Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **511,** 4350–4359 (2022) https://doi.org/10.1093/mnras/stac347
Advance Access publication 2022 February 9

The effect of spin mismodelling on gravitational-wave measurements of the binary neutron star mass distribution

Sylvia Biscoveanu [1,2]★ Colm Talbot [1,2] and Salvatore Vitale [1,2]

[1]*LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany St, Cambridge, MA 02139, USA*
[2]*Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA*

Accepted 2022 February 4. Received 2022 February 1; in original form 2021 November 26

ABSTRACT

The binary neutron star (BNS) mass distribution measured with gravitational-wave observations has the potential to reveal information about the dense matter equation of state, supernova physics, the expansion rate of the Universe, and tests of general relativity. As most current gravitational-wave analyses measuring the BNS mass distribution do not simultaneously fit the spin distribution, the implied population-level spin distribution is the same as the spin prior applied when analysing individual sources. In this work, we demonstrate that introducing a mismatch between the implied and true BNS spin distributions can lead to biases in the inferred mass distribution. This is due to the strong correlations between the measurements of the mass ratio and spin components aligned with the orbital angular momentum for individual sources. We find that applying a low-spin prior that excludes the true spin magnitudes of some sources in the population leads to significantly overestimating the maximum neutron star mass and underestimating the minimum neutron star mass at the population level with as few as six BNS detections. The safest choice of spin prior that does not lead to biases in the inferred mass distribution is one that allows for high spin magnitudes and tilts misaligned with the orbital angular momentum.

Key words: gravitational waves – methods: data analysis – stars: neutron – neutron star mergers.

1 INTRODUCTION

The growing catalogue of compact binary mergers detected in gravitational waves (Abbott et al. 2021b) has provided a novel means to probe the properties of black holes and neutron stars (Abbott et al. 2021c; Bouffanais et al. 2021; Roulet et al. 2021; Wong et al. 2021; Zevin et al. 2021). The neutron star mass distribution has the potential to independently yield information on the dense matter equation of state (EoS) via the maximum mass, $M_{\rm TOV}$, beyond which the internal pressure of the neutron star can no longer support it against gravitational collapse to a black hole (Miller, Chirenti & Lamb 2019; Chatziioannou & Farr 2020; Landry, Essick & Chatziioannou 2020; Legred et al. 2021). The value of the maximum mass depends on the yet-unknown EoS and the neutron star spin (Lasota, Haensel & Abramowicz 1996), although astrophysical processes may prevent the formation of neutron stars with mass up to $M_{\rm TOV}$ in some scenarios. The mass distribution can also be used to constrain the supernova physics leading to neutron star formation (Pejcha, Thompson & Kochanek 2012; Vigna-Gómez et al. 2018), the astrophysical stochastic gravitational-wave background (Zhu et al. 2013; Abbott et al. 2018b), the rate of expansion of the Universe (Chernoff & Finn 1993; Finn 1996; Taylor & Gair 2012; Taylor, Gair & Mandel 2012), and alternative theories of gravity beyond general relativity (Finke et al. 2021).

Initial measurements of the mass distribution of Galactic neutron stars detected electromagnetically suggested that their component masses follow a narrow Gaussian distribution (Thorsett & Chakrabarty 1999), particularly for those found in binary systems with another neutron star (Ozel et al. 2012; Kiziltan et al. 2013). More recent measurements reveal that the overall distribution of neutron star masses, including those found in binary systems with other types of companions, is broader (Kiziltan et al. 2013; Özel & Freire 2016) and better described by a double Gaussian (Antoniadis et al. 2016; Tauris et al. 2017; Alsing, Silva & Berti 2018; Shao et al. 2020). Including the first binary neutron star (BNS) system observed in gravitational waves, GW170817 (Abbott et al. 2017), Farrow, Zhu & Thrane (2019) also find weak evidence for bimodality in the mass distribution of the double neutron star population alone.

Gravitational-wave observations of compact binary mergers involving at least one neutron star offer a complementary means to probe the neutron star mass distribution at extragalactic distances. While only two BNS systems have been detected in gravitational waves, these observations already suggest that there may be a distinction between the Galactic population accessible as pulsars and the gravitational-wave population (Pankow 2018; Safarzadeh, Ramirez-Ruiz & Berger 2020; Abbott et al. 2021c; Galaudage et al. 2021). GW190425 represents the most massive BNS system ever detected, with a total mass five standard deviations away from the mean of the Galactic population (Abbott et al. 2020b). When analysing the full population of neutron stars detected in gravitational waves – including neutron star–black hole mergers (Abbott et al. 2021e) – Landry & Read (2021) find that the masses are consistent

★ E-mail: sbisco@mit.edu

with being uniformly distributed, although with significantly more support for high neutron star masses compared with the Galactic population. Thus, accurate and precise measurements of the neutron star mass distribution from gravitational-wave observations have the potential to elucidate the formation channels of these systems.

Such measurements utilize the framework of hierarchical Bayesian inference (e.g. Mandel 2010; Thrane & Talbot 2019), where the properties of the population as a whole are determined while taking into account the uncertainty on the parameters of individual sources and the selection effect introduced by the varying sensitivity of the detector to sources with different properties (Loredo 2004; Mandel, Farr & Gair 2019; Vitale et al. 2020). This requires both unbiased parameter estimates for individual sources and physically realistic models for the population and detector sensitivity. One potential source of systematic error is the correlation between measurements of the intrinsic parameters describing BNS systems, including the masses, spins, and tidal deformabilities of the components. Wysocki et al. (2020) and Golomb & Talbot (2021) recently demonstrated that unphysical assumptions about the EoS (or tidal deformability) when measuring the mass distribution independently can lead to biases with as few as 37 events, emphasizing the importance of fitting these distributions simultaneously.

The mass ratio and spin components aligned with the orbital angular momentum are particularly correlated for individual sources (Cutler & Flanagan 1994; Hannam et al. 2013; Berry et al. 2015; Farr et al. 2016; Ng et al. 2018). This is due to the fact that for a given chirp mass, $\mathcal{M} \equiv (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$, binaries with larger spins aligned with the orbital angular momentum will merge more slowly (Campanelli, Lousto & Zlochower 2006), while binaries with more unequal mass ratios will merge more quickly, introducing a degenerate effect on the waveform. Gravitational-wave analyses of individual BNS sources typically assume two different priors for the spin distributions, a 'low-spin' prior and a 'high-spin' prior (Abbott et al. 2017, 2020b). The 'low-spin' prior restricts the maximum dimensionless spin magnitude – defined as $|\boldsymbol{\chi}| = c|\boldsymbol{S}|/(Gm^2)$, where $\boldsymbol{S}$ is the spin vector of the neutron star with mass m – to $|\boldsymbol{\chi}| \leq 0.05$, informed by the spins of Galactic double neutron stars that will merge within a Hubble time (Lorimer 2008; Lo & Lin 2011; Zhu et al. 2018). The 'high-spin' prior extends to $|\boldsymbol{\chi}| \lesssim 0.89$, as allowed by available waveform models without making any assumptions about the consistency of the system with the observed Galactic population. It is worth noting that pulsars have been observed with spins as high as $|\boldsymbol{\chi}| \lesssim 0.4$, even in binary systems (Hessels et al. 2006).

In this work, we demonstrate that mismodelling the spin distribution of BNSs can lead to a bias in the recovered mass distribution. We find that for a population of sources with moderate aligned spins extending out to $|\boldsymbol{\chi}| \leq 0.4$, using individual-event mass estimates obtained with the low-spin prior without simultaneously fitting the spin distribution hierarchically leads to significant bias in both the inferred mass ratio distribution and the maximum mass. Conversely, in the case of a low-aligned-spin population with $|\boldsymbol{\chi}| \leq 0.05$, the inferred mass ratio distribution under the high-spin prior is also biased, although this effect can be ameliorated by allowing for the spins to be misaligned with the orbital angular momentum. In Section 2, we describe our methodology and simulated BNS populations. In Section 3, we present the results of our hierarchical inference and conclude with a discussion of the implications of our findings in Section 4. We also include a demonstration of the importance of obtaining unbiased inferences for individual sources via the fallibility of the commonly used 'probability–probability plot' as a test of sampler performance in Appendix A.

