Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **484,** 4008–4023 (2019)
Advance Access publication 2019 January 22

doi:10.1093/mnras/stz225

Digging the population of compact binary mergers out of the noise

Sebastian M. Gaebel,[1]★ John Veitch,[2] Thomas Dent[3,4] and Will M. Farr[1,5,6]

[1]*Institute for Gravitational Wave Astronomy, School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*
[2]*School of Physics and Astronomy, Institute for Gravitational Research, University of Glasgow, Glasgow G12 8QQ, UK*
[3]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), D-30167 Hannover, Germany*
[4]*Galician Institute for High Energy Physics (IGFAE), Campus Vida, Universidade de Santiago de Compostela, Santiago de Compostela, E-15782, Spain*
[5]*Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA*
[6]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA*

Accepted 2019 January 18. Received 2019 January 18; in original form 2018 September 14

ABSTRACT

Coalescing compact binaries emitting gravitational wave (GW) signals, as recently detected by the Advanced LIGO–Virgo network, constitute a population over the multidimensional space of component masses and spins, redshift, and other parameters. Characterizing this population is a major goal of GW observations and may be approached via parametric models. We demonstrate hierarchical inference for such models with a method that accounts for uncertainties in each binary merger's individual parameters, for mass-dependent selection effects, and also for the presence of a second population of candidate events caused by detector noise. Thus, the method is robust to potential biases from a contaminated sample and allows us to extract information from events that have a relatively small probability of astrophysical origin.

Key words: gravitational waves – methods: statistical.

1 INTRODUCTION

Since the first detection of gravitational waves (GWs) in 2015 (Abbott et al. 2016c), the Advanced LIGO and Advanced Virgo detectors have observed the coalescence of multiple compact binary systems, and have begun to reveal the population of coalescing compact objects (Abbott et al. 2016a). This population is enabling studies in fields from probing alternative theories of gravity to constraining models of stellar evolution. These tend to be interested either in individual, preferably loud, signals or in the population of sources as a whole. The latter type of population analysis tries to estimate the parameters governing the distribution of sources in the Universe; their masses and spins and the value of the astrophysical merger rate (Abbott et al. 2016g) are of particular interest. As the sensitivity of detectors improves over the coming years, the detected number of sources is expected to grow at an accelerated pace, rapidly increasing the amount of information available for population studies (Abbott et al. 2018).

When undertaking population analyses, one has to consider that real detectors may produce noise transients that cannot in all cases be distinguished from astrophysical GW sources. To avoid population inferences being biased by noise events, one might consider only events with a much higher probability to be of astrophysical origin than to be caused by noise artefacts. In templated searches for compact binaries, the relative probability of astrophysical versus noise origin for a candidate event is a function of a detection statistic calculated for each event by a search analysis pipeline (see e.g. Abbott et al. 2016h; Usman et al. 2016; Cannon et al. 2012; Cannon, Hanna & Peoples 2015; Messick et al. 2017; Nitz et al. 2017). Typically, a candidate event is generated as a local maximum in matched filter signal-to-noise ratio (SNR) above a search threshold; the detection statistic value then incorporates the matched filter SNR, as well as other goodness-of-fit tests to reject non-Gaussian instrumental noise transients (Allen 2005; Nitz 2018).

At low SNR, the population of events is dominated by the noise 'background', whereas at high SNR (or in general for events assigned high-statistic values) the astrophysical 'foreground' dominates.[1] To limit possible pollution of the sample used for population inference, one may place a minimum threshold on the detection statistic; any event above threshold is then assumed to be astrophysical, whereas all other events are discarded as potential noise transients. Note that the choice of threshold value requires an empirical estimate of the rate and distribution of background events (Capano et al. 2017) since the rate, strength, and morphologies of detector noise artefacts are not known a priori (Abbott et al. 2016b).

A simple strategy of thresholding is sub-optimal for two reasons. First, discarding events below the threshold will almost certainly discard information from some number of quiet but still identifiable signals (Kovetz et al. 2017); secondly, there is still a finite chance

★ E-mail: sgaebel@star.sr.bham.ac.uk

[1]We will loosely refer to the detection statistic as 'SNR' when discussing the distinction between instrumental noise events and astrophysical events.

that the resulting 'signal' set is nevertheless contaminated by noise, leading to potentially biased inferences. SNR threshold requires a trade-off between these two considerations and depends on the intended use. One also has to take into consideration any bias in the observed population produced by the effect of source parameters on the loudness of a signal, and thus its chance of exceeding an SNR threshold (Fairhurst & Brady 2008); we expect potentially major observation selection effects for binary mass(es) (Tiwari 2017) and component spins (Ng et al. 2018).

Here, we propose a method that alleviates the issues associated with simple thresholding by applying a hierarchical mixture model under which each event is considered to originate from either a foreground (astrophysical) or a background (noise) population. For each event, the probability of either case naturally defines a weight for its contribution to inferences on population parameters.

This method combines the processes of estimating the expected number of events in either class (Farr et al. 2015; the number of foreground events being a proxy for the astrophysical merger rate density) and estimating parameters of the underlying populations, which have previously been performed separately. It avoids being biased through the inclusion of background events, while being able to use events with a non-negligible probability of noise origin, which would be discarded by thresholding. In theory, this method allows the SNR threshold to be reduced to an arbitrarily low value, though in practice we are still limited by the computational resources required to extract the source parameters from each event under consideration.[2]

Our method is applicable to any hierarchical model of a source population, examples of which have been explored in the literature. This includes analyses that combine information from multiple events to infer a parameter common to all, such as deviations from general relativity (Li et al. 2012) or a parametrized neutron star equation of state (Del Pozzo et al. 2013). The use of a mixture model with astrophysical and noise populations is particularly useful when the model of interest has a strong effect on the detectability of sources, i.e. the detected events are unrepresentative of the underlying population. A good example is the mass distribution of sources, which we consider next.

For a compact binary in its inspiral phase, the frequency-domain amplitude of the GW in the stationary phase approximation is proportional to $\mathcal{M}^{5/6}$, where $\mathcal{M} = (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ is the chirp mass of the signal (Cutler & Flanagan 1994). Since the GW detectors are sensitive to the signal amplitude, more massive sources will produce a larger SNR for a fixed position relative to the detectors. For a search that counts signals above a particular threshold, more massive signals will be overrepresented in the selected events by a factor $\approx\mathcal{M}^{5/2}$, neglecting cosmological redshifting of the source and assuming a constant source rate per unit volume at all distances.

Messenger & Veitch (2013) considered the problem of selection effects in mass distribution inference by dividing the observing time into discrete chunks, which each contain zero or one sources, and computing population likelihoods while accounting for false alarms and false dismissals from an idealized noise distribution. Farr et al. (2015) derived an equivalent formalism for rate inference that allows a population shape function to be estimated alongside. Our derivation in Section 2 follows similar lines.

The selection function for masses was important in estimating the astrophysical event rates in the first Advanced LIGO observing run (O1), which inferred rates using a mixture model, for fixed choices of population shape (i.e. mass distribution; Abbott et al. 2016d,g). A separate analysis also estimated the slope of a power-law model of the mass distribution function considering the detected events, described in Abbott et al. (2016a; and updated in Abbott et al. 2017).

The selection function in the form of a sensitivity-weighted measure of the space–time volume VT surveyed for signals above a certain SNR threshold is also important when considering searches that do not make a clear detection. There, the loudest background event, or a nominal detection threshold, is used to set an upper limit on the astrophysical rate of a fiducial source population, for example, limits on the rate of mergers of binary neutron stars and neutron star–black hole binaries in O1 (Abbott et al. 2016f). The sensitive-volume approach has been in use since the initial detector era (Biswas et al. 2009; Abadie et al. 2011), and continues to be refined to incorporate mass- and spin-dependent selection effects (O'Shaughnessy et al. 2010; Dominik et al. 2015; Ng et al. 2018) and cosmological effects, as well as to improve the accuracy of measurement (Tiwari 2017).

As the number of detections increases, determination of the population of coalescing compact binaries is expected to provide insight into the astrophysics of black hole and neutron star binary formation (Kalogera et al. 2007; Abbott et al. 2016e; Mandel & Farmer 2018). Population synthesis models can describe the masses and spins of coalescing compact binaries under a variety of formation scenarios (see e.g. Belczynski et al. 2016b; Belczynski et al. 2016a; 2017; Spera, Mapelli & Bressan 2015). Comparison of these predictions to the observed distribution can be used to constrain the uncertainties in parametrized models of source populations (Barrett et al. 2017; Zevin et al. 2017). This has motivated the development of methods to determine the mass-dependent coalescence rate in the absence of false alarms, using both specific parametrized models (Talbot & Thrane 2018; Taylor & Gerosa 2018; Wysocki, Lange & O'Shaughnessy 2018a; Wysocki et al. 2018b) and non-parametric methods (Mandel et al. 2017). The alignment of black hole spins is expected to be a key distinguishing feature between binaries formed in the field or through dynamical interactions (see e.g. Mandel & O'Shaughnessy 2010; Gerosa et al. 2013; Stevenson, Ohme & Fairhurst 2015; Farr et al. 2017; Fishbach, Holz & Farr 2017; Gerosa & Berti 2017; O'Shaughnessy, Gerosa & Wysocki 2017; Stevenson, Berry & Mandel 2017; Talbot & Thrane 2017; Vitale et al. 2017), which also requires an understanding of the spin-selection function (O'Shaughnessy et al. 2010; Dominik et al. 2015; Ng et al. 2018; Tiwari, Fairhurst & Hannam 2018). This is caused by the 'orbital hang-up' effect (Campanelli, Lousto & Zlochower 2006), where binaries with component spins aligned with the total orbital angular momentum tend to inspiral (i.e. reduce orbital radius) more slowly than those with anti-aligned spins. This leads to an increase in the radiation emitted at specific frequencies in the sensitive band of ground-based detectors, thus increasing the detectability of these sources.

The work presented here is complementary to these studies, as it aims to incorporate an astrophysical distribution model as part of a mixture with a noise component. As the observed population is limited by the sensitivity of Advanced ground-based detectors, the population of candidate sources at the greatest

[2] An analysis that effectively removes all SNR thresholds, applying Bayesian analysis to the entirety of the GW data set rather than restricting to data close to events triggered on SNR maxima, is proposed in Smith & Thrane (2018); its application appears at present to be still more limited by computational cost.

distances (lowest detection significance) will be contaminated with background events. We expect our method, using information from such sources, to improve both the precision and accuracy of merger rate and population parameter estimates; though as we will see, the degree of improvement depends on how easily the foreground and background populations can be separated by existing analyses.

We start by defining our notation and deriving the general form of the model in Section 2. Section 3 describes its application to a toy model of mass distribution inference in the presence of noise, and shows its application to a range of simple analytic population models. In Section 5, we consider a more realistic simulated data set derived from an engineering run prior to the start of Advanced LIGO observations in 2015. We conclude in Section 6.

2 DERIVATION OF THE GENERIC MODEL

We consider a mixture of two populations, the astrophysical 'foreground' and terrestrial noise 'background': quantities defined analogously for both populations will be distinguished by the subscripts F or B, respectively. Quantities without subscript then refer to the total population that is the union of foreground and background.

