RAA 2017 Vol. **17** No. **2**, 19 (14pp) doi: 10.1088/1674–4527/17/2/19

http://www.raa-journal.org *http://iopscience.iop.org/raa*

Research in Astronomy and Astrophysics

Statistical analyses for NANOGrav 5-year timing residuals

Yan Wang[1,2,3], James M. Cordes[4], Fredrick A. Jenet[2,3], Shami Chatterjee[4], Paul B. Demorest[5], Timothy Dolch[6], Justin A. Ellis[7], Michael T. Lam[4], Dustin R. Madison[5], Maura A. McLaughlin[8], Delphine Perrodin[9], Joanna Rankin[10], Xavier Siemens[11] and Michele Vallisneri[7]

[1] MOE Key Laboratory of Fundamental Physical Quantities Measurements, School of Physics, Huazhong University of Science and Technology, Wuhan 430074, China; *ywang12@hust.edu.cn*

[2] Center for Advanced Radio Astronomy, University of Texas at Brownsville, 1 West University Boulevard, Brownsville, TX 78520, USA

[3] Department of Physics and Astronomy, University of Texas at Brownsville, 1 West University Boulevard, Brownsville, TX 78520, USA

[4] Department of Astronomy, Cornell University, Ithaca, NY 14853, USA

[5] National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903, USA

[6] Department of Physics, Hillsdale College, 33 E. College Street, Hillsdale, MI 49242, USA

[7] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91106, USA

[8] Department of Physics, West Virginia University, P.O. Box 6315, Morgantown, WV 26505, USA

[9] INAF-Osservatorio Astronomico di Cagliari, Via della Scienza 5, 09047 Selargius (CA), Italy

[10] Department of Physics, University of Vermont, Burlington, VT 05405, USA

[11] Center for Gravitation, Cosmology and Astrophysics, Department of Physics, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201, USA

Received 2016 September 9; accepted 2016 October 27

Abstract In pulsar timing, timing residuals are the differences between the observed times of arrival and predictions from the timing model. A comprehensive timing model will produce featureless residuals, which are presumably composed of dominating noise and weak physical effects excluded from the timing model (e.g. gravitational waves). In order to apply optimal statistical methods for detecting weak gravitational wave signals, we need to know the statistical properties of noise components in the residuals. In this paper we utilize a variety of non-parametric statistical tests to analyze the whiteness and Gaussianity of the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) 5-year timing data, which are obtained from Arecibo Observatory and Green Bank Telescope from 2005 to 2010. We find that most of the data are consistent with white noise; many data deviate from Gaussianity at different levels, nevertheless, removing outliers in some pulsars will mitigate the deviations.

Key words: pulsar timing array: general — statistical tests

1 INTRODUCTION

Pulsar timing is a powerful technique and has achieved many of the most important science results in pulsar astronomy. The timing of single pulsars has been used as a probe of the dispersive interstellar medium (Cordes & Lazio 2002), to test theories of gravitation in the strong field regime (Damour & Taylor 1992; Stairs 2003; Kramer et al. 2006), to discover the first extrasolar planetary system (Wolszczan & Frail 1992), and to constrain the nuclear equation of state of a neutron star (Demorest et al. 2010; Lattimer & Prakash 2007, 2010). It has provided the first evidence of the existence of gravitational waves (GWs) (Taylor & Weisberg 1982, 1989). Timing a number of pulsars and analyzing the data coherently can be used to search for irregularities in terrestrial time standards and develop a timescale based on pulsars (Hobbs et al. 2012), and to deepen our understanding of solar system dynamics (Champion et al. 2010). Amazingly, it can be operated as a Galactic scale detector for very-low frequency GWs (Hellings & Downs 1983; Foster & Backer 1990; Jenet et al. 2005).

A pulsar timing array (PTA) is an experiment that regularly observes a set of millisecond pulsars (MSPs). Currently, three PTAs, i.e., the North American Nanohertz Observatory for Gravitational Waves (NANOGrav, Demorest et al. (2013)), the Parkes Pulsar Timing Array (PPTA, Manchester et al. (2013)) and the European Pulsar Timing Array (EPTA, Ferdman et al. (2010)), have started to produce astrophysically important results. These PTAs compose the International Pulsar Timing Array (IPTA, Manchester & IPTA (2013); McLaughlin (2014)) with approximately 50 pulsars that are regularly monitored. The first data combination has been released (Verbiest et al. 2016).

A PTA is sensitive to very low frequency (10^{-9}–10^{-7} Hz) GWs, and is complementary to ground based interferometric detectors (e.g., LIGO (Abbott et al. 2009) and Virgo (Accadia et al. 2011)) running in the high frequency band ($10-10^3$ Hz), and space based laser rangers (e.g., eLISA (eLISA Consortium et al. 2013) and TianQin (Luo et al. 2016)) proposed for the low frequency band (0.1 mHz–0.1 Hz). Potential sources of GWs in very low frequencies include supermassive black hole binaries (Jaffe & Backer 2003; Wyithe & Loeb 2003; Sesana et al. 2008), cosmic strings (Damour & Vilenkin 2005; Ölmez et al. 2010), and relic GWs (Grishchuk 2005).

At the current timing precision, it is very likely that different kinds of noise are the dominant components of timing residuals (Jenet et al. 2006; van Haasteren et al. 2011; Demorest et al. 2013; Shannon et al. 2013; Arzoumanian et al. 2014). On one hand, to improve timing precision at the longest timescale, it is very important to have a comprehensive understanding of the sources (e.g., radiometer, pulse phase jitter, diffractive interstellar scintillation) and the characteristics of noise in terms of times of arrival (TOAs), and identify mitigation methods to reduce the noise (Cordes & Shannon 2010; Wang 2015). On the other hand, many data analysis methods designed for detecting weak GW signals (Corbin & Cornish 2010; Babak & Sesana 2012; Ellis et al. 2012; Wang et al. 2014, 2015; Zhu et al. 2015, 2016) are usually geared to work well for data having some specific statistical properties. Blindly applying these data analysis strategies and pipelines without checking the presumptions may lead to nonsensical results (Tiburzi et al. 2016). In this paper, as a first step in noise characterization, we implement a suite of robust non-parametric statistical tools to test the most important noise properties, namely the whiteness and Gaussianity, of the NANOGrav 5-year (2005–2010) data published in Demorest et al. (2013).

Using these tools, we found that most of the frequency separated data are individually consistent with the whiteness assumption, except the high frequency data from PSR J2145–0754 and J2317+1439 which show mild deviations. However, combining data from different frequencies as one set causes significant deviations for PSR J0613–0200, J1455–3330, J1744–1134, J1909–3744, J1918–0642 and J2317+1439. We found that this may be due to the minute inaccuracy of DM estimation for these pulsars with the current observation strategy. In terms of Gaussianity, most of the data show different levels of deviation, however, removing outliers in some pulsars would reduce the deviations.

The rest of the paper is organized as follows. In Section 2, a brief description of the observation and data set is given. We use the zero-crossing method to test the whiteness of the data in Section 3, and use descriptive statistics and hypothesis testings to check the Gaussianity in Section 4. Demonstrations of these analyses on three pulsars are presented in Section 5. The paper is concluded in Section 6.

2 OBSERVATIONS AND DATA

The NANOGrav collaboration has conducted observations with the Arecibo Observatory (AO) and the Green Bank Telescope (GBT), two of the largest single dish radio telescopes to date. Currently 37 MSPs (Arzoumanian et al. 2015) have been regularly timed by NANOGrav. The first five years of data (2005–2010) for 17 MSPs along with an upper limit on the GW stochastic background have been published in Demorest et al. (2013). In order to precisely analyze the time dependent dispersion measure (DM) and frequency dependent pulse shape, two receivers operating at 1.4 GHz and 430 MHz for AO and 1.4 GHz and 820 MHz for GBT have been used in most of the observations. Observations using two different receivers were not simultaneous. At AO, the observations from the two bands were obtained within 1 hour; whereas at GBT, the separation could be up to a week. All observations during this 5-year period have been carried out with identical pulsar backends, i.e. the Astronomical Signal Processor (ASP) at AO and the Green Bank Astronomical Signal Processor (GASP) at GBT, in which the input analog signal is split into

324 MHz channels (sub-bands). Due to the limitation imposed by the real-time computation load or the receiver bandpass, typically 16 channels would be processed in most observations. The cadence between observation sessions is typically 4–6 weeks. There is a gap in the observations of all pulsars in 2007 due to maintenance at both telescopes.