2 METHODS

In addition to the component masses, spins, and tidal deformabilities, quasi-circular BNS mergers are characterized by seven extrinsic parameters, including the distance, sky location, time of coalescence, and inclination angle between the orbital angular momentum and observer line of sight. We simulate two populations of BNS systems with distinct spin distributions. Both have spins aligned with the orbital angular momentum drawn from the implied distribution on χ_z assuming that the magnitudes are distributed uniformly on [0, $\chi_{\rm max}$] and the directions are isotropic. The first population allows for medium spins consistent with the maximum observed neutron star spin (Hessels et al. 2006), $\chi_{\rm max} = 0.4$, while the second population is restricted to $\chi_{\rm max} = 0.05$, following the observed spins of Galactic double neutron stars (Lorimer 2008; Lo & Lin 2011). The mass distribution is chosen based on Farrow et al. (2019); the mass ratio is drawn from a narrow truncated Gaussian with mean $\mu = 1$, width $\sigma = 0.1$, and lower limit $q_{\rm min} = 0.4$. The total mass distribution is a power law with index $\alpha = -2.5$ between $M_{\rm tot,\,min} = 2.3\,{\rm M}_\odot$ and $M_{\rm tot,\,min} = 4.3\,{\rm M}_\odot$ with low-mass smoothing (see equations 7–8 of Talbot & Thrane 2018) over $\delta M_{\rm tot} = 0.4\,{\rm M}_\odot$. We choose to parametrize the mass distribution in terms of total mass and mass ratio since the latter is particularly sensitive to correlations with the spin parameters (Pürrer, Hannam & Ohme 2016; Ng et al. 2018). For the extrinsic parameters, we assume the sources are distributed uniformly in comoving volume between luminosity distances of $d_{\rm L} = 10\,{\rm Mpc}$ and $d_{\rm L} = 300\,{\rm Mpc}$ and isotropically on the sky. Standard distributions are chosen for the remaining binary parameters (see e.g. Romero-Shaw et al. 2020).

For both the populations described previously, we generate 100 events that are detectable with a LIGO Hanford–Livingston detector network (Aasi et al. 2015) operating at the sensitivity achieved during the third observing run (Abbott et al. 2018a, 2020a). We consider an event to be detectable if it is observed with a network optimal signal-to-noise ratio (SNR) $\rho^{\rm net}_{\rm opt} \geq 9$. For each event, we perform Bayesian parameter estimation to obtain samples from the posterior probability distributions for the binary parameters, $\boldsymbol{\theta}$, describing an individual BNS system, i:

$$p(\boldsymbol{\theta}_i|d_i) \propto \mathcal{L}(d_i|\boldsymbol{\theta}_i)\pi_{\rm PE}(\boldsymbol{\theta}_i), \tag{1}$$

where $\mathcal{L}(d_i|\boldsymbol{\theta}_i)$ is the likelihood of observing data d_i given the binary parameters $\boldsymbol{\theta}_i$ (Veitch & Vecchio 2010; Romano & Cornish 2017):

$$\mathcal{L}(d_i|\boldsymbol{\theta}_i) \propto \exp\left(-\sum_k \frac{2|d_k - h_k(\boldsymbol{\theta}_i)|^2}{T\,S_k}\right). \tag{2}$$

Here, $h(\boldsymbol{\theta}_i)$ represents the gravitational waveform for the BNS signal with parameters $\boldsymbol{\theta}_i$, T is the duration of the analysed data segment, S_k is the noise power spectral density (PSD) characterizing the sensitivity of the detector, and k indicates the frequency dependence of the data, waveform, and PSD. The prior in equation (1) is denoted by $\pi_{\rm PE}(\boldsymbol{\theta}_i)$, with the 'PE' subscript indicating that this is the prior assumed during the initial parameter estimation step for each individual event.

We use the reduced order quadrature (ROQ) implementation (Smith et al. 2016) of the IMRPhenomPv2 waveform model (Hannam et al. 2014; Husa et al. 2016; Khan et al. 2016) in the aforementioned likelihood in order to curtail the computational cost of the individual-event parameter estimation. As such, we assume that the neutron stars we simulate are point masses with no tidal deformability. While previous works have demonstrated the importance of simultaneously inferring the mass and tidal deformability distributions at the risk of introducing biases when

Table 1. Summary of the different combinations of true spin distributions and priors applied during the initial parameter estimation step for each of the scenarios we explore in this work. The $p_{\rm draw}$ column describes the distribution assumed for the sensitivity injections used to incorporate selection effects as detailed in Appendix B.

$\pi_{\rm pop}(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2)$	$\pi_{\rm PE}(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2)$	$p_{\rm draw}(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2)$
Aligned, $\chi_{\rm max} = 0.4$	Aligned, $\chi_{\rm max} = 0.8$	Aligned, $\chi_{\rm max} = 0.8$
Aligned, $\chi_{\rm max} = 0.4$	Aligned, $\chi_{\rm max} = 0.4$	Aligned, $\chi_{\rm max} = 0.4$
Aligned, $\chi_{\rm max} = 0.4$	Aligned, $\chi_{\rm max} = 0.05$	Aligned, $\chi_{\rm max} = 0.05$
Aligned, $\chi_{\rm max} = 0.05$	Aligned, $\chi_{\rm max} = 0.8$	Aligned, $\chi_{\rm max} = 0.8$
Aligned, $\chi_{\rm max} = 0.05$	Precessing, $\chi_{\rm max} = 0.8$	Aligned, $\chi_{\rm max} = 0.8$

they are fitted independently (Wysocki et al. 2020; Golomb & Talbot 2021), we limit the scope of this work to mass–spin correlations and leave a full exploration of the potential correlations between the three intrinsic parameters of each neutron star to future studies. However, we comment on the effects of correlations with tides in Section 4. We use the PyMultiNest (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009; Buchner et al. 2014; Feroz et al. 2019) and dynesty (Speagle 2020) nested samplers as implemented in the Bilby (Ashton et al. 2019; Romero-Shaw et al. 2020) parameter estimation package to obtain samples from the posterior distributions for each event.

We choose priors $\pi_{\rm PE}(\boldsymbol{\theta}_i)$ that are uniform in chirp mass with a width of $0.2\,{\rm M}_\odot$ centred on the true value of the chirp mass for each event and uniform in mass ratio over [0.125, 1]. This choice of prior and mass parametrization is more convenient for sampling. The priors on the extrinsic parameters are the same as those from which the populations were drawn. We adopt a number of different choices for the spin priors, as described subsequently and summarized in Table 1.

Once we have obtained posterior samples for each of the individual BNS events, we can combine them to measure the underlying mass distribution of our simulated population. In this case, we are no longer interested in the individual binary parameters, $\boldsymbol{\theta}_i$, but rather in a set, $\boldsymbol{\Lambda}$, of hyperparameters that describe the population distribution, $\pi_{\rm pop}(\boldsymbol{\theta}|\boldsymbol{\Lambda})$, which is also referred to as the hyperprior. For our choice of population distribution, $\boldsymbol{\Lambda} = \{\alpha, M_{\rm tot,\,min}, M_{\rm tot,\,max}, \delta M_{\rm tot}, \mu, \sigma\}$. The likelihood of observing a set of individual events $\{d\}$ given the hyperparameters $\boldsymbol{\Lambda}$ is obtained by marginalizing over the binary parameters for each event and multiplying the resultant marginal likelihoods:

$$\mathcal{L}(\{d\}|\boldsymbol{\Lambda}) = \prod_i \int \mathcal{L}(d_i|\boldsymbol{\theta}_i)\pi_{\rm pop}(\boldsymbol{\theta}_i|\boldsymbol{\Lambda}){\rm d}\boldsymbol{\theta}_i. \tag{3}$$

This joint likelihood can be constructed from the individual-event posterior samples obtained in the first parameter estimation step via the 'recycling' method (e.g. Thrane & Talbot 2019)

$$\mathcal{L}(\{d\}|\boldsymbol{\Lambda}) \propto \prod_i \sum_j \frac{\pi_{\rm pop}(\boldsymbol{\theta}_{i,j}|\boldsymbol{\Lambda})}{\pi_{\rm PE}(\boldsymbol{\theta}_{i,j})}, \tag{4}$$

so that the joint likelihood is the ratio of the hyperprior and the original PE prior, where the subscript j denotes a sum over the individual posterior samples for each event.