The model is also hierarchical: each event, if assumed astrophysical, has a set of intrinsic properties such as component masses and spins, which we collectively denote γ. The distribution of these properties over each population is assumed to have a form described by a set of hyper-parameters. We do not have access to the 'true' values of properties for each event, only to a set of samples from an (typically Bayesian) estimate based on data around the event. These samples are derived under the assumption that the event is astrophysical, thus events that are in fact background will also be assigned parameter estimates.[3]

We then define the core quantities used in the following derivation as

(i) ρ_i, $\{\rho\}$: ranking statistic for one, respectively, for all events in a given data set,

(ii) $N_{\rm obs}$, $N_{\rm F,obs}$, $N_{\rm B,obs}$: observed number of events above a threshold $\rho_i > \rho_{\rm thr}$,

(iii) $N_{\rm exp}$, $N_{\rm F,\,exp}$, $N_{\rm B,\,exp}$: expected number of events with $\rho_i > \rho_{\rm thr}$, when modelling these as a Poisson process,

(iv) $\theta_{\rm F}$, $\theta_{\rm B}$: hyper-parameters that describe the shape of the foreground and background populations,

(v) η_i, $\{\eta\}$: indicator variable showing whether any given event, respectively all events, belong(s) to the astrophysical ($\eta = F$) or to the noise population ($\eta = B$), and

(vi) $\vec{\gamma}_i$, $\{\vec{\gamma}\}$: vector of samples representing the parameter estimates (masses, spins, etc.) of one, respectively all events, under the assumption that events are astrophysical.

We wish to infer the joint posterior probability distribution of rates and population parameters for the two populations, given some events for which $\{\rho\}$ and $\{\vec{\gamma}\}$ have been determined by the search and parameter-estimation stages of data analysis:

$$p(N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}|\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs}). \tag{1}$$

[3]We do not, of course, know with certainty that any given event is background.

Using Bayes' theorem, we can express the posterior distribution (1) in terms of prior and likelihood functions:

$$p(N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}|\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs}) = \frac{p(\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs}|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B})p(N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B})}{p(\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs})}. \tag{2}$$

We drop the normalization constant $p(\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs})$ and factor out the likelihood for $N_{\rm obs}$ as being independent of the population hyper-parameters $\theta_{\rm F}$ and $\theta_{\rm B}$:

$$\begin{aligned} &p(\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs}|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}) \\ &= p(N_{\rm obs}|N_{\rm F,exp}, N_{\rm B,exp})p(\{\rho\}, \{\vec{\gamma}\}|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}) \\ &= \frac{N_{\rm exp}^{N_{\rm obs}} e^{-N_{\rm exp}}}{N_{\rm obs}!} p(\{\rho\}, \{\vec{\gamma}\}|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}), \end{aligned} \tag{3}$$

where we use a Poisson likelihood for $N_{\rm obs}$ with a total expected number of events $N_{\rm exp} = N_{\rm F,\,exp} + N_{\rm B,\,exp}$. The second term, $p(\{\rho\}, \{\vec{\gamma}\}|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B})$, is the likelihood for the observed SNRs and parameter estimates, for the mixture model. We assume each event is conditionally independent given the population parameters, and so the joint likelihood is just the product of the likelihood for each one:

$$\begin{aligned} &p(\{\rho\}, \{\vec{\gamma}\}|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}) \\ &= \prod_i p(\rho_i, \vec{\gamma}_i|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}). \end{aligned} \tag{4}$$

Now, we can split each of these into terms for the astrophysical and noise sub-models by introducing an indicator variable $\eta_i \in \{F, B\}$, whose probability will depend on the rate parameters $N_{\rm F,exp}$ and $N_{\rm B,exp}$:

$$\begin{aligned} &p(\rho_i, \vec{\gamma}_i|N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}) \\ &= p(\rho_i, \vec{\gamma}_i|\theta_{\rm F}, \eta_i = {\rm F})p(\eta_i = {\rm F}|N_{\rm F,exp}, N_{\rm B,exp}) \\ &\quad + p(\rho_i, \vec{\gamma}_i|\theta_{\rm B}, \eta_i = {\rm B})p(\eta_i = {\rm B}|N_{\rm F,exp}, N_{\rm B,exp}) \\ &= p(\rho_i, \vec{\gamma}_i|\theta_{\rm F}, \eta_i = {\rm F})\frac{N_{\rm F,exp}}{N_{\rm exp}} + p(\rho_i, \vec{\gamma}_i|\theta_{\rm B}, \eta_i = {\rm B})\frac{N_{\rm B,exp}}{N_{\rm exp}}, \end{aligned} \tag{5}$$

where the probability of each class is just the expected fraction of the total number. Since this is a sum of probability densities, special care must be taken to ensure all terms are properly normalized, such that

$$\int_{\rho_{\rm thr}}^{\infty} {\rm d}\rho \int {\rm d}\vec{\gamma}\; p(\rho, \vec{\gamma}|\theta_\eta, \eta_i) = 1, \tag{6}$$

for $\eta = {\rm F}$ and $\eta = {\rm B}$, where $\rho_{\rm thr}$ is a minimum SNR value for which events are considered, either as a result of the event generation method or as a choice to limit computational costs. Neglecting this normalization would introduce an artificial preference for one component over the other. An extension to further sub-populations is simply achieved by including additional classes with their own rate and hyper-parameters.

Recombining the pieces, we can write the desired posterior in equation (1) as

$$\begin{aligned} &p(N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B}|\{\rho\}, \{\vec{\gamma}\}, N_{\rm obs}) \\ &\propto p(N_{\rm F,exp}, N_{\rm B,exp}, \theta_{\rm F}, \theta_{\rm B})e^{-N_{\rm exp}} \\ &\quad \times \prod_i \left[\, p(\rho_i, \vec{\gamma}_i|\theta_{\rm F}, \eta_i = {\rm F})N_{\rm F,exp} + p(\rho_i, \vec{\gamma}_i|\theta_{\rm B}, \eta_i = {\rm B})N_{\rm B,exp} \,\right]. \end{aligned} \tag{7}$$

This expression is similar to equation (21) from Farr et al. (2015) with an explicitly added dependence on source parameter estimates.

This implies that our formalism reduces to the Farr et al. (2015) result as used by the LIGO–Virgo Collaboration to estimate binary black hole merger rates (Abbott et al. 2016a,g), if the event distribution over mass or similar parameters is not free to vary.

The dependence on event parameters arises through the use of samples $\vec{\gamma}_i$, $i = 1 \ldots n$, drawn from the likelihood function of the data d for a given point in parameter space $p(\mathrm{d}|\vec{\gamma})$. These allow us to evaluate the population likelihood function via marginalization over the unknown true parameters, using the n samples to perform a Monte Carlo integral as in Mandel (2010):

$$\begin{aligned} p(\mathrm{d}|\theta_\mathrm{F}) &= \int p(\mathrm{d}|\vec{\gamma})p(\vec{\gamma}|\theta_\mathrm{F})\mathrm{d}\vec{\gamma} \\ &= \langle p(\vec{\gamma}_i|\theta_\mathrm{F})\rangle_{p(\mathrm{d}|\vec{\gamma}_i)} \\ &\approx n^{-1}\sum_j p(\vec{\gamma}_j|\theta_\mathrm{F}). \end{aligned} \tag{8}$$

Samples from the likelihood therefore serve as a useful intermediate representation of the raw interferometer data d. To obtain a quantity directly relevant for an astrophysical interpretation, the expected number can be transformed into the local merger rate R using the observing time T and the sensitive volume $V(\gamma)$:

$$R = \frac{N_\mathrm{F,exp}}{T\int \mathrm{d}\vec{\gamma}\; V(\vec{\gamma})p(\vec{\gamma}|\theta_\mathrm{F})}, \tag{9}$$

where the integral marginalizes over the space of source parameters $\vec{\gamma}$. In practice, $V(\vec{\gamma})$ is estimated for a particular data set by a Monte Carlo campaign, adding ('injecting') a large number of simulated signals to the data and counting the resulting number of events above threshold.[4]

An additional quantity, which is not directly used in the derivation of our model but is important for an astrophysical interpretation, is the probability of any given event originating from the astrophysical foreground, p_astro:

$$\begin{aligned} &p(\eta_i = \mathrm{F}, N_\mathrm{F,exp}, N_\mathrm{B,exp}, \theta_\mathrm{F}, \theta_\mathrm{B}|\rho_i, \vec{\gamma}_i) \\ &\quad = \frac{p(\rho_i, \vec{\gamma}_i|\theta_\mathrm{F}, \eta_i = \mathrm{F})N_\mathrm{F,exp} \times p(N_\mathrm{F,exp}, N_\mathrm{B,exp}, \theta_\mathrm{F}, \theta_\mathrm{B})}{p(\rho_i, \vec{\gamma}_i|\theta_\mathrm{F}, \eta_i = \mathrm{F})N_\mathrm{F,exp} + p(\rho_i, \vec{\gamma}_i|\theta_\mathrm{B}, \eta_i = \mathrm{B})N_\mathrm{B,exp}}, \end{aligned} \tag{10}$$

$$\begin{aligned} p_{\mathrm{astro},i} = &\iiiint p(\eta_i = \mathrm{F}, N_\mathrm{F,exp}, N_\mathrm{B,exp}, \theta_\mathrm{F}, \theta_\mathrm{B}|\rho_i, \vec{\gamma}_i) \\ &\mathrm{d}N_\mathrm{F,exp}\,\mathrm{d}N_\mathrm{B,exp}\,\mathrm{d}\theta_\mathrm{F}\,\mathrm{d}\theta_\mathrm{B}\,, \end{aligned} \tag{11}$$

where we marginalized over the population parameters. The integration range for both $N_\mathrm{F,exp}$ and $N_\mathrm{B,exp}$ is $(0,\infty)$, while the population parameters θ_F and θ_B are integrated over their respective domains.

3 TOY MODEL

We construct a simple toy model of the Universe to test our inference framework in various ways. The toy model allows us to generate a large number of realizations from the same underlying parameters, and to be certain that we use the correct model when analysing these realizations. For simplicity, we consider a static, flat, and finite universe. Events are characterized completely by the distance r to the source and a single mass parameter m, which takes the place of the γ used in the previously derived expressions. This mass parameter can be thought of as similar to the chirp mass. Additional effects such as inclination, spins, mass ratio, or antenna patterns are ignored.

For our detection statistic ρ, we simply use (a simplified proxy for) the SNR, whose expected value ρ_true is determined by r and m as

$$\rho_\mathrm{true} = K\frac{m}{r}, \tag{12}$$

where K is an arbitrary constant that quantifies the detector sensitivity. We also model the uncertainty in the estimation of the mass parameter as

$$\sigma_\mathrm{PE} \propto \frac{m}{\rho}, \tag{13}$$

which is a simplification of the relation given in Cutler & Flanagan (1994).