The data product from an observation epoch is the pulse TOA which is the time that is recorded for radio emission, from a fiducial rotation phase of a pulsar, arriving at a telescope. The standard TOA estimation includes polarization calibration, pulse profile folding, profile template creation and TOA measurement by correlating the folded profile and the profile template. Those steps are integrated in the package PSRCHIVE (Hotan et al. 2004) and ASPFitsReader (Ferdman 2008). Both packages are used for cross-checking of errors which otherwise would hardly be targeted.

The next step in timing analysis is to fit the observed TOAs of each pulsar to a timing model. The timing model contains a set of physical parameters which account for the pulsar's rotation (spin period, spin period derivatives), astrometry (position, proper motion), interstellar medium (DM), binary orbital dynamics, etc. This procedure is executed in the standard timing analysis package TEMPO2 (Hobbs et al. 2006; Edwards et al. 2006) via a weighted least squares fitting. The so-called post-fit timing residuals are the differences between the measured TOAs and the TOAs predicted by this model. A positive residual means that the observed pulse arrives later than expected. The timing residuals potentially contain stochastic noise from various sources and physical effects that are not included in the timing model. One can refer to Demorest et al. (2013) for a thorough account of the NANOGrav observation strategy and related timing analysis.

We can generate multiple timing residuals, denoted as $r(t,\nu)$, from timing analysis of the NANOGrav data set, where t is the time of observation of a pulse in Modified Julian Date (MJD) and ν is the central frequency of a channel. To simplify the study of timing effects induced by achromatic physical processes (e.g. a GW), we can average the timing residuals from the TOAs recorded from the same rotation phase of the pulsar. If there are only TOAs from one pulsar rotation in an observation epoch (true for most observations), this averaged residual will equal to the daily averaged residual used in figure 1 of Demorest et al. (2013) and in Perrodin et al. (2013). The averaged residual r_I for the I-th observation epoch in the data of a pulsar is

$$r_I = \frac{\sum_{i=1}^{N_I} r_{Ii}\sigma_{Ii}^{-2}}{\sum_{i=1}^{N_I} \sigma_{Ii}^{-2}}\,, \tag{1}$$

where $r_{Ii} = r(t_I, \nu_i)$ is the post-fit multi-frequency timing residual from the i-th frequency channel at the I-th observation epoch, N_I is the number of frequency channels and σ_{Ii} is the uncertainty for the corresponding TOA. The uncertainty associated with the averaged residual is

$$\sigma_I = \sqrt{\left(\sum_{i=1}^{N_I} \sigma_{Ii}^{-2}\right)^{-1} \frac{1}{N_I - 1} \sum_{i=1}^{N_I} (r_{Ii} - r_I)^2 \sigma_{Ii}^{-2}}\,. \tag{2}$$

Equation (2) is the standard deviation of Equation (1) with correction for underestimation of errors in TOA. This estimator is suitable when σ_{Ii} does not include all the noise sources associated with TOAs. In addition, since we have not used two independent receivers simultaneously, we will separate the low frequency and high frequency averaged residuals and test them independently in our analysis.

The averaged timing residuals can be used as inputs for the GW detection pipelines. One advantage of averaging is that it reduces the random noise components across different frequency channels while keeping the achromatic GW signals intact. Moreover, the averaged residuals provide a quantitative way to compare with data from EPTA (Ferdman et al. 2010; Lentati et al. 2015) and PPTA (Manchester et al. 2013), which have routinely produced a single TOA per observation epoch.

3 WHITENESS TEST

In this section, we test the consistency of the averaged timing residuals with the white noise assumption for each pulsar. A white noise time series is statistically uncorrelated in time, but the distribution of its associated values does not necessarily adhere to any specific probability distribution (Gaussian, Poisson, etc.). If evenly sampled, we can use Fourier analysis to calculate the power spectrum of the time series and to check whether it is consistent with a flat spectrum in the interested frequency range. However, the pulsar timing data are usually not evenly sampled, i.e. the observation cadence varies, so that this conventional spectral analysis is not applicable. The Lomb-Scargle periodgram (Lomb 1976; Scargle 1982) which is designed for unevenly sampled data suffers from occasional large gaps between observations (see fig. 1 in Demorest et al. 2013), as well as limited data volume for each pulsar (see Table 1 for detailed numbers).

Table 1 Results for the inferential statistical tests on Gaussianity. The numbers represent the sample size. For post-fit residuals, if $p > 0.05$, the data are consistent with Gaussianity (labeled by 'Y'), if $0.05 > p > 10^{-3}$, the data mildly deviate from Gaussianity ('n'), and if $p < 10^{-3}$, the data strongly deviate from Gaussianity ('N'). For averaged residuals, the criterion intervals are set to be $p > 0.1$ ('Y'), $0.1 > p > 2 \times 10^{-3}$ ('n') and $p < 2 \times 10^{-3}$ ('N') respectively. 'NA' appears when the test is not applicable to such a small sample size.

Source	P	DM	Averaged timing residuals						Post-fit
	(ms)	(pc cm^{-3})	327 MHz	430 MHz	820 MHz	1.4 GHz	2.3 GHz	Comb.	
J0030+0451	4.87	4.33	–	24 N	–	26 n	–	50 N	545 Y
J0613−0200	3.06	38.78	–	–	40 Y	40 Y	–	80 N	1113 n
J1012+5307	5.26	9.02	–	–	47 N	63 n	–	110 N	1678 n
J1455−3330	7.99	13.57	–	–	41 n	45 Y	–	86 N	1100 n
J1600−3053	3.60	52.33	–	–	22 n	26 n	–	48 n	625 Y
J1640+2224	3.16	18.43	–	34 N	-	32 N	–	68 N	631 N
J1643−1224	4.62	62.42	–	–	47 N	50 Y	–	97 N	1266 N
J1713+0747	4.57	15.99	–	–	38 N	84 N	31 Y	153 N	2368 N
J1744−1134	4.07	3.14	–	–	48 N	60 Y	–	108 N	1617 N
J1853+1308	4.09	30.57	–	–	–	41 Y	2 NA	43 Y	497 Y
B1855+09	5.36	13.30	–	37 N	–	32 Y	–	69 n	702 N
J1909−3744	2.95	10.39	–	–	35 n	33 Y	–	68 N	1001 N
J1910+1256	4.98	34.48	–	–	–	31 Y	6 Y	37 Y	525 Y
J1918−0642	7.65	26.60	–	–	40 Y	54 n	–	94 N	1306 Y
B1953+29	6.13	104.50	–	–	–	23 Y	2 NA	25 Y	208 Y
J2145−0750	16.05	9.03	–	–	22 N	24 Y	–	46 n	675 n
J2317+1439	3.45	21.90	43 n	41 n	–	–	–	84 N	458 n

It turns out that after subtracting the mean value, number of zero-crossing Z_W in a white noise time series is a Gaussian random variable,

$$Z_W \sim \mathcal{N}(\mu_{Z_W}, \sigma^2_{Z_W}), \tag{3}$$

with expected value $\mu_{Z_W} = (N-1)/2$ and standard deviation $\sigma_{Z_W} = \sqrt{N-1}/2$. The zero-crossing test checks how large the number of zero-crossings is for a time series compared to the expectation. It is designed to operate in the time domain, and is thus applicable to unevenly sampled data with gaps. This test is not sensitive to any non-stationarity associated with statistics of white noise, such as the case where the white noise has a jump in variance at some epoch because of a change in instrumentation. It assumes that the white noise is "dense" which means that all data values are non-zero and stochastic (Papoulis 1984). Other kinds of white noise, such as shot noise with a low shot rate, cannot be analyzed with the zero-crossing test described here.

In Table 2, we show the results of the zero-crossing test for the frequency separated averaged residuals as well as the total averaged residuals (by combining the high and low frequency averaged residuals and sorting them in ascending order of corresponding TOAs) for 17 pulsars. N_{crs} is the actual number of zero crossings for the data and Δ is the difference between μ_{Z_W} and N_{crs}. The significance of the test is measured by how large Δ is compared with σ_{Z_W}. If $|\Delta| < 2\sigma_{Z_W}$ (>5% in terms of p-value [1]), the data are said to be consistent with white noise (labeled by 'Y'); if $3\sigma_{Z_W} > |\Delta| > 2\sigma_{Z_W}$, they are said to mildly deviate from white noise ('n'); and if $|\Delta| > 3\sigma_{Z_W}$, they are said to strongly deviate from white noise ('N'). We defer a detailed discussion and possible interpretation of these results to Section 5 in order to consolidate with results from the Gaussianity tests.