This likelihood assumes that the individual events included in the observed population are an unbiased sample of the true population found in nature. However, we know that this is not the case for observed gravitational-wave events, since the detectors are more sensitive to high-mass sources (Fishbach & Holz 2017). This selection effect needs to be accounted for in the likelihood in order to obtain unbiased estimates of the hyperparameters describing the astrophysical, rather than the observed, distribution (Loredo 2004; Mandel et al. 2019; Thrane & Talbot 2019; Vitale et al. 2020)

$$\mathcal{L}(\{d\}|\boldsymbol{\Lambda}) = \prod_i \frac{\int \mathcal{L}(d_i|\boldsymbol{\theta}_i)\pi_{\rm pop}(\boldsymbol{\Lambda}|\boldsymbol{\theta}_i){\rm d}\boldsymbol{\theta}_i}{\alpha(\boldsymbol{\Lambda})}, \tag{5}$$

$$\alpha(\boldsymbol{\Lambda}) = \int {\rm d}\boldsymbol{\theta}_i\, p_{\rm det}(\boldsymbol{\theta}_i)\pi_{\rm pop}(\boldsymbol{\theta}_i|\boldsymbol{\Lambda}). \tag{6}$$

The function $p_{\rm det}(\boldsymbol{\theta}_i)$ gives the probability that an individual event with parameters $\boldsymbol{\theta}_i$ will be detected. We evaluate $\alpha(\boldsymbol{\Lambda})$ using a Monte Carlo integral over a set of simulated signals drawn from the distribution $p_{\rm draw}(\boldsymbol{\theta})$ following the approach described in Farr (2019). More details on the simulated population used for determining the detection probability can be found in Appendix B. We evaluate the corrected likelihood in equation (5) to obtain samples from the posterior distributions of the hyperparameters using the nestle sampler (Barbary et al. 2021) and the GWPopulation package (Talbot et al. 2019). The priors on the hyperparameters are all uniform over the ranges given in Table 2.

Table 2. Hyperparameters describing the mass distribution and the maximum and minimum values allowed in the prior applied during hierarchical inference. The priors on all parameters are uniform.

Symbol	Parameter	Minimum	Maximum
α	Total mass power-law index	0	4
$M_{\rm tot,min}$	Minimum total mass ($\rm M_\odot$)	2	3
$M_{\rm tot,max}$	Maximum total mass ($\rm M_\odot$)	3.2	5
$\delta M_{\rm tot}$	Smoothing parameter ($\rm M_\odot$)	0	1
μ	Mass ratio mean	0.4	1
σ	Mass ratio standard deviation	0.01	0.5

So far we have not mentioned the spins in the hierarchical inference step and have restricted the hyperparameters to include only those governing the mass distribution. For any parameters that are not explicitly included in the ratio in equation (4), it is implicitly assumed that the PE prior and the hyperprior are the same (see section 5.4 of Vitale et al. 2020). Even if we are not interested in measuring the hyperparameters of the spin distribution, if there is a mismatch between the underlying population distribution and the prior applied during the initial sampling, this can introduce biases in the hyperparameters that are being measured if there are significant correlations between those binary parameters, as is the case for the mass ratio and spins for BNS systems, for example. To probe how such mismatches between the assumed and true population distributions for the spins can introduce biases in the mass distribution, we deliberately apply spin priors during the PE step that differ from those used to generate the population. This is consistent with what is currently done in population analyses, as will be described in detail in Section 4.

For the medium-spin population, we apply both high- and low-spin priors with $\chi_{\rm max} = 0.8$ and 0.05, respectively, as limited by the range of validity of the ROQ. We also perform parameter estimation and hierarchical inference on the mass distribution with a PE prior that matches the true population distribution with $\chi_{\rm max} = 0.4$ as a sanity check to ensure that the resultant biases we see are indeed due to a spin-prior mismatch. An example corner plot showing the posteriors on the mass ratio and component spins obtained under each of these priors for one individual event is shown in Fig. 1. While the true values of all three parameters are included within the prior range for all three choices of spin prior, the q posterior peaks at lower values when analysed with the low-spin prior compared with the other two due to the correlation between mass ratio and spin. For the low-spin population, we apply the high-spin prior first assuming aligned spins and then relax this assumption to allow for precessing spins. The

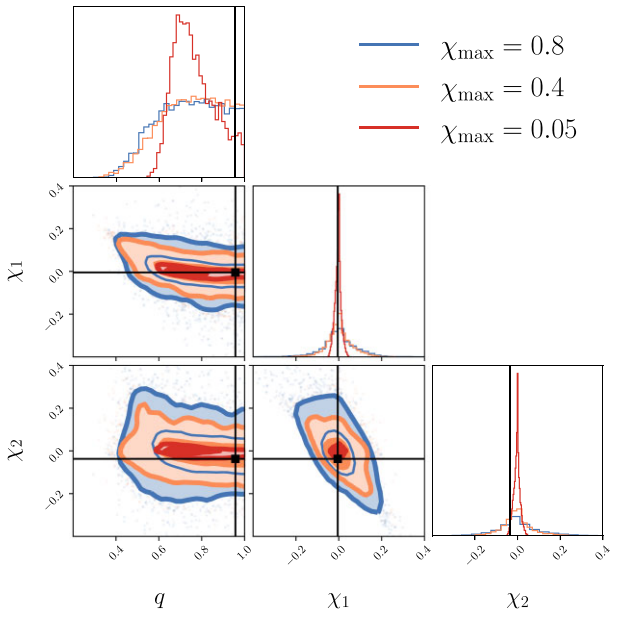

Figure 1. Corner plot showing the 1D posteriors and 50 and 90 per cent contours for the mass ratio and component spin magnitudes for one event recovered with aligned-spin priors for three different choices of $\chi_{\rm max}$. The true values are indicated with the black lines and are $q = 0.96$, $\chi_1 = -0.005$, and $\chi_2 = -0.037$, within the prior range for all three choices of prior.

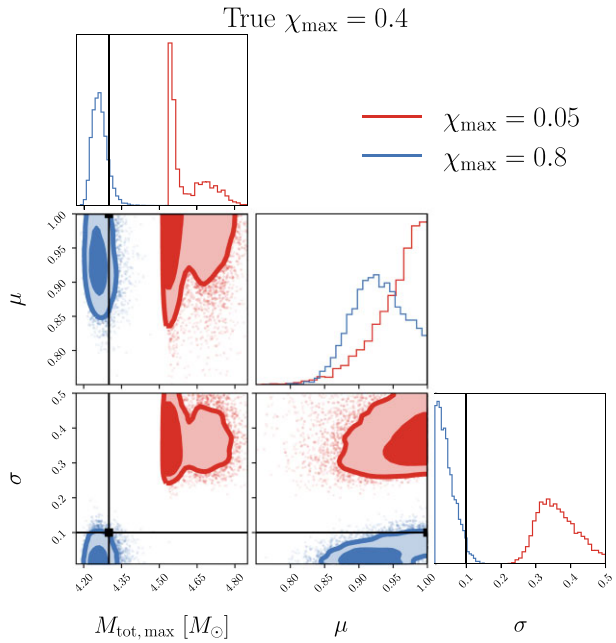

Figure 2. Hyperparameter posteriors on the maximum total mass and mean and width of the mass ratio distribution for two different choices of initial sampling prior applied to the medium-spin population. While the sources were drawn from an aligned-spin distribution with $\chi_{\rm max} = 0.4$, the red and blue contours show the hierarchical inference results when they are recovered with priors with $\chi_{\rm max} = 0.05$ and 0.8, respectively. The black lines denote the true values of the hyperparameters. The value for $\mu = 1$ lies at the edge of the prior, and hence the black line is not visible for this parameter. The high-spin prior results are consistent with the true population, while the low-spin prior results favour a much wider mass ratio distribution and a higher maximum total mass. The true values are excluded at $>3\sigma$ confidence.

combinations of true population and PE prior are summarized in Table 1.

3 MASS–SPIN CORRELATIONS

The hyperparameter posteriors for $M_{\rm tot,max}$, μ, and σ obtained when analysing the medium-spin population with true $\chi_{\rm max} = 0.4$ with both the low- and high-spin priors are shown in the corner plot in Fig. 2. The blue contours show the results obtained with the high-spin prior. The true hyperparameter values are all recovered within the 90 per cent credible region of the posterior, demonstrating that analysing a population with $\chi_{\rm max} = 0.4$ with a prior out to $\chi_{\rm max} = 0.8$ does not lead to biases in the inferred mass distribution. This is consistent with the fact that the analyses with the medium- and high-spin priors shown in Fig. 1 do not lead to significant differences in the mass ratio posteriors for individual events.

Conversely, the low-spin results shown in red are significantly biased in both σ and $M_{\rm tot,max}$. This can be explained by the correlations shown in Fig. 1, which are exacerbated by the low-spin prior and push the mass ratio posteriors for individual events towards lower values. This leads to a preference for wider distributions in the hierarchical inference step, since there is more support for extreme mass ratios in the population. Underestimating the mass ratio results in overestimating the total mass, since the chirp mass of the system

$$\mathcal{M} = \left(\frac{q}{(1+q)^2}\right)^{3/5} M_{\rm tot} \tag{7}$$

is still well constrained, propagating into the overestimation of the maximum total mass at the population level.

