Discriminating between different scenarios for the formation and evolution of massive black holes with LISA

Alexandre Toubiana,[1,2] Kaze W.K. Wong,[3] Stanislav Babak,[1,4] Enrico Barausse,[5,6]
Emanuele Berti,[3] Jonathan R. Gair,[7,8] Sylvain Marsat,[1] and Stephen R. Taylor[9]

[1]*Université de Paris, CNRS, Astroparticule et Cosmologie, F-75006 Paris, France*
[2]*Institut d'Astrophysique de Paris, CNRS & Sorbonne Universités, UMR 7095, 98 bis bd Arago, 75014 Paris, France*
[3]*Department of Physics and Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland 21218, USA*
[4]*Moscow Institute of Physics and Technology, Dolgoprudny, Moscow region, Russia*
[5]*SISSA, Via Bonomea 265, 34136 Trieste, Italy & INFN, Sezione di Trieste*
[6]*IFPU - Institute for Fundamental Physics of the Universe, Via Beirut 2, 34014 Trieste, Italy*
[7]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am M¨uhlenberg 1, Potsdam-Golm, 14476, Germany*
[8]*School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom*
[9]*Department of Physics & Astronomy, Vanderbilt University, 2301 Vanderbilt Place, Nashville, Tennessee 37235, USA*

Electromagnetic observations have provided strong evidence for the existence of massive black holes in the center of galaxies, but their origin is still poorly known. Different scenarios for the formation and evolution of massive black holes lead to different predictions for their properties and merger rates. LISA observations of coalescing massive black hole binaries could be used to reverse engineer the problem and shed light on these mechanisms. In this paper, we introduce a pipeline based on hierarchical Bayesian inference to infer the mixing fraction between different theoretical models by comparing them to LISA observations of massive black hole mergers. By testing this pipeline against simulated LISA data, we show that it allows us to accurately infer the properties of the massive black hole population as long as our theoretical models provide a reliable description of the Universe. We also show that measurement errors, including both instrumental noise and weak lensing errors, have little impact on the inference.

I. INTRODUCTION

The detection of gravitational waves in the 10-1000 Hz band over the last six years by the LIGO/Virgo collaboration [1–3] has allowed us to infer for the first time the population of stellar-mass black hole (BH) binaries in the Universe [4, 5], shedding some light on their possible formation channels (see e.g. [6–22]). Scheduled for 2034, the Laser Interferometer Space Antenna (LISA) [23] will be sensitive to gravitational waves in the mHz band, and will reveal a virtually unexplored population of compact binaries. Some of the anticipated sources include Galactic binaries, which will be so numerous that they will form a stochastic foreground dominating over instrumental noise but should also include $\approx 10^4$ individually resolvable binaries [24, 25], and massive black hole binaries (MBHBs) with total mass in the range 10^4-$10^9\ M_\odot$ [26–31].

Electromagnetic observations indicate that massive BHs (MBHs) are present in the centers of most galaxies in the local universe [32–36], including our own Galaxy [37–40] and M87 [41], and that their properties are correlated with those of their host galaxies, suggesting a synergistic growth [33, 42–45]. Unfortunately, these observations are sensitive only to active MBHs up to $z \sim 7$ (cf. e.g. [46]), or local ones for which we can observe the gas/stellar dynamics. Gravitational waves will allow us to probe much more distant MBHs: LISA will be capable of detecting MBHBs up to $z \sim 20$, provided that they exist at such high redshift [23]. In this paper, we address the question of how these observations can help constrain scenarios for the formation and subsequent evolution of MBHs.

The population of MBHBs that LISA will observe is the result of a complex evolutionary path, whose details are still largely unknown. Two open issues, of particular importance for LISA, can be highlighted. First, which astrophysical mechanisms provided the seeds that grew into MBHs? Several scenarios have been proposed, suggesting seed masses ranging from 10^2 to $10^5\ M_\odot$, forming at $z \sim 15 - 20$ (see e.g. [47] for a review). Once these intermediate mass BHs form, they are thought to grow via gas accretion and successive mergers. Following the merger of two galaxies hosting a BH at their center, dynamical friction drives the BHs to the center of the newly formed galaxy, where they may form a bound binary system [48] (see however Ref. [49] for the possibility that a significant fraction of galaxy mergers may never produce a bound MBHB). If this happens (at $\sim$ pc separation for systems of $\sim 10^8 M_\odot$), dynamical friction becomes inefficient and other processes take over to control the binary's evolution, including three body interactions with stars (stellar hardening)[50, 51], gas-driven migration [52–58] or interactions with other MBHs [29, 30, 59]. The efficiency of these processes is uncertain, but they are crucial because it is not until $\sim 10^{-2}$ pc separations that gravitational wave emission is sufficient to make the binary coalesce within a Hubble time. Whether MBHBs can transition efficiently from pc to sub-pc separation is therefore still uncertain, which is usually referred to in the literature as the "last parsec problem" [48]. The physics of BH seeding at high redshift and the last parsec problem significantly affect the properties of the population of events that LISA will observe, such as the component masses

and spins, the redshift, and the rates themselves. Thus, by accumulating observations with LISA, one can in principle reverse engineer the problem, and shed light on these mechanisms.

We focus here on the ability of LISA to distinguish between different seeding scenarios. We improve upon Refs. [27, 60] in a number of ways. We use a more refined treatment of selection effects; we use updated astrophysical models, with improved treatment of the baryonic physics, of the formation of MBH pairs, of the hardening of MBHBs and of the effect of SN winds and accretion on MBH evolution; and we use more realistic assumptions about the LISA data, including an up to date model of the LISA instrument, and more realistic models for the gravitational waveforms generated by merging MBHs. We use the predictions of the semianalytic model of Ref. [61] (with updates described in Refs. [29, 30, 62–64]) for the evolution of galaxies and MBHs to simulate LISA data. This model has light seed (LS) and heavy seed (HS) variants, differing in the prescription for the initial masses of BHs. We consider the possibility that the population of MBHs is described by a mixture between the LS and HS scenarios. We treat the mixing fraction between models as a *hyperparameter* controlling the population, and estimate it from simulated datasets using a hierarchical Bayesian framework. We test the robustness of our analysis by using the predictions of different semianalytic simulations to generate data, and assess the impact of measurement errors (due to detector noise and weak lensing) on our inference of the MBHB population.

This paper is organized as follows. In Sec. II we explain how LISA data is simulated and how we perform parameter estimation. Sec. III describes the astrophysical models used for the population of MBHs and our mixing procedure. In Secs. IV and V we review the main aspects of the hierarchical Bayesian analysis and how to combine it with results from numerical simulations. We present our main results in Sec. VI and our conclusions in Sec. VII.

II. DATA SIMULATION AND PARAMETER ESTIMATION

LISA will observe the last stages of the coalescence of MBHBs, where higher harmonics can be comparable in amplitude to the $(2, \pm 2)$ harmonics [65–69]. Therefore, we use the phenomenological approximant PhenomHM [70] to generate the signal and perform parameter estimation. In this work we consider, for simplicity, quasicircular binaries with component spins aligned or antialigned with the orbital angular momentum (we comment on this in Sec. III). We compute the full LISA response and parametrize MBHBs as described in [69, 71]. Denoting by m_1 and χ_1 the mass and spin of the heaviest BH in a binary and by m_2 and χ_2 those of its companion, we define the chirp mass as $\mathcal{M}_c = (m_1 m_2)^{3/5}/(m_1+m_2)^{1/5}$, the mass ratio as $q = m_1/m_2 \geq 1$ and the symmetric mass ratio as $\eta = q/(1+q)^2$. We also introduce the effective spin χ_+ and the corresponding antisymmetric combination χ_-, defined as $\chi_{+,-} = (m_1\chi_1 \pm m_2\chi_2)/(m_1+m_2)$. We adopt the cosmological parameters reported by the Planck mission (2018) [72] to compute the luminosity distance D_L from the cosmological redshift z. Recall that source-frame (subscript s) and detector-frame (subscript d) masses are related via $m_d = (1+z)m_s$. We use the SciRDv1 noise curve [73], including the confusion noise due to Galactic binaries [74], and assume a low-frequency cutoff of 10^{-5} Hz in the LISA noise power spectral density. We assume a mission duration of four to ten years and an ideal 100% duty cycle.

For our purposes we will not need state-of-the-art MBHB parameter estimation, but just realistic error estimates for the intrinsic parameters of the source and for the luminosity distance. Therefore, we work in the zero-noise approximation [75] and simply compute the Fisher information matrix [76] to obtain the errors on source parameters, and more specifically we use the extended Fisher formalism of Ref. [77]. A more complete parameter estimation study is in preparation. As shown in Fig. 1, astrophysical models predict some events with large mass ratios and/or large spins, far outside the range of validity of current waveform models. Again, for simplicity, we will use PhenomHM for our calculations.

The chirp mass is the best measured parameter, and because we can observe the late inspiral and the merger-ringdown with high signal-to-noise ratio (SNR) up to thousands, we can measure the mass ratio and the spin of the primary quite accurately. For the heaviest systems, we can also measure the spin of the secondary. As for distance measurements, the error due to weak lensing dominates over the statistical error at high redshifts. We use the (pessimistic) model of [78], which estimates that the error due to lensing goes as

$$\frac{\sigma_{D_L,\text{lensing}}}{D_L} = 0.066\left[\frac{1-(1+z)^{-0.25}}{0.25}\right]^{1.8}. \tag{2.1}$$

We include this error by convolving the measured LISA posterior distribution with a Gaussian of width $\sigma_{D_L,\text{lensing}}$. The error due to weak lensing propagates into the determination of source-frame masses.

III. MASSIVE BLACK HOLES CATALOGUES

A. Semianalytic models

To describe the expected population of MBHBs detectable by LISA, we utilize the semianalytic galaxy formation model of Ref. [61], with updates described in Refs. [29, 30, 62–64]. Our model relies on dark matter halo merger trees produced with an extended Press-Schechter formalism [79], modified to reproduce the results of N-body simulations [80]. Baryonic structures contained in the halos are evolved along the branches and through the nodes of these merger trees. These structures include: a diffuse intergalactic medium with primordial metallicity, which accretes onto the halos either by getting shock-heated to the halo virial temperature (in large low-redshift systems) or along cold flows (at high redshift and/or small systems) [81–83]; a cold interstellar medium where star formation takes place, and which we assume to be in the form of disks and/or bulges; stellar disks and bulges; and nuclear compact configurations, i.e. nuclear star clusters

and MBHs. The latter, which are obviously of crucial importance for this work, are assumed to grow from high-redshift seeds by accretion – thus shining as quasars and active galactic nuclei (AGNs) – and coalescences. The model also accounts for AGN feedback (i.e., the effect of AGN jets, disk winds and radiation) and supernova feedback (i.e., supernova explosions). Both processes can affect the evolution of baryonic structures, quenching star formation (mainly in large and small systems, respectively), ejecting/heating up nuclear gas, and also suppressing accretion onto MBHs. In order to minimize the uncertainties, the model is calibrated to a number of observations at both galactic and nuclear scales [30, 61–64, 84, 85]. Nevertheless, as already mentioned, the predictions for LISA are crucially dependent on the assumptions made about two poorly understood processes: the formation of the high-redshift seeds and the "delays" with which MBHs come together and eventually coalesce after a galaxy merger.