4 GAUSSIANITY TEST

In this section, we first use descriptive statistics, namely histograms and Quantile-Quantile (Q-Q) plots, to visually inspect the general features of the data. Then we implement a suite of inferential statistical tests to quantitatively measure the deviations from Gaussianity. The observation conditions changed during the 5 years that observations were acquired due to a number of factors, for instance, radiometer noise, interstellar scintillation and factors inherent to the instruments. Therefore, the underlying random variables representing noise at different frequencies and epochs are heteroscedastic. These changes are reflected in variations of the error bars (e.g., for the averaged residuals shown in Figs. 1, 4 and 7).

[1] p-value gives the probability of obtaining a test statistic (N_{crs}) at least as extreme as the one that was actually observed, assuming that the presumption (e.g., whiteness) is true.

Table 2 Results for the Zero-crossing Test. Consistent with whiteness – 'Y', mildly deviate – 'n' and strongly deviate – 'N'.

Source	Low-frequency band					High-frequency band					Combined				
	N_{crs}	μ_{Z_W}	Δ	σ_{Z_W}	Y/n/N	N_{crs}	μ_{Z_W}	Δ	σ_{Z_W}	Y/n/N	N_{crs}	μ_{Z_W}	Δ	σ_{Z_W}	Y/n/N
J0030+0451	12	11.5	–0.5	2.4	Y	14	12.5	–1.5	2.5	Y	27	24.5	–2.5	3.5	Y
J0613−0200	20	19.5	–0.5	3.1	Y	19	19.5	0.5	3.1	Y	58	39.5	–18.5	4.4	N
J1012+5307	26	23	–3	3.4	Y	34	31	–3	3.9	Y	62	54.5	–7.5	5.2	Y
J1455−3330	22	20	–2	3.2	Y	23	22	–1	3.3	Y	58	42.5	–15.5	4.6	N
J1600−3053	13	10.5	–2.5	2.3	Y	13	12.5	–0.5	2.5	Y	33	23.5	–9.5	3.4	n
J1640+2224	18	16.5	–1.5	2.9	Y	14	15.5	1.5	2.8	Y	35	32.5	–2.5	4.0	Y
J1643−1224	26	23	–3	3.4	Y	22	24.5	2.5	3.5	Y	62	48	–14	4.9	n
J1713+0747	19	18.5	–0.5	3	Y	46	41.5	–4.5	4.6	Y	–	–	–	–	–
J1713+0747	–	–	–	–	–	18	15	–3	2.7	Y	94	76	–18	6.2	n
J1744−1134	25	23.5	–1.5	3.4	Y	33	29.5	–3.5	3.8	Y	70	53.5	–16.5	5.2	N
J1853+1308	21	20	–1	3.2	Y	–	–	–	–	–	19	21	2	3.2	Y
B1855+09	17	18	1	3.0	Y	17	15.5	–1.5	2.8	Y	46	34	–12	4.1	n
J1909−3744	17	17	0	2.9	Y	17	16	–1	2.8	Y	52	33.5	–18.5	4.1	N
J1910+1256	13	15	2	2.7	Y	3	2.5	–0.5	1.1	Y	21	18	–3	3	Y
J1918−0642	20	19.5	–0.5	3.1	Y	29	26.5	–2.5	3.6	Y	64	46.5	–17.5	4.8	N
B1953+29	14	11	–3	2.3	Y	–	–	–	–	–	16	12	-4	2.4	Y
J2145−0750	15	10.5	–4.5	2.3	Y	18	11.5	–6.5	2.4	n	31	22.5	–8.5	3.4	n
J2317+1439	26	21	–5	3.2	Y	27	20	–7	3.2	n	56	41.5	–14.5	4.6	N

This aspect is treated here by a simple normalization, so that the tested time series are from the same underlying distribution. For the averaged residuals, each residual is normalized according to Equation (1) by its associated uncertainty calculated in Equation (2). In addition, each of the multi-frequency residuals are normalized by the uncertainty associated with its TOA.

Inferential statistics based on statistical hypothesis testing theory argues against a *null hypothesis* (Gaussianity) analogous to mathematical proof by contradiction. First, the data are summarized into a single number called the *test statistic*, which follows a certain probability distribution. Second, a p-value is calculated based on this distribution assuming that the null hypothesis is true. The lower the p-value is, the smaller the chance that the sample comes from a Gaussian distribution.

One often rejects the null hypothesis when the p-value is less than a pre-assigned significance level which is usually 0.05. However, the power of the tests decreases as the sample size decreases. It is a common practice to set the significance level at higher values, such as 0.1 or 0.2 for a small sample in order to detect possible deviation that may be present. This is an important point since the data sets that we will test vary greatly in size (cf. Table 1).

To avoid possible bias in different tests, the significance of the Gaussianity test is measured by the averaged p-value of five tests, among which the Shapiro-Wilk test (S-W) and Shapiro-Francia test (S-F) are order statistics; the Anderson-Darling test (A-D), Cramér-von Mises test (C-vM) and Lilliefors test (Lillie) are based on the empirical distribution function (EDF). The results are summarized in Table 1. For multi-frequency residuals, if $p > 0.05$, the data are said to be consistent with Gaussianity (labeled by 'Y'); if $0.05 > p > 10^{-3}$, the data are said to mildly deviate from Gaussianity ('n'); and if $p < 10^{-3}$, the data are said to strongly deviate from Gaussianity ('N'). For averaged residuals, the criterion intervals are set to be $p > 0.1$ ('Y'), $0.1 > p > 2 \times 10^{-3}$ ('n') and $p < 2 \times 10^{-3}$ ('N'), respectively. The p-values of all tests are only shown for three pulsars in the legends of Figures 3, 6 and 9.

5 RESULTS

The results for the whiteness and Gaussianity tests are summarized in Tables 1 and 2 respectively. Here, we describe in detail how results from these tests can be applied to three pulsars.

5.1 PSR J0613-0200

The frequency separated averaged timing residuals of PSR J0613–0200 are shown in Figure 1. The red asterisks with error bars represent high-frequency (1.4 GHz) residuals, and blue short-bars with error bars represent low frequency (820 MHz) residuals. Apparently, the high frequency residuals have larger variances than the low frequency residuals, and the high frequency error bars for this pulsar are a factor of a few larger than the low fre-

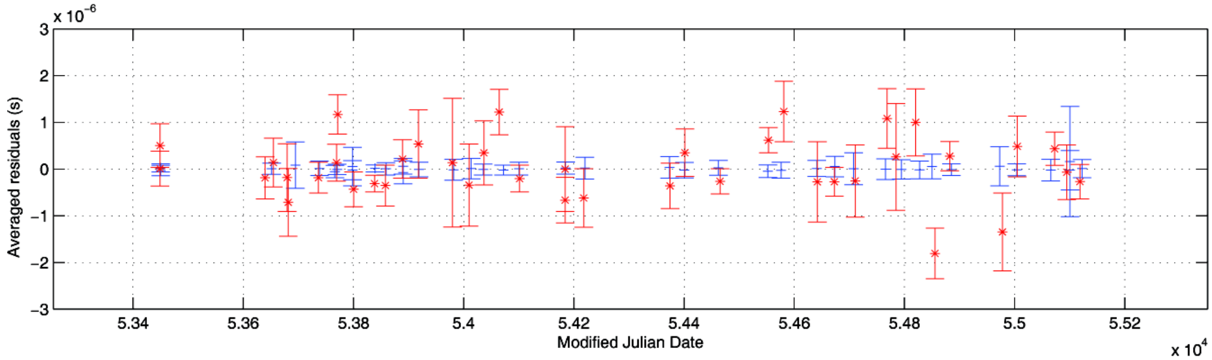

Fig. 1 Frequency separated averaged timing residuals with error bars for J0613–0200. The *red asterisks* represent high frequency data, while the *blue short-bars* represent low frequency data.