The inferred mass ratio and total mass distributions under the low-spin prior are shown in Fig. 3. These are represented by the posterior population distribution (PPD), which is the astrophysical distribution of the binary parameters $\boldsymbol{\theta}$ implied by the inferred hyperparameters $\boldsymbol{\Lambda}$ (Abbott et al. 2021d):

$$p(\boldsymbol{\theta}|\{d\}) = \int \mathrm{d}\boldsymbol{\Lambda}\, p(\boldsymbol{\Lambda}|\{d\})\pi_{\rm pop}(\boldsymbol{\theta}|\boldsymbol{\Lambda}). \tag{8}$$

The recovery of a wider mass ratio distribution and higher maximum total mass results in overestimating the maximum mass of the primary neutron star and underestimating the minimum mass of the secondary, as can be seen in the bottom panel of Fig. 3. These biases can have profound implications for both single and binary neutron star formation mechanisms and their EoS. Few equations of state are able to support neutron stars with $m \geq 2.5\,\mathrm{M}_\odot$, where 7 per cent of the probability lies for the recovered PPD on m_1. Additionally, it is difficult to form neutron stars with $m \lesssim 1\,\mathrm{M}_\odot$ under current stellar evolution models (Vigna-Gómez et al. 2018), a region of parameter space that contains 10 per cent of the probability for the inferred PPD on m_2.

The evolution of the bias in the hyperparameters as a function of the number of individual events included in the analysis is shown in blue in Fig. 4. With six events, denoted by the first vertical grey line, the true value of σ is already excluded from the 3σ credible region. The same occurs with 13 events for the maximum total mass parameter. We also show the evolution of the bias in the maximum and minimum component masses, which are represented by the 99th percentile of the primary mass distribution and the first percentile of the secondary mass distribution, respectively, since we do not directly parametrize the population in terms of the component masses. We find similar constraints on $m_{\rm max}$ and $m_{\rm min}$ to those presented in Chatziioannou & Farr (2020), but these parameters become significantly biased with

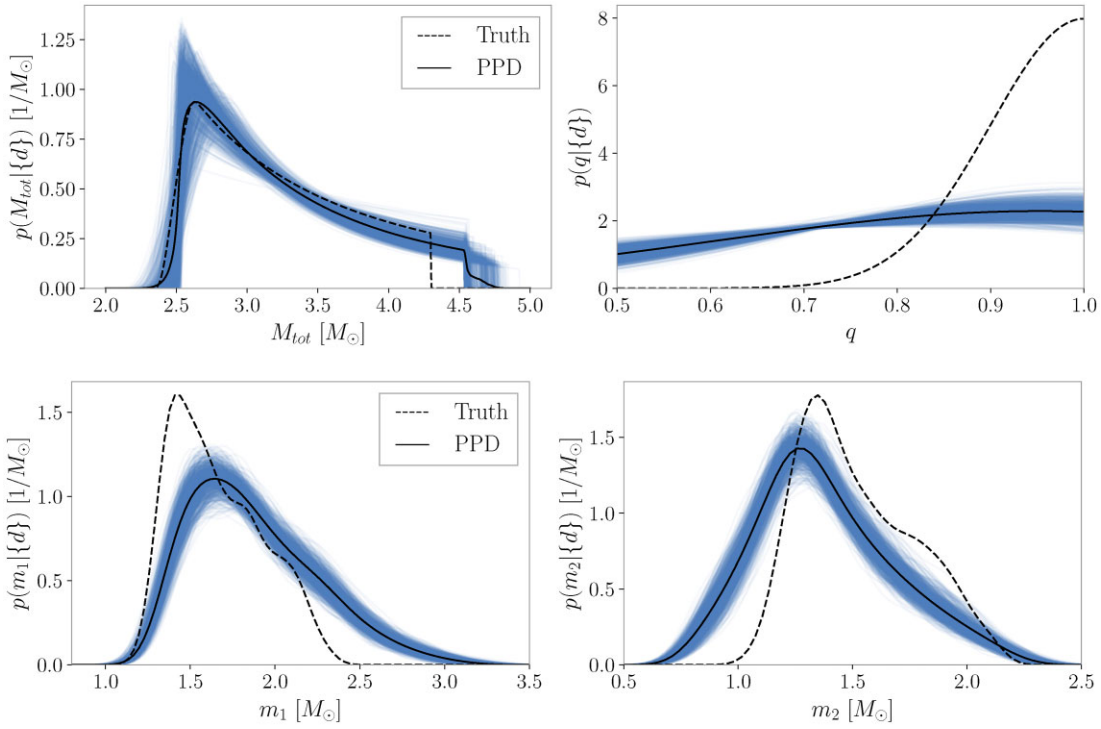

Figure 3. Inferred mass posterior population distributions (solid black lines) when a spin-prior mismatch is introduced for a population where the true spin follows the aligned-spin distribution out to $\chi_{\rm max} = 0.4$, but the assumed population only allows $\chi_{\rm max} = 0.05$. The dashed black lines show the true distributions, while the light blue lines show individual draws from the hyperparameter posterior. Top: total mass (left) and mass ratio (right) distributions. Bottom: primary (left) and secondary (right) mass distributions.

a similar number of events to the σ and $M_{\rm tot,max}$ hyperparameters. The bias in $M_{\rm tot,max}$ is driven by four events in particular, which correspond to the distinguishable upward jumps in the top panel of Fig. 4. These events all have true values of $|\boldsymbol{\chi_1}| > 0.05$, which is outside the range allowed by the low-spin prior. This leads to a significant bias in the mass ratio posterior towards low values, driving the total mass upwards to keep the chirp mass constant. We emphasize that the posteriors for these individual events do not exhibit railing against the prior edges in either the spin or mass parameters and are well converged. As such, they would not immediately be identified as problematic if they corresponded to real events. These results demonstrate that choosing a population model for the spins (even implicitly) that does not include all the sources in the population can significantly bias the BNS mass distribution. The width of the 90 per cent credible interval (CI) for the same parameters in the case of no spin-prior mismatch is shown in the black dotted lines in Fig. 4 for comparison. The true values of all the hyperparameters lie within the 90 per cent CI, demonstrating that there is no bias when the implied and true spin distributions match.

While we have found that analysing the population of sources with $\chi_{\rm max} = 0.4$ with a prior going up to $\chi_{\rm max} = 0.8$ does not introduce a bias on the mass distribution, we now seek to investigate whether the same conclusion holds for the low-spin population with $\chi_{\rm max} = 0.05$. The hyperparameter posteriors for the low-spin population analysed with high-spin priors assuming both aligned and precessing spins are shown in the corner plot in Fig. 5. The true values of the hyperparameters are not always contained within the 90 per cent credible region for the high-aligned-spin prior shown in red, indicating a hint of a bias when 100 individual events are included in the population. The much larger difference in the prior volume between the low- and high-spin priors in this case leads to a bias even when all the observed events have spins within the allowed prior region. Allowing for the tilts to be misaligned to the orbital angular momentum introduces additional degrees of freedom that break the strong degeneracy between q and χ_z, alleviating this bias. The results obtained under the high-precessing-spin prior shown in blue in Fig. 5 include the true hyperparameter values within the 90 per cent credible region. Thus, we conclude that using the high, precessing-spin prior is the safest choice for BNS systems if the mass and spin distributions are not modelled simultaneously during hierarchical inference.

4 CONCLUSIONS

In this work, we have demonstrated that introducing a mismatch between the true, underlying spin distribution for BNS systems observed in gravitational waves and that assumed when characterizing individual systems can lead to a bias in the inferred mass distribution at the population level. If the mass and spin distributions are not fitted simultaneously, the implied population model for the spin distribution is the same as the prior used when conducting parameter estimation for individual sources. To investigate the effects of such a mismatch, we simulated two distinct populations of BNS sources, one with medium aligned spins out to $\chi_{\rm max} = 0.4$ and the other with low aligned spins out to $\chi_{\rm max} = 0.05$.

The mass distribution inferred for the medium-spin population was significantly biased when the individual sources were analysed with a low-aligned-spin prior but unbiased when analysed with a high-aligned-spin prior. The bias is due to the degeneracy between the mass ratio and aligned spin components for BNS systems, which pushes the mass ratio posteriors for individual sources out towards more extreme values and also drives the total mass towards higher values.