As our fiducial astrophysical scenario, we adopt *Model-delayed* of [29], of which we consider two variants, with either LSs or HSs. In the LS model, MBHs grow from the remnants of Pop III stars at $z \gtrsim 15$ [86]. We seed large halos collapsing from the 3.5σ peaks of the primordial density field, and to describe the Pop III stellar mass function we use a log-normal distribution centered at $300M_\odot$ and with rms of 0.2 dex (with an exclusion region between 140 and 260 $M_\odot$ to account for pair instability supernova explosions). The mass of the seed MBH is then assumed to be $\sim 2/3$ of the initial Pop III star mass, to account for the mass loss during the supernova explosion. In the HS model, MBHs form instead with masses already $\sim 10^5\ M_\odot$. In more detail, we use the model of Ref. [87], in which seeds form from the collapse of protogalactic disks as a result of bar instabilities, at $z \gtrsim 15$ and in halos with spin parameter and virial temperature below critical threshold values. The latter are given by Eq. (4) – with $Q_c = 2.5$ – and Eq. (5) of Ref. [87], and we use Eq. (3) of the same work to set the seed mass. As for the delays between galaxy/halo and BH mergers, Ref. [29] accounts for the dynamical friction between the dark matter halos (including the effect of tidal disruption and evaporation); for the timescales associated (on much smaller $\sim$ pc scales) to stellar hardening[1], gas-induced migration and interactions with additional MBHs (brought in by later galaxy mergers); and finally for the gravitational-wave driven evolution timescale at sub-pc separations. The timescale associated to the binary's evolution at ∼kpc separations is instead neglected in Refs. [29, 61], on the premise that it should be negligible when compared to the other timescales involved. Recently, however, large scale cosmological simulations have challenged this notion [49], i.e. they have found that evolution timescales on those large separations can be significant. This prompted Ref. [30] to include an additional timescale in the semianalytic model of Refs. [29, 61] to account for the binary's evolution at ∼kpc separations. Moreover, Ref. [30] also modified the supernova feedback model of Refs. [29, 61] to account for the possibility that supernova winds may quench not only star formation, but also accretion onto MBHs in low-mass, high-redshift galaxies [88]. We implement this effect by assuming that the growth of the gas reservoir off which the MBH accretes is curtailed in systems with escape velocity (from the bulge) lower than 270 km/s [88]. We refer to the model including these additional ingredients (delays on scales of hundreds of pc and SN feedback on BH accretion) as *SN-delays*, adopting the same designation as in Refs. [30, 31].

We use the semianalytic model to produce simulated populations of MBHBs, including information on their masses, spins and redshift. It is worth noting that the eccentricity of a binary and the degree of alignment of the component spins depend on the mechanism that triggers the merger. For instance, triple/quadruple interactions between MBHs can lead to large eccentricities as a result of Kozai-Lidov resonances [89, 90] and/or chaotic interactions [91, 92]. Binaries merging in a gas-rich environment tend to have aligned spins, because of the Bardeen-Petterson effect [93, 94], i.e. the gravitomagnetic torques exerted by the circumbinary disk. We also stress that the evolution of the spin under accretion is described in our model by neither coherent nor chaotic accretion, but by the *hybrid* model of Ref. [62]. The latter incorporates Bardeen-Petterson torques, is intermediate between chaotic and coherent accretion, and reproduces the sample of spin measurements from iron Kα lines.

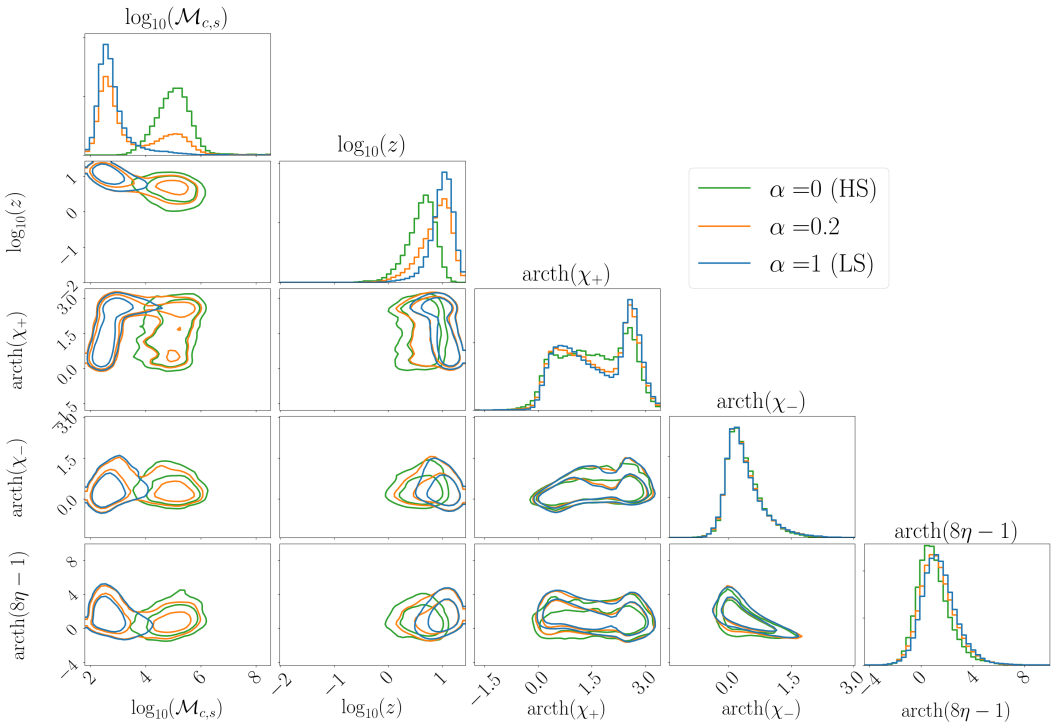
These effects are included in our semianalytic model (cf. in particular Refs. [29, 61, 62]), with the final remnant mass and spin produced by the MBH merger computed via fitting formulas reproducing the results of numerical-relativity simulations [95, 96]. However, the information on spin alignment and eccentricity is not fully exploited in the analysis performed for this paper. Indeed, because PhenomHM covers only quasicircular binaries with component spins aligned or antialigned with the orbital angular momentum, we simply take the projection of spins along the orbital angular momentum and neglect the eccentricity. Nevertheless, the information on the spin alignment is partially contained in the effective spin of the binary. To complete the set of parameters θ needed to describe LISA events, we draw the sky location uniformly on the sphere, the phase at coalescence and the polarization uniformly in $[0, 2\pi]$, and the inclination angle $\cos\iota$ uniformly in $[-1, 1]$. We assume a time to coalescence of at most one year, and we do not consider the part of the signal below 10^{-5} Hz.

B. Population properties

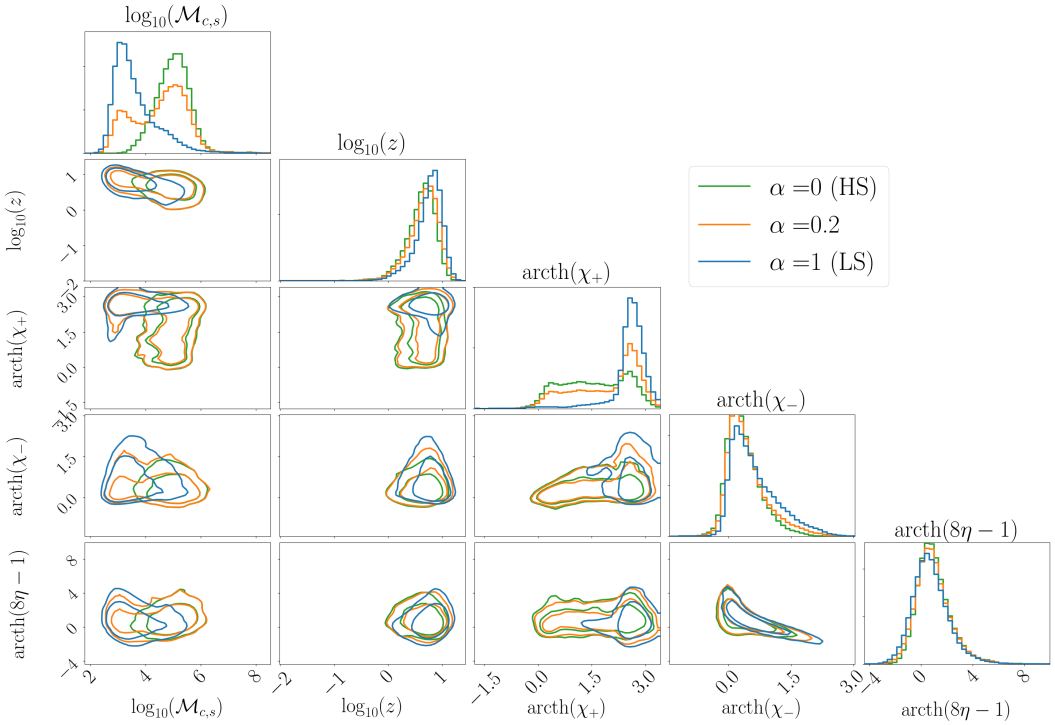
When running the simulations, we use only one of the seeding prescriptions. However, the population of MBHs in the Universe is unlikely to be described by any of these "pure" models, but rather by a mixture of models. Following [27], we introduce a mixing fraction α between the LS and HS scenario and define the full (unnormalized) MBHB population distribution to be

$$N_{\rm pop}(\theta|\alpha) = \alpha N_{\rm pop}(\theta|{\rm LS}) + (1-\alpha) N_{\rm pop}(\theta|{\rm HS}). \tag{3.1}$$

[1] As suggested by N-body simulations [51], the stellar hardening timescales are computed from the density at the mass influence radius of the binary, i.e. the radius at which the enclosed stellar mass is twice the binary mass.



(a) Without SNR threshold.



(b) With an SNR threshold of 10.

FIG. 1. Normalized population distribution for different values of the mixing fraction between the fiducial LS and HS models. We show the 68% and 90% confidence intervals. The (source-frame) chirp mass distribution is the most sensitive to α. The redshift distributions of detectable events look much more similar, unlike the effective spin distributions, as discussed in the main text.

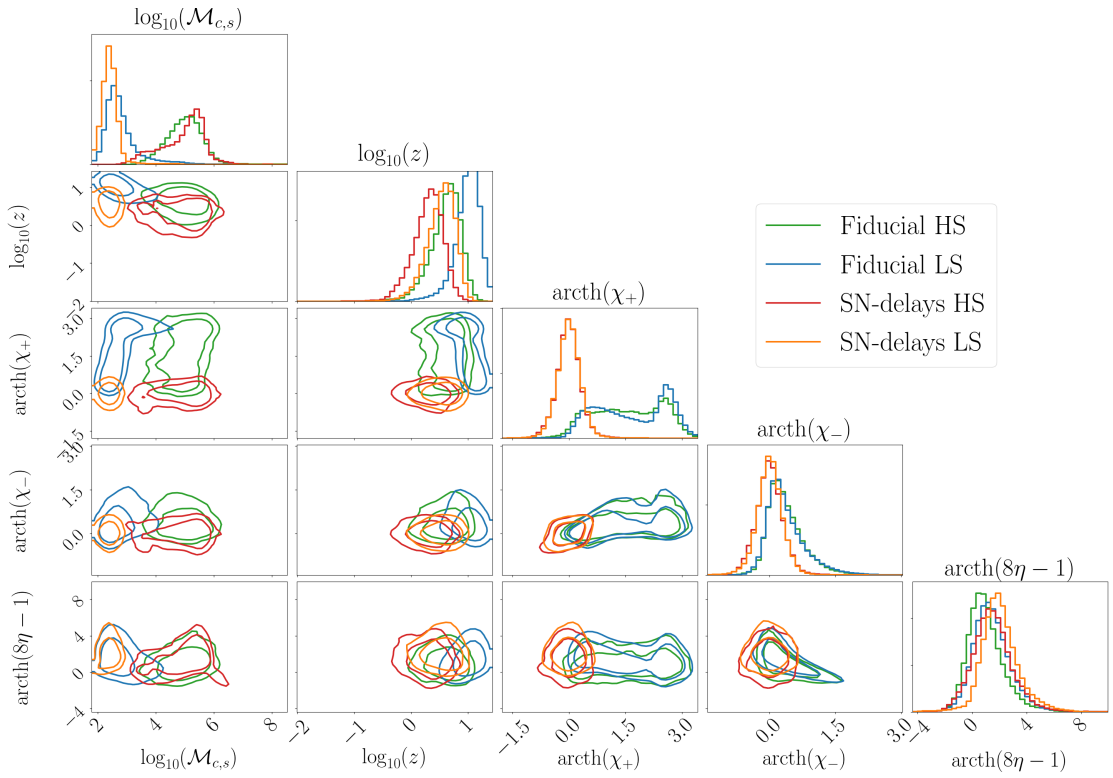


FIG. 2. Normalized population distributions predicted by our fiducial model and the SN-delays model, both in the LS and HS scenarios. We show the 68% and 90% confidence intervals. While the chirp mass distributions in the two models are quite similar, the redshift and effective spin distributions are not.

