows that the detection of spatial correlations (unsurprisingly) is much more robust evidence for the existence of the GWB than the CRN is, in terms of false-detection rates. Our analysis also provides further support for the superior performance of the noise-marginalized optimal statistic over the standard maximum-likelihood method for mitigating biases in the search for spatial correlations.

There have been recent and ongoing efforts to characterise the CRN more robustly and efficiently (e.g. Goncharov et al. 2022; Taylor et al. 2022; Johnson et al. 2022; Hazboun et al. 2020). These works are crucial as PTA collaborations move toward obtaining the first positive detections of the nanohertz GWB. Our results here sound a strong cautionary note of the perverse influence that independent pulsar noise terms and choice of priors can have in the efforts to detect the GWB. Further development is required if we are indeed in the “intermediate” S/N regime of GWB detection (Romano et al. 2021).

Figure 10. Distribution of S/N ratios $\rho$ for HD, dipole, and monopole correlations, for the maximum-likelihood (top) and mean noise-marginalized (bottom) optimal statistic, over the full ($\Delta \log_{10} P_0, \Delta y$) parameter space. The black dashed line indicates a S/N threshold of 3.

Figure 11. Noise-marginalized optimal statistic (top) and S/N distributions (bottom) for HD, dipolar, and monopolar correlations from a simulation containing only independent pulsar timing noise with ($\Delta \log_{10} P_0, \Delta y$) = (1.4, 0.8). The squared fixed-slope common noise spectrum amplitude is shown in grey in the top panel. Dashed vertical lines indicate the posterior mean for the corresponding PDF (marginalized over the timing noise parameters), whereas the dash-dotted lines indicate the optimal statistic amplitude and S/N at the maximum-likelihood timing noise parameter values.

ACKNOWLEDGEMENTS

We would like to acknowledge Bill Coles, Paul Baker, Jeff Hazboun, Nihan Pol, and the IPTA gravitational-wave analysis working group for useful discussion during the development of this project. We thank the anonymous reviewer for their constructive comments, and Valentina di Marco for reviewing the text. This work has been carried out by the Parkes Pulsar Timing Array, which is part of the International Pulsar Timing Array. The Parkes radio telescope (Murriyang) is part of the Australia Telescope, which is funded by the Commonwealth Government for operation as a National Facility managed by CSIRO. We acknowledge the Wiradjuri people as the traditional owners of the Parkes observatory site. Parts of this research were carried out on the traditional lands of the Wajarri people. This paper includes archived data obtained through the CSIRO Data Access Portal (https://data.csiro.au). Parts of this research were conducted by the Australian Research Council (ARC) Centre of Excellence for Gravitational Wave Discovery (OzGrav), through project number CE170100004. R.M.S. acknowledges support through ARC future fellowship FT190100155. BG is supported by the Italian Ministry of Education, University and Research within the PRIN 2017 Research Program Framework, n. 2017SYRTCN. Work at NRL is
Figure 12. Optimal statistic-derived spatial covariances for a realization with \((\Delta \log_{10} P, \Delta \nu) = (1.4, 0.8)\) having a noise-marginalized S/N of 3.2 for HD correlations. Pulsar pairs are grouped into bins according to their angular offsets before computing the average cross-correlated power per angular bin. The corresponding noise-marginalized optimal statistic and S/N distribution are shown in Figure 11. The expected covariance for HD, dipole, and monopole offsets before computing the average cross-correlated power per angular bin.

supported by NASA. This work made use of the following software: ASTROPy (Astropy Collaboration et al. 2013, 2018), ENTERPRISE (Ellis et al. 2019), LBSTEMPO (Vallisneri 2020), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), PTMCMCSAMPLER (Ellis & van Haasteren 2017), PTASIMULATE (https://bitbucket.org/psrsoft/ptasimulate), TEMPO2 (Edwards et al. 2006).

DATA AVAILABILITY

Simulated datasets and code for this work are available at https://github.com/andrewzic/gwb_crn_sims.

