A Flexible Gaussian Process Reconstruction Method and the Mass Function of the Coalescing Binary Black Hole Systems

Yin-Jie Li (李银杰)1,2, Yuan-Zhu Wang (王远瞩)1, Ming-Zhe Han (韩明哲)1,2, Shao-Peng Tang (唐少鹏)1,2, Qiang Yuan (袁强)1,2, Yi-Zhong Fan (范一中)1,2, and Da-Ming Wei (韦大明)1,2

1 Key Laboratory of Dark Matter and Space Astronomy, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023, People’s Republic Of China
2 School of Astronomy and Space Science, University of Science and Technology of China, Hefei, Anhui 230026, People’s Republic Of China

Received 2021 April 4; revised 2021 June 7; accepted 2021 June 7; published 2021 August 12

Abstract

We develop a new method based on Gaussian processes to reconstruct the mass distribution of binary black holes (BBHs). Instead of prespecifying the formalisms of mass distribution, we introduce a more flexible and nonparametric model with which the distribution can be mainly determined by the observed data. We first test our method with simulated data and find that it can well recover the injected distribution. Then, we apply this method to analyze the data of BBHs’ observations from LIGO-Virgo Gravitational-Wave Transient Catalog 2. By reconstructing the chirp mass distribution, we find that there is a peak or a platform located at 20–30\( M_\odot \) rather than a single-power-law-like decrease from low mass to high mass. Moreover, one or two peaks in the chirp mass range of \( M < 20 \ M_\odot \) may be favored by the data. Assuming a mass-independent mass ratio distribution of \( p(q) \propto q^{-1.4} \), we further obtain a distribution of primary mass and find that there is a feature located in the range of \( (30, 40) \ M_\odot \), which can be related to BROKEN POWER LAW and POWER LAW + PEAK distributions described in Collaboration et al. Besides, the merger rate of BBHs is estimated to be \( \mathcal{R} = 26.29^{+14.21}_{-8.96} \) Gpc\(^{-3} \) yr\(^{-1} \), supposing that there is no redshift evolution.

Unified Astronomy Thesaurus concepts: Astrophysical black holes (98); Gravitational wave astronomy (675); Gaussian Processes regression (1930)

1. Introduction

The third observing run (O3) of Advanced LIGO/Virgo has already been accomplished, lasting about 1 yr from 2019 April to 2020 March, and data from the first half of the third observing run (O3a) have been released in the LIGO-Virgo Gravitational-Wave Transient Catalog 2 (GWTC-2; Abbott et al. 2021a). Together with the events observed in the first two observing runs (O1 and O2; Abbott et al. 2019a), the whole catalog consists of about 50 confident gravitational-wave (GW) signals from compact binary coalescences. In addition to the first detected binary black hole (BBH) merger (GW150914; Abbott et al. 2016) and binary neutron star merger (GW170817; Abbott et al. 2017, 2019b), there are some interesting candidates (e.g., GW190426, GW190425, or GW190814) that may have originated from the neutron star–BH mergers (Han et al. 2020; Li et al. 2020; Abbott et al. 2021a, 2021b, 2020). The observations of very massive binaries such as GW190521 (Abbott et al. 2020b) and the extreme mass ratio event GW190814 have challenged the theories of stellar formation and evolution.

Investigating the population of a growing number of BBH events plays an important role in revealing the physical mechanisms associated with the formation/evolution of stars. Some evolutionary channels may produce BBH mergers that can be observed by Advanced LIGO and Virgo, like common envelope evolution (Portegies Zwart & Yungelson 1998), chemically homogeneous evolution (Marchant et al. 2016), isolated binary evolution via the remnants of Population III stars (Inayoshi et al. 2017), and dynamical formation (Kulkarni et al. 1993). Several parameterization models have been proposed and applied to reconstruct the mass distribution of BBH events (Abbott et al. 2019c; Abbott et al. 2021b). They found that the primary mass distribution is not a single power law with an abrupt cutoff, and there must be a feature located at \( \sim 37 \ M_\odot \).

Are there any additional features in the mass spectrum of BBHs? Since we do not know the formation channels of BBHs very well, it may be hard to characterize the mass function of BBHs exactly. Therefore, we propose a new nonparametric method and obtain model-independent results that are fully faithful to the basic observation data. This nonparametric method is based on Gaussian processes (GPs), which are used to reconstruct the chirp mass distribution and estimate the merger rate of BBHs’ population. Our method is distinct from the widely used hierarchical formalism modeling the population properties of BBHs with parametric functions (Thrane & Talbot 2019; Ding et al. 2020; Abbott et al. 2021b; Tiwari & Fairhurst 2021; Wang et al. 2021). Since the chirp mass, which determines the leading-order GW emission during the inspiral, is much better measured than the other intrinsic parameters, we here mainly reconstruct the chirp mass distribution and transform it to the distribution of component mass (primary mass) by assuming a fixed mass ratio distribution of \( p(q) \propto q^{-1.4} \) (Abbott et al. 2021b). Though physical parameters (i.e., component masses and the spins) of BBHs directly relate to the formation/evolution process of their progenitor stars, their measurements always degenerate with each other when performing the parameter estimation on an individual event. Thus, modeling the population properties of chirp mass is more robust (Tiwari & Fairhurst 2021).
2. Method

2.1. Gaussian Processes

GPs are useful tools in GW astronomy and have been implemented in many works, such as the modeling of gravitational waveforms (Doctor et al. 2017; Huerta et al. 2018) and electromagnetic counterpart light curves (Coughlin et al. 2018), the optimization of parameter estimation strategies (Moore & Gair 2014; Moore et al. 2016; Lange et al. 2018), the investigation of the binary stellar evolution (Taylor & Gerosa 2018), and the inference of the equation of state of neutron star matter (Landry & Essick 2019). For the first time, we introduce this approach to investigate the BH mass functions.