To apply the generic form found in equation (7) to a specific problem, we need to evaluate the terms $p(\rho_i, \vec{m}_i|\theta_\mathrm{F}, \eta_i = \mathrm{F})$ and $p(\rho_i, \vec{m}_i|\theta_\mathrm{B}, \eta_i = \mathrm{B})$. This involves finding a functional form for the selection effects. For sources distributed uniformly in our static universe, we can derive the needed expression directly by manipulating the joint distribution of masses and observed (detection) SNRs ρ_obs. We use the SNR relation defined above in equation (12), which defines a mass-dependent lower cut-off $\rho_\mathrm{cut-off}(m) = K\; m\; r_U^{-1}$ to the ρ_true possible in our toy universe with radius r_U. Additionally, we use the fact that the euclidean distances r of sources distributed uniformly in volume follow a r^2-distribution. Therefore ,

$$\begin{aligned} &p(\rho_\mathrm{obs}, m) \\ &= \frac{1}{V_\mathrm{U}}\int_0^{r_\mathrm{U}} \mathrm{d}r\; 4\pi r^2 p(\rho_\mathrm{obs}, m|r) \\ &= \frac{1}{V_\mathrm{U}}\int_0^{r_\mathrm{U}} \mathrm{d}r \int_{\rho_\mathrm{cutoff}(m)}^{\infty} \mathrm{d}\rho_\mathrm{true}\; 4\pi r^2 p(\rho_\mathrm{obs}|\rho_\mathrm{true}) \\ &\quad \times p(\rho_\mathrm{true}|m, r)p(m) \\ &= \frac{4\pi p(m)}{V_U}\int_{\rho_\mathrm{cutoff}(m)}^{\infty} \mathrm{d}\hat{\rho} \int_{\rho_\mathrm{cutoff}(m)}^{\infty} \mathrm{d}\rho_\mathrm{true}\; p(\rho_\mathrm{obs}|\rho_\mathrm{true}) \\ &\quad \times \frac{(Km)^3}{\hat{\rho}^4}\delta(\hat{\rho} - \rho_\mathrm{true}) \\ &= \frac{4\pi K^3 m^3 p(m)}{V_\mathrm{U}}\int_{\rho_\mathrm{cutoff}(m)}^{\infty} \mathrm{d}\rho_\mathrm{true}\; \rho_\mathrm{true}^{-4}\; p(\rho_\mathrm{obs}|\rho_\mathrm{true}) \\ &\quad \propto m^3 p(m)\int_{\rho_\mathrm{cutoff}(m)}^{\infty} \mathrm{d}\rho_\mathrm{true}\; \rho_\mathrm{true}^{-4}\; p(\rho_\mathrm{obs}|\rho_\mathrm{true}), \end{aligned} \tag{14}$$

where V_U is the volume of our universe, and $\rho_\mathrm{cut-off}(m) = \frac{Km}{r_\mathrm{U}}$ is the lower SNR cut-off defined by a source of mass m being placed at the maximum allowed distance r_U. Thus, the SNR distribution for astrophysical sources is $p(\rho) \propto \rho^{-4}$, and we expect the observed mass distribution to be biased by a factor of m^3. The mass and SNR components of equation (14) are generally connected via the mass-dependent SNR cut-off. The term $p(\rho_\mathrm{obs}|\rho_\mathrm{true})$ accounts for the shift in search SNR relative to the expected value due to detector noise:

$$p(\rho_\mathrm{obs}|\rho_\mathrm{true}) = \chi_\mathrm{NC}(\rho_\mathrm{obs}; \lambda = \rho_\mathrm{true}, k = 2), \tag{15}$$

where χ_NC is the non-central chi distribution with a non-centrality of $\lambda = \rho_\mathrm{true}$ and $k = 2$ degrees of freedom.

For real binary merger events, we can pursue an analogous derivation, though the resulting relation differs as the SNR is a more complex function of event parameters than equation (12). Additional complications arise if the detector is sensitive to events

[4] If the data set already contains a number of detectable astrophysical signals, then the expected number of such GW events above threshold should be subtracted from the simulated event count. Alternatively, a data set that is, to a good approximation, empty of astrophysical signals may be used for the sensitivity estimate.

at cosmological distances, causing the observed masses to be redshifted by a distance-dependent amount.

The search and parameter-estimation analyses that produce our events can only cover a finite range of masses, of which we denote the limits as $m_{\rm min}, m_{\rm max}$. We will assume that all astrophysical foreground events have masses lying within these limits; in practice one should take sufficiently wide limits that the density of foreground events at these limits becomes vanishingly small.

For the mass distribution of the astrophysical foreground, we consider two types of population distribution: a truncated power law

$$p(m|\theta_{\rm F}) \equiv p(m|\alpha, m_{\rm low}, m_{\rm high}) \propto \begin{cases} m^{\alpha} & \text{if } m_{\rm low} < m < m_{\rm high} \\ 0 & \text{else} \end{cases} \quad (16)$$

with three free parameters, the slope α, lower mass cut-off $m_{\rm low}$, and high-mass cut-off $m_{\rm high}$. The two mass cut-offs are constrained by the mass range considered as $m_{\rm min} \leq m_{\rm low} < m_{\rm high} \leq m_{\rm max}$. The second population is a Gaussian

$$p(m|\theta_{\rm F}) = p(m|\mu, \sigma) \propto \begin{cases} \mathcal{N}(m; \mu, \sigma) & \text{if } m_{\rm min} < m < m_{\rm max} \\ 0 & \text{else} \end{cases} \quad (17)$$

with two free parameters, the mean μ and standard deviation σ. Strictly, this distribution is a truncated Gaussian, however, in practice we consider parameter ranges such that $p(m|\theta_{\rm F}) \ll 1$ at the boundaries. In contrast to the explicit differentiation between the true and observed SNR, Bayesian parameter estimation provides us with samples from the probability distribution of the true mass that are used directly as in equation (8), which eliminates the need to introduce a variable representing an observed mass.

In the background case, there are no selection effects, and we assume the noise characteristics are such that there is no correlation between the mass distribution and the SNR distribution. As a result, $p(\rho_i, \vec{m}_i|\theta_{\rm B}, \eta_i = {\rm B})$ decomposes as

$$p(\rho_i, \vec{m}_i|\theta_{\rm B}, \eta_i = {\rm B}) = p(\rho_i|\theta_{\rm B}, \eta_i = {\rm B}) p(\vec{m}_i|\theta_{\rm B}, \eta_i = {\rm B}). \quad (18)$$

Note that in realistic data the SNR distribution of background events may be strongly dependent on the mass (and other template parameters; Abbott et al. 2016h), so this decomposition is not necessarily valid.

The expected rate and distribution of background events caused by instrumental noise can, in practice, be measured to high precision using techniques such as time-shifted analyses (Usman et al. 2016; Capano et al. 2017; see also Cannon, Hanna & Keppel 2013). For our artificial universe, we have the freedom to choose the SNR and mass distributions, though this choice was informed by observed distributions in real data. We choose a power law with slope -12 in SNR; the mass posteriors are of constant width with their central values distributed uniformly between $m_{\rm min}$ and $m_{\rm max}$:

$$p(\rho|\eta = {\rm B}) \propto \rho^{-12}, \quad (19)$$

$$p(m|\eta = {\rm B}) \propto 1. \quad (20)$$

More realistic choices would include the effect of template bank density (Dent & Veitch 2014) and transient noise glitches (Nitz et al. 2017; Nitz 2018) on the distribution of noise triggers over mass space. Note that our inference of the foreground mass distribution is expected to become more precise the more distinct the foreground and background are, especially in SNR. Here, both SNR distributions are falling power laws, however, background drops off much more rapidly than foreground.

Finally, we combine the mass distribution with equations (14)–(15). Using (16) for the truncated power law, we obtain

$$p(\rho, m|\alpha, m_{\rm low}, m_{\rm high}, \eta = {\rm F}) \propto \int_{\rho_{\rm cutoff}(m)}^{\infty} {\rm d}\rho_{\rm true}\, \rho_{\rm true}^{-4}\, \chi_{\rm NC}(\rho; \lambda = \rho_{\rm true}, k = 2) \times \begin{cases} m^{\alpha+3} & \text{if } m_{\rm low} < m < m_{\rm high} \\ 0 & \text{else} \end{cases}. \quad (21)$$

Using (17) for the Gaussian,

$$p(\rho, m|\mu, \sigma, \eta = {\rm F}) \propto \int_{\rho_{\rm cutoff}(m)}^{\infty} {\rm d}\rho_{\rm true}\, \rho_{\rm true}^{-4}\, \chi_{\rm NC}(\rho; \lambda = \rho_{\rm true}, k = 2) \times \begin{cases} m^3 \mathcal{N}(m; \mu, \sigma) & \text{if } m_{\rm min} < m < m_{\rm max} \\ 0 & \text{else} \end{cases}.$$

The background model does not involve selection effects and yields

$$p(\rho, m|\eta = {\rm B}) \propto \rho^{-12}. \quad (23)$$

In general, normalizing these expressions requires an integral over ρ and m, which can be difficult or computationally expensive. In our model, this simplifies somewhat as the integrand $\rho_{\rm true}^{-4}\, \chi_{\rm NC}(\rho; \lambda = \rho_{\rm true}, k = 2)$ happens to assume values very close to zero for the $\rho_{\rm cut\text{-}off}$ values of [0.25,4] allowed by our prior mass range of [5,80] therefore we are able to approximate $\rho_{\rm cut\text{-}off} = 1$.

To generate the artificial data sets, we draw a total number of foreground and background events from a Poisson distribution around the true values determined by the intrinsic rate. Each of those events corresponds necessarily to a local maximum of signal likelihood over time, mass, and, in general, other parameters – we generally approximate this local maximum as a multivariate Gaussian distribution. For each foreground event, we then draw the true mass and distance, from which we can uniquely determine the intrinsic SNR. We then simulate the impact of noise on the measurement of both SNR and mass by drawing a value from a non-central chi distribution around the true SNR to obtain the observed SNR, and drawing the maximum likelihood value from a normal distribution around the true mass with a width as determined by equation (13). We use a uniform in mass prior therefore the posterior samples for the mass estimate are drawn from a Gaussian with the same width around the maximum likelihood value. For background events, the observed SNR is drawn directly from a power law as the background SNR distribution is determined empirically from the observed SNR values, while the posterior samples for the mass are drawn from a constant-width Gaussian around a central value drawn from a uniform distribution between $m_{\rm min}$ and $m_{\rm max}$ for each event.

Our method does not require us to make strong assumptions about the shape or width of the mass likelihood, however, it is important to generate these artificial results carefully as negligence can have unexpected consequences. In practice, we have found the scaling and width of the posteriors to have little effect on our results when the posteriors are of smaller scale than the population.

4 TOY MODEL RESULTS

We applied our method to a large number of realizations for each choice of foreground distribution, though the figures in the following section only show results for a single realization. The results across realizations will be given in text only. The

mass limits chosen for all toy model results[5] were $m_{\rm min} = 5$, $m_{\rm max} = 80$. The total expected number of events above an SNR of 8, the lowest threshold considered, is 1600, with 95 per cent contamination due to background events. The chosen slope of -12 for the background SNR distribution is less steep than in typical LIGO–Virgo analyses for stellar mass compact binary mergers; our choice exaggerates the transition region in which the chances of an event belonging to either foreground or background are comparable.

To simulate the limitations due to the computational costs of the analysis, we impose an SNR threshold on events, assessing its influence on our inferences by varying its value between 8 and 30. Most of these SNR values would typically be considered as sub-threshold since an SNR of $\approx$13.7 is required for an event to have a $p_{\rm astro}$ value of 50 per cent, and to reach $p_{\rm astro} = 99$ per cent an SNR of $\approx$24 is needed. These numbers are meaningful only in relation to this simulation. The actual relationship between SNR and $p_{\rm astro}$ varies between detection pipelines as they typically use additional information in their detection statistics to reduce the significance of background events. The number of detectors used in the network, as well as their sensitivities and the actual characteristics of the foreground and background distributions will also have an effect. A more realistic application is given in Section 5. Lastly, the free parameters in equation (12) and equation (13) are chosen such that an event of true mass $m = 30$ at a notional distance of 400 Mpc has a mass posterior with width $\sigma_{\rm PE} \approx 1$, and has an SNR of $\rho \approx 50$. The width of the mass posteriors of background events is set to the constant value of 3.2, typical of foreground events at the lowest SNR considered. In reality, the mass posterior distributions for background events are mass dependent and often irregular.

The priors chosen are flat in all hyper-parameters, with two exceptions: the width of Gaussian populations, where the prior was flat-in-log, and the expected number of astrophysical foreground events, where we used a Jeffreys prior:

$$\text{prior}(\sigma) \propto \frac{1}{\sigma}, \tag{24}$$

$$\text{prior}(N_{\rm F,exp}) \propto \frac{1}{\sqrt{N_{\rm F,exp}}}. \tag{25}$$

Parameter estimation was performed using the `emcee` (Foreman-Mackey et al. 2013) implementation of an Affine Invariant Markov chain Monte Carlo Ensemble sampler (Goodman & Weare 2010).