In the following, we will denote the normalized population distribution by $p_{\rm pop}(\theta|\alpha)$ and the predicted rate by $R_{\rm ev}$ (in yr^{-1}), such that $N_{\rm pop}(\theta|\alpha) = R_{\rm ev}(\alpha)p_{\rm pop}(\theta|\alpha)$, with similar definitions for the LS and HS models. The rate for a given value of the mixing fraction is

$$R_{\rm ev}(\alpha) = \int N_{\rm pop}(\theta|\alpha){\rm d}\theta = \alpha R_{\rm ev}(LS)+(1-\alpha)R_{\rm ev}(HS), \quad (3.2)$$

where $R_{\rm ev}(LS) = \int N_{\rm pop}(\theta|\alpha){\rm d}\theta$ is the rate for the LS model, and similarly for HS.

For a given SNR threshold, we denote by $R_{\rm det}(\alpha, {\rm SNR})$ the number of events (per year) above this threshold. In Table I we provide the annual rates for the LS and HS scenarios[2], as well as the number of detectable events by LISA assuming an SNR threshold of 10, which we use in the remaining of the paper. For comparison, we also give the results for an SNR threshold of 20. The LS scenario predicts more merger events, but many of these have low SNR and are not detectable by LISA. On the contrary, almost all events in the HS scenario are detectable.

In Fig. 1 we show the normalized population distribution for different values of α in a "corner plot" [97]. In the lower panel we show only events that have an SNR above 10. We use

		LS	HS
Fiducial	$R_{\rm ev}$ (yr^{-1})	234.3	23.98
	$R_{\rm det}(10)$ (yr^{-1})	53.01	23.89
	$R_{\rm det}(20)$ (yr^{-1})	29.85	23.67
SN-delays	$R_{\rm ev}$ (yr^{-1})	11.82	5.94
	$R_{\rm det}(10)$ (yr^{-1})	1.11	5.92
	$R_{\rm det}(20)$ (yr^{-1})	0.29	5.73

TABLE I. Number of events per year $N_{\rm ev}$ and number of detectable events per year with LISA with two different SNR thresholds, $R_{\rm det}(10)$ and $R_{\rm det}(20)$. The LS scenario predicts more events than the HS one, but many of them are not detectable by LISA. Rates in the SN-delays models (bottom) are substantially lower than in our fiducial model (top).

"transformed" parameters (e.g. $\log_{10}\mathcal{M}_{c,s}$, $\mathrm{arcth}\,\chi_+$) to make the salient features of the distributions more evident. As expected, the HS model predicts binaries with higher masses than the LS model. When mixing between them, we get a double-peaked distribution, whose relative weights depend on the value of α. After imposing an SNR cut, lighter events are suppressed, and the relative weights change due to the fact that many LS events are not detectable. The effect of the SNR cut

[2] Note that we use a different noise curve and SNR threshold than [29, 30], hence the difference in the rates of detectable events.

can be clearly seen in the redshift distribution: high-redshift events predicted in the LS scenario are not detectable, and as a consequence the LS and HS redshift distributions after the cut look much more similar. On the contrary, the effective spin distributions are easier to distinguish after imposing the SNR cut. This is because of the correlation between effective spin, redshift and chirp mass, which can be seen in the upper panel. The physical explanation is that the events that survive the SNR cut in the LS scenario tend to be closer and more massive (both because of the SNR threshold and because the BHs had more time to grow via accretion and mergers). Accretion also leads to larger spins for this subset of the population. Moreover, the presence of gas around binaries tends to align the spins through the Bardeen-Petterson effect, which in turn translates into larger values of the effective spin.

In Fig. 2 we compare the normalized population distribution predicted by the SN-delays model to our fiducial model, both in the LS and HS cases, without any SNR threshold. Notice that the chirp mass distributions of the fiducial and SN-delays models are reasonably similar, but the redshift and effective spin ones are very different. The glaring difference in redshift distributions is due to the additional delays included in the SN-delays model, whereas the one in spin distributions is due to supernova feedback, which expels the gas surrounding the BHs in shallow potential wells, resulting in binaries with more isotropic spin orientations and smaller component spin magnitudes.

In Table I we also provide the rates predicted by the SN-delays model. We see that the rates not only differ substantially between the LS and HS scenarios, but also between the fiducial and SN-delays model. A simple way to provide robustness to this rate variation is to introduce an additional parameter into the model, allowing both the mixing fraction α and the total number of events over the observation period N_α to be hyperparameters that we constrain using the observed events. Although we will ultimately marginalize over the number of observations and focus on the mixing parameter, this approach ensures that our inference will be robust as long the model can match the parameter distribution of events, even if the total number of events varies significantly from the semianalytic model predictions.

IV. HIERARCHICAL BAYESIAN ANALYSIS

Assuming that MBHB events are distributed following the mixing prescription of Eq. (3.1), and introducing the overall number of events as an additional parameter characterizing the population, as described in the previous section, the population distribution is described by two hyperparameters, α and N_α. By observing many events, we will measure the distribution of MBHB parameters θ (such as masses, spins and redshifts), and from this we will be able to infer the hyperparameters. Working in a Bayesian framework, our goal is to estimate the posterior distribution of the hyperparameters from a set of observed MBHB events, $\mathbf{d}$. To do so, we use a similar approach to the "top-down" derivation of [98]. We assume that each MBHB event is independently drawn from the population distribution $p_{\rm pop}(\theta|\alpha, N_\alpha)$. Independence is a highly nontrivial assumption for LISA, since the data stream will contain many signals at the same time, from sources of different types, including extreme mass ratio inspirals, Galactic binaries and MBHBs. However, given the expected event rates for LISA sources (see Table I) and the long duration of the LISA mission, these sources are unlikely to have significant overlap with one another. As a result each source will be sensitive to an independent set of components of the instrumental noise. This means that it should be reasonable to treat each MBHB observation as independent.

Under this assumption the probability that, in a certain observation period, a total of N_t events occur in the Universe, with parameters $\boldsymbol{\theta}$, and producing associated strain data, $\mathbf{d}$, in the detector, is given by

$$p(\mathbf{d}, \boldsymbol{\theta}, N_t|\alpha, N_\alpha) = p(\mathbf{d}|\boldsymbol{\theta}, N_t) p_{\rm pop}(\boldsymbol{\theta}, N_t|\alpha, N_\alpha). \tag{4.1}$$

Assuming that the population of MBHBs is described by a mixture between two independent populations, the second term can be modeled as a Poisson distribution

$$p(\boldsymbol{\theta}, N_t|\alpha, N_\alpha) \propto N_\alpha^{N_t} e^{-N_\alpha} \prod_{k=1}^{N_t} \Big[f(\alpha) p_{\rm pop}(\theta_k|{\rm LS}) + (1 - f(\alpha)) p_{\rm pop}(\theta_k|{\rm HS}) \Big], \tag{4.2}$$

where

$$f(\alpha) = \frac{\alpha R_{\rm ev}({\rm LS})}{\alpha R_{\rm ev}({\rm LS}) + (1-\alpha) R_{\rm ev}({\rm HS})} \tag{4.3}$$

is the expected fraction of events in the Universe that come from the LS population.

Not all the N_t events that occur are detectable. Whether the k'th event is detectable is a property of the associated data, d_k, only. As shown in [98], assuming the events are statistically independent, substituting Eq. (4.2) into Eq. (4.1) and marginalizing over the unobserved data yields the following joint likelihood for the detected events:

$$p(\mathbf{d}, \boldsymbol{\theta}, N_t|\alpha, N_\alpha) \propto \exp\{-N_\alpha (f(\alpha)\Xi({\rm LS}) + (1 - f(\alpha))\Xi({\rm HS})\} \times N_\alpha^{N_{\rm obs}} \prod_{i=1}^{N_{\rm obs}} p(d_i|\theta_i) \Big(f(\alpha) p_{\rm pop}(\theta_i|{\rm LS}) + (1 - f(\alpha)) p_{\rm pop}(\theta_i|{\rm HS}) \Big), \tag{4.4}$$

where $N_{\rm obs}$ is the number of above threshold events observed and $\Xi({\rm LS}) = R_{\rm det}({\rm LS})/R_{\rm ev}({\rm LS})$ is the fraction of events in the LS population expected to be detectable, which is given by

$$\begin{aligned} \Xi({\rm LS}) &= \int {\rm d}\theta \; p_{\rm pop}(\theta|{\rm LS}) \int_{d,{\rm detectable}} {\rm d}d \; p(d|\theta) \\ &= \int {\rm d}\theta \; p_{\rm pop}(\theta|{\rm LS}) p_{\rm det}(\theta), \end{aligned} \tag{4.5}$$

where the last equality defines $p_{\rm det}(\theta)$, the probability of detecting an event with parameters θ. The quantity $\Xi({\rm HS})$ is defined in an analogous way for the HS population. In this work

we use the SNR to quantify detectability and assume that an event, d, is detectable if $\mathrm{SNR}[d] > \mathrm{SNR}_{\mathrm{threshold}}$. Since we work in the zero-noise approximation, we evaluate this using the optimal SNR to determine the detectability of each source. The selection function, $\Xi(\mathrm{LS})$, is equal to the fraction of events in the population that have SNR above the threshold.

The final form of the posterior distribution on α and N_α is obtained by marginalization over the parameters of the individual events, $\boldsymbol{\theta}$, in Eq. (4.9) and using Bayes' theorem. After some rearrangement we obtain

$$\begin{aligned} p(\alpha, N_\alpha|\mathbf{d}) &= \frac{p(\mathbf{d}|\alpha, N_\alpha)p(\alpha, N_\alpha)}{p(\mathbf{d})} \\ &\propto \frac{p(\alpha, N_\alpha)\prod_{i=1}^{N_{\mathrm{obs}}} p(d_i)}{p(\mathbf{d})} N_\alpha^{N_{\mathrm{obs}}} \exp[-N_\alpha \Xi(\alpha)] \\ &\quad \times \prod_{i=1}^{N_{\mathrm{obs}}} \int d\theta_i \frac{p(\theta_i|d_i)p_{\mathrm{pop}}(\theta_i|\alpha)}{p_i(\theta_i)}, \end{aligned} \tag{4.6}$$

in which $p(\theta_i|d_i) = p(d_i|\theta_i)p_i(\theta_i)/p(d_i)$, $p_i(\theta_i)$ denotes the prior used to obtain some posterior samples in an initial analysis of event-i, and we have introduced

$$\Xi(\alpha) = f(\alpha)\,\Xi(\mathrm{LS}) + (1 - f(\alpha))\Xi(\mathrm{HS}) \tag{4.7}$$

$$p_{\mathrm{pop}}(\theta|\alpha) = f(\alpha)p_{\mathrm{pop}}(\theta|\mathrm{LS}) + (1 - f(\alpha))p_{\mathrm{pop}}(\theta|\mathrm{HS}). \tag{4.8}$$

In an analysis of LISA data we would construct this posterior on both hyperparameters. However, the parameter of most interest is the mixing fraction α, and so we will focus on this here. We proceed by marginalizing over the rate parameter, N_α. We first specify that the hyperprior is separable, $p(\alpha, N_\alpha) = p(\alpha)p(N_\alpha)$, and then assume a scale-invariant prior on the rate, $p(N_\alpha) \propto 1/N_\alpha$. The scale-invariant $1/N_\alpha$ prior is natural when the order of magnitude of the rate is uncertain, as is the case here. After this marginalization we obtain

$$\begin{aligned} p(\alpha|\mathbf{d}) &= \frac{p(\mathbf{d}|\alpha)p(\alpha)}{p(\mathbf{d})} \\ &\propto \frac{p(\alpha)\prod_{i=1}^{N_{\mathrm{obs}}} p(d_i)}{p(\mathbf{d})} \prod_{i=1}^{N_{\mathrm{obs}}} \int d\theta_i \frac{p(\theta_i|d_i)p_{\mathrm{pop}}(\theta_i|\alpha)}{p_i(\theta_i)\Xi(\alpha)}. \end{aligned} \tag{4.9}$$

If N_i posterior samples have been obtained for event i using the reference prior $p_i(\theta_i)$, these can be used to obtain a Monte Carlo approximation to the integrals in the preceding equation

$$\begin{aligned} p(\alpha|\mathbf{d}) &= \prod_{i=1}^{N_{\mathrm{obs}}} \left[\frac{1}{N_i}\sum_{j=1}^{N_i} \frac{p_{\mathrm{pop}}(\theta_{i,j}|\alpha)}{p_i(\theta_{i,j})\Xi(\alpha)}\right] \\ &\quad \times p(\alpha)\frac{\prod_{i=1}^{N_{\mathrm{obs}}} p(d_i)}{p(\mathbf{d})}, \end{aligned} \tag{4.10}$$

where $\theta_{i,j}$ is the parameter vector for the j'th sample for source i. The individual event and overall evidences, $p(d_i)$ and $p(\mathbf{d})$, are useful for model selection but merely enter as a normalization constant when the interest is on parameter estimation, as here. Therefore, we discard all evidence terms from our analysis. For the prior on α, we take a flat distribution in $[0, 1]$.