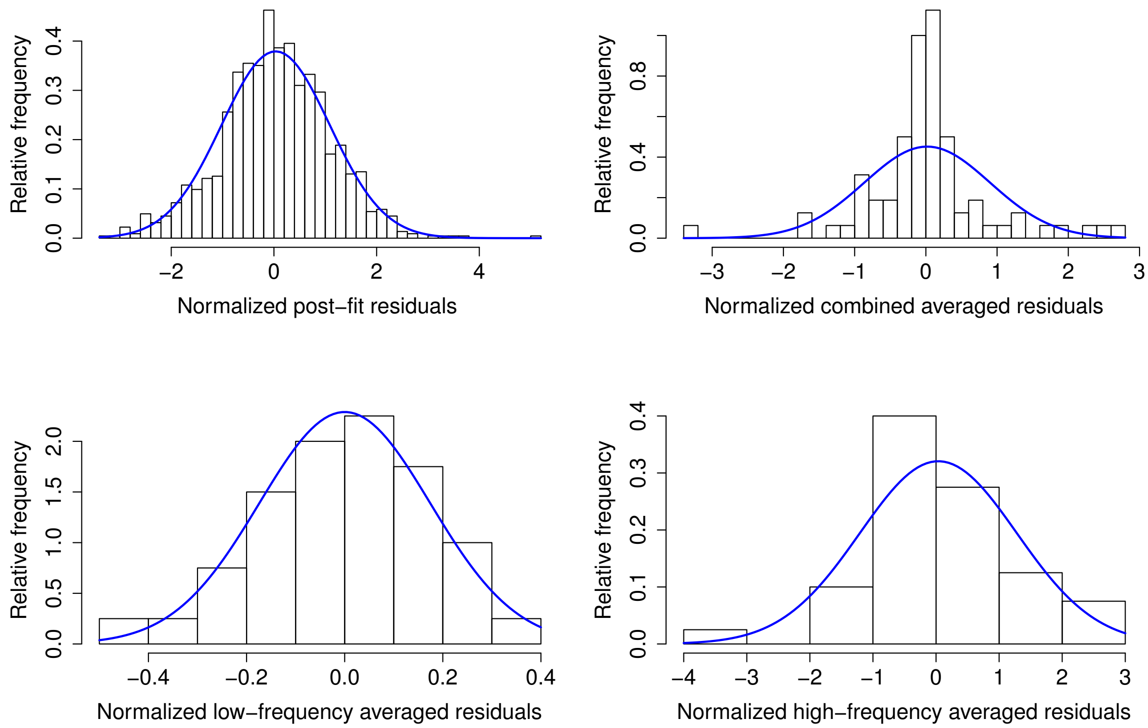

Fig. 2 Histograms for the post-fit multi-frequency residuals (*top-left*), total (*top-right*), low-frequency (*bottom-left*) and high-frequency (*bottom-right*) averaged residuals of J0613–0200. The *blue curve* is a Gaussian distribution with the same mean and variance as the data.

quency error bars. This is mainly due to the fact that the mean flux density at high frequency is lower than that at low frequency according to the power-law spectrum of the flux density. For similar integration time, this will result in a larger uncertainty in the measurement of TOA by correlating a lower S/N folded pulse with the template pulse profile (Taylor 1992).

From Table 2 we can see that the low frequency and high frequency residuals are individually consistent with the white noise assumption. However when they are combined into a single time series, the total residuals show more zero crossings than expected, and the deviation for this pulsar is more than 4σ. We found that the excess of zero crossings is caused by the error in estimated

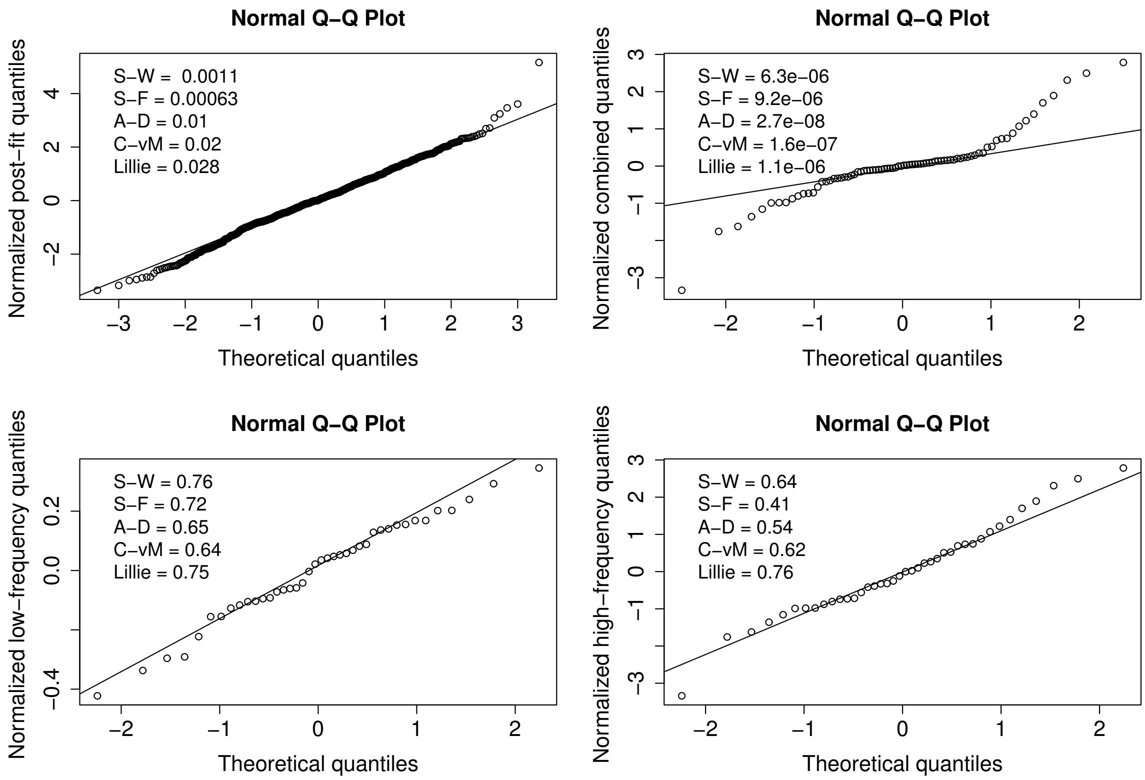

Fig. 3 Q-Q plots for the post-fit multi-frequency residuals (*top-left*), total (*top-right*), low-frequency (*bottom-left*) and high-frequency (*bottom-right*) averaged residuals of J0613–0200. The p-values of individual tests are listed in the legends.

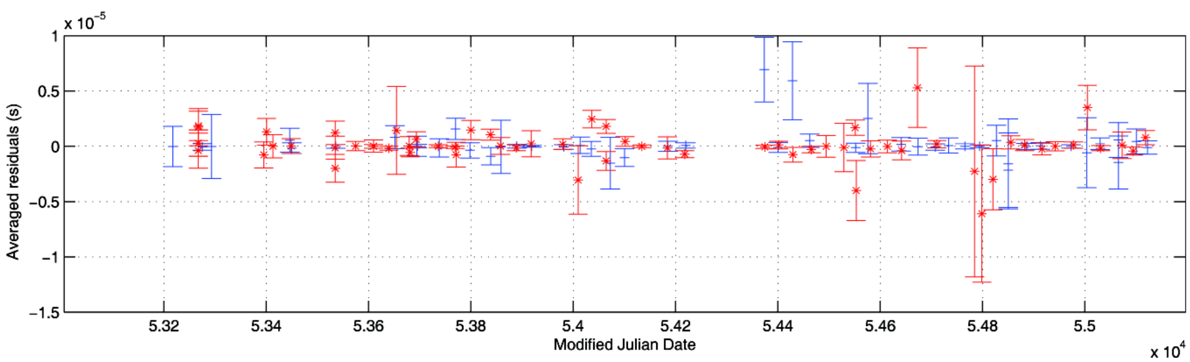

Fig. 4 Frequency separated averaged timing residuals with error bars for J1012+5307. The *red asterisks* represent high frequency data, while the *blue short-bars* represent low frequency data.

values of time dependent DM with the observation strategy adopted in the NANOGrav 5-year data. However, we defer a detailed discussion on this topic to Section 6.