4.1 Power-law distribution

The first population considered was the truncated power law, which was inspired by the idea that black hole masses may be distributed analogously to the initial mass function of their progenitor stars. We add parameters $m_{\rm min}$ and $m_{\rm max}$ to define the lower and upper limits of the power-law distribution. This is motivated by the desire to determine whether there are gaps in the astrophysical black hole mass distribution: at the low end to compare with the apparent lower limit of black hole mass in X-ray binaries (Farr et al. 2011), and at the high end to determine the maximum mass above which a pair-instability supernova completely disrupts the star (Barkat, Rakavy & Sack 1967). In our simulation, we chose the power-law slope to be -2.4, and the cut-off values to be 12 and 64.

[5]Since we do not claim a specific link to astrophysics in the toy model, the mass units are arbitrary.

Our primary results, the estimates of model parameters and their correlations, are shown in Fig. 1. These results use an SNR threshold 8, the lowest value for which we run our analysis, as we would expect this to yield the best possible parameter estimates. Notable features are the large spread of possible merger rate densities (abbreviated as 'Rate') and their correlation with the lower mass cut-off. This is a consequence of the fact that the power-law slope is effectively increased by 3 due to selection effects, thus detected events are described by a positive slope. The detection bias towards high mass means that fewer events are available to constrain the lower cut-off value, and low-mass events that are observed tend to have lower SNR values. As the total rate is still dominated by low-mass (and low-amplitude) events, the large uncertainty of the low-mass cut-off yields a high uncertainty on the rate. The estimated fraction of foreground events in the sample of observed events couples linearly to the merger rate density, but is less significant than the lower mass cut-off.

To assess our method in the light of its main goal of avoiding bias while lowering the SNR threshold, a single analysis result is insufficient. Therefore, we analyse the same data with a range of different SNR thresholds to observe the change in the hyper-parameter estimates. Fig. 2 shows the marginalized posteriors for the rate and the three population parameters as a function of SNR threshold. We can observe the posteriors growing wider as the SNR threshold is increased and information from fewer events is considered. The result from one single realization is, however, not necessarily representative of the general behaviour. Combining the results from multiple realization shows there is no noticeable bias regardless of the threshold chosen, and estimates improve as the threshold is lowered. Between SNR thresholds of 8 and 24, the width of the 90 per cent credible intervals decreases on average by factors of 2.4 for the power-law slope, 1.3 and 1.6 for the lower and upper mass cut-offs, respectively, and 1.7 for the log of the inferred merger rate density.

Given the estimates of population parameters, we can also compute an estimate of the underlying mass distribution, which we show in Fig. 3. Here, we show the 50 per cent and 90 per cent confidence bands, as defined by computing the percentiles of $p(m|\theta)$ across all samples θ from the posterior for any given mass m. We observe that the true distribution is contained well within the credible interval and deviations are generally caused by an underestimated lower cut-off. In general, there is a trade-off between expanding the bounds of the mass distribution to include additional events, and shrinking it to increase the value of the probability density function (PDF) for highly significant events. The lower cut-off tends to have more freedom of movement as there are fewer high-SNR events at low mass to constrain it.

As a final result for this population, Fig. 4 shows the estimated probability of any given event to have an astrophysical origin $p_{\rm astro}$, and how it compares to an SNR-only estimate indicated by the black dash–dotted line. While this figure does not show quantitative results, we do observe that foreground events are largely located above the dash–dotted black line, indicating that our confidence in them being real has increased, while background events tend to be located below and are often on the $p_{\rm astro} = 0$ line when their masses are outside the hard cut-offs of the truncated power-law population. Thus, we find, as expected, that the discrimination between signal and noise populations is improved with the incorporation of information about their mass distributions (Dent & Veitch 2014). We determine that the percentage point difference between $p_{\rm astro}$ using our method and the SNR-only approach to be $\approx$2 per cent, although

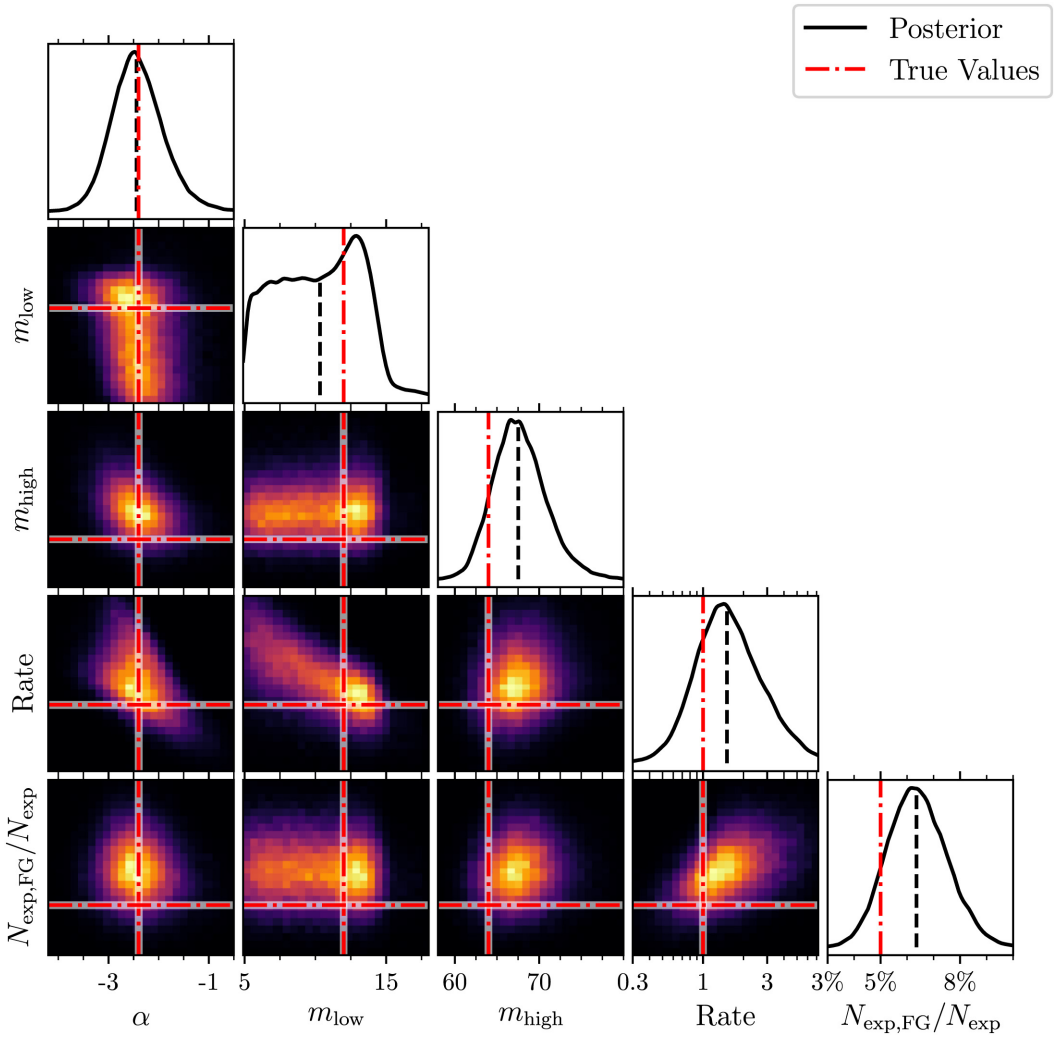

Figure 1. Parameter estimates for a single realization of the toy model described in Section 4.1. The foreground population model is a truncated power law with slope α, and cut-offs $m_{\rm min}$ and $m_{\rm max}$. The expected number of events above the SNR threshold of 8 are 1600, 5 per cent of which are expected to be foreground events. The black lines show the kernel density estimate of the posterior (solid) and its median value (dashed). The red dash–dotted line indicates the true value for the underlying population.

this includes events with tiny absolute shifts due to them being very close to either 0 per cent or 100 per cent in the first place.

4.2 Gaussian distribution

The second core population considered is a simple Gaussian with a very small width. This population was chosen to test the inference on hard-to-infer parameters and to see the effect a very distinctive distribution has on the discriminating power of our method. We chose the width to be very narrow with a standard deviation of 1.6 around a mean mass of 27. The population is narrower than the individual posteriors, which have a typical width of $\approx 2-3$. Therefore, we expect the population to require a relatively large number of events to resolve. How many events are needed to resolve these features generally depends on the population, in our case we find that $\approx 10-20$ events are needed to consistently constrain the population width to be smaller than individual posteriors. On the other hand, the discriminating power of using information from the mass estimates should be much greater than for a wide distribution, such as the truncated power law used in the previous section.

The parameter estimates for a single realization are shown in Fig. 5. We observe that true width of the population σ is contained comfortably within the inferred posterior, though the uncertainty is rather large. It is generally overestimated slightly. Similarly, the mean of the population is found well with an uncertainty comparable to the population width. The rate is constrained much better than in case of the truncated power law as this model lacks the degeneracy between the rate and a poorly constrained population parameter. The lack of a strong correlation between a population parameter and the merger rate density also highlights its linear relation to the estimated number of foreground events contained within the sample.

When lowering the SNR threshold from 24 down to 8, the sizes of the 90 per cent confidence intervals of the population parameters and merger rate density decrease by factors of 4.9, 2.0, 3.4 for the mean, the log of the width, and the log of the merger rate density, respectively. This is illustrated in Fig. 6 for one specific realization.

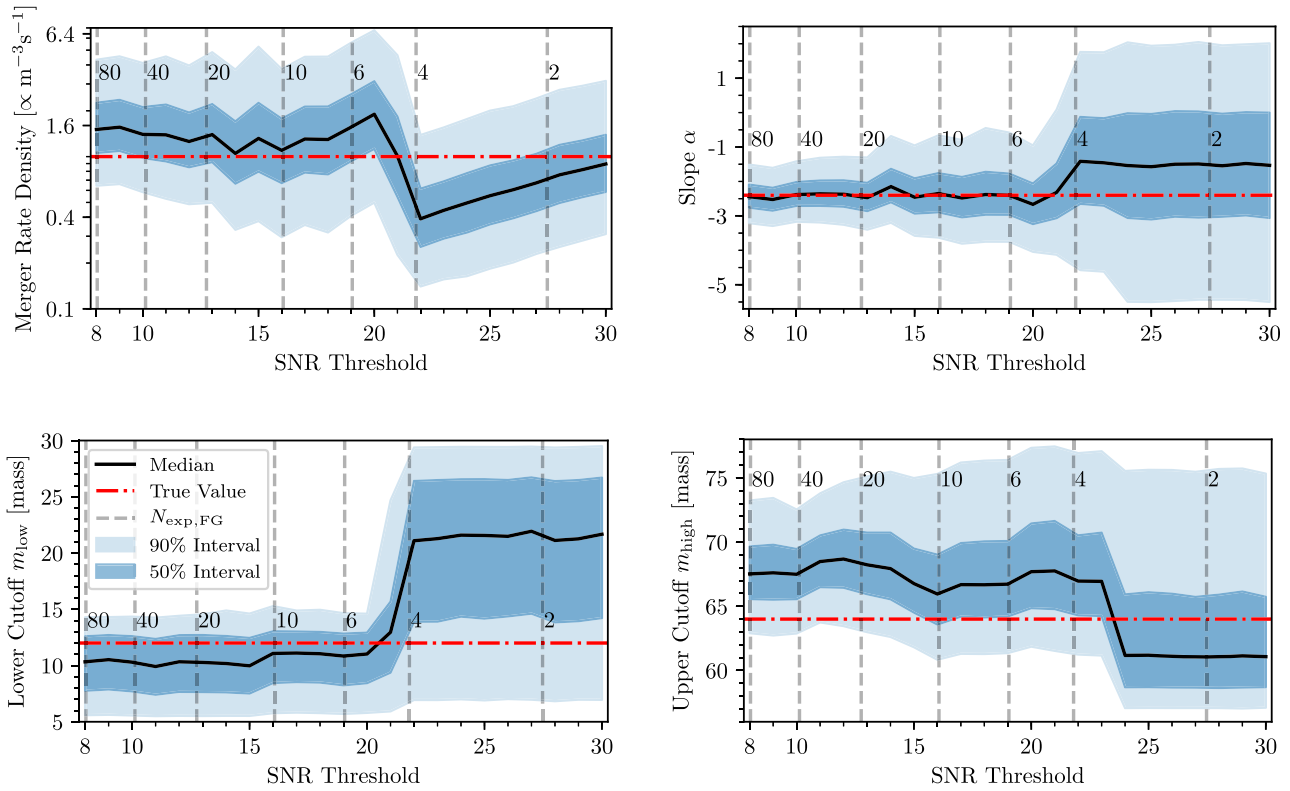

Figure 2. Confidence intervals for individual parameters of one realization of the truncated power-law model (see Section 4.1), as a function of SNR threshold. The parameters shown are the inferred astrophysical merger rate (upper left) and power-law slope (upper right), as well as the low (lower left) and high (lower right) mass cut-offs. The red dash–dotted line indicates the true value for the underlying population. The dashed grey lines indicate the expected number of foreground events at the given SNR threshold.