REFERENCES

Anholm M., Ballmer S., Creighton J. D. E., Price L. R., Siemens X., 2009, Phys. Rev. D, 79, 084030
Antoniadis J., et al., 2022, MNRAS, 510, 4873
Arzoumanian Z., et al., 2016, ApJ, 821, 13
Arzoumanian Z., et al., 2018, ApJ, 859, 47
Arzoumanian Z., et al., 2020, ApJ, 905, L34
Arzoumanian Z., et al., 2021, Phys. Rev. Lett., 127, 251302
Astropy Collaboration et al., 2013, A&A, 558, A33
Astropy Collaboration et al., 2018, AJ, 156, 123
Bailes M., et al., 2020, Publ. Astron. Soc. Australia, 37, e028
Bonetti M., Sesana A., Barausse E., Haardt F., 2018, MNRAS, 477, 2599
Carlin B. P., Chib S., 1995, Journal of the Royal Statistical Society. Series B (Methodological), 57, 473
Chamberlin S. J., Creighton J. D. E., Siemens X., Demorest P., Ellis J., Price L. R., Romano J. D., 2015, Phys. Rev. D, 91, 044048
Chen S., et al., 2021, MNRAS, 508, 4970
Coles W., Hobbs G., Champion D. J., Manchester R. N., Verbiest J. P. W., 2011, MNRAS, 418, 561
Cordes J. M., 2013, Classical and Quantum Gravity, 30, 224002
Demorest P. B., et al., 2013, ApJ, 762, 94
Dvorkin I., Barausse E., 2017, MNRAS, 470, 4547
Edwards R. T., Hobbs G. B., Manchester R. N., 2006, MNRAS, 372, 1549
Ellis J., van Haasteren R., 2017, Jellis18/Ptmcmcsampler: Official Release, Zenodo, doi:10.5281/zenodo.1037579
Ellis J., van Haasteren R., 2019, PTMCMCSampler: Parallel tempering MCMC sampler package written in Python, Astrophysics Source Code Library, record ascl:1912.017 (ascl:1912.017)
Ellis J. A., Vallisneri M., Taylor S. R., Baker P. T., 2019, ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust Pulsar Inference Suite, Astrophysics Source Code Library, record ascl:1912.015 (ascl:1912.015)
Foster R. S., Backer D. C., 1990, ApJ, 361, 300
Goncharov B., et al., 2021a, MNRAS, 502, 478
Goncharov B., et al., 2021b, ApJ, 917, L19
Goncharov B., et al., 2022, ApJ, 932, L22
Grishchuk L. P., 1976, Soviet Journal of Experimental and Theoretical Physics Letters, 23, 293
Groth E. J., 1975, ApJS, 29, 453
Harris C. R., et al., 2020, Nature, 585, 357
Haboub J. S., Simon J., Siemens X., Romano J. D., 2020, ApJ, 905, L6
Hee S., Handley W. J., Hobson M. P., Lasenby A. N., 2016, MNRAS, 455, 2461
Hellings R. W., Downs G. S., 1983, ApJ, 265, L39
Hobbs G., et al., 2010a, Classical and Quantum Gravity, 27, 084013
Hobbs G., Lyne A. G., Kramer M., 2010b, MNRAS, 402, 1027
Hunter J. D., 2007, Computing in Science & Engineering, 9, 90
Jenet F. A., Hobbs G. B., Lee K. J., Manchester R. N., 2005, ApJ, 625, L123
Joachimi B., Köhlinger F., Handley W., Lemos P., 2021, A&A, 647, L5
Johnson A. D., Vigeland S. J., Siemens X., Taylor S. R., 2022, ApJ, 932, 105
Joshi B. C., et al., 2018, Journal of Astrophysics and Astronomy, 39, 51
Kerr M., et al., 2020, Publ. Astron. Soc. Australia, 37, e020
Kobahkhide A., Lagger C., Manning A., Yue J., 2017, European Physical Journal C, 77, 570
Kramer M., Champion D. J., 2013, Classical and Quantum Gravity, 30, 224009
Lasky P. D., Melatos A., Ravi V., Hobbs G., 2015, MNRAS, 449, 3293
Lasky P. D., et al., 2016, Physical Review X, 6, 011035
Lee K. J., 2016, in Qian L., Li D., eds, Astronomical Society of the Pacific Conference Series Vol. 502. Frontiers in Radio Astronomy and FAST Early Sciences Symposium 2015. p. 19
Lentati L., et al., 2015, MNRAS, 453, 2576
Linde A. D., 1982, Physics Letters B, 108, 389
Lyne A., 1999, in Arzoumanian Z., Van der Hooft F., van den Heuvel E. P. J., eds, Pulsar Timing, General Relativity and the Internal Structure of Neutron Stars. p. 141
Manchester R. N., et al., 2013, Publ. Astron. Soc. Australia, 30, e017
McLaughlin M. A., 2013, Classical and Quantum Gravity, 30, 224008
Meyers P. M., Melatos A., O’Neill N. J., 2021, MNRAS, 502, 3113
Mingarelli C. M. F., Mingarelli A. B., 2018, Journal of Physics Communications, 2, 105002
Ölmez S., Mandic V., Siemens X., 2010, Phys. Rev. D, 81, 104028
Parthasarathy A., et al., 2019, MNRAS, 489, 3810
Phinney E. S., 2001, arXiv e-prints, pp astro-ph/0108028
Pol N. S., et al., 2021, ApJ, 911, L34
Romano J. D., Hazboun J. S., Siemens X., Archibald A. M., 2021, Phys. Rev. D, 103, 063027
Rosado P. A., Sesana A., Gair J. 2015, MNRAS, 451, 2417
Sesana A., 2013, MNRAS, 433, L1
Shannon R. M., Cordes J. M., 2010, ApJ, 725, 1607
Shannon R. M., et al., 2015, Science, 349, 1522
Starobinsky A. A., 1980, Physics Letters B, 91, 99
Taylor S. R., Vallisneri M., Ellis J. A., Mingarelli C. M. F., Lazio T. J. W., van Haasteren R., 2016, ApJ, 819, L6
Taylor S. R., Lentati L., Babak S., Brem P., Gair J. R., Sesana A., Vecchio A., 2017, Phys. Rev. D, 95, 042002
Taylor S. R., van Haasteren R., Sesana A., 2020, Phys. Rev. D, 102, 084039
Taylor S. R., Simon J., Schult L., Pol N., Lamb W. G., 2022, Phys. Rev. D, 105, 084049
Tiburzi C., et al., 2016, MNRAS, 455, 4339
Vallisneri M., 2020, libstempo: Python wrapper for Tempo2, Astrophysics Source Code Library, record ascl:2002.017 (ascl:2002.017)
Vigeland S. J., Islo K., Taylor S. R., Ellis J. A., 2018, Phys. Rev. D, 98, 044003
Wyithe J. S. B., Loeb A., 2003, ApJ, 590, 691

MNRAS 000, 1–12 (2022)
Xue X., et al., 2021, Phys. Rev. Lett., 127, 251303

This paper has been typeset from a TeX/LaTeX file prepared by the author.