After normalizing the averaged residuals by their associated error bars, we notice that in Figures 2 and 3 the standard deviation of the low frequency residuals is significantly smaller than unity, which hints that the error bars calculated for the low frequency averaged data are overestimated. Therefore, the combined residuals deviate from a Gaussian distribution, even if the low frequency averaged residuals and high frequency averaged residuals are both consistent with a Gaussian distribution individually. This may suggest that in order to properly combine the data from different frequency bands in GW detection algorithms, we may need to add a scaling parameter for each frequency band that is similar to the EFAC[2] parameter

[2] A multiplication factor for TOA error bars of each pulsar.

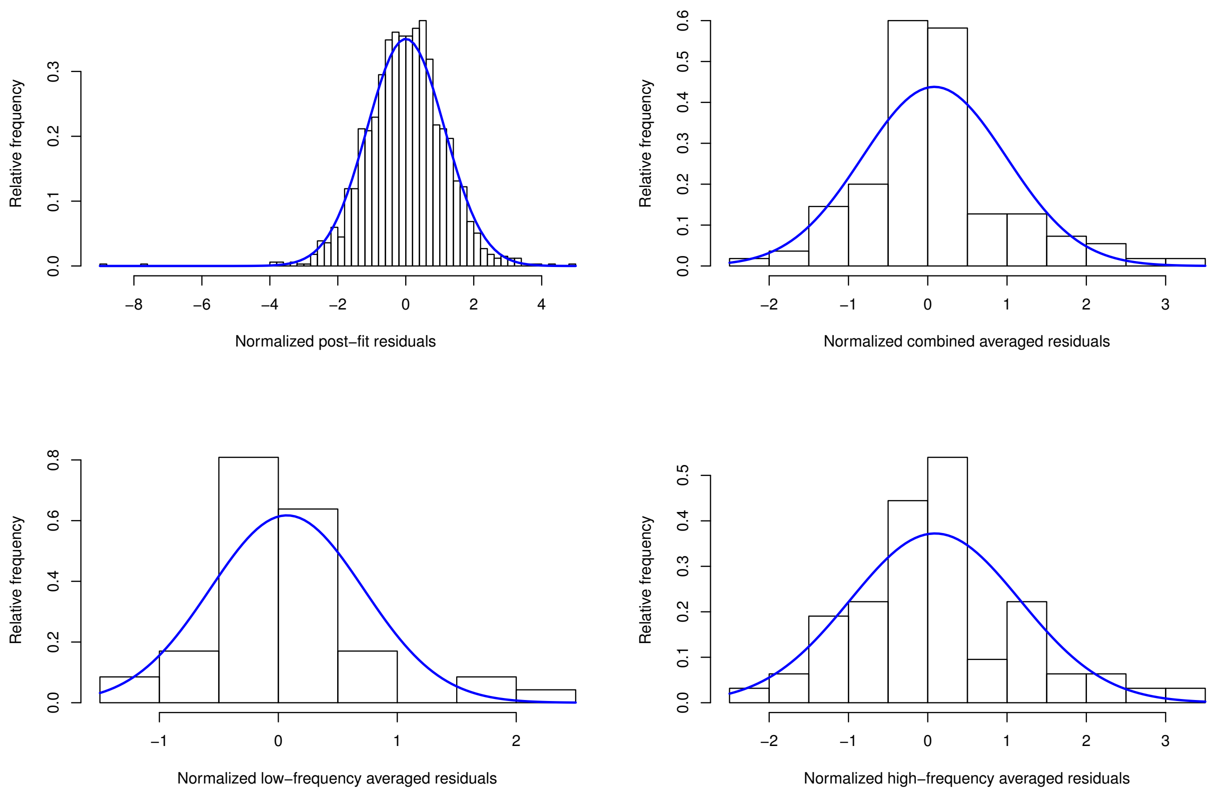

Fig. 5 Histograms for the post-fit residuals (*top-left*), total (*top-right*), low-frequency (*bottom-left*) and high-frequency (*bottom-right*) averaged residuals of J1012+5307. The *blue curve* is a Gaussian distribution with the same mean and variance as the data.

eter used in timing analysis. The post-fit multi-frequency residuals mildly deviate from a Gaussian distribution.

5.2 PSR J1012+5307

From Table 2, we can see that the low-frequency, high-frequency and total averaged residuals are all consistent with the white noise assumption. From Table 1, we can see that the high frequency averaged and post-fit multi-frequency residuals mildly deviate from a Gaussian distribution, while the low frequency averaged and total averaged residuals strongly deviate from a Gaussian distribution.

We notice from Figure 6 that for the post-fit residuals the results from the order statistic tests (strong deviation) are not consistent with the EDF tests (mild deviation). This is ascribed to the fact that the order statistic tests are sensitive to outliers, which can be identified from Figures 5 and 6. After removing two outliers in the residuals, the results from the order statistics are improved rapidly (S-W = 8.3×10^{-4}, S-F = 4.5×10^{-4}), and become more consistent with the other tests.

5.3 PSR J1713+0747

PSR J1713+0747 is the only one among the 17 pulsars that has been observed by both the AO and the GBT. Currently, it is the best timed pulsar in NANOGrav. It has been observed extensively in three frequency bands, i.e. 820 MHz, 1.4 GHz and 2.3 GHz [3], which are marked by blue short-bars, red asterisks and black squares respectively in Figure 7. (There are actually two sessions conducted in 2.7 GHz at the AO, which are not included in this analysis.)

The three frequency separated averaged residuals are all consistent with the white noise assumption individually, whereas the total averaged residuals mildly deviate from it. Except for the residuals from the 2.3 GHz band, the residuals from the other two bands all strongly devi-

[3] High frequency observations from 2.3 GHz, see Section 5.3 for details.

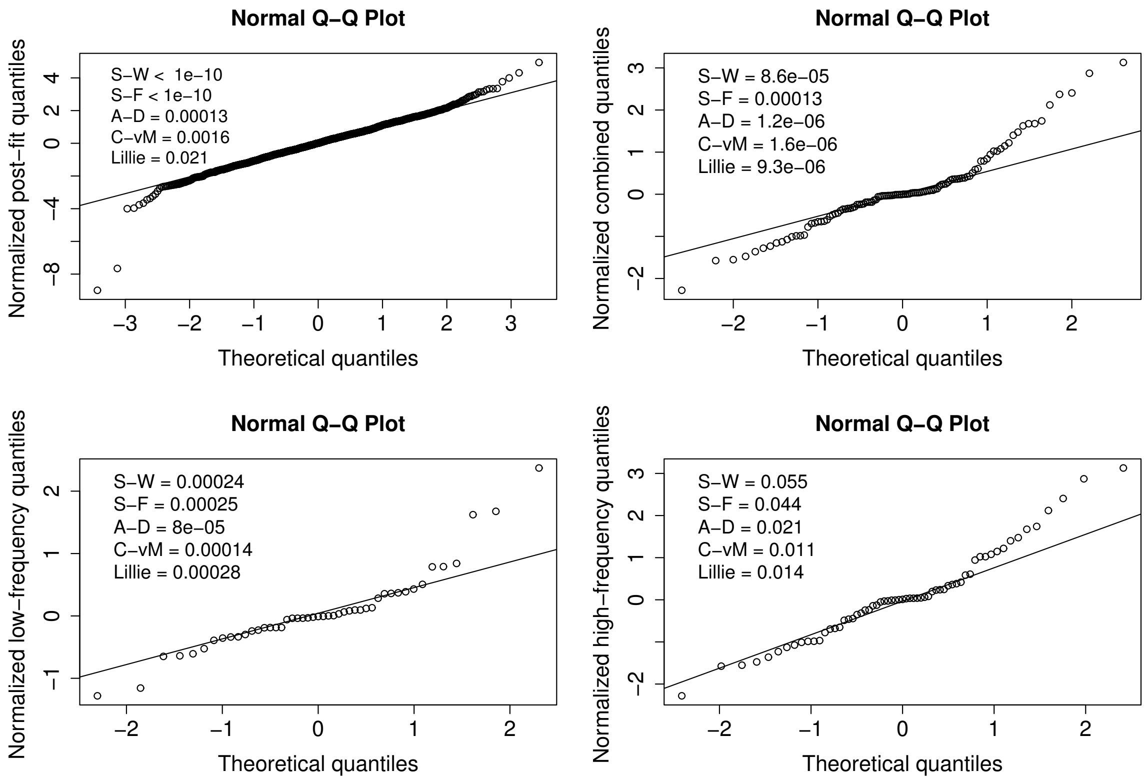

Fig. 6 Q-Q plots for the post-fit residuals (*top-left*), total (*top-right*), low-frequency (*bottom-left*) and high-frequency (*bottom-right*) averaged residuals of J1012+5307. The p-values of individual tests are listed in the legends.