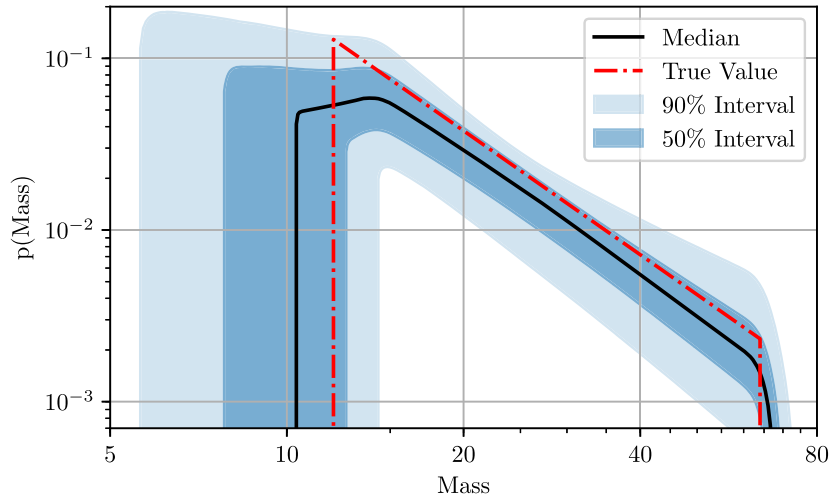

Figure 3. The inferred mass distribution of the foreground population using the truncated power-law model and simulated data (see Section 4.1). The bands indicate the given percentiles in the probability density at any given mass across all posterior samples.

The uncertainty on inferred population parameters grows rapidly as the number of events decreases. The most noticeable change is observed at SNR thresholds of 19, 20, and 21. This is caused by two events with SNRs between 19 and 20 whose removal decreases the number of events to 5, one of which is a low-mass outlier. This outlier has an SNR slightly above 20, and is therefore removed when the SNR threshold is set to 21.The true mass distribution is well within the confidence interval shown in Fig. 7, though the true distribution is somewhat more narrow than inferred as seen previously in Fig. 5.

The comparison of the estimated $p_{\rm astro}$ as shown in Fig. 8 shows how important the inclusion of masses is for this population. We can clearly identify the band of foreground events for which $1 - p_{\rm astro}$ is smaller by factors of a few up to 10 compared to the SNR-based estimate. In this specific realization, only 6 of 76 foreground events lost any $p_{\rm astro}$, across multiple simulations on

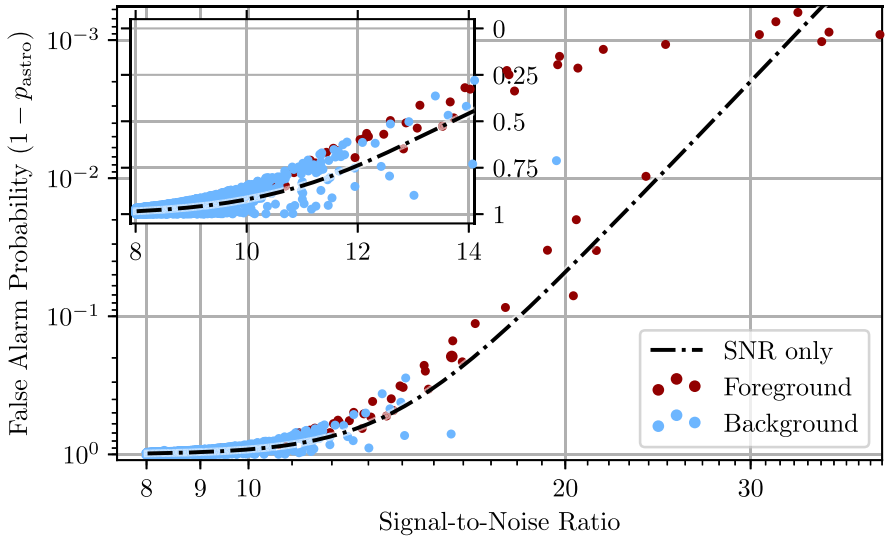

Figure 4. The probability that an event is caused by detector noise rather than being of astrophysical origin, $1 - p_{\mathrm{astro}}$, versus SNR. The foreground population model is a truncated power-law distribution. The blue and red dots represent foreground and background events, respectively. The dash–dotted line shows the probability that would be inferred by an SNR-based estimate, assuming the relative number of expected foreground and background events is known perfectly. The inset focuses on the region with background events and emphasizes events that are unlikely to be astrophysical using a linear scale.

average 97 per cent of foreground events saw an increase in p_{astro}. In the case of background events, ≈20 per cent saw an increase in their p_{astro} of up to 10 per cent, though most are demoted and often down to effectively 0.

4.3 Incorrect models – No background component

Previous analyses of GW populations (such as the power-law model used in Abbott et al. 2017) use a high threshold to ensure a high probability that the events used are of astrophysical origin, in effect neglecting the possibility of background. Here, we investigate the behaviour of our toy model with the background component disabled, corresponding to such a scenario. This shows the results one would obtain if simply fitting the foreground model to a contaminated data set. The underlying population is a truncated power law identical to the one used in the first set of results presented in Section 4.1.

The results are shown in Fig. 9, where we observe the inferred distribution to be very different from the true one when the lowest SNR threshold of 8 is used and the data set is 95 per cent polluted (left-hand panel). The mass cut-offs are extended to the edges of the prior ranges to incorporate noise events at those values. The confidence interval includes the true value as long as the SNR threshold is sufficiently high since the number of background events is negligible, but trends towards −3 as the threshold is lowered. This is expected since the background dominates the low-SNR region and has a power-law slope of 0 in mass. This slope corresponds to an actual slope of −3 when selection effects are considered. In the right-hand panel, we see the effect on the estimation of the power-law slope as the threshold is varied. Once the SNR threshold reaches ≈12, the statistical uncertainty of the slope becomes large enough that the systematic bias is not noticeable.

4.4 Incorrect models – Neglected selection effects

The second kind of error we considered was the neglect to properly account for the mass dependence of selection effects. In case of a power-law distribution this is trivial, as it simply adds +3 to the inferred value of the slope. Therefore, we chose a Gaussian as the population, and we increased the width to 9 to highlight the impact of selection effects on the inferred population. Fig. 10 shows that the selection effects effectively shift the distribution towards higher masses. This is a general feature as the m^3 term strongly favours high-mass events in the observed set of events. Depending on the population, this may also affect the width of the population, which happened to be a very minor effect in this case.

Together with the previous Section 4.3, this illustrates that population inference can be made impossible even when the model matches the underlying distribution. Accounting for the presence of noise and selection effects is essential for correct inference and to avoid bias when attempting to lower the SNR threshold.

5 ADVANCED LIGO ENGINEERING DATA SIMULATION

After the successful tests using the simple toy model, we also applied this method in a more realistic context, within an end-to-end analysis of simulated GW strain data where event candidates are identified by one of the compact binary detection pipelines currently used to search LIGO–Virgo data, `PyCBC` (Dal Canton et al. 2014; Abbott et al. 2016h; Usman et al. 2016; Nitz et al. 2018). We searched semirealistic simulated data from the fourth advanced LIGO engineering run (ER4): LIGO Hanford Observatory (LHO) noise data were simulated as Gaussian noise using the LIGO design sensitivity noise spectrum with an angle-averaged range of ∼1600 Mpc for a $30 + 30\,\mathrm{M}_{\odot}$ black hole coalescence signal (Abbott et al. 2018), while LIGO Livingston (LLO) data were derived from an instrumental channel monitoring the input laser power, recoloured to the same average target spectrum. The ER4 data contained a non-trivial population of high-amplitude noise transients, mostly arising from the LLO laser channel (LHO simulated data were also not entirely free of artefacts from data generation and transmission). However, these ‘glitches’ generally did not have similar morphology to binary merger signals, and

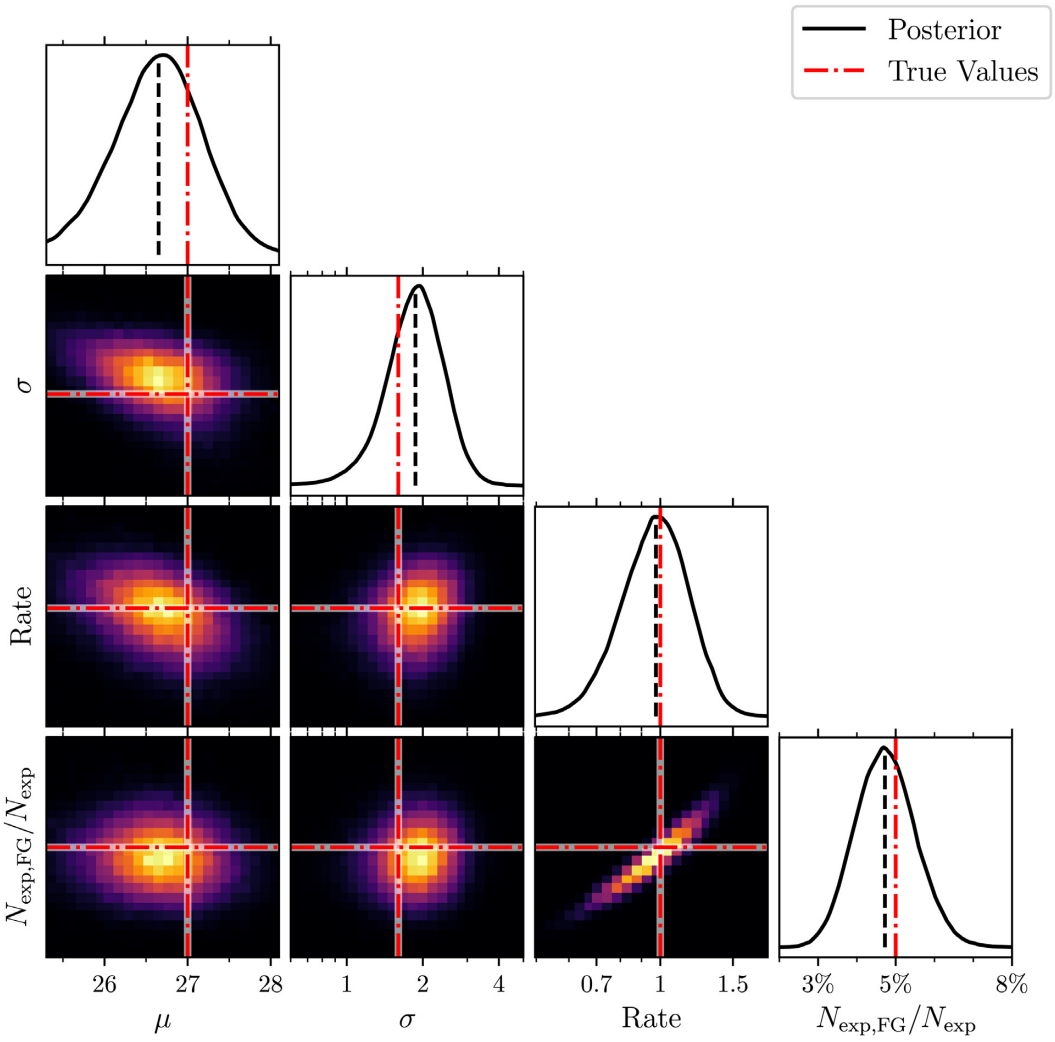

Figure 5. Parameter estimates for a single realization of the Gaussian mass distribution described in Section 4.2. The foreground population model is a Gaussian with two free parameters, the mean μ and width σ. The true expected number of events above an SNR threshold 8 is 1600, 5 per cent of which are expected to be foreground events. The black lines show the kernel density estimate of the posterior (solid) and its median value (dashed). The red dash–dotted lines indicate the true values for the underlying population.