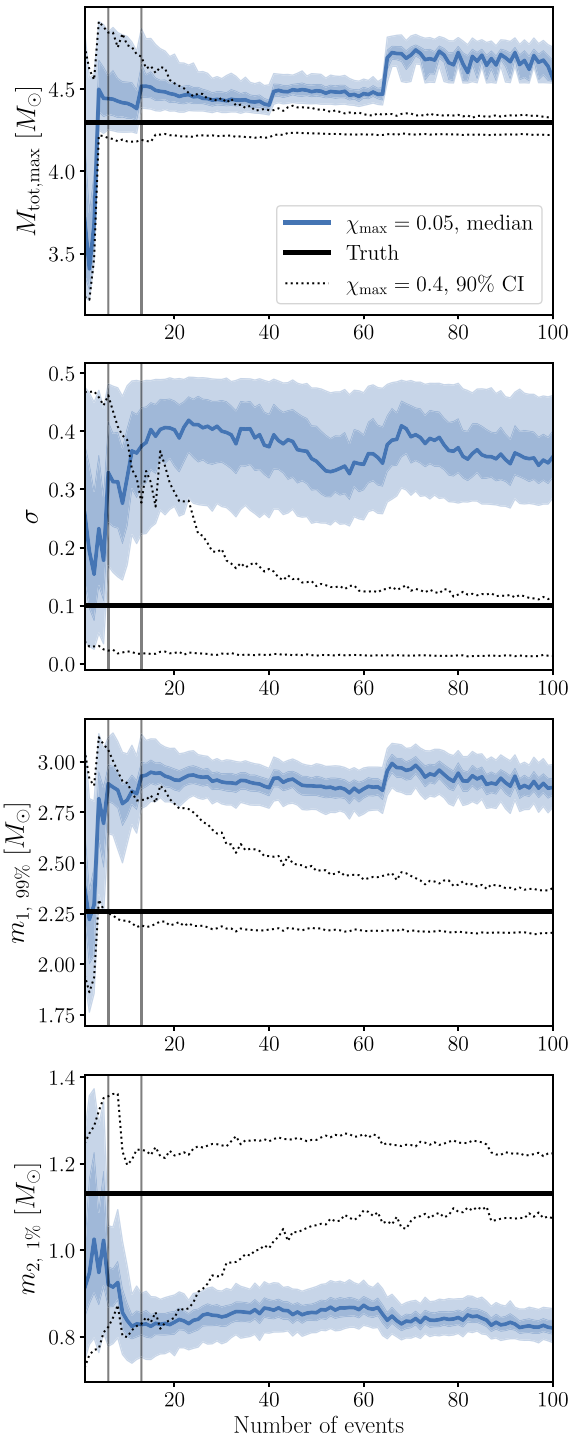

Figure 4. The top two panels show the evolution of the hyperparameters $M_{\rm tot,max}$ and σ as a function of the number of events included in the hierarchical analysis. The bottom two panels show the evolution of the maximum and minimum inferred component masses, represented by the 99th percentile of the m_1 distribution and the first percentile of the m_2 distribution, respectively. The solid black line shows the true value for each hyperparameter, while the solid blue line shows the median obtained when applying the low-spin prior to the medium-spin population. The blue shading gives the 50 per cent and 90 per cent CIs. The dotted black lines denote the 90 per cent credible region when there is no spin-prior mismatch applied to the medium-spin population. The vertical grey lines show the 6th and 13th events, where the true values of σ and $M_{\rm tot,max}$ are excluded at $>3\sigma$, respectively.

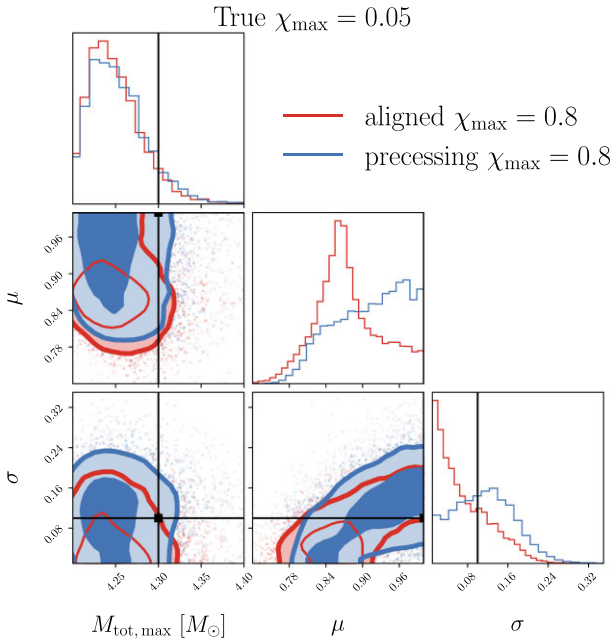

Figure 5. Hyperparameter posteriors on the maximum total mass and mean and width of the mass ratio distribution for two different choices of initial sampling prior applied to the low-spin population. The sources were drawn from an aligned-spin distribution with $\chi_{\rm max} = 0.05$, but both priors assume $\chi_{\rm max} = 0.8$. However, the red contours show the hierarchical inference results when an aligned-spin prior is applied, while the blue allows for spin precession. While the high-aligned-spin results are only marginally consistent with the true hyperparameter values, allowing for precession breaks the strong degeneracy between q and χ_z and ameliorates the bias.

This translates into an overestimation of the maximum neutron star mass and an underestimation of the minimum mass, with adverse implications for both the inference of the nuclear EoS and supernova mechanisms. The most massive neutron stars with posterior support in Fig. 3 are only supported by the stiffest equations of state. The illusion of these high-mass neutron stars in the population would falsely populate the putative lower mass gap between the heaviest neutron stars and lightest black holes. Additionally, the false presence of a significant subpopulation of sub-solar-mass compact objects would affect the inferred contribution of primordial sub-solar-mass black holes to the dark matter density (Abbott et al. 2018c, 2019c, 2021a; Nitz & Wang 2021), as neutron stars with such low masses are not expected to form theoretically, and ground-based gravitational-wave detectors are not sensitive to the gravitational radiation from less massive compact objects such as white dwarfs. While in principle tidal effects could be used to distinguish sub-solar-mass black holes from neutron stars, in practice gravitational-wave constraints on the tidal deformability are often weak (e.g. Abbott et al. 2020b, 2021e).

The mass distribution inferred for the low-spin population demonstrated a hint of bias with the high-aligned-spin prior, but this was alleviated when the individual sources were analysed with a prior allowing for misaligned spin tilts. The extra degrees of freedom introduced by the precessing-spin model break the strong degeneracy between q and χ_z. These biases demonstrate the importance of fitting the mass and spin distributions simultaneously, to avoid implicitly mismodelling the spin distribution. However, if the mass distribution must be analysed independently, we conclude that using a high-precessing-spin prior for the individual sources is the safest choice.

We note that the unbiased results we obtain in this demonstration are only robust if the choice of parametrized mass model used during the hierarchical inference step is physically realistic. If the assumed shape of the mass distribution does not match the underlying population, further biases can be introduced. However, this sort of mismatch is unlikely to affect the inferred maximum and minimum neutron star masses as significantly as the spin-prior mismatch, so long as the parametrized population model covers the full range of allowed neutron star masses. This potential problem can be further ameliorated by fitting the mass distribution with several different hierarchical models, including for example the bimodal models favoured by current observations and simulations (e.g. Antoniadis et al. 2016; Tauris et al. 2017; Alsing et al. 2018; Vigna-Gómez et al. 2018; Farrow et al. 2019), and comparing the statistical evidence obtained between them to determine which provides the best fit. Alternatively, a more flexible model that does not impose a specific functional form on the mass distribution could be used (e.g. Mandel et al. 2019; Wong, Contardo & Ho 2020; Li et al. 2021b; Rinaldi & Del Pozzo 2021; Sadiq, Dent & Wysocki 2021; Tiwari 2021).

Correlations between the tidal parameters – which we have set to zero in our analysis – and the masses and spins can also introduce systematic errors in the inferred mass distribution if not accounted for. For the posteriors of individual events, more extreme values of mass ratio allow for smaller values of the tidal parameter that enters the gravitational waveform at leading order, $\tilde{\Lambda}$ (Wade et al. 2014; Abbott et al. 2019a). Since the high-spin prior typically provides more support for more unequal mass ratios, changing the spin prior can also affect the inferred tidal parameters. Comparing our results with those of Golomb & Talbot (2021), who demonstrate the effect of mismodelling the tidal parameters on the inferred mass distribution, we conclude that enforcing a low-spin prior when there are larger spins in the population introduces a much more significant bias in the inferred mass distribution. However, both types of mismatches lead to increased support for higher neutron star masses.

In addition to the mismatches between the true and implied population models for the spins, masses, and tides discussed earlier, inadequate sampler performance when conducting parameter estimation for individual events can also manifest as a bias in the inferred population properties when multiple events are combined, as demonstrated in Appendix A. However, this sort of bias can be diagnosed and addressed by performing hierarchical inference on a simulated population. Another potential source of systematic error is the accuracy limitations of the waveform models used to infer the properties of individual events. For the SNRs expected with the current generation of detectors, this effect should be small compared with the bias introduced by mismodelling the spin distribution (Dudi et al. 2018; Abbott et al. 2019b; Dietrich et al. 2019; Messina et al. 2019; Abbott et al. 2020b).