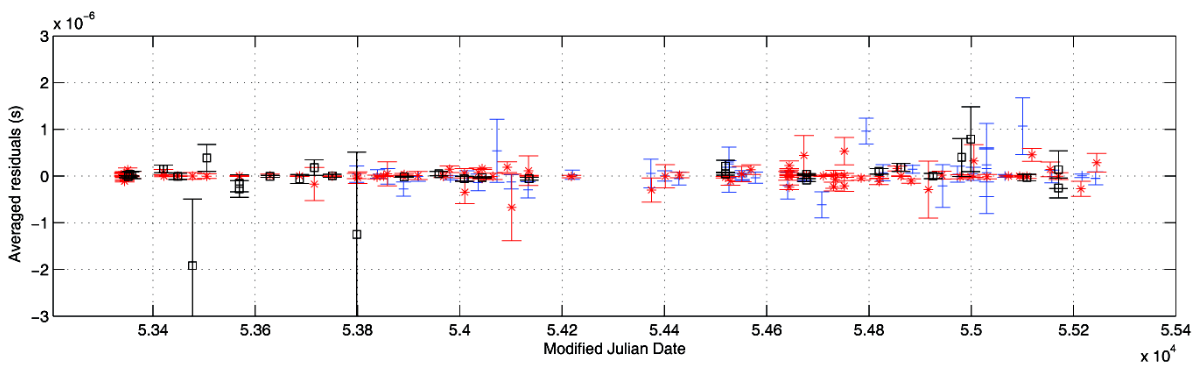

Fig. 7 Frequency separated averaged timing residuals with error bars for J1713+0747. The *blue short-bars* represent 820 MHz data, the *red asterisks* represent 1.4 GHz data and *black squares* represent 2.3 GHz data.

ate from a Gaussian distribution; removing a few outliers improves the statistics considerably.

6 SUMMARY AND DISCUSSIONS

In this paper we utilized a set of non-parametric statistical tests to analyze the NANOGrav 5-year timing residuals for 17 pulsars. Zero crossing has been used to test the whiteness assumption for averaged timing residuals. The results are summarized in Table 2. Both descriptive and inferential statistical methods have been used to test Gaussianity for the post-fit multi-frequency and averaged timing residuals. The results are summarized in Table 1. The histogram and Q-Q plots for three pulsars are shown for demonstration purposes.

We found that for most cases, except the high frequency averaged residuals of J2145–0750 and

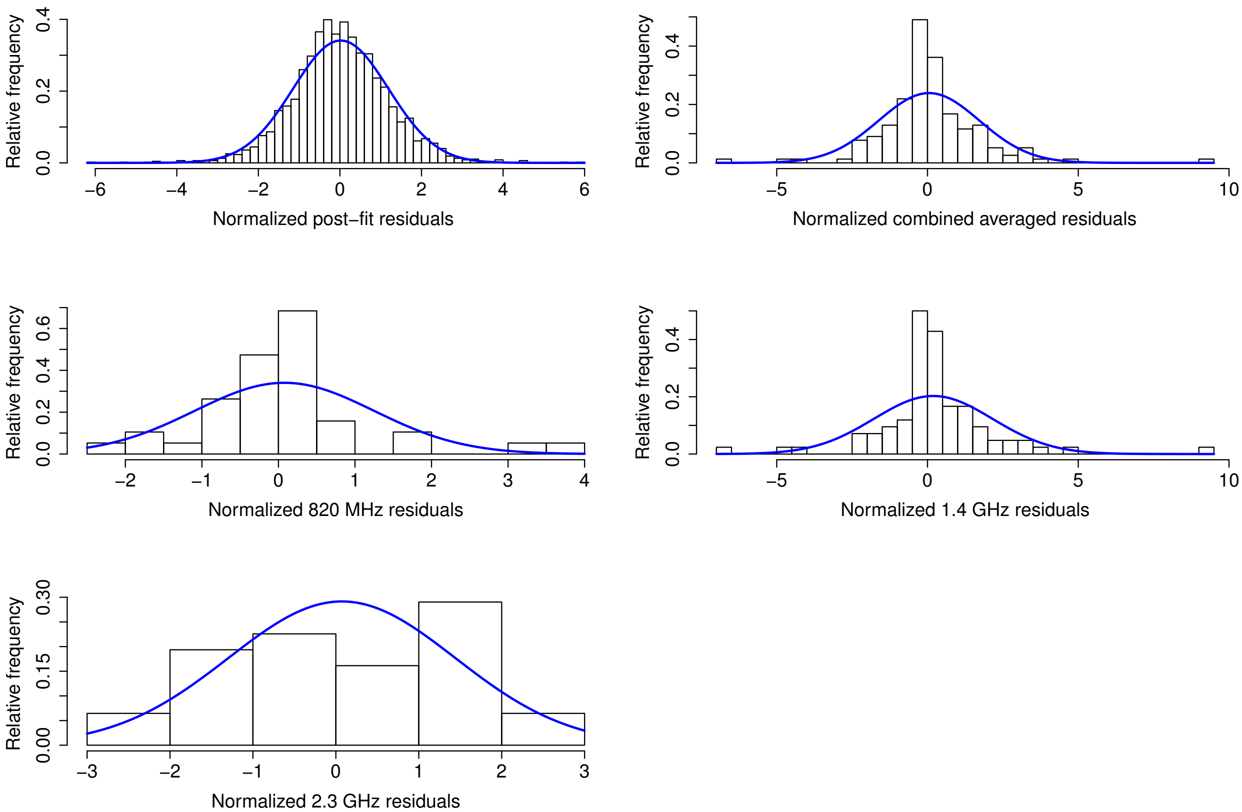

Fig. 8 Histograms for the post-fit residuals (*top-left*), total (*top-right*), 820 MHz (*middle-left*), 1.4 GHz (*middle-right*) and 2.3 GHz (*bottom*) averaged residuals of J1713+0747. The *blue curve* is the Gaussian distribution with the same mean and variance as the data.

J2317+1439, the frequency separated averaged residuals are consistent with the white noise assumption. However, when they are combined, the total averaged residuals of five pulsars show strong deviations from whiteness ('N') and four pulsars show mild deviations from whiteness ('n').

In principle, the total averaged residuals can be modeled by combining two time series $x_1(t_i)$ and $x_2(t'_j)$ representing the low frequency and high frequency averaged residuals respectively, where t_i $(i = 1, 2, 3, ..., N_1)$ and t'_j $(j = 1, 2, 3, ..., N_2)$ are not necessarily identical or evenly spaced. If the two time series are separately drawn from white noise processes, then the number of zero crossings of the combined time series (sorted in ascending order of the union of $\{t_i\}$ and $\{t'_j\}$) is a Gaussian random variable with expectation equal to $(N1+N2-1)/2$ and variance equal to $(N1 + N2 - 1)/4$.

The cumulative number of zero crossings for total averaged residuals with low and high frequency averaged residuals for PSR J1012+5307 and J0613–0200 are shown in Figure 10 and Figure 11 respectively. Asterisks represent the number of zero crossings (y-axis) added up to a given time (x-axis). It is equivalent to the number of zero crossings for the data within an enlarging time window with the left end fixed at the beginning of the time series and the right end sliding to the time of this data point. The solid curves are the expected numbers of zero crossings of the data size within the window and the dash-dotted lines are 1σ contours. They are all monotonic functions of time. Red, black and blue represent the low frequency, high-frequency and total averaged residuals respectively.