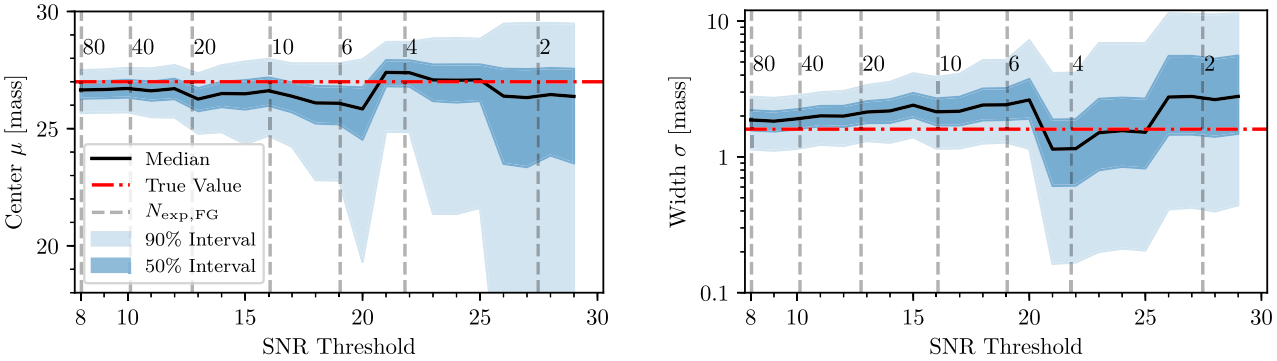

Figure 6. Confidence intervals for individual parameters of one realization of the Gaussian population (see Section 4.2), as a function of SNR threshold. The parameters shown are the inferred mean (left) and the width of the distribution (right). The red dash–dotted line indicates the true value for the underlying population. The dashed grey lines indicate the expected number of foreground events at the given SNR threshold. The plot terminates at an SNR threshold of 29 as there are no events with SNR 30 or higher in this realization.

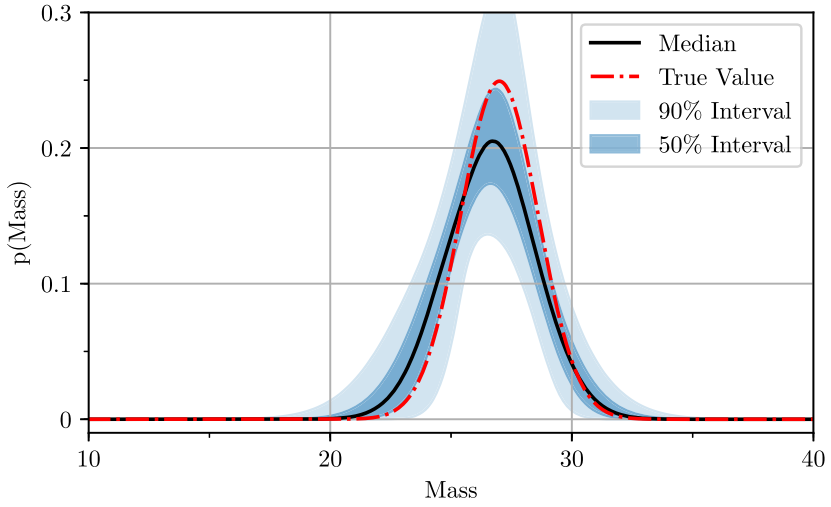

Figure 7. The inferred mass distribution of the foreground population using the Gaussian model and simulated data set (see Section 4.2). The bands indicate the given percentiles in the probability density at any given mass across all posterior samples.

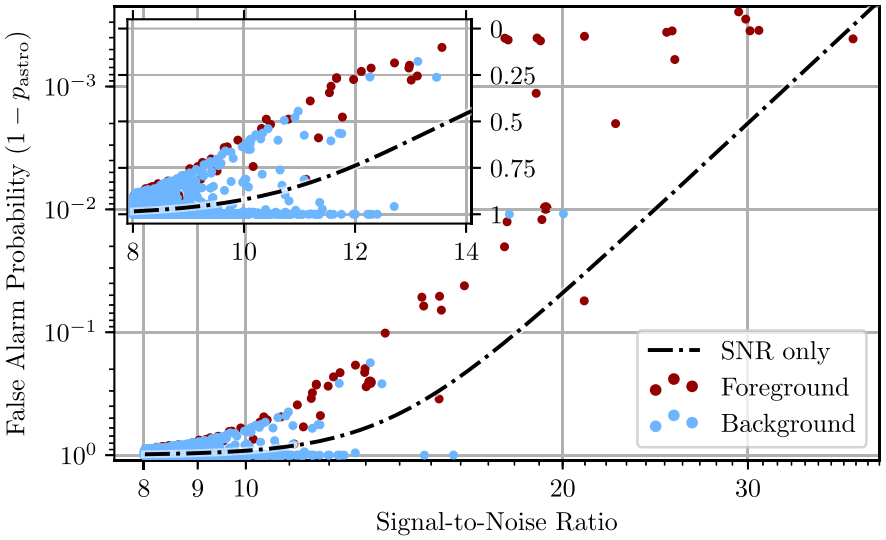

Figure 8. The probability of an event being caused by detector noise rather than being of astrophysical origin, $1 - p_{\mathrm{astro}}$, versus SNR. The foreground population model is a narrow Gaussian distribution (see Section 4.2). The blue and red dots represent foreground and background events, respectively. The dash–dotted line shows the probability that would be inferred by an SNR-based estimate, assuming the relative number of expected foreground and background events is known perfectly. The inset focuses on the region with background events and emphasizes events that are unlikely to be astrophysical using a linear scale.

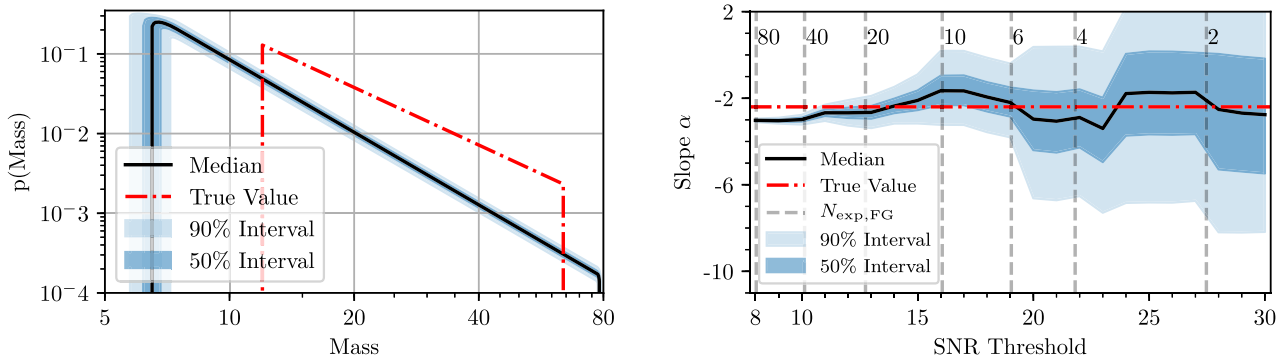

Figure 9. Results for the power-law model when we neglect to account for contamination due to noise, as described in Section 4.3. Left: Inferred mass distribution at the lowest SNR threshold of 8, the PDF percentiles are calculated across the population posterior at any given mass. Right: Inferred power-law slope as a function of SNR threshold. The vertical dashed lines indicate the expected number of true events above a particular SNR.

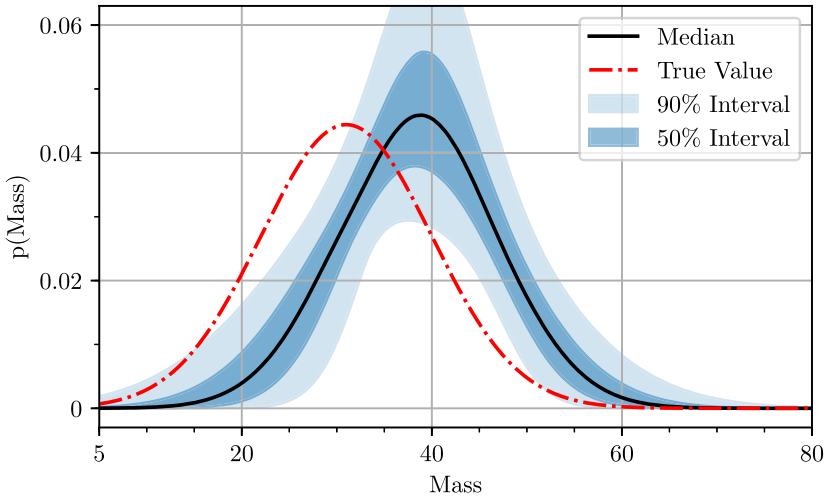

Figure 10. The inferred mass distribution for the Gaussian model without compensating for the mass dependence of selection effects (see Section 4.4), using the lowest SNR threshold of 8. The PDF percentiles are calculated across the population posterior at any given mass.

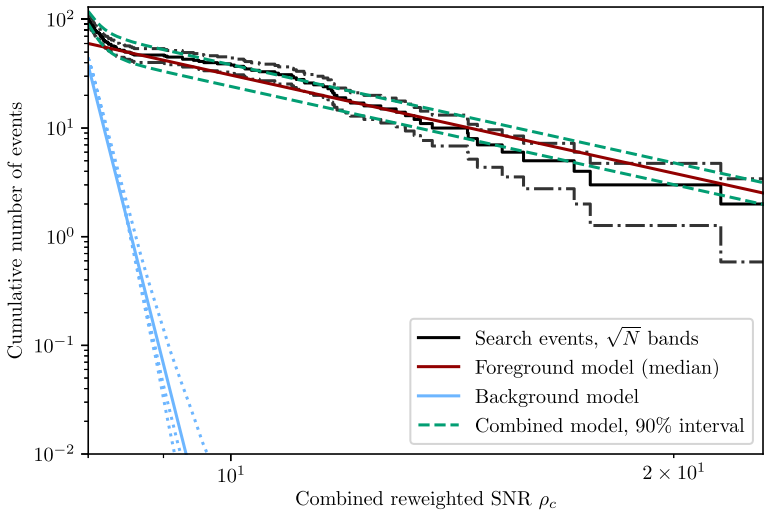

Figure 11. Cumulative number of (simulated) events detected in ≈37 d of LIGO ER4 engineering run data recolored to the 'ZDHP' design spectrum, versus threshold search-detection statistic ρ_c. The black steps indicate the search results, with $\pm\sqrt{N}$ bands indicating expected counting fluctuations. The dark red and light blue lines indicate power-law models of signal and noise distributions, respectively. The dotted light blue lines indicate empirical estimates of the noise distribution from each of three disjoint analysis periods, showing that the background model $p(\rho_c|\eta_i = \mathrm{B}) \propto \rho_c^{-54.8}$ is sufficiently accurate in the range of interest. The dark green dashed lines show the total expected number of events (signal+noise) as a 90 per cent credible band.

did not give rise to a long-tailed background distribution, such as those occurring for some ranges of candidate parameters in recent Advanced detector data (Abbott et al. 2016b,h; Nuttall 2018).