The upcoming fourth observing run of the LIGO and Virgo detectors is expected to add tens of new BNS sources to the catalogue of compact binaries detected in gravitational waves (Abbott et al. 2018a). Based on our results, a bias in the mass distribution could be imposed on the observed population with as few as four additional BNS detections if a low-spin prior is applied to a population with higher spins. We note that while current population studies do not model the BNS mass distribution independently due to the paucity of detections, a variety of models consider BNS sources as part of the compact object population as a whole (Mandel et al. 2017; Fishbach, Essick & Holz 2020; Abbott et al. 2021c; Farah et al. 2021) or as part of the population of neutron-star-containing systems (Abbott et al. 2021c; Landry & Read 2021; Li et al. 2021a). Most of these models and analyses assume the BNS spin distribution is uniform in magnitude on the interval [0, 1) with isotropic tilts, corresponding to the safe choice identified in Section 3. Both the MULTI SOURCE model presented in Abbott et al. (2021c) and the analysis of Li et al. (2021a) fit the mass and spin distributions of the BNS subpopulation simultaneously. The MULTI SOURCE model assumes that the spin magnitudes follow a beta distribution with $\chi_{\rm max} = 0.05$ and the tilts are isotropically distributed. The analysis of Li et al. (2021a) takes a similar approach, assuming the neutron star spin magnitudes follow a truncated Gaussian out to $\chi_{\rm max} = 0.05$ and fitting the tilts as a mixture model between aligned and isotropic distributions. While assuming small spins will not lead to a bias for the two BNS events currently detected, this assumption should be relaxed to avoid introducing a bias in the mass distribution with the first few events detected during the next observing run.

ACKNOWLEDGEMENTS

SB, CT, and SV acknowledge support of the National Science Foundation (NSF) and the LIGO Laboratory. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the NSF and operates under cooperative agreement PHY-0757058. SB is also supported by the NSF Graduate Research Fellowship under grant no. DGE-1122374. SV is also supported by NSF PHY-2045740. The authors would like to thank Amanda Farah and Phil Landry for insightful comments on the manuscript. The authors are grateful for computational resources provided by the LIGO Lab and supported by NSF grants PHY-0757058 and PHY-0823459. This paper carries LIGO document number LIGO-P2100426.

DATA AVAILABILITY

A public repository with example code to reproduce the results in this manuscript is available on GITHUB. Only simulated data were analysed using the publicly-available programs BILBY (Ashton et al. 2019; Romero-Shaw et al. 2020), PYMULTINEST (Feroz & Hobson 2008; Feroz et al. 2009, 2019; Buchner et al. 2014), DYNESTY (Speagle 2020), NESTLE (Barbary et al. 2021), and GWPOPULATION (Talbot et al. 2019)

REFERENCES

Aasi J. et al., 2015, Class. Quant. Grav., 32, 074001
Abbott B. P. et al., 2017, Phys. Rev. Lett., 119, 161101
Abbott B. P. et al., 2018a, Living Rev. Relativ., 21, 3
Abbott B. P. et al., 2018b, Phys. Rev. Lett., 120, 091101
Abbott B. P. et al., 2018c, Phys. Rev. Lett., 121, 231103
Abbott B. P. et al., 2019a, Phys. Rev. X, 9, 011001
Abbott B. P. et al., 2019b, Phys. Rev. X, 9, 031040
Abbott B. P. et al., 2019c, Phys. Rev. Lett., 123, 161102
Abbott B. P. et al., 2020a, Noise Curves Used for Simulations in the Update of the Observing Scenarios Paper, Available at: https://dcc.ligo.org/LIGO-T2000012/public
Abbott B. P. et al., 2020b, ApJ, 892, L3
Abbott R. et al., 2021a, preprint (arXiv:2109.12197)
Abbott R. et al., 2021b, preprint (arXiv:2111.03606)
Abbott R. et al., 2021c, preprint (arXiv:2111.03634)
Abbott R. et al., 2021d, ApJ, 913, L7
Abbott R. et al., 2021e, ApJ, 915, L5
Alsing J., Handley W., 2021, MNRAS, 505, L95
Alsing J., Silva H. O., Berti E., 2018, MNRAS, 478, 1377
Antoniadis J., Tauris T. M., Ozel F., Barr E., Champion D. J., Freire P. C. C., 2016, preprint (arXiv:1605.01665)
Ashton G. et al., 2019, ApJS, 241, 27

Barbary K. et al., 2021, Nestle, Available at: https://github.com/kbarbary/nestle
Berry C. P. L. et al., 2015, ApJ, 804, 114
Bouffanais Y., Mapelli M., Santoliquido F., Giacobbo N., Di Carlo U. N., Rastello S., Artale M. C., Iorio G., 2021, MNRAS, 507, 5224
Buchner J. et al., 2014, A&A, 564, A125
Campanelli M., Lousto C. O., Zlochower Y., 2006, Phys. Rev. D, 74, 041501
Chatziioannou K., Farr W. M., 2020, Phys. Rev. D, 102, 064063
Chernoff D. F., Finn L. S., 1993, ApJ, 411, L5
Cook S. R., Gelman A., Rubin D. B., 2006, J. Comput. Graph. Stat., 15, 675
Cutler C., Flanagan E. E., 1994, Phys. Rev. D, 49, 2658
Dietrich T. et al., 2019, Phys. Rev. D, 99, 024029
Dudi R., Pannarale F., Dietrich T., Hannam M., Bernuzzi S., Ohme F., Brügmann B., 2018, Phys. Rev. D, 98, 084061
Farah A. M., Fishbach M., Essick R., Holz D. E., Galaudage S., 2021, preprint (arXiv:2111.03498)
Farr B. et al., 2016, ApJ, 825, 116
Farr W. M., 2019, Research Notes of the AAS, 3, 66
Farrow N., Zhu X.-J., Thrane E., 2019, ApJ, 876, 18
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2019, Open J. Astrophys., 2, 10
Finke A., Foffa S., Iacovelli F., Maggiore M., Mancarella M., 2021, preprint (arXiv:2108.04065)
Finn L. S., 1996, Phys. Rev. D, 53, 2878
Fishbach M., Holz D. E., 2017, ApJ, 851, L25
Fishbach M., Essick R., Holz D. E., 2020, ApJ, 899, L8
Galaudage S., Adamcewicz C., Zhu X.-J., Stevenson S., Thrane E., 2021, ApJ, 909, L19
Golomb J., Talbot C., 2022, ApJ, 926, 79
Hannam M., Brown D. A., Fairhurst S., Fryer C. L., Harry I. W., 2013, ApJ, 766, L14
Hannam M., Schmidt P., Bohé A., Haegel L., Husa S., Ohme F., Pratten G., Pürrer M., 2014, Phys. Rev. Lett., 113, 151101
Hessels J. W. T., Ransom S. M., Stairs I. H., Freire P. C. C., Kaspi V. M., Camilo F., 2006, Science, 311, 1901
Husa S., Khan S., Hannam M., Pürrer M., Ohme F., Jiménez Forteza X., Bohé A., 2016, Phys. Rev. D, 93, 044006
Khan S., Husa S., Hannam M., Ohme F., Pürrer M., Jiménez Forteza X., Bohé A., 2016, Phys. Rev. D, 93, 044007
Kiziltan B., Kottas A., De Yoreo M., Thorsett S. E., 2013, ApJ, 778, 66
Landry P., Read J. S., 2021, ApJ, 921, L25
Landry P., Essick R., Chatziioannou K., 2020, Phys. Rev. D, 101, 123007
Lasota J.-P., Haensel P., Abramowicz M. A., 1996, ApJ, 456, 300
Legred I., Chatziioannou K., Essick R., Han S., Landry P., 2021, Phys. Rev. D, 104, 063003
Li Y.-J., Tang S.-P., Wang Y.-Z., Yuan Q., Fan Y.-Z., Wei D.-M., 2021a, ApJ, 923, 97
Li Y.-J., Wang Y.-Z., Han M.-Z., Tang S.-P., Yuan Q., Fan Y.-Z., Wei D.-M., 2021b, ApJ, 917, 33
Lo K.-W., Lin L.-M., 2011, ApJ, 728, 12
Loredo T. J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, AIP Conf. Proc. Vol. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 195
Lorimer D. R., 2008, Living Rev. Relativ., 11, 8
Mandel I., 2010, Phys. Rev. D, 81, 084029
Mandel I., Farr W. M., Colonna A., Stevenson S., Tiňo P., Veitch J., 2017, MNRAS, 465, 3254
Mandel I., Farr W. M., Gair J. R., 2019, MNRAS, 486, 1086
Messina F., Dudi R., Nagar A., Bernuzzi S., 2019, Phys. Rev. D, 99, 124051
Miller M. C., Chirenti C., Lamb F. K., 2019, ApJ, 888, 12
Mukherjee P., Parkinson D., Liddle A. R., 2006, ApJ, 638, L51
Ng K. K. Y., Vitale S., Zimmerman A., Chatziioannou K., Gerosa D., Haster C.-J., 2018, Phys. Rev. D, 98, 083007
Nitz A. H., Wang Y.-F., 2021, Phys. Rev. Lett., 127, 151101
Özel F., Freire P., 2016, ARA&A, 54, 401
Ozel F., Psaltis D., Narayan R., Villarreal A. S., 2012, ApJ, 757, 55
Pankow C., 2018, ApJ, 866, 60
Pejcha O., Thompson T. A., Kochanek C. S., 2012, MNRAS, 424, 1570
Pürrer M., Hannam M., Ohme F., 2016, Phys. Rev. D, 93, 084042
Rinaldi S., Del Pozzo W., 2021, MNRAS, 509, 5454
Romano J. D., Cornish N. J., 2017, Living Rev. Relativ., 20, 2
Romero-Shaw I. M. et al., 2020, MNRAS, 499, 3295
Roulet J., Chia H. S., Olsen S., Dai L., Venumadhav T., Zackay B., Zaldarriaga M., 2021, Phys. Rev. D, 104, 083010
Sadiq J., Dent T., Wysocki D., 2021, preprint (arXiv:2112.12659)
Safarzadeh M., Ramirez-Ruiz E., Berger E., 2020, ApJ, 900, 13
Shao D.-S., Tang S.-P., Jiang J.-L., Fan Y.-Z., 2020, Phys. Rev. D, 102, 063006
Shaw R., Bridges M., Hobson M. P., 2007, MNRAS, 378, 1365
Smith R., Field S. E., Blackburn K., Haster C.-J., Pürrer M., Raymond V., Schmidt P., 2016, Phys. Rev. D, 94, 044031
Speagle J. S., 2020, MNRAS, 493, 3132
Talbot C., Thrane E., 2018, ApJ, 856, 173
Talbot C., Smith R., Thrane E., Poole G. B., 2019, Phys. Rev. D, 100, 043030
Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A., 2018, preprint (arXiv:1804.06788)
Tauris T. M. et al., 2017, ApJ, 846, 170
Taylor S. R., Gair J. R., 2012, Phys. Rev. D, 86, 023502
Taylor S. R., Gair J. R., Mandel I., 2012, Phys. Rev. D, 85, 023535
Thorsett S. E., Chakrabarty D., 1999, ApJ, 512, 288
Thrane E., Talbot C., 2019, Publ. Astron. Soc. Aust., 36, e010
Tiwari V., 2021, Class. Quantum Gravity, 38, 155007
Veitch J., Vecchio A., 2010, Phys. Rev. D, 81, 062003
Vigna-Gómez A. et al., 2018, MNRAS, 481, 4009
Vitale S., Gerosa D., Farr W. M., Taylor S. R., 2020, preprint (arXiv:2007.05579)
Wade L., Creighton J. D. E., Ochsner E., Lackey B. D., Farr B. F., Littenberg T. B., Raymond V., 2014, Phys. Rev. D, 89, 103012
Wong K. W. K., Contardo G., Ho S., 2020, Phys. Rev. D, 101, 123005
Wong K. W. K., Breivik K., Kremer K., Callister T., 2021, Phys. Rev. D, 103, 083021
Wysocki D., O'Shaughnessy R., Wade L., Lange J., 2020, preprint (arXiv:2001.01747)
Zevin M. et al., 2021, ApJ, 910, 152
Zhu X.-J., Howell E. J., Blair D. G., Zhu Z.-H., 2013, MNRAS, 431, 882
Zhu X., Thrane E., Osłowski S., Levin Y., Lasky P. D., 2018, Phys. Rev. D, 98, 043002