We note that the quantity $f(\alpha)$ is directly interpretable as the fraction of events in the Universe that are drawn from the LS model, while the mixing fraction α, as we have defined it, is not. However, these are related by the simple transformation given in Eq. (4.3), and so the posterior for $f(\alpha)$ can readily be derived from that for α and vice versa.

After inferring a posterior distribution on α, we can construct the posterior predictive distribution (PPD) for the parameters of future observed events

$$\mathrm{PPD}(\theta|\mathbf{d}) = \int \mathrm{d}\alpha\; p_{\mathrm{pop}}(\theta|\alpha)p(\alpha|\mathbf{d}). \tag{4.11}$$

When performing simulations, comparing the PPD with the population distribution used to generate the data provides a guide to the quality of the inference.

V. ESTIMATING THE PROBABILITY DENSITY FUNCTION

From Eq. (4.10), we can see that the hierarchical Bayesian analysis requires being able to evaluate the probability density function of the population distribution. However, semianalytic models only provide samples from the population distribution, not the analytic probability density function. In this work, we use a kernel density estimator (KDE) [99, 100] to approximate the population probability density function from the samples. More specifically, we use the Gaussian KDE implementation of `scipy` [101]. In Appendix A, we provide additional details on how the KDE is computed.

The required accuracy on the estimation of the probability density function increases with the number of observed events. The accuracy of the KDE is limited by the number of simulation points at our disposal, in particular for the HS variant of our fiducial astrophysical model (~ 2500 points). This leads to a systematic error, which dominates over statistical errors when increasing the number of observed events, and leads to systematic biases in the hierarchical Bayesian analysis. Similarly, from Eq. (4.10) it can be seen that the error on $\ln(p(\alpha|\mathbf{d}))$ due to a misevaluation of the selection function increases linearly with the number of observed events. In our case, the accuracy to which the selection function is computed depends on the accuracy of the selection function for the LS and HS models: cf. Eq. (4.7). In Appendix B, we show that using too few points to compute these terms also leads to systematic biases. To mitigate these issues, we make an approximation: we take the probability density function computed from the KDE to be the "true" probability density function of our fiducial astrophysical model, and use it to generate mock data. By doing this, the data generation process is fully consistent with the probability density function used in the hierarchical Bayesian analysis, avoiding systematic biases. We compute the selection function for the LS and HS variants of our fiducial astrophysical model by generating many ($\sim 10^6$) events from the KDE and computing the fraction of detectable events. We then use Eq. (4.7) to evaluate the selection function for any value of α. This approximation should be seen as the limit where we have enough simulation points to build

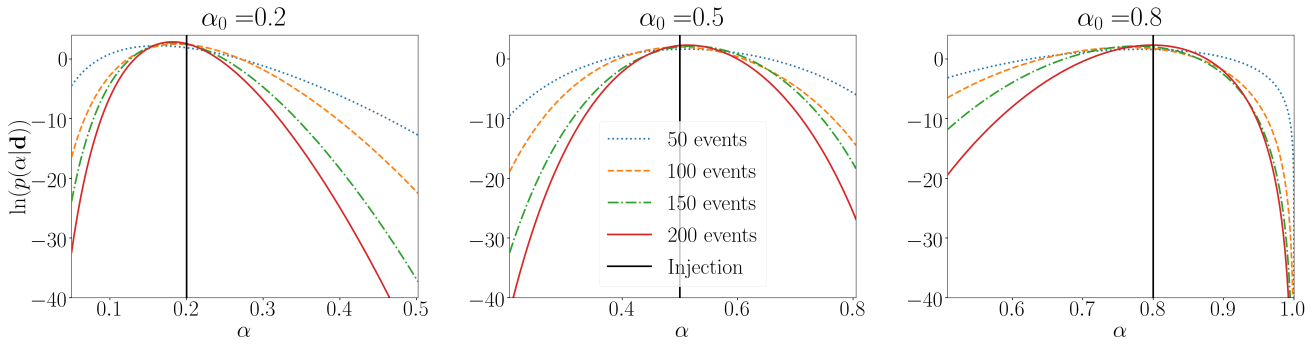


FIG. 3. Posterior distribution on α for observation sets with an increasing number of observed events, generated using different values of the mixing fraction α_0: $\alpha_0 = 0.2$ (left), $\alpha_0 = 0.5$ (middle) and $\alpha_0 = 0.8$ (right). The posteriors peak near the true value and become narrower as we increase the number of events.

very accurate KDEs and compute the selection function to high precision. In Appendix C we compare the population distribution of the LS and HS variants of the fiducial astrophysical model computed from numerical simulations to the one obtained from the KDEs, computed as described in Appendix A. Note that, when building the KDE that will serve as our fiducial astrophysical model, we use $\text{arcth}\,\chi_{1,2}$ instead of $\text{arcth}\,\chi_{+,-}$ to make sure that the spins are in the physically allowed range. The distributions are overall in very good agreement, so we expect that our results should not depend much on this approximation.

VI. RESULTS

We start by testing our pipeline in the limit where the parameters of the source are perfectly measured by LISA, and we perform two experiments. In the first one (Sec. VI A) we generate mock observation sets using the predictions of our fiducial astrophysical model, as computed from the KDE, and use this same model in the hierarchical Bayesian analysis. In the second experiment (Sec. VI B) we use the SNdelays model to generate mock observation sets, but still use our fiducial astrophysical model in the hierarchical Bayesian analysis. The goal of this second experiment is to test if we could still draw meaningful conclusions if the population of MBHBs in the Universe were different from the one used in the data analysis pipeline. In Sec. VI C we discuss the impact of measurement errors in the analysis. In all cases we use an SNR threshold of 10 to define detectability of a source.

A. Model-consistent inference

We start by investigating how the inference on α improves with the number of observed events. Although we do not use information on the rates in the inference, we make sure that the number of events in the datasets is realistic for a LISA mission duration of four to ten years, given the predicted rates (see Table I). In Fig. 3, we plot the log-posterior on α for observation sets with an increasing number of observed events. In the left panel, the dataset was generated with a mixing fraction $\alpha_0 = 0.2$ between the LS and HS variants of our fiducial astrophysical model, in the middle panel with $\alpha_0 = 0.5$, and in the right panel with $\alpha_0 = 0.8$. The posteriors peak near the true value and become narrower as we increase the number of events. We observe a sharp drop in the posterior close to the extremal values. This is because as $\alpha \to 0$ ($\alpha \to 1$) the resulting population is no longer compatible with the lightest (heaviest) events. Moreover, due to our choice of mixing prescription in Eq. (3.1) and to the higher event rate of the LS variant, the population distribution varies faster for small values of α, so the posterior is narrower for $\alpha_0 \simeq 0$ than for $\alpha_0 \simeq 1$.

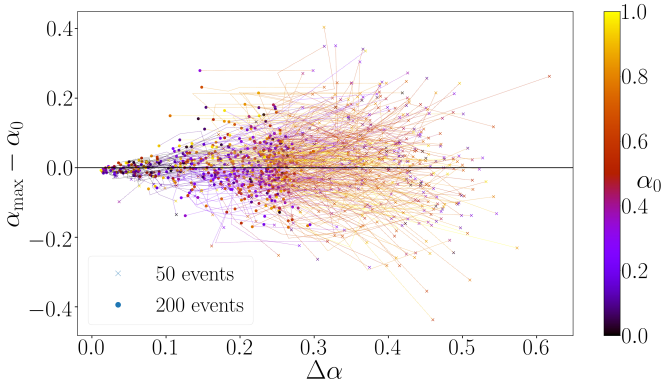
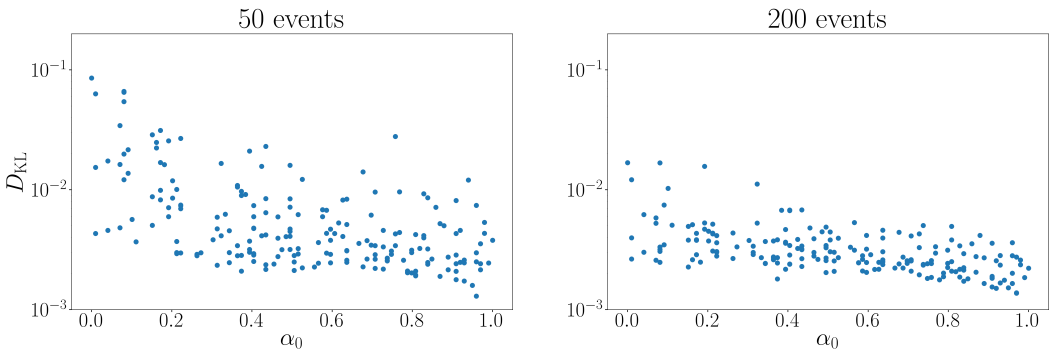


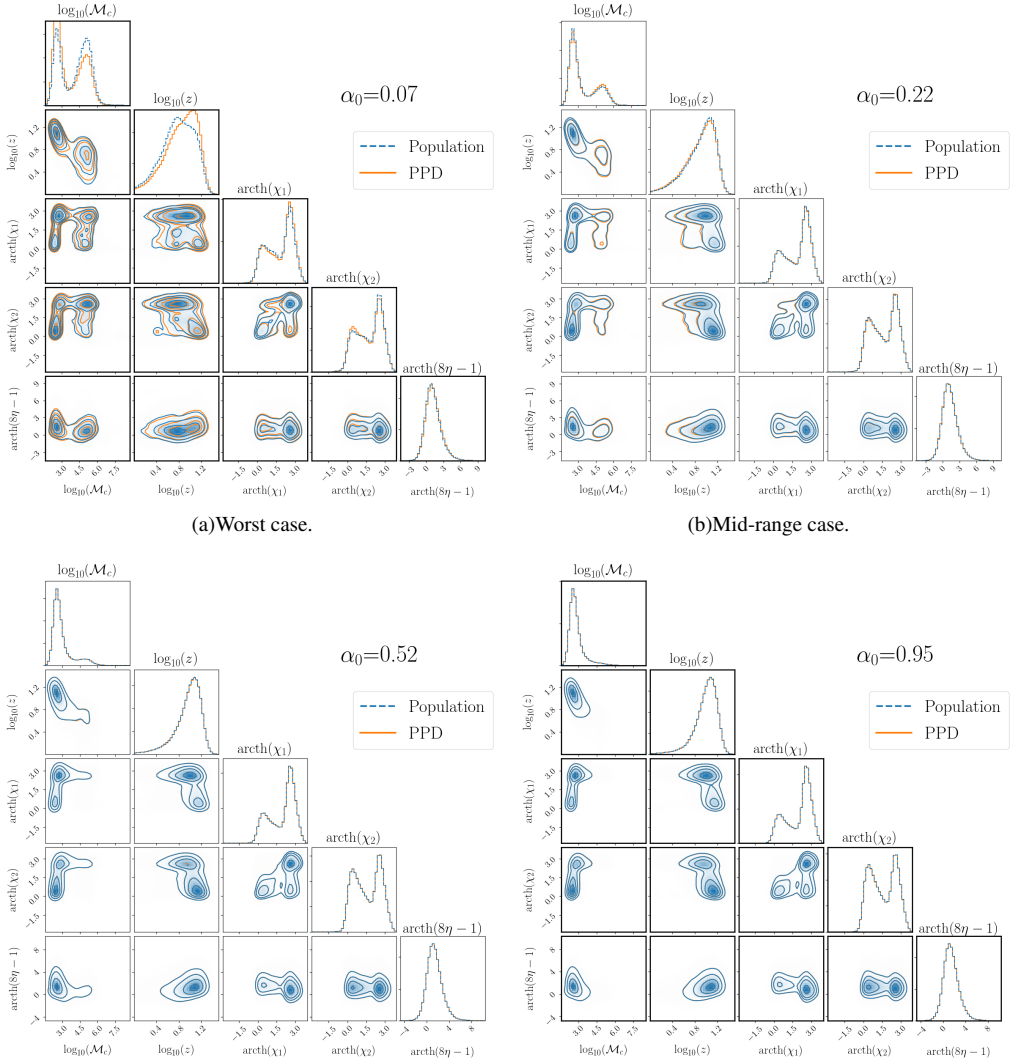
FIG. 4. Evolution of the shift and the error on α (90% confidence interval) with the number of observed events. We consider two sets of observations, with 50 events (crosses) and 200 events (dots). The color scale indicates the value of α_0. As expected, they tend do decrease as we observe more events. The fact that the points are equally distributed on both sides of the $\alpha_{\rm max} = \alpha_0$ line indicates that there is little systematic bias in our analysis.