Simulated GW signals were added ('injected') to the noise data streams before they were stored and broadcast to the collaboration's computing grid. The injected population of binary black hole mergers was chosen to be uniform in both component masses between limits of 5 and 20 $\mathrm{M}_\odot$, and uniform in volume, with no cosmological (redshift) effects included. The `EOBNRv2HM` approximant tuned to numerical relativity (Pan et al. 2011) was used to simulate binary black hole mergers including non-dominant GW emission modes, for non-spinning binary components. The intended astrophysical rate corresponding to the injected merger signals was 5 $\mathrm{Gpc}^{-3}\mathrm{yr}^{-1}$.[6]

Our `PyCBC` search covered binary mergers of non-spinning components with masses between 3 and 50 $\mathrm{M}_\odot$; this range also defined the prior for parameter estimation performed on each event using `LALinference` (Veitch et al. 2015). The search-detection statistic for candidate events, ρ_c, is the quadrature sum of χ^2-reweighted SNRs $\hat{\rho}_{H,L}$ over single-detector events having consistent component masses and times of arrival between the two detectors (Abadie et al. 2011; Babak et al. 2013). The number of events we chose to analyse is limited by computational cost; we impose a threshold $\rho_c > 8$ leaving us with 100 events in ≈37 d of LHO–LLO coincident observing time; 51 of these events corre-

[6]Due to a software error the amplitude of injected signals was a factor 2 higher than intended, effectively simulating a true merger rate of 40 $\mathrm{Gpc}^{-3}\mathrm{yr}^{-1}$; however, in the results presented here, we rescale our rate estimates to compensate for this error.

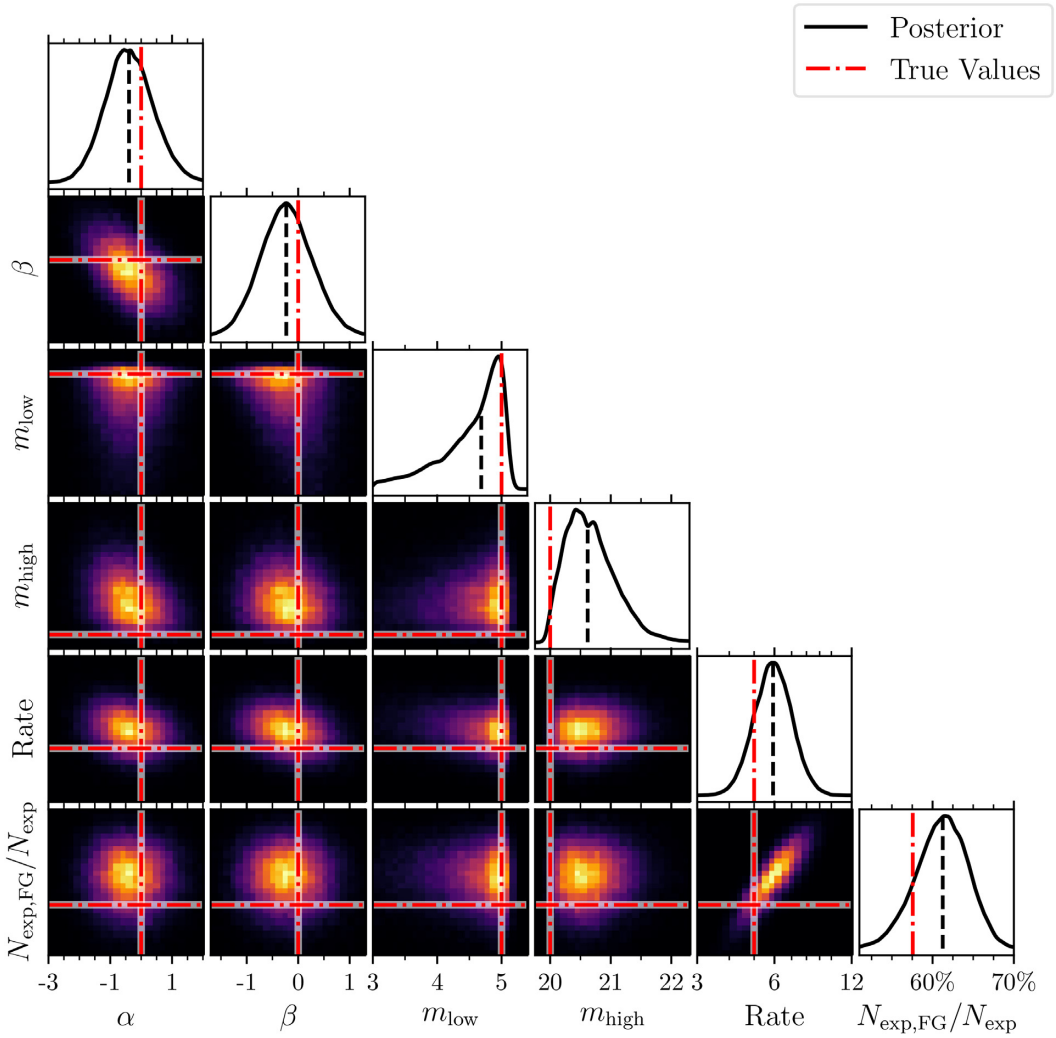

Figure 12. Parameter estimates for the ER4 data set as summarized in Fig. 11 and Section 5. The foreground population model is power law in both component masses with separate slopes α and β and shared cut-offs $m_{\rm low}$ and $m_{\rm high}$. The cut-off masses and merger rate density are given in units of $M_\odot$ and $\mathrm{Gpc^{-3}yr^{-1}}$, respectively. The black lines show the kernel density estimate of the posterior (solid) and its median value (dashed). The red dash–dotted line indicates the true value.

spond to known injected signals, with the remainder due to noise fluctuations.[7]

We first determine the rates of signal and noise events and the relative probabilities of signal versus noise origin for each event (Farr et al. 2015; Abbott et al. 2016a,g), given only the ρ_c value of each event, models of the signal, and noise event distributions over ρ_c, and an estimate of the total rate of noise events derived from time-shifted analyses (Babak et al. 2013; Nitz et al. 2018). The result of this estimate is summarized in Fig. 11.

We find 53 events with a signal probability $p_{\rm astro}$ above 50 per cent, of which 47 have $p_{\rm astro} > 90$ per cent. This analysis is comparable to those used to estimate the rate and $p_{\rm astro}$ for binary black hole mergers in the first Advanced LIGO Observing period (Abbott et al. 2016a), and does not use information about the mass distributions of signal or noise events, besides the assumption that the signal population is contained within the analysis mass limits.

[7]In reality, we will not have access to an independent record listing all true signals!

We now turn to our analysis, which estimates the rate and population model parameters simultaneously. The population model used here is a power law in each component mass, of the form

$$p(m_1, m_2|\theta_{\rm F}, \eta_i = {\rm F}) \propto \begin{cases} m_1^\alpha m_2^\beta & \text{if } m_{\rm low} < m_2 \leq m_1 < m_{\rm high}, \\ 0 & \text{else} \end{cases} \quad (26)$$

where α and β are the two power-law slopes, both with true values equal to 0. The mass cut-offs $m_{\rm low}$ and $m_{\rm high}$ are shared between both power laws, resulting in four free parameters in our population model. The selection effects were simulated numerically using the `LALsimulation` implementation of the IMRPhenomPv2 waveform (Hannam et al. 2014; Khan et al. 2016) to implement the method described by Finn & Chernoff (1993).

As the background model, we used a power-law fit to the distribution of ρ_c values from the `PyCBC` time-shifted analysis, giving a slope of ≈ -54.8 for $\rho_c > 8$. In a separate step, we constructed 2D fit to the distribution of component masses in the search results. We find empirically that the background distribution can be approximated as the product of a power law in chirp mass and an exponential distribution in mass ratio. Using the masses as determined by the search is not strictly the correct approach,

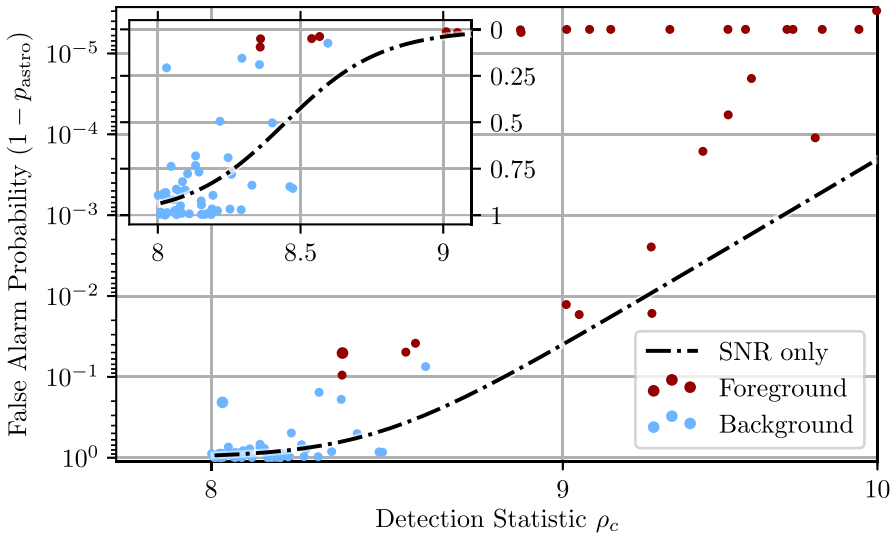

Figure 13. The probability of an event recovered from the ER4 data being caused by noise instead of corresponding to an astrophysical event, which is $1 - p_{\rm astro}$. The blue and red dots represent foreground and background events, respectively. The dash–dotted line shows the probability that would be inferred by an SNR-based estimate. The inset focuses on the region with background events and emphasizes events that are unlikely to be astrophysical using a linear scale.

which would be to run the time-shifted data through the same LALInference analysis as used for the zero-lag events: this was computationally infeasible. Therefore, the search results serve as a proxy for the optimal analysis.

We have found empirically that small changes to the background mass distribution do not have a strong effect on the result, which is expected given the dissimilarity of foreground and background distributions. (Note that real interferometer data containing chirp-like and blip-like features may be less forgiving in terms of the separability of foreground and background, as illustrated in Abbott et al. 2016b,h). We do not include any additional uncertainty on the background model rate or mass distribution.

The results of estimating the population parameters as displayed in Fig. 12 show that we successfully recover the true population parameters. The slopes are underestimated slightly that causes the inferred merger rate density to be elevated, although the true value is still encompassed. The mass cut-offs are found well with some tails to lower or higher masses for $m_{\rm low}$ and $m_{\rm high}$, respectively, since the uniform distribution of injections covers all regions of the mass range without major gaps, and the cut-offs are shared between both component masses.