APPENDIX A: THE INSUFFICIENCY OF P–P PLOTS AS A DIAGNOSTIC TEST

In order to obtain unbiased posteriors for the hyperparameters describing a population of events using the likelihood in equation (5), the individual-event posteriors must also be unbiased. A common diagnostic tool for evaluating the performance of a stochastic sampler is a 'probability–probability plot', or a P–P plot (Cook, Gelman & Rubin 2006; Talts et al. 2018). This provides a graphical way to verify that the true parameter values are recovered within a certain credible interval for the expected fraction of events in a population. If the likelihood correctly describes the distribution of data, the true parameter values should be recovered within the 5 per cent CI 5 per cent of the time, the 95 per cent CI 95 per cent of the time, etc. For the Whittle likelihood in equation (2) used when performing parameter estimation on individual gravitational-wave sources, this is satisfied if the data are Gaussian about the assumed noise PSD. In the case of unbiased individual-event posterior samples for a particular parameter, the P–P plot should be approximately diagonal, as it shows the fraction of events for which the true value of a given parameter falls within the given credible interval as a function of that credible interval.

An example P–P plot obtained using the PyMultiNest sampler for a population of 100 simulated BNS sources is shown in Fig. A1.

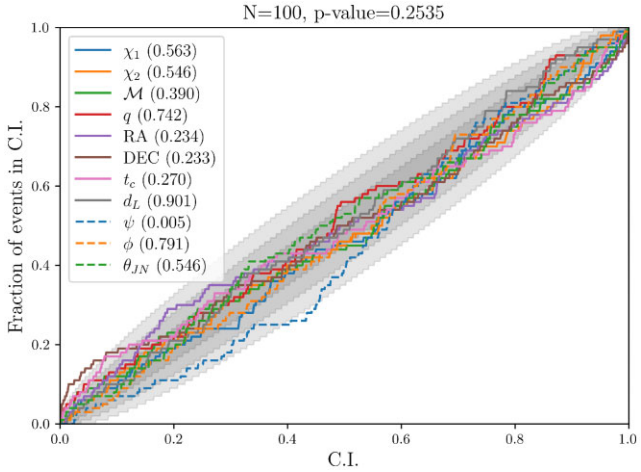

Figure A1. P–P plot showing the fraction of events for which the true value is recovered within a certain credible interval as a function of that credible interval for a population of 100 simulated BNS sources, sampled with the PyMultiNest package. The lines for individual parameters stay within the 3σ credible region, shaded in light grey, and the probability values quoted in the legend are consistent with passing the P–P test. The other grey-shaded areas show the 1 and 2σ credible regions.

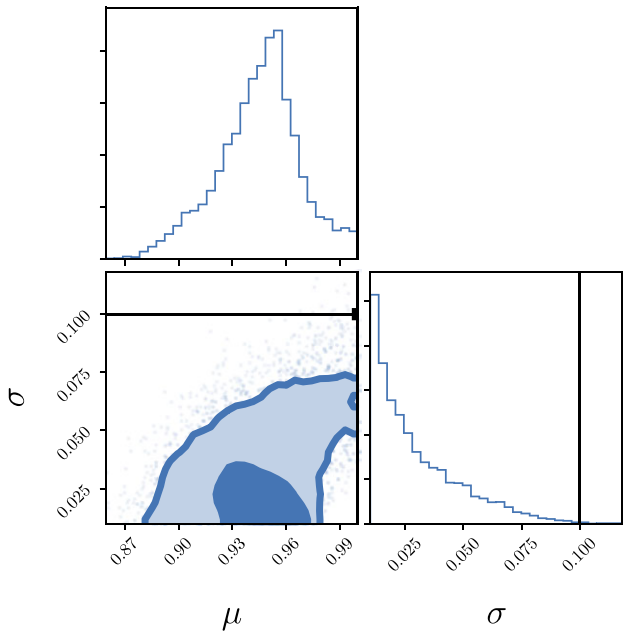

Figure A2. Corner plot for the inferred mass ratio hyperparameters using the 100 events used to generate the P–P plot in Fig. A1.

The extrinsic parameters are drawn from the same prior distributions described in the main text, and the spins follow the low-aligned-spin prior. The mass ratio is also drawn from the same population explored earlier in the main text, namely a narrow truncated Gaussian with $\mu = 1$ and $\sigma = 0.1$. The chirp masses are drawn from a uniform prior between 1.52 and 1.70 $M_\odot$. For the P–P plot to be unbiased, the distributions from which the events are drawn must match the priors applied during sampling, meaning that there is no prior mismatch and no cut based on the SNR of the individual events, as has been the case in the rest of this work. The legend shows the probability for the fraction of events within each credible interval to be drawn from a uniform distribution for individual parameters, as expected from Gaussianity. For the mass ratio, this value is $p = 0.742$, passing the P–P test (where the threshold is $p > 1/11 = 0.09$). The probability that the individual-parameter probabilities are drawn from a uniform distribution is 0.254, consistent with random chance for 11 parameters and indicative of unbiased sampling across the 11 parameters drawn from and recovered with this particular set of priors.