In order to have a more global view, we generate several observation sets with an increasing number of events, drawing the mixing fraction uniformly in $[0, 1]$. We estimate the shift on α as the difference between the maximum-posterior point $\alpha_{\rm max}$ and the injection value α_0, and the error on the mixing fraction $\Delta\alpha$ as the 90% confidence interval centered around the median value. In Fig. 4 we plot these quantities for two



(a) 50 events. (b) 200 events.

FIG. 5. Kullback-Leibler divergence between the PPD and the population distribution for different observation sets generated with different values of α_0. On the left (right) panel the observation sets contain 50 (200) observed events. The smaller the KL divergence, the better our inference of the population distribution. Increasing the number of events tends to improve the inference, as expected.



(a)Worst case. (b)Mid-range case.

(c)Mid-range case. (d)Best case.

FIG. 6. Population distribution and PPD for four sets of observations generated with different values of α_0. Each observation set contains 100 events. On the upper-left and lower-right panels we show the cases that yield the largest and smallest values of the KL divergence among the cases shown in Fig. 5. The other two panels show cases yielding mid-range values of the KL divergence.

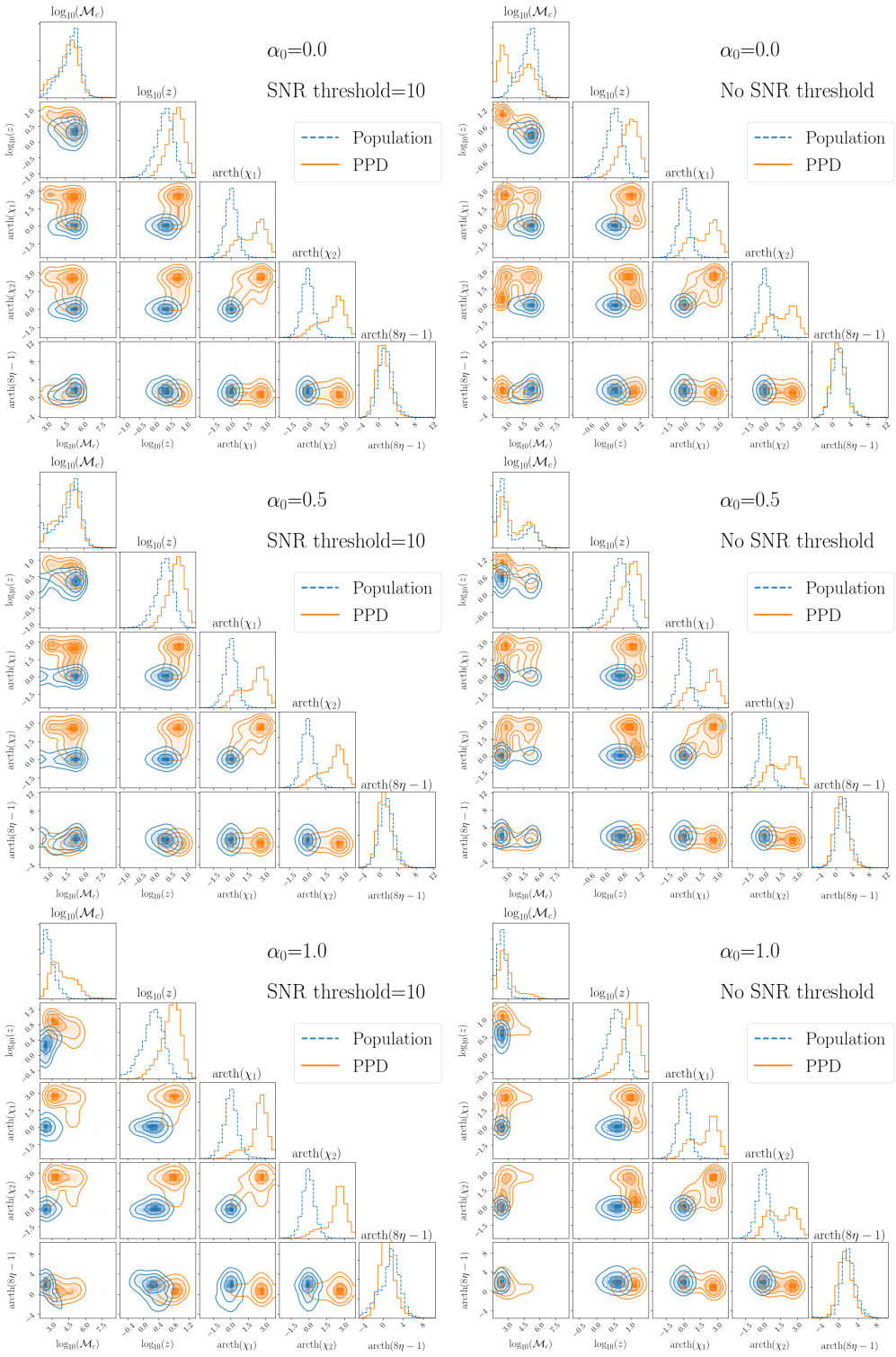


FIG. 7. Comparison between the population distribution for the SN-delays model and the PPD for an observation set containing 20 events from the same catalogue. Different rows refer to the HS variant ($\alpha_0 = 0$, top), a mixing fraction $\alpha_0 = 0.5$ between the HS and LS variants (middle), and the LS variant ($\alpha_0 = 1$, bottom). Panels on the left refer to the detectable population; panels on the right, to the intrinsic population.

selected values of the number of observed events. The color scale indicates the value of the injected mixing fraction α_0 for each observation set. As expected, both tend to decrease as we observe more events. Also, note that the points are equally distributed on both sides of the $\alpha_{\rm max} = \alpha_0$ line, indicating that there is little systematic bias in our analysis, as we would expect given that the models used to generate and analyze the data are consistent. We find that the error on α tends to be smaller for injected values close to 0 or 1, with even smaller errors in the former case, in agreement with our discussion on the shape of the posterior above.

Next, we assess our ability to infer the population distribution from an observed dataset, using the PPD defined in Eq. (4.11). In order to make a quantitative comparison, we compute the Kullback-Leibler (KL) divergence [102] between them, defined as

$$D_{\rm KL} = \sum_{\theta} p_1(\theta) \ln\left(\frac{p_1(\theta)}{p_2(\theta)}\right), \tag{6.1}$$

with p_1 and p_2 the distributions we wish to compare. In Fig. 5, we plot the KL divergence between the PPD and the population distribution for datasets of 50 and 200 observed events, taking the population distribution as the reference distribution (p_1). Given the similarity between the distributions (as indicated by the smallness of the KL divergence), the results would not be significantly altered had we chosen the PPD as the reference distribution. The KL divergence tends to be smaller for larger datasets, meaning that our inference on the population distribution improves. As a trend, the largest values of the KL divergence correspond to $\alpha_0 \sim 0$. This is because the population distribution varies faster for small α, so even small (statistical) deviations in the estimation of the mixing fraction lead to larger discrepancies between the PPD and the population distribution for $\alpha_0 \sim 0$. As an illustration, in Fig. 6 we compare the PPD obtained from four simulated LISA datasets of 100 observed events generated with different values of α_0 to the corresponding population distribution. Those realizations are chosen to span the range of values of KL divergences. As can be seen in the upper-left panel, even in the worst case (the largest value of the KL divergence among the cases shown in Fig. 5) we can reconstruct the population distribution reasonably well. The other panels show the comparison between the PPD and the population distribution for datasets of 100 events yielding mid-range values of the KL divergence and for the dataset yielding the smallest one. Overall, this pipeline allows us to infer the population distribution accurately when the model used to generate the data is the same as the one used in the pipeline. We will now test the robustness of this pipeline by using different models in the two stages.

B. Robustness

We mix the HS and LS variants of the SN-delays model as described in Eq. (3.1), and generate datasets of 20 observed events for $\alpha_0 = 0$, $\alpha_0 = 0.5$ and $\alpha_0 = 1$. We run our pipeline on these observation sets, still using our fiducial astrophysical model in the hierarchical Bayesian analysis and compare the PPD to the population distribution. The results are shown in Fig. 7. In each case, we show both the intrinsic distribution and the detected one (where detection is defined by imposing an SNR threshold of 10). For $\alpha_0 = 0$ (top panels), we can reproduce reasonably well the chirp mass distribution of the detectable population, but we overestimate the fraction of small-$\mathcal{M}_c$ events in the intrinsic population. This is because the HS variant of the SN-delays model has a tail extending to lighter values than the HS variant of the fiducial model, as can be seen on Fig. 2. Our pipeline compensates for this by adding events from the LS variant, and since only $\sim 25\%$ of LS events are detectable, the fraction of light events in the intrinsic population is overestimated. Similarly, for $\alpha_0 = 0.5$ (middle panels) the PPD agrees reasonably well with the population distribution of the chirp mass for detectable events, but this time the fraction of light events in the intrinsic population is underestimated. This is due to the difference in the fraction of detectable events between the LS variant of our fiducial model and the SN-delays model (see Table I). For a given number of detected light events, the latter predicts twice as many light events in the intrinsic population as our fiducial model. Finally, for $\alpha_0 = 1$ (bottom panels) even the chirp mass distribution of detectable events is badly estimated. This is due to a tail of heavy events predicted by the LS variant of the SN-delays model, which causes our pipeline to estimate α_0 to be different from 1. In all three cases, due to the differences in the fiducial and SN-delays population, redshift and spin distributions are poorly reconstructed.

These results show that this pipeline would lead to erroneous predictions if the population of MBHBs is too different from the one predicted by our astrophysical models. Note that in the LS SN-delays model we do not expect to observe 20 events even for a ten-year mission duration, but this does not change our previous conclusion.

C. Including measurement errors

We now wish to consider two sources of error: weak lensing and statistical errors due to detector noise. They are accounted for with the following procedure. For each event predicted by the model:

(1) we draw a new value of the luminosity distance from a Gaussian distribution centered at the original value with variance given by the lensing error of Eq. (2.1), keeping the detector-frame mass constant;

(2) from that new event, we draw a shifted event from a multinormal Gaussian distribution with covariance given by the Fisher information matrix at that point;

(3) if this new event has SNR above the threshold, we perform parameter estimation;

(4) we broaden the posterior distribution of the luminosity distance (and therefore of the redshift and the source frame mass) with the lensing error of Eq. (2.1).

For step (3), we use the Fisher information matrix instead of doing a full Bayesian analysis in order to speed up computations. Some events from the LS variant have very low SNR of order unity, and in those cases the Fisher information matrix is poorly conditioned. For this reason, events with such low SNRs might end up with large enough SNRs to be detected after applying the Fisher matrix shift of step (2). This is not physically realistic, since the detector noise is unlikely to make such events detectable, and therefore between steps (2) and (3) we discard all events that have SNR below 5 *before* the shift.