Since the slope of the background SNR distribution is much larger than the slope of the astrophysical foreground, the transition region where events may belong to either source category with comparable probabilities is quite small. This means there are few events that fall in between the region of certain background and certain foreground, limiting the gains that can be made by our method in this case. Nevertheless, the fact that the foreground mass distribution is quite distinct from the background causes a significant increase in $p_{\rm astro}$ relative to the ρ_c-based approach, as can be seen in Fig. 13.

As with the previous ranking statistic based analysis, we can count the events found with a $p_{\rm astro}$ value above some given threshold. We find 56 events above a threshold of 50 per cent, which is three events more than before. The number of included events that we later identified as noise triggers rises from 4 to 5. When the threshold is set to $p_{\rm astro} > 90$ per cent, the number of events increases from 5 to 52, though it now includes one noise trigger. Thus, we find that applying our new method to realistic data not only reproduces the results obtained using established methods, but identifies additional foreground events. Since, in real Advanced detector data, the background distributions tend to be *less* steeply falling than those obtained here, we expect that the average (fractional) increase in number of signals found with high $p_{\rm astro}$ may be larger in the real-data case.

6 OUTLOOK

In this work, we have derived a new technique for simultaneous estimation of parameters defining the shape of two or more sub-populations and their expected contribution to the overall number of events. This technique allows us to extract information from formerly sub-threshold events without biasing the result due to uncertainties in classifying their origin. The method is agnostic to the specific choice of threshold, and lowering the threshold will improve the result by allowing information from events of an uncertain nature to be included.

Such improvements will eventually diminish in the regime where events have a high probability of background (noise) origin, though the point where further improvements become negligible depends on the specific characteristics of the two components. Fig. 1 of Farr et al. (2015) shows a simple example, using events for which only a single data value is measured, where improvements in foreground rate measurement become negligible in the background-dominated regime. In any case, digging deep into the background will not be detrimental; doing so can be especially worthwhile when the background model has some uncertainty, for instance following a specific functional form with partially unknown parameters. Additional events with low probability of astrophysical origin will reduce uncertainty on the background parameters, which in turn reduces the uncertainty on the foreground model.

The greatest gains over existing methods are found when there is a large number of events for which the source classification gives comparable probabilities for at least two categories, while

the distributions in secondary parameters are very distinct. This behaviour near the transition between populations is likely to be especially useful in the characterization of weak event populations, such as unresolvable binary mergers at cosmological distances. This is of particular interest when determining whether the source population evolves with redshift. Conversely, gains are expected to be small when the primary source classification is very potent and population models are uncertain; in this case our method converges to that with a single population.

While such thresholded analyses that ignore the possible presence of background cannot be guaranteed free of systematic bias, the expected size of bias can be bounded by considering the rate of background events above threshold, as well as the degree of divergence between foreground and background distributions over the parameters of interest. Controlling the bias of a thresholded analysis thus still requires accurate background estimation. In particular, for the small number of high-significance events thus far detected by LIGO–Virgo, possible biases on population inference due to neglecting background contamination are expected to be well below statistical errors.

Furthermore, we find that selection effects must be included in the analysis to avoid systematic error of the population parameters. Our example study illustrates that this is particularly important for the mass distribution of binary black hole mergers.

We have successfully tested this new model on different binary merger mass distributions in an artificial universe, as well as to synthetic LIGO data from an engineering run. This demonstrates the feasibility of applying this method to existing and future LIGO–Virgo observing runs, which should allow a better joint determination of source event rates and distributions. Challenges in application to real data over a broad signal parameter space include adequately modelling the complex distribution of noise events over binary masses and spins (Abbott et al. 2016b,h; Nitz et al. 2017). The method itself is, however, not limited to the realm of GW astronomy, and can be useful whenever a set of data points contains multiple populations.

ACKNOWLEDGEMENTS

We are grateful to the LIGO Scientific Collaboration for allowing the use of LIGO software engineering run data hosted on the Caltech CIT cluster, and access to LIGO Data Grid computing resources. We are grateful to Chris Pankow for useful discussions and assistance in the ER4 analysis. We would also like to thank Jonathan Gair and Richard O'Shaughnessy for their helpful comments and feedback. This work was supported by the Science and Technology Facilities Council (STFC) grants ST/K005014/2 and ST/M004090/2. TD acknowledges support from the Maria de Maeztu Unit of Excellence MDM-2016-0692.

REFERENCES

Abadie J. et al., 2011, Phys. Rev. D, 85, 082002
Abbott B. P. et al., 2016a, Phys. Rev. X, 6, 041015
Abbott B. P. et al., 2016b, Class. Quantum Gravity, 33, 134001
Abbott B. P. et al., 2016c, Phys. Rev. Lett., 116, 061102
Abbott B. P. et al., 2016d, Astrophys. J. Suppl., 227, 14
Abbott B. P. et al., 2016e, Astrophys. J. Lett., 818, L22
Abbott B. P. et al., 2016f, ApJ, 832, L21
Abbott B. P. et al., 2016g, ApJ, 833, L1
Abbott B. P. et al., 2016h, Phys. Rev., D93, 122003
Abbott B. P. et al., 2017, Phys. Rev. Lett., 119, 141101
Abbott B. P. et al., 2018, Living Rev. Relativ., 21, 3
Allen B., 2005, Phys. Rev. D, 71, 062001
Babak S. et al., 2013, Phys. Rev. D, 87, 024033
Barkat Z., Rakavy G., Sack N., 1967, Phys. Rev. Lett., 18, 379
Barrett J. W., Gaebel S. M., Neijssel C. J., Vigna- Gómez A., Stevenson S., Berry C. P. L., Farr W. M., Mandel I., 2018, MNRAS, 477, 4685
Belczynski K., Holz D. E., Bulik T., O'Shaughnessy R., 2016a, Nature, 534, 512
Belczynski K. et al., 2016b, Astron. Astrophys., 594, A97
Belczynski K. et al., 2017, preprint(arXiv:1706.07053)
Biswas R., Brady P. R., Creighton J. D. E., Fairhurst S., 2009, Class. Quant. Grav., 26, 175009
Campanelli M., Lousto C. O., Zlochower Y., 2006, Phys. Rev., D74, 041501
Cannon K. et al., 2012, ApJ, 748, 136
Cannon K., Hanna C., Keppel D., 2013, Phys. Rev. D, 88, 024025
Cannon K., Hanna C., Peoples J., 2015, preprint(arXiv:1504.04632)
Capano C., Dent T., Hanna C., Hendry M., Messenger C., Hu Y.-M., Veitch J., 2017, Phys. Rev. D, 96, 082002
Cutler C., Flanagan E. E., 1994, Phys. Rev., D49, 2658
Dal Canton T. et al., 2014, Phys. Rev., D90, 082004
Del Pozzo W., Li T. G. F., Agathos M., Van Den Broeck C., Vitale S., 2013, Phys. Rev. Lett., 111, 071101
Dent T., Veitch J., 2014, Phys. Rev. D, 89, 062002
Dominik M. et al., 2015, ApJ, 806, 263
Fairhurst S., Brady P., 2008, Class. Quant. Grav., 25, 105002
Farr W. M., Gair J. R., Mandel I., Cutler C., 2015, Phys. Rev. D, 91, 023005
Farr W. M., Sravan N., Cantrell A., Kreidberg L., Bailyn C. D., Mandel I., Kalogera V., 2011, ApJ, 741, 103
Farr W. M., Stevenson S., Miller M. C., Mandel I., Farr B., Vecchio A., 2017, Nature, 548, 426
Finn L. S., Chernoff D. F., 1993, Phys. Rev. D, 47, 2198
Fishbach M., Holz D. E., Farr B., 2017, ApJ, 840, L24
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J ., 2013, PASP, 125, 306
Gerosa D., Berti E., 2017, Phys. Rev., D95, 124046
Gerosa D., Kesden M., Berti E., O'Shaughnessy R., Sperhake U., 2013, Phys. Rev. D, 87, 104028
Goodman J., Weare J., 2010, Commun. Appl. Math. Comput. Sci., 5, 65
Hannam M., Schmidt P., Bohé A., Haegel L., Husa S., Ohme F., Pratten G., Pürrer M., 2014, Phys. Rev. Lett., 113, 151101
Kalogera V., Belczynski K., Kim C., O'Shaughnessy R., Willems B., 2007, Phys. Rep., 442, 75
Khan S., Husa S., Hannam M., Ohme F., Pürrer M., Forteza X. J., Bohé A., 2016, Phys. Rev. D, 93, 044007
Kovetz E. D., Cholis I., Breysse P. C., Kamionkowski M., 2017, Phys. Rev., D95, 103010
Li T. G. F. et al., 2012, Phys. Rev.D, 85, 082003
Mandel I., 2010, Phys. Rev., D81, 084029
Mandel I., Farmer A., 2018, preprint(arXiv:e-prints)
Mandel I., O'Shaughnessy R., 2010, Class. Quantum Gravity, 27, 114007
Mandel I., Farr W. M., Colonna A., Stevenson S., Tiňo P., Veitch J., 2017, MNRAS, 465, 3254
Messenger C., Veitch J., 2013, New J. Phys., 15, 053027
Messick C. et al., 2017, Phys. Rev. D, 95, 042001
Ng K. K. Y., Vitale S., Zimmerman A., Chatziioannou K., Gerosa D., Haster C.-J., 2018, Phys. Rev. D, 98, 83007
Nitz A. et al., 2018, gwastro/pycbc: Post-O2 release 12C. Available at:https://doi.org/10.5281/zenodo.1313589
Nitz A. H., 2018, Class. Quantum Gravity, 35, 035016
Nitz A. H., Dent T., Dal Canton T., Fairhurst S., Brown D. A., 2017, ApJ, 849, 118
Nuttall L. K., 2018, Phil. Trans. R. Soc. A, 376, 20170286

O'Shaughnessy R., Vaishnav B., Healy J., Shoemaker D., 2010, Phys. Rev. D, 82, 104006
O'Shaughnessy R., Gerosa D., Wysocki D., 2017, Phys. Rev. Lett., 119, 011101
Pan Y., Buonanno A., Boyle M., Buchman L. T., Kidder L. E., Pfeiffer H. P., Scheel M. A., 2011, Phys. Rev. D, 84, 124052
Smith R., Thrane E., 2018, Phys. Rev. X, 8, 021019
Spera M., Mapelli M., Bressan A., 2015, MNRAS, 451, 4086
Stevenson S., Ohme F., Fairhurst S., 2015, ApJ, 810, 58
Stevenson S., Berry C. P. L., Mandel I., 2017, MNRAS, 471, 2801
Talbot C., Thrane E., 2017, Phys. Rev. D, 96, 023012
Talbot C., Thrane E., 2018, ApJ, 856, 173
Taylor S. R., Gerosa D., 2018, Phys. Rev. D, 98, 83017
Tiwari V., Peters W. K., Jonas D. M., 2017, J. Chem. Phys., 147, 154308
Tiwari V., Fairhurst S., Hannam M., 2018, ApJ, 868, 140
Usman S. A. et al., 2016, Class. Quantum Gravity, 33, 215004
Veitch J. et al., 2015, Phys. Rev. D, 91, 042003
Vitale S., Lynch R., Sturani R., Graff P., 2017, Class. Quantum Gravity, 34, 3LT01
Wysocki D., Lange J., O'Shaughnessy R., 2018a, preprint(arXiv:e-prints)
Wysocki D., Gerosa D., O'Shaughnessy R., Belczynski K., Gladysz W., BertiBerti E., Kesden M., Holz D. E., 2018b, Phys. Rev. D, 97, 043014
Zevin M., Pankow C., Rodriguez C. L., Sampson L., Chase E., Kalogera V., Rasio F. A., 2017, ApJ, 846, 82

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