However, when the same 100 individual events are analysed hierarchically to recover the true values of μ and σ, the results are biased at the 3σ level, as shown in Fig. A2. This demonstrates a case where the sampling algorithm passes the P–P test for a particular population but still yields biased hierarchical inference results, highlighting the insufficiency of P–P plots as a diagnostic tool for individual-event parameter estimation. In this case, the recovered hyperparameters favour a narrowly distributed population peaking away from $\mu = 1$, indicating that the sampler is unable to thoroughly explore the edge of the prior space where most of the probability lies for the nearly equal-mass events included in the population. This is due to the adapted simultaneous ellipsoidal nested sampling method used by PyMultiNest (Mukherjee, Parkinson & Liddle 2006; Shaw, Bridges & Hobson 2007; Feroz & Hobson 2008; Feroz et al. 2009), which bounds the iso-likelihood contours around clusters of live points with N-dimensional ellipsoids. Because the probability for mass ratio rails against the edge of the prior and the algorithm is inefficient at sampling near edges, the peak of the distribution at equal mass is undersampled.

PyMultiNest internally works with a uniform prior for samples from the unit hypercube, which must then be scaled to the physical parameter space such that the scaled samples are drawn from the desired physical prior distribution. In order to improve the convergence near equal masses, we propose to use a two-stage mapping that shifts the peak of the probability at $q = 1$ away from the edge of the prior in the frame of the sampler. Typically, samples from the unit cube are rescaled on to the appropriate prior distribution for a given parameter via the inverse of the prior's cumulative distribution function, such that for a sample from the unit cube, x, $q(x) = \mathrm{CDF}^{-1}(x)$ (although other methods have been proposed, e.g. Alsing & Handley 2021). Here, we propose to add an intermediate step

$$u = 2\min(x, 1 - x), \tag{A1}$$

$$q = \mathrm{CDF}^{-1}(u), \tag{A2}$$

where u still takes on values within the unit interval, but instead of $q(x = 1) = 1$, $q(x = 0.5) = 1$. This transformation maps equal mass – where the peak of the probability lies – to the centre of the sampled space rather than the edge, which is more difficult to sample. In Fig. A3, we show the mass ratio PPDs for the medium-spin population described in the main text with no spin-prior mismatch obtained with and without this modified rescaling method. Similarly to the corner plot in Fig. A2, without the modified rescaling, the PPD shown in red peaks at lower mass ratios and is more narrowly distributed. The true hyperparameters are excluded from the recovered posteriors at $>3\sigma$ credibility. Once the modified rescaling is implemented for the individual-event parameter estimation, the hierarchical inference becomes unbiased, as shown in blue. This stealth bias introduced by the stochastic sampling algorithm that is not caught with a P–P test demonstrates the importance of verifying hierarchical inference analyses with synthetic populations where

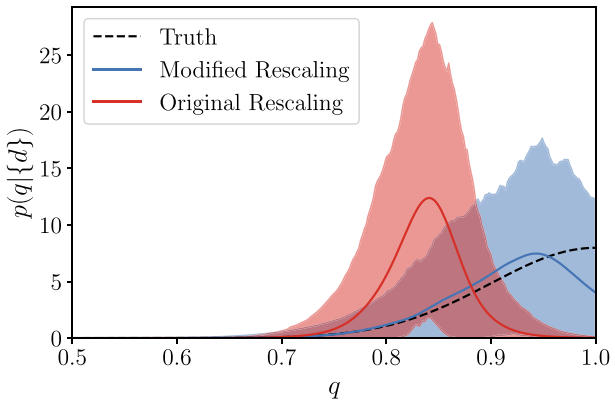

Figure A3. Inferred-mass-ratio PPDs when applying the original inverse CDF rescaling method (red) and the modified rescaling method described in equation (A2) (blue) to the medium-spin population described in the main text analysed with no spin-prior mismatch. The dotted line shows the true distribution, and the shading shows the 90 per cent credible region. Without the modified rescaling, the mass ratio distribution is biased towards lower q values and more narrowly peaked.

the true hyperparameter values are known and controllable before conducting the analysis on real data.

APPENDIX B: SELECTION EFFECTS

In order to evaluate the selection function in equation (6), we calculate the detection probability, $p_{\rm det}(\boldsymbol{\theta}_i)$, using an injection campaign. $\alpha(\boldsymbol{\Lambda})$ gives the fraction of signals that will be detected drawn from a population model with hyperparameters $\boldsymbol{\Lambda}$. We generate 194 953 total simulated signals, calculating the network optimal SNR, $\rho^{\rm net}_{\rm opt}$ for each. Of these, 40 000 are above the threshold for detection, $\rho^{\rm net}_{\rm opt} \geq 9$. The distribution from which the injections are drawn, $p_{\rm draw}(\boldsymbol{\theta})$, is uniform in total mass over the range $[2, 5]\,{\rm M}_\odot$, and uses the same prior distributions described in the main text for all the extrinsic parameters. The mass ratio distribution is

$$p_{\rm draw}(q) = 0.5\left(1.67 + \frac{1}{\sigma}\frac{\mathcal{N}(q|\mu,\sigma)}{\Phi\left(\frac{q_{\rm max}-\mu}{\sigma}\right) - \Phi\left(\frac{q_{\rm min}-\mu}{\sigma}\right)}\right),$$
$$q_{\rm min} < q < q_{\rm max} \tag{B1}$$

which is the normalized superposition of a uniform distribution and truncated Gaussian distribution with $\mu = 1$, $\sigma = 0.1$ between $q_{\rm min} = 0.4$ and $q_{\rm max} = 1$. The truncated Gaussian is added to enhance the number of injections with nearly equal-mass ratios, since this is the part of the parameter space that should have the most support given the true distribution we used for the simulated populations described in Section 2. The spins are drawn following the aligned-spin prior with $\chi_{\rm max} = 0.99$. The injections are reweighted so that $\chi_{\rm max}$ matches the corresponding value used during the first parameter estimation step for each of the spin-prior mismatches we consider earlier. The different combinations of true population distribution, PE prior, and injection distribution are summarized in Table 1.

We can then estimate the detection probability as follows:

$$\alpha(\boldsymbol{\Lambda}) = \int {\rm d}\boldsymbol{\theta}\, p_{\rm det}(\boldsymbol{\theta})\pi_{\rm pop}(\boldsymbol{\theta}|\boldsymbol{\Lambda}), \tag{B2}$$

$$\alpha_{\rm draw} = \int {\rm d}\boldsymbol{\theta}\, p_{\rm det}(\boldsymbol{\theta})p_{\rm draw}(\boldsymbol{\theta}) \approx \frac{N_{\rm found}}{N_{\rm draw}} \tag{B3}$$

$$p_{\rm found}(\boldsymbol{\theta}) = \frac{p_{\rm draw}(\boldsymbol{\theta})p_{\rm det}(\boldsymbol{\theta})}{\alpha_{\rm draw}}, \tag{B4}$$

$$\alpha(\boldsymbol{\Lambda}) = \alpha_{\rm draw}\int {\rm d}\boldsymbol{\theta}\frac{p_{\rm found}(\boldsymbol{\theta})}{p_{\rm draw}(\boldsymbol{\theta})}\pi_{\rm pop}(\boldsymbol{\theta}|\boldsymbol{\Lambda}), \tag{B5}$$

$$\alpha(\boldsymbol{\Lambda}) \approx \frac{1}{N_{\rm found}}\sum_j \frac{\pi_{\rm pop}(\boldsymbol{\theta}_j|\boldsymbol{\Lambda})}{p_{\rm draw}(\boldsymbol{\theta}_j)}. \tag{B6}$$

We account for the uncertainty in the Monte Carlo integral in equation (B6) following the method in Farr (2019) and reject parts of the hyperparameter space during sampling that do not have enough injections to meet the accuracy requirements therein. We note that our population model, $\pi_{\rm pop}(\boldsymbol{\theta}_j|\boldsymbol{\Lambda})$, in equation (B6) includes only the total mass distribution, and not the mass ratio distribution, as the latter has a negligible effect on the detectability of the source and would require a much higher number of injections to meet the accuracy requirements for our choice of narrow truncated Gaussian population distribution.

This paper has been typeset from a TEX/LATEX file prepared by the author.