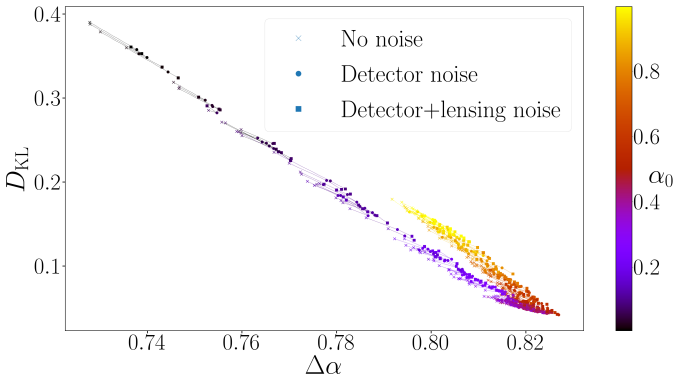


FIG. 8. Error on α and KL divergence between the rescaled posterior distribution of α and the (flat) prior. As we include the different sources of error, $\Delta\alpha$ tends to increase and D_{KL} tends to decrease, reflecting a degradation in the measurement of α. Note that these are the errors and KL divergences for the rescaled posterior, i.e. we artificially bring the number of detected events to 1, as detailed in the main text.

In order to assess the impact of measurement errors, we generate datasets of 500 events (before applying the detectability criterion) for α_0 drawn randomly in $[0, 1]$, and consider three scenarios:

(i) there is no noise, i.e., none of the steps above are applied;

(ii) there is only detector noise, i.e., only steps (2) and (3) are applied;

(iii) there is both detector noise and lensing noise, i.e. all four steps are applied.

Note that steps (1) and (2) modify the number of detectable events, therefore we have to include these effects in the computation of the selection function. Moreover, increasing the number of observed events tends to narrow the posterior distribution, so in order to scale out this effect and allow for a fair comparison between the three different scenarios, we define a "rescaled" posterior distribution $\tilde{p}(\alpha|\mathrm{d}) = p(\alpha|\mathrm{d})^{1/N_{\rm obs}}$. In Fig. 8 we plot on the x-axis the error on α (obtained from the rescaled posterior) and on the y-axis the KL divergence between the rescaled posterior distribution of α and the (flat) prior on α, for different datasets and in the three scenarios. The color scale indicates the value of α_0. The larger the KL divergence, the more information we gain from the dataset. As expected, including the different sources of error tends to decrease the KL divergence and increase $\Delta\alpha$. The dotted lines going from the top-left to bottom-right link simulations with the same underlying populations, and show (slight) degradation in the measurement of α. Note that the KL divergence is larger and the error smaller for $\alpha_0 \sim 0$ and also for $\alpha_0 \sim 1$, in agreement with the discussion on the shape of the posterior in the previous subsection. Finally, we do not observe the appearance of systematic biases when including measurement errors.

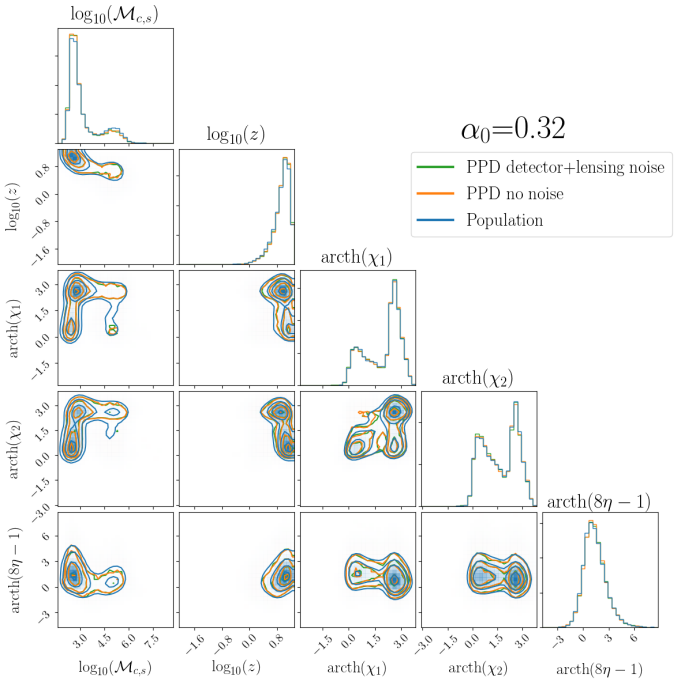


FIG. 9. Population distribution and PPDs obtained in the no-noise and detector+lensing noise scenarios for a representative case. The dataset contains 500 events (before applying the detectability criterion). Including measurement errors barely affects our ability to infer the population distribution.

Although the determination of α gets slightly worse when including the different sources of error, this barely affects our inference of the population distribution of MBHBs, as can be seen in Fig. 9. There, we compare for a representative case the population distribution to the PPDs obtained in the no-noise and in the detector+lensing noise scenarios, which look very similar.

Finally, we performed a last test: we generated datasets including both sources of noise in steps (1) and (2), but we did not include the effect of lensing in the hierarchical analysis, i.e. step (4). Moreover, we used the selection function obtained when accounting only for the detector noise. Our goal is to assess how our analysis would be biased if we did not properly model the effect of lensing. We observe a tendency to bias the measurement of α toward higher values, but no real impact on the PPD. This could be an artifact of our simplistic model, and should be verified through further work.

VII. CONCLUSIONS

In this paper we discussed of the ability of LISA to distinguish between different astrophysical models for the formation and evolution of MBHs by inferring the population of MBHBs. We introduced a mixing fraction between astrophysical models to account for the possibility that the population of MBHBs in the Universe cannot be described by one single model. More specifically, we mixed between two variants of the same model: one that predicts that MBHs form from LSs and another from HSs. We built a pipeline based on the hierarchical Bayesian framework to measure the mixing fraction from LISA observations, and infer the population of MBHs. We have shown that this pipeline allows us to reconstruct accurately the population of MBHBs if it is similar to the one used in the pipeline, but not if the populations are too different.

This problem could be mitigated by introducing more flexibility in the population model, at the cost of having greater uncertainty in the inferred population distribution. One approach would potentially be to include additional mixing fractions: one could in principle mix between as many models as desired. However, given the large uncertainty surrounding astrophysical models, we believe a better alternative is to use a theory-agnostic approach. We are currently working on a simplified astrophysical model for the formation and evolution of MBHs where the population of MBHBs depends on physically meaningful hyperparameters controlling the initial mass distribution, the delay between dark matter halo mergers and MBHB mergers, etcetera. We could then perform a hierarchical Bayesian analysis to infer these hyperparameters from LISA observations.

We have shown that measurement errors due to lensing and detector noise will not significantly impact our ability to infer the MBHB population. On the other hand, mismodelling the effect of weak lensing could lead to biases in our analysis. In our model, this bias has a negligible impact on our inference of the population of MBHBs, but this could be due to the simplicity of our model and will have to be further verified for different models. Finally, we commented on an important aspect: analyses based on results from numerical simulations, such as ours, require a large number of points in order to properly evaluate the probability density function of the theoretical model and the selection function, and thus avoid systematic biases. We estimate that at least a few tens of thousands of points are needed.

Concerning our astrophysical model, we mixed the distributions *a posteriori*, i.e. with the results obtained by running numerical simulations with LSs and HSs independently. Therefore, our model cannot account for mergers between BHs formed from LSs and HSs and how they impact the population distribution. This could be included by mixing the seeding prescriptions *a priori*, when running the simulations. We could then use these results to assess the validity of our *a posteriori* approach. We leave this for future work.

ACKNOWLEDGMENTS

A.T. is thankful to Johns Hopkins University for their hospitality during the early stages of this work. S.B. and A.T. acknowledge support by CNES, in the framework of the LISA mission. This work has been supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 690904. E. Barausse acknowledges financial support provided under the European Union's H2020 ERC Consolidator Grant "GRavity from Astrophysical to Microscopic Scales" grant agreement no. GRAMS-815673. E. Berti and K.W.K. Wong are supported by NSF Grants No. PHY-1912550 and No. AST-2006538, NASA ATP Grants No. 17-ATP17-0225 and No. 19-ATP19-0051, NSF-XSEDE Grant No. PHY-090003, and NSF Grant No. PHY-20043. This research project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC). S. Taylor is supported by NSF Grant No. AST-2007993 and PHY-2020265. The authors would like to acknowledge networking support from the COST Action CA16104.

Appendix A: Kernel density estimation

From a set of n_s samples drawn from the distribution $p_{\rm pop}(\theta|\alpha)$, the KDE approximates its probability density function as

$$\hat{p}_{\rm pop}(\theta|\alpha) = \frac{1}{n_s}\sum_{i=1}^{n_s} K_H(\theta-\theta_i), \tag{A1}$$

where K_H is the *kernel function*. We choose to work with Gaussian KDEs, where, denoting by n_d the dimensionality of the parameter space,

$$K_H(y) = \frac{1}{(2\pi)^{n_d/2}}[\det(H)]^{-1/2} e^{-\frac{1}{2}y^T H^{-1} y}. \tag{A2}$$

In the Gaussian KDE implementation of `scipy` [101], H is taken to be proportional to the identity matrix. The proportionality constant is called the *bandwidth* of the KDE, and is a very important parameter, since it defines the smoothing scale of the approximation to the target probability density function. In Fig. 10 we show the approximations to the population probability density function of $\log_{10}(\mathcal{M}_{c,s})$ that we obtain using different values of the bandwidth (noted bw).

For too large values of the bandwidth, we cannot resolve the features of the distribution, and for too small values, the resulting probability density function is not smooth. We deal with this issue by choosing the bandwidth that minimizes the integrated squared error $\int {\rm d}\theta(p_{\rm pop}(\theta|\alpha) - \hat{p}_{\rm pop}(\theta|\alpha))^2$. In practice, it is estimated by using a Monte Carlo averaging, and the quantity we seek to minimize is [103]

$$\int {\rm d}\theta \hat{p}_{\rm pop}(\theta|\alpha)^2 - \frac{2}{n_s}\sum_{i=1}^{n_s} \hat{p}_{\rm pop,-i}(\theta_i|\alpha), \tag{A3}$$

where the sum runs over the n_s samples drawn from $p_{\rm pop}(\theta|\alpha)$ used to approximate the integral, and $\hat{p}_{\rm pop,-i}(\theta|\alpha)$ is the KDE

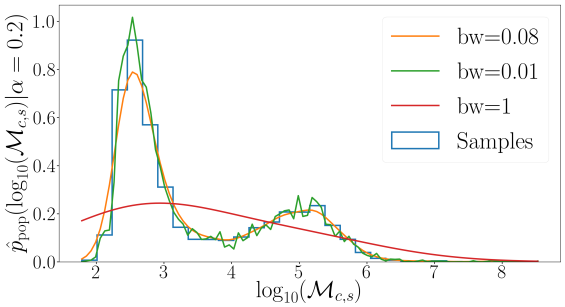


FIG. 10. Comparison between different KDE approximations to the population probability density function of $\log_{10}\mathcal{M}_{c,s}$, using different values of the bandwidth. If the bandwidth is too small the KDE is not smooth, and if it is too large we cannot resolve the features of the distribution. For the case shown here, a bandwidth of 0.08 is a good choice. This value was obtained by minimizing the integrated squared error, as described in the main text.