For J1012+5307, the cumulative number of zero-crossings for low frequency, high frequency and total averaged residuals closely follow the expected values within a 1σ contour. This is exactly what is expected for a combination of two white noise time series. By contrast, for J0613–0200, although the low frequency and high frequency zero crossings closely follow the expectations as in J1012+5307, the combined data show strong

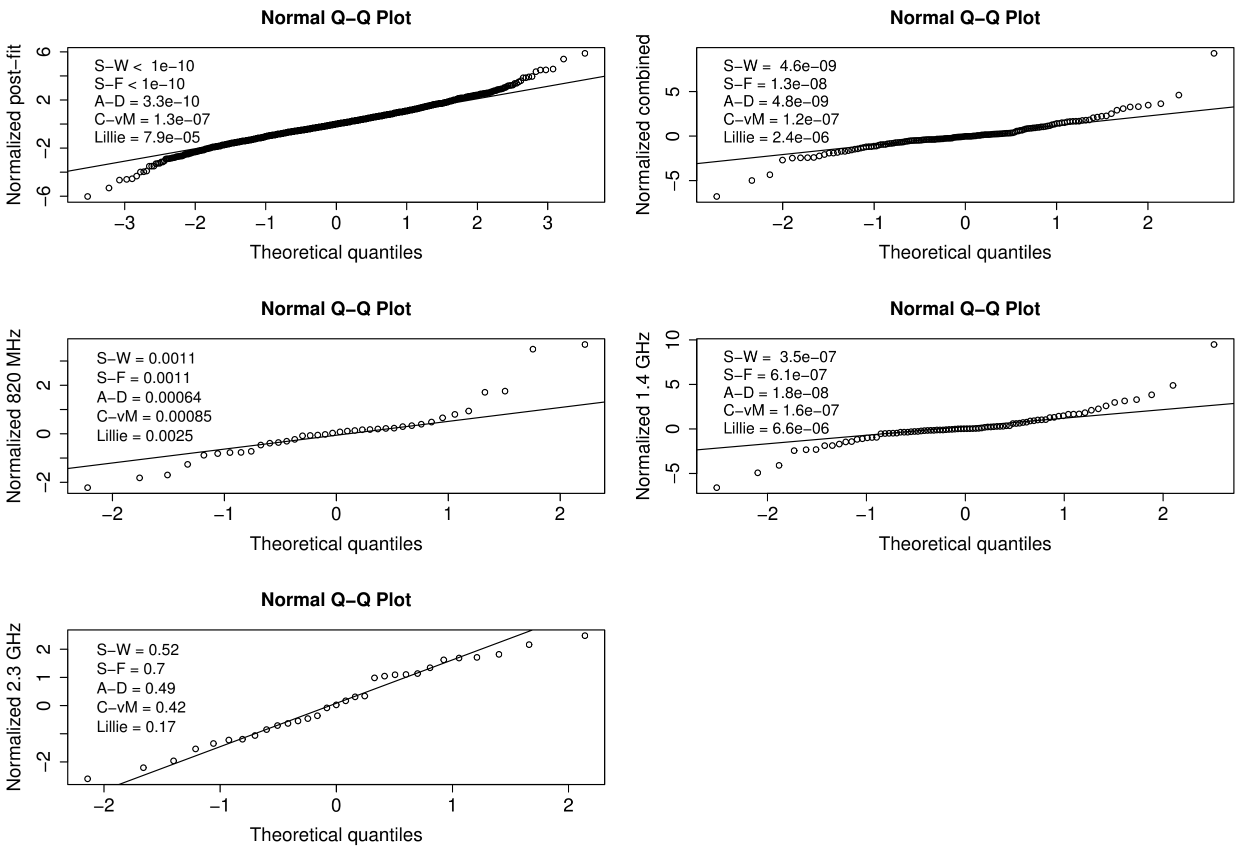

Fig. 9 Q-Q plots for the post-fit residuals (*top-left*), total (*top-right*), 820 MHz (*middle-left*), 1.4 GHz (*middle-right*) and 2.3 GHz (*bottom*) averaged residuals of J1713+0747. The p-values of individual tests are listed in the legends.

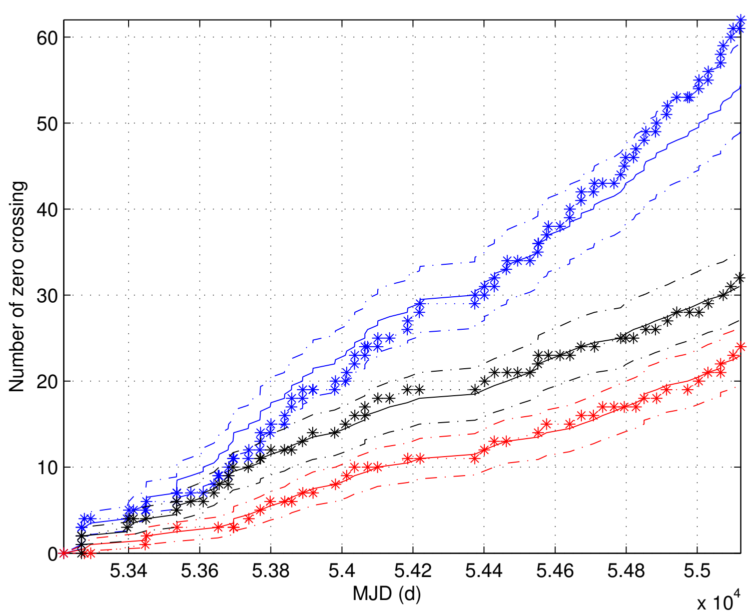

Fig. 10 Cumulative number of zero crossings for PSR J1012+5307. The y-axis is the number of zero crossings and the x-axis is the MJD on which the observations are made; see text for details.

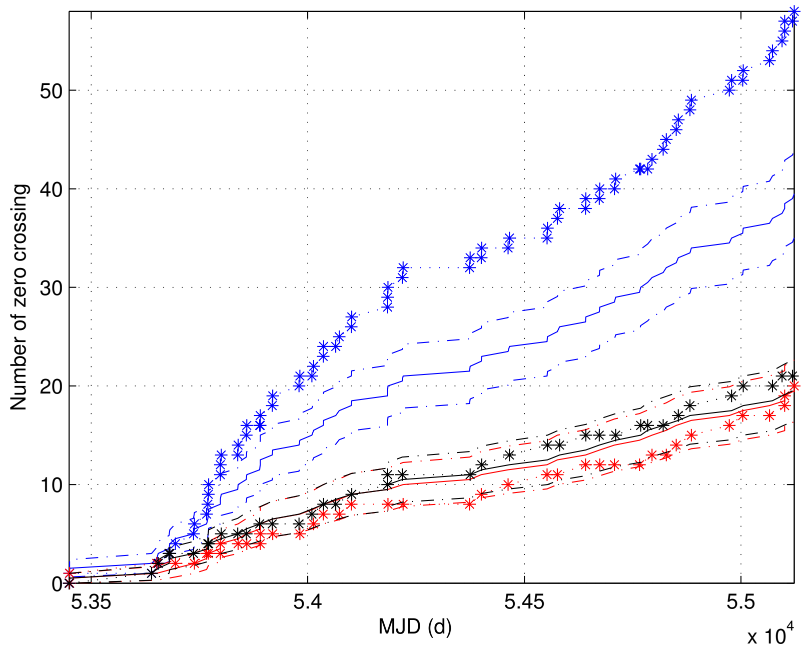

Fig. 11 Cumulative number of zero crossings for PSR J0613–0200. The y-axis is the number of zero crossings and the x-axis is the MJD on which the observations are made; see text for details.

deviation from their expectation. In Figure 11, they start to deviate from the beginning of the time series, and the final deviation is more than 4σ. As stated, this strong deviation for the combined averaged residuals also appears in other pulsars in Table 2.

In fact, the apparent excess of zero crossings is mainly due to the strategy of fitting the physical parameters, especially the time-variable DM in the timing analysis. It is general practice in the NANOGrav 5-year data timing analysis to include a piecewise-constant DM(t) function in the fitting model along with other parameters (rotation, astrometry, binary dynamics and pulse profile evolution). The window for a constant DM value is typically 15 days which include a couple of observations conducted at high and low frequencies. However, any fluctuation of DM within this window or inaccuracy of the DM fit will introduce an additional error between the adjacent averaged timing residuals from two widely separated bands as follows,

$$\delta t \simeq 4.15 \times 10^6\,\mathrm{ms} \times \delta\mathrm{DM} \times \left(f_1^{-2} - f_2^{-2}\right) . \quad (4)$$

Here, f is measured in MHz. For J0613–0200, the uncertainty in the DM ($\sim\ 10^{-4}\,\mathrm{cm}^{-3}\,\mathrm{pc}$) can produce a fluctuation of several hundred nanoseconds between low (820 MHz) and high (1.4 GHz) frequency. This is comparable with the RMS of averaged timing residuals reported in table 2 of Demorest et al. (2013). The minute error of DM will cause low frequency TOAs to tend to be advanced and high frequency TOAs to be delayed or vice versa (the DM fit will tend to move the two sets of TOAs from low and high frequencies, so that their averaged residual is zero) which will produce extra zero crossing between low and high frequency timing residuals within a DM fitting window. This effect is expected to be seen more clearly in the GBT observed pulsars, since the time separation between two observation bands is much larger and the frequency coverage (crucial for the DM) is significantly smaller than the AO. We found that all five pulsars which have total averaged residuals strongly deviate from whiteness, whereas frequency separated averaged residuals are all consistent, as observed by the GBT.