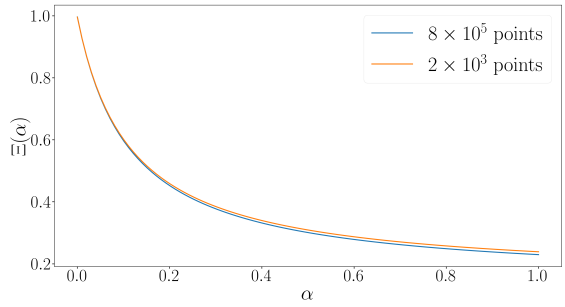
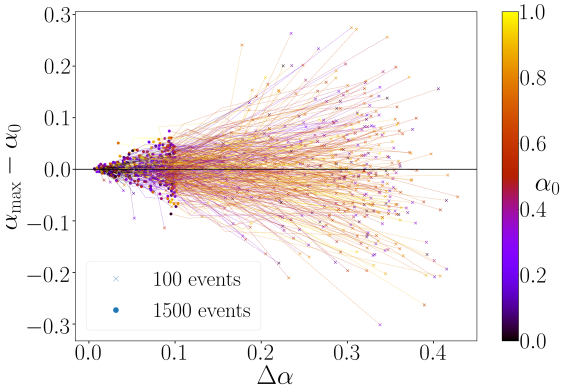


FIG. 11. Comparison between the selection functions obtained using different numbers of points.

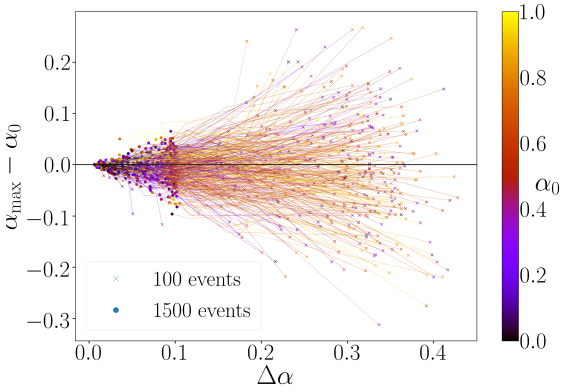
obtained using all n_s samples but the i^{th} one. The value of 0.08 used in Fig. 10 was obtained with this method. We also apply it to compute the bandwidth of the KDE for the LS and HS population distributions.

Appendix B: Systematic biases due to misevaluation of the selection function

The selection function used to obtain the results of this paper was computed with Eq. (4.7). We generated 8×10^5 events for the LS and HS variants from the KDE and computed the terms Ξ(LS) and Ξ(HS) individually. In Fig. 11 we compare this selection function with one obtained using only 2×10^3 points to compute each term. There is a clear discrepancy between the two functions, which reflects on the population inference as can be seen in Fig. 12. There we compare the shift versus error on α plots obtained using each of these selection functions. Clearly, using too few points to compute the selection function leads to systematic biases, as can be seen by the fact that many more points are below the $\alpha_{\rm max}=\alpha_0$ line than above. We do not expect to observe thousands of MBHBs with LISA, but we have chosen this large number of events to emphasize this effect. Even for fewer events we could be biased due to misevaluation of the selection function, and a large number of points from numerical simulations will be needed to mitigate this effect (see also [104]). Moreover, third generation ground-based detectors are expected to detect thousands of events, and will face this same problem. In our study, this systematic bias becomes negligible when using $O(10^5)$ points for each model.



(a)We use 8×10^5 points to evaluate the selection function of the LS and HS variants.



(b)We use 2×10^3 points to evaluate the selection function of the LS and HS variants.

FIG. 12. Evolution of the bias and the error on α using the selection function in blue in Fig. 11 (top) and the one in orange (bottom). We can clearly observe a systematic bias in the latter case due to misevaluation of the selection function.

Appendix C: Comparison between KDE and the population obtained from simulations

In Fig. 13 we compare the population distribution predicted from numerical simulations to the one obtained from building a KDE on it.

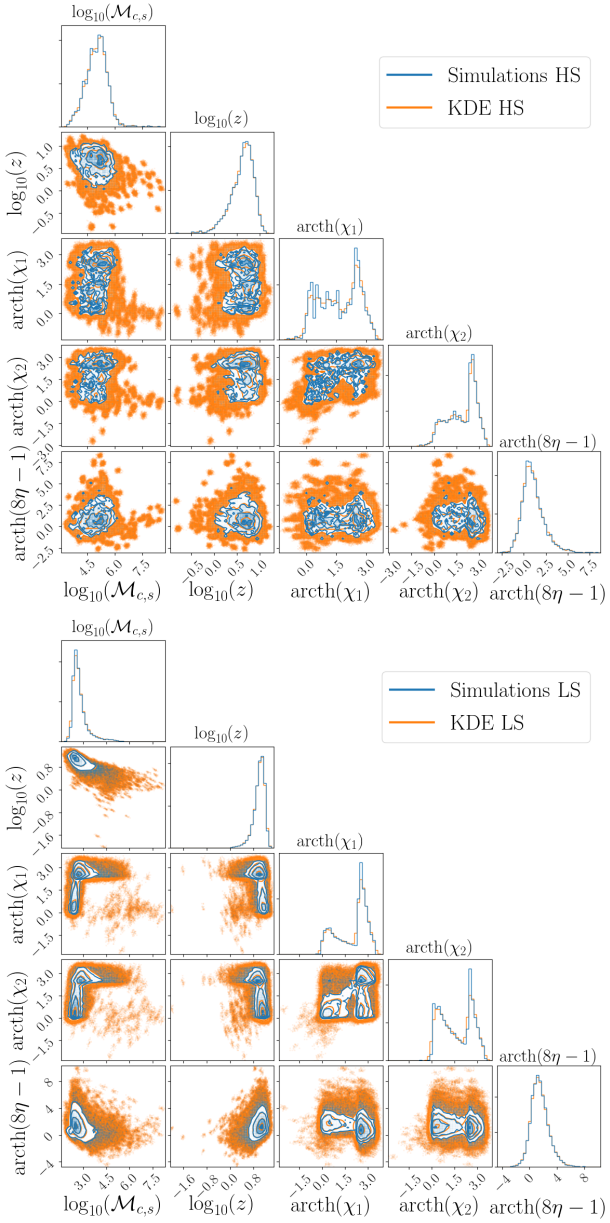


FIG. 13. Comparison between the population distributions obtained from numerical simulations and the KDE we build from it. We purposefully did not smooth the corner plot in order to reflect the real level of agreement between the two distributions. The top and bottom panels refers to the LS and HS variants, respectively. The "bumpy" histograms for the HS variant (in particular for the spin) highlight that we do not have enough points to build an accurate enough KDE for our purposes. However the two distributions are overall in good agreement, and therefore we expect that the approximation of using the KDE as our "true" fiducial astrophysical model should not sensibly affect our results.

[1] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].

[2] B. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. X **9**, 031040 (2019), arXiv:1811.12907 [astro-ph.HE].

[3] R. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. X **11**, 021053 (2021), arXiv:2010.14527 [gr-qc].

[4] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. Lett. **882**, L24 (2019), arXiv:1811.12940 [astro-ph.HE].

[5] R. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. Lett. **913**, L7 (2021), arXiv:2010.14533 [astro-ph.HE].

[6] M. Zevin, C. Pankow, C. L. Rodriguez, L. Sampson, E. Chase, V. Kalogera, and F. A. Rasio, Astrophys. J. **846**, 82 (2017), arXiv:1704.07379 [astro-ph.HE].

[7] C. Talbot and E. Thrane, Phys. Rev. D **96**, 023012 (2017), arXiv:1704.08370 [astro-ph.HE].

[8] K. Belczynski *et al.*, Astron. Astrophys. **636**, A104 (2020), arXiv:1706.07053 [astro-ph.HE].

[9] C. Talbot and E. Thrane, Astrophys. J. **856**, 173 (2018), arXiv:1801.02699 [astro-ph.HE].

[10] J. Roulet and M. Zaldarriaga, Mon. Not. Roy. Astron. Soc. **484**, 4216 (2019), arXiv:1806.10610 [astro-ph.HE].

[11] Y. Bouffanais, M. Mapelli, D. Gerosa, U. N. Di Carlo, N. Giacobbo, E. Berti, and V. Baibhav, Astrophys. J. **886** (2019), 10.3847/1538-4357/ab4a79, arXiv:1905.11054 [astro-ph.HE].

[12] V. Baibhav, D. Gerosa, E. Berti, K. W. K. Wong, T. Helfer, and M. Mould, Phys. Rev. **D102**, 043002 (2020), arXiv:2004.00650 [astro-ph.HE].

[13] J. Roulet, T. Venumadhav, B. Zackay, L. Dai, and M. Zaldarriaga, Phys. Rev. **D102**, 123022 (2020), arXiv:2008.07014 [astro-ph.HE].

[14] A. Hall, A. D. Gow, and C. T. Byrnes, Phys. Rev. **D102**, 123524 (2020), arXiv:2008.13704 [astro-ph.CO].

[15] K. W. K. Wong, K. Breivik, K. Kremer, and T. Callister, Phys. Rev. D **103**, 083021 (2021), arXiv:2011.03564 [astro-ph.HE].

[16] C. Kimball *et al.*, Astrophys. J. Lett. **915**, L35 (2021), arXiv:2011.05332 [astro-ph.HE].

[17] M. Zevin, S. S. Bavera, C. P. L. Berry, V. Kalogera, T. Fragos, P. Marchant, C. L. Rodriguez, F. Antonini, D. E. Holz, and C. Pankow, Astrophys. J. **910**, 152 (2021), arXiv:2011.10057 [astro-ph.HE].

[18] Y. Bouffanais, M. Mapelli, F. Santoliquido, N. Giacobbo, U. N. Di Carlo, S. Rastello, M. C. Artale, and G. Iorio, (2021), arXiv:2102.12495 [astro-ph.HE].

[19] V. De Luca, G. Franciolini, P. Pani, and A. Riotto, JCAP **05**, 003 (2021), arXiv:2102.03809 [astro-ph.CO].

[20] V. Gayathri, Y. Yang, H. Tagawa, Z. Haiman, and I. Bartos, (2021), arXiv:2104.10253 [gr-qc].

[21] G. Franciolini, V. Baibhav, V. De Luca, K. K. Y. Ng, K. W. K. Wong, E. Berti, P. Pani, A. Riotto, and S. Vitale, (2021), arXiv:2105.03349 [gr-qc].

[22] T. A. Callister, C.-J. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, (2021), arXiv:2106.00521 [astro-ph.HE].

[23] P. Amaro-Seoane *et al.* (LISA), (2017), arXiv:1702.00786 [astro-ph.IM].

[24] G. Nelemans, L. R. Yungelson, and S. F. Portegies Zwart, Astron. Astrophys. **375**, 890 (2001), arXiv:astro-ph/0105221.

[25] V. Korol, E. M. Rossi, P. J. Groot, G. Nelemans, S. Toonen, and A. G. Brown, Mon. Not. Roy. Astron. Soc. **470**, 1894 (2017), arXiv:1703.02555 [astro-ph.HE].

[26] A. Sesana, M. Volonteri, and F. Haardt, Mon. Not. Roy. Astron. Soc. **377**, 1711 (2007), arXiv:astro-ph/0701556.

[27] A. Sesana, J. Gair, E. Berti, and M. Volonteri, Phys. Rev. D **83**, 044036 (2011), arXiv:1011.5893 [astro-ph.CO].

[28] A. Klein *et al.*, Phys. Rev. **D93**, 024003 (2016), arXiv:1511.05581 [gr-qc].

[29] M. Bonetti, A. Sesana, F. Haardt, E. Barausse, and M. Colpi, Mon. Not. Roy. Astron. Soc. **486**, 4044 (2019), arXiv:1812.01011 [astro-ph.GA].

[30] E. Barausse, I. Dvorkin, M. Tremmel, M. Volonteri, and M. Bonetti, Astrophys. J. **904**, 16 (2020), arXiv:2006.03065 [astro-ph.GA].

[31] E. Barausse and A. Lapi, (2020), arXiv:2011.01994 [astro-ph.GA].

[32] T. Gehren, J. Fried, P. A. Wehinger, and S. Wyckoff, ApJ **278**, 11 (1984).

[33] J. Kormendy and D. Richstone, Ann. Rev. Astron. Astrophys. **33**, 581 (1995).

[34] A. E. Reines, G. R. Sivakoff, K. E. Johnson, and C. L. Brogan, Nature **470**, 66 (2011), arXiv:1101.1309 [astro-ph.CO].

[35] A. E. Reines, J. E. Greene, and M. Geha, Astrophys. J. **775**, 116 (2013), arXiv:1308.0328 [astro-ph.CO].

[36] V. F. Baldassare, M. Geha, and J. Greene, Astrophys. J. **896**, 10 (2020), arXiv:1910.06342 [astro-ph.HE].

[37] M. J. Reid, A. C. S. Readhead, R. C. Vermeulen, and R. N. Treuhaft, ApJ **524**, 816 (1999), arXiv:astro-ph/9905075 [astro-ph].

[38] R. Schodel *et al.*, Nature **419**, 694 (2002), arXiv:astro-ph/0210426 [astro-ph].

[39] M. J. Reid, K. M. Menten, R. Genzel, T. Ott, R. Schodel, and A. Eckart, The Astrophysical Journal **587**, 208 (2003).

[40] S. Gillessen, F. Eisenhauer, S. Trippe, T. Alexander, R. Genzel, F. Martins, and T. Ott, Astrophys. J. **692**, 1075 (2009), arXiv:0810.4674 [astro-ph].

[41] K. Akiyama *et al.* (Event Horizon Telescope), Astrophys. J. **875**, L1 (2019), arXiv:1906.11238 [astro-ph.GA].

[42] L. Ferrarese and D. Merritt, Astrophys. J. Lett. **539**, L9 (2000), arXiv:astro-ph/0006053.

[43] N. J. McConnell and C.-P. Ma, Astrophys. J. **764**, 184 (2013), arXiv:1211.2816 [astro-ph.CO].

[44] M. Schramm and J. D. Silverman, Astrophys. J. **767**, 13 (2013), arXiv:1212.2999 [astro-ph.CO].

[45] J. Kormendy and L. C. Ho, Ann. Rev. Astron. Astrophys. **51**, 511 (2013), arXiv:1304.7762 [astro-ph.CO].

[46] F. Wang, J. Yang, X. Fan, J. F. Hennawi, A. J. Barth, E. Banados, F. Bian, K. Boutsia, T. Connor, F. B. Davies, and et al., Astrophys. J. **907**, L1 (2021).

[47] M. A. Latif and A. Ferrara, Publ. Astron. Soc. Austral. **33**, e051 (2016), arXiv:1605.07391 [astro-ph.GA].

[48] M. C. Begelman, R. D. Blandford, and M. J. Rees, Nature **287**, 307 (1980).

[49] M. Tremmel, F. Governato, M. Volonteri, T. R. Quinn, and A. Pontzen, "Mon. Not. Roy. Astron. Soc." **475**, 4967 (2018), arXiv:1708.07126.

[50] G. D. Quinlan, New Astron. **1**, 35 (1996), arXiv:astro-ph/9601092.

[51] A. Sesana and F. M. Khan, Mon. Not. Roy. Astron. Soc. **454**, L66 (2015), arXiv:1505.02062 [astro-ph.GA].

[52] A. I. Macfadyen and M. Milosavljevic, Astrophys. J. **672**, 83 (2008), arXiv:astro-ph/0607467.

[53] J. Cuadra, P. J. Armitage, R. D. Alexander, and M. C. Begelman, Mon. Not. Roy. Astron. Soc. **393**, 1423 (2009), arXiv:0809.0311 [astro-ph].

[54] G. Lodato, S. Nayakshin, A. R. King, and J. E. Pringle, Monthly Notices of the Royal Astronomical Society **398**, 1392–1402 (2009).

[55] C. Roedig, M. Dotti, A. Sesana, J. Cuadra, and M. Colpi, Monthly Notices of the Royal Astronomical Society **415**, 3033–3041 (2011).

[56] C. J. Nixon, P. J. Cossins, A. R. King, and J. E. Pringle, Mon. Not. Roy. Astron. Soc. **412**, 1591 (2011), arXiv:1011.1914 [astro-ph.HE].

[57] P. C. Duffell, D. D'Orazio, A. Derdzinski, Z. Haiman, A. MacFadyen, A. L. Rosen, and J. Zrake, Astrophys. J. **901**, 25 (2020), arXiv:1911.05506 [astro-ph.SR].

[58] D. J. Muñoz, R. Miranda, and D. Lai, Astrophys. J. **871**, 84 (2019), arXiv:1810.04676 [astro-ph.HE].

[59] M. Bonetti, A. Sesana, E. Barausse, and F. Haardt, Mon. Not. Roy. Astron. Soc. **477**, 2599 (2018), arXiv:1709.06095 [astro-ph.GA].

[60] J. R. Gair, A. Sesana, E. Berti, and M. Volonteri, *Hadron collider physics. Proceedings, 22nd Conference, HCP 2010, Toronto, Canada, August 23-27, 2010*, Class. Quant. Grav. **28**, 094018 (2011), arXiv:1009.6172 [gr-qc].

[61] E. Barausse, Mon. Not. Roy. Astron. Soc. **423**, 2533 (2012), arXiv:1201.5888 [astro-ph.CO].

[62] A. Sesana, E. Barausse, M. Dotti, and E. M. Rossi, Astrophys. J. **794**, 104 (2014), arXiv:1402.7088 [astro-ph.CO].

[63] F. Antonini, E. Barausse, and J. Silk, Astrophys. J. Lett. **806**, L8 (2015), arXiv:1504.04033 [astro-ph.GA].

[64] F. Antonini, E. Barausse, and J. Silk, Astrophys. J. **812**, 72 (2015), arXiv:1506.02050 [astro-ph.GA].

[65] K. G. Arun, B. R. Iyer, B. S. Sathyaprakash, S. Sinha, and C. Van Den Broeck, Phys. Rev. D **76**, 104016 (2007), [Erratum: Phys.Rev.D 76, 129903 (2007)], arXiv:0707.3920 [astro-ph].

[66] M. Trias and A. M. Sintes, Phys. Rev. D **77**, 024030 (2008), arXiv:0707.4434 [gr-qc].

[67] E. K. Porter and N. J. Cornish, Phys. Rev. D **78**, 064005 (2008), arXiv:0804.0332 [gr-qc].

[68] S. T. McWilliams, J. I. Thorpe, J. G. Baker, and B. J. Kelly, Phys. Rev. D **81**, 064014 (2010), arXiv:0911.1078 [gr-qc].

[69] S. Marsat, J. G. Baker, and T. Dal Canton, Phys. Rev. D **103**, 083011 (2021), arXiv:2003.00357 [gr-qc].

[70] L. London, S. Khan, E. Fauchon-Jones, C. García, M. Hannam, S. Husa, X. Jiménez-Forteza, C. Kalaghatgi, F. Ohme, and F. Pannarale, Phys. Rev. Lett. **120**, 161102 (2018), arXiv:1708.00404 [gr-qc].

[71] S. Marsat and J. G. Baker, (2018), arXiv:1806.10734 [gr-qc].

[72] N. Aghanim *et al.* (Planck), Astron. Astrophys. **641**, A6 (2020), [Erratum: Astron.Astrophys. 652, C4 (2021)], arXiv:1807.06209 [astro-ph.CO].

[73] LISA Science Study Team, "LISA Science Requirements Document," https://www.cosmos.esa.int/documents/678316/1700384/SciRD.pdf/25831f6b-3c01-e215-5916-4ac6e4b306fb?t=1526479841000 (2018).

[74] A. Mangiagli, A. Klein, M. Bonetti, M. L. Katz, A. Sesana, M. Volonteri, M. Colpi, S. Marsat, and S. Babak, Phys. Rev. D **102**, 084056 (2020), arXiv:2006.12513 [astro-ph.HE].

[75] C. L. Rodriguez, B. Farr, V. Raymond, W. M. Farr, T. B. Littenberg, D. Fazi, and V. Kalogera, Astrophys. J. **784**, 119 (2014), arXiv:1309.3273 [astro-ph.HE].

[76] M. Vallisneri, Phys. Rev. **D77**, 042001 (2008), arXiv:gr-qc/0703086 [GR-QC].

[77] A. Toubiana, S. Marsat, S. Babak, J. Baker, and T. Dal Canton, Phys. Rev. D **102**, 124037 (2020), arXiv:2007.08544 [gr-qc].

[78] C. M. Hirata, D. E. Holz, and C. Cutler, Phys. Rev. D **81**, 124046 (2010), arXiv:1004.3988 [astro-ph.CO].

[79] W. H. Press and P. Schechter, Astrophys. J. **187**, 425 (1974).

[80] H. Parkinson, S. Cole, and J. Helly, "Mon. Not. Roy. Astron. Soc." **383**, 557 (2008), arXiv:0708.1382.

[81] A. Dekel and Y. Birnboim, Mon. Not. Roy. Astron. Soc. **368**, 2 (2006), arXiv:astro-ph/0412300.

[82] A. Cattaneo, A. Dekel, J. Devriendt, B. Guiderdoni, and J. Blaizot, Mon. Not. Roy. Astron. Soc. **370**, 1651 (2006), arXiv:astro-ph/0601295.

[83] K. Schaal, V. Springel, R. Pakmor, C. Pfrommer, D. Nelson, M. Vogelsberger, S. Genel, A. Pillepich, D. Sijacki, and L. Hernquist, Mon. Not. Roy. Astron. Soc. **461**, 4441 (2016), arXiv:1604.07401 [astro-ph.CO].

[84] E. Barausse, F. Shankar, M. Bernardi, Y. Dubois, and R. K. Sheth, Mon. Not. Roy. Astron. Soc. **468**, 4782 (2017), arXiv:1702.01762 [astro-ph.GA].

[85] C. Guépin, K. Kotera, E. Barausse, K. Fang, and K. Murase, Astron. Astrophys. **616**, A179 (2018), [Erratum: Astron.Astrophys. 636, C3 (2020)], arXiv:1711.11274 [astro-ph.HE].

[86] P. Madau and M. J. Rees, Astrophys. J. Lett. **551**, L27 (2001), arXiv:astro-ph/0101223.

[87] M. Volonteri, G. Lodato, and P. Natarajan, Mon. Not. Roy. Astron. Soc. **383**, 1079 (2008), arXiv:0709.0529 [astro-ph].

[88] M. Habouzit, M. Volonteri, and Y. Dubois, "Mon. Not. Roy. Astron. Soc." **468**, 3935 (2017), arXiv:1605.09394 [astro-ph.GA].

[89] Y. Kozai, Astron. J. **67**, 591 (1962).

[90] M. L. Lidov, Planet. Space Sci. **9**, 719 (1962).

[91] M. Bonetti, F. Haardt, A. Sesana, and E. Barausse, "Mon. Not. Roy. Astron. Soc." **461**, 4419 (2016), arXiv:1604.08770.

[92] M. Bonetti, F. Haardt, A. Sesana, and E. Barausse, "Mon. Not. Roy. Astron. Soc." **477**, 3910 (2018), arXiv:1709.06088.

[93] J. M. Bardeen and J. A. Petterson, Astrophys. J. Lett. **195**, L65 (1975).

[94] T. Bogdanovic, C. S. Reynolds, and M. C. Miller, Astrophys. J. Lett. **661**, L147 (2007), arXiv:astro-ph/0703054.

[95] E. Barausse, V. Morozova, and L. Rezzolla, Astrophys. J. **758**, 63 (2012), [Erratum: Astrophys.J. 786, 76 (2014)], arXiv:1206.3803 [gr-qc].

[96] F. Hofmann, E. Barausse, and L. Rezzolla, Astrophys. J. Lett. **825**, L19 (2016), arXiv:1605.01938 [gr-qc].

[97] D. Foreman-Mackey, The Journal of Open Source Software **24** (2016), 10.21105/joss.00024.

[98] I. Mandel, W. M. Farr, and J. R. Gair, Mon. Not. Roy. Astron. Soc. **486**, 1086 (2019), arXiv:1809.02063 [physics.data-an].

[99] E. Parzen, The Annals of Mathematical Statistics **33**, 1065 (1962).

[100] M. Rosenblatt, The Annals of Mathematical Statistics **27**, 832 (1956).

[101] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors,

Nature Methods **17**, 261 (2020).

[102] S. Kullback and R. A. Leibler, Ann. Math. Statist. **22**, 79 (1951).
[103] S.-T. Chiu, The Annals of Statistics **19**, 1883 (1991).
[104] W. M. Farr, Research Notes of the AAS **3**, 66 (2019).