Gaussianity is one of the fundamental assumptions used in most if not all GW detection methods. The tests here suggest that many of the NANOGrav pulsars show deviations from a Gaussian distribution at different levels. The deviations in some data, as well as averaged and multi-frequency post-fit residuals, can be mitigated by removing a few outliers. This strategy is consistent with so called robust statistics (Allen et al. 2002, 2003) which is used to confront the non-Gaussianity in GW data analysis by clipping samples with values located in the outlying parts of a probability distribution. It is robust in the sense that it is close to optimal for Gaussian noise but far less sensitive to large excess events than conventional statistics. Moreover, for the purpose of detection, coherent methods (e.g., Wang et al. (2014, 2015)) have been shown to be robust against non-Gaussianity for detecting deterministic GW signals (Finn 2001). An alternative method in the wavelet domain has also been explored for searching stochastic GW signals in non-Gaussian and non-stationary noise (Klimenko et al. 2002) with ground based GW detectors. The results here suggest that these methods should be investigated for GW detection with a PTA in the future.

Acknowledgements This work was supported by the National Science Foundation (NSF) under PIRE grant 0968296. We are grateful to the NANOGrav members for helpful comments and discussions. Y.W. acknowledges support by the National Natural Science Foundation of China (Grant Nos. 11503007, 91636111 and 11690021). D.R.M. acknowledges partial support through the New York Space Grant Consortium. J.A.E. acknowledges support by NASA through the Einstein Fellowship grant PF4-150120. M.V. acknowledges support from the JPL RTD program. Data for this project were collected using the facilities of the National Radio Astronomy Observatory and the Arecibo Observatory. The National Radio Astronomy Observatory is a facility of the NSF operated under cooperative agreement by Associated Universities, Inc. The Arecibo Observatory is operated by SRI International under a cooperative agreement with the NSF (AST-1100968), and in alliance with Ana G. Méndez-Universidad Metropolitana and the Universities Space Research Association.

References

Abbott, B. P., Abbott, R., Adhikari, R., et al. 2009, Reports on Progress in Physics, 72, 076901
Accadia, T., Acernese, F., Antonucci, F., et al. 2011, Classical and Quantum Gravity, 28, 114002
Allen, B., Creighton, J. D., Flanagan, É. É., & Romano, J. D. 2002, Phys. Rev. D, 65, 122002
Allen, B., Creighton, J. D., Flanagan, É. É., & Romano, J. D. 2003, Phys. Rev. D, 67, 122002
Arzoumanian, Z., Brazier, A., Burke-Spolaor, S., et al. 2014, ApJ, 794, 141
Arzoumanian, Z., Brazier, A., The NANOGrav Collaboration 2015, ApJ, 813, 65
Babak, S., & Sesana, A. 2012, Phys. Rev. D, 85, 044034
Champion, D. J., Hobbs, G. B., Manchester, R. N., et al. 2010, ApJ, 720, L201
Corbin, V., & Cornish, N. J. 2010, arXiv:1008.1782
Cordes, J. M., & Lazio, T. J. W. 2002, astro-ph/0207156
Cordes, J. M., & Shannon, R. M. 2010, arXiv:1010.3785
Damour, T., & Taylor, J. H. 1992, Phys. Rev. D, 45, 1840
Damour, T., & Vilenkin, A. 2005, Phys. Rev. D, 71, 063510
Demorest, P. B., Pennucci, T., Ransom, S. M., Roberts, M. S. E., & Hessels, J. W. T. 2010, Nature, 467, 1081
Demorest, P. B., Ferdman, R. D., Gonzalez, M. E., et al. 2013, ApJ, 762, 94
eLISA Consortium: Seoane, P. A., et al. 2013, arXiv:1305.5720
Edwards, R. T., Hobbs, G. B., & Manchester, R. N. 2006, MNRAS, 372, 1549
Ellis, J. A., Siemens, X., & Creighton, J. D. E. 2012, ApJ, 756, 175
Ferdman, R. D. 2008, Binary Pulsar Systems: Evolution and Fundamental Physics, PhD Thesis, University of British Columbia, Vancouver
Ferdman, R. D., van Haasteren, R., Bassa, C. G., et al. 2010, Classical and Quantum Gravity, 27, 084014
Finn, L. S. 2001, Phys. Rev. D, 63, 102001
Foster, R. S., & Backer, D. C. 1990, ApJ, 361, 300
Grishchuk, L. P. 2005, Physics Uspekhi, 48, 1235
Hellings, R. W., & Downs, G. S. 1983, ApJ, 265, L39
Hobbs, G. B., Edwards, R. T., & Manchester, R. N. 2006, MNRAS, 369, 655
Hobbs, G., Coles, W., Manchester, R. N., et al. 2012, MNRAS, 427, 2780
Hotan, A. W., van Straten, W., & Manchester, R. N. 2004, PASA, 21, 302
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., Hobbs, G. B., van Straten, W., et al. 2006, ApJ, 653, 1571
Klimenko, S., Mitselmakher, G., & Sazonov, A. 2002, gr-qc/0208007
Kramer, M., Stairs, I. H., Manchester, R. N., et al. 2006, Science, 314, 97
Lattimer, J. M., & Prakash, M. 2007, Phys. Rep., 442, 109
Lattimer, J. M., & Prakash, M. 2010, arXiv:1012.3208
Lentati, L., Taylor, S. R., Mingarelli, C. M. F., et al. 2015, MNRAS, 453, 2576
Lomb, N. R. 1976, Ap&SS, 39, 447
Luo, J., Chen, L.-S., Duan, H.-Z., et al. 2016, Classical and Quantum Gravity, 33, 035010
Manchester, R. N., & IPTA. 2013, Classical and Quantum Gravity, 30, 224010
Manchester, R. N., Hobbs, G., Bailes, M., et al. 2013, PASA, 30, e017
McLaughlin, M. A. 2014, arXiv:1409.4579
Ölmez, S., Mandic, V., & Siemens, X. 2010, Phys. Rev. D, 81, 104028

Papoulis, A. 1984, Probability, Random Variables and Stochastic Processes (New York: McGraw-Hill)
Perrodin, D., Jenet, F., Lommen, A., et al. 2013, arXiv:1311.3693
Scargle, J. D. 1982, ApJ, 263, 835
Sesana, A., Vecchio, A., & Colacino, C. N. 2008, MNRAS, 390, 192
Shannon, R. M., Ravi, V., Coles, W. A., et al. 2013, Science, 342, 334
Stairs, I. H. 2003, Living Reviews in Relativity, 6, 5
Taylor, J. H. 1992, Philosophical Transactions of the Royal Society of London Series A, 341, 117
Taylor, J. H., & Weisberg, J. M. 1982, ApJ, 253, 908
Taylor, J. H., & Weisberg, J. M. 1989, ApJ, 345, 434
Tiburzi, C., Hobbs, G., Kerr, M., et al. 2016, MNRAS, 455, 4339
van Haasteren, R., Levin, Y., Janssen, G. H., et al. 2011, MNRAS, 414, 3117
Verbiest, J. P. W., Lentati, L., Hobbs, G., et al. 2016, MNRAS, 458, 1267
Wang, Y. 2015, Journal of Physics Conference Series, 610, 012019
Wang, Y., Mohanty, S. D., & Jenet, F. A. 2014, ApJ, 795, 96
Wang, Y., Mohanty, S. D., & Jenet, F. A. 2015, ApJ, 815, 125
Wolszczan, A., & Frail, D. A. 1992, Nature, 355, 145
Wyithe, J. S. B., & Loeb, A. 2003, ApJ, 590, 691
Zhu, X.-J., Wen, L., Xiong, J., et al. 2016, MNRAS, 461, 1317
Zhu, X.-J., Wen, L., Hobbs, G., et al. 2015, MNRAS, 449, 1650