We model the population properties of BBHs’ chirp masses in the range of (4.5, 87) $M_{\odot}$. The lower bound (4.5 $M_{\odot}$) is a conservative choice, which is lower than the minimum posterior sample of the lightest BBH observation (Abbott et al. 2021a), while the upper bound (87 $M_{\odot}$) corresponds to an equal-mass system with component masses of 100 $M_{\odot}$. Then, this target range is divided into 30 bins, which have equal length in the logarithm space. We find that 30 bins are reasonably enough to recover the chirp mass distribution of BBHs using current observation data. We have analyzed the data with 50 bins and find that the improvement is minor, i.e., the results of 30 bins and 50 bins are highly consistent with each other. Thus, the distribution of chirp mass $M$ can be given by $r = (r_1, \ldots, r_{N_{\text{bin}}})$, where $r_i$ is the merger rate per unit mass for the BBHs in the $i$th bin and $N_{\text{bin}}$ is the number of bins ($N_{\text{bin}} = 30$). In order to avoid the fluctuation in each bin that is caused by the currently limited number of observations, a flexible GP to model the distribution of chirp masses is

$$\ln r = \begin{bmatrix} \ln r_1 \\ \ln r_2 \\ \vdots \\ \ln r_{N_{\text{bin}}} \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma),$$

where the elements of the covariance matrix $\Sigma$ are determined by a covariance kernel, i.e., $\Sigma_{ij} = K(x_i, x_j) = \text{Cov}(\ln \lambda_i, \ln \lambda_j)$. As for the choice of the kernel type, we take the Matern-3/2 kernel$^4$ $K_{\text{Matern-3/2}}(x, l)$ (Rasmussen & Williams 2006) as the covariance kernel in this work and use the gpflow (Matthews et al. 2017; van der Wilk et al. 2020) to calculate the covariance matrix. We have also tested other kernels, such as the squared-exponential kernel, but they are not better than the Matern-3/2 kernel for this work$^5$. Therefore, the covariance kernel is $K(x_i, x_j) = \sigma^2 K_{\text{Matern-3/2}}(|x_i - x_j|, l)$, where $x_i$ corresponds to the median point of the $i$th bin in logarithmic space, and the hyperparameters $\sigma$ and $l$ determine the functions’ behavior; $\sigma$ sets the amplitude of the function we modeled, while $l$ governs the length scale over which the correlations occur. Longer length scales cause long-range correlations, whereas for short length scales function values are strongly correlated only if their respective inputs are very close to each other.

2.2. Bayesian Inference

2.2.1. Hyperparameters

The prior of chirp mass function $r = (r_1, \ldots, r_{N_{\text{bin}}})$ is generated by the GP: $\pi(r|\mu, \sigma, l)$. In principle, one could choose the mean values $\mu = (\mu_1, \ldots, \mu_{N_{\text{bin}}})$ and hyperparameters $(\sigma, l)$ depending on the prior information of the chirp mass distribution. In our implementation, we set $\mu = 0$, which is a general option in GP. As for $\sigma$, it will potentially affect the smoothness of the function though, while its effect is weaker than that caused by length scale; thus, we fix it in our inference. We find that $\sigma = 5$ is appropriate to generate a prior in our Bayesian inference.$^6$

Another hyperparameter $l$ is the critical parameter for our model, which mainly shapes the function. The models with larger $l$ will ignore the details of the mass distribution, while the models with smaller $l$ will show more details of the mass distribution but may give rise to overfitting. For completeness, we perform several inferences with different length scales and then compare the related evidences.

2.2.2. Likelihood

The construction of the likelihood function in this work is different from the standard hierarchical Bayesian inference (Adams et al. 2012; Thrane & Talbot 2019), widely used in studying the population properties of GW sources (Abbott et al. 2019c; Abbott et al. 2021b; Tiwari & Fairhurst 2021). Given the detection numbers in all bins $k = (k_1, \ldots, k_{N_{\text{bin}}})$, the likelihood can be expressed as the function of expected detection numbers in all bins $\lambda = (\lambda_1, \ldots, \lambda_{N_{\text{bin}}})$, which is

$$\mathcal{L}(k|\lambda) = \prod_{i=1}^{N_{\text{bin}}} e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!}.$$ (2)

However, there are uncertainties in the measurements of chirp mass for individual events, i.e., $k$ remains uncertain; thus, we should integrate the likelihood over the probability of $k$. We define $C_k$ as the set of $k$, and $P_k$ as the probability distribution of $k$, i.e., $k \sim P_k$, so the integrated likelihood becomes

$$\mathcal{L}(P_k|\lambda) = \sum_{k \in C_k} \mathcal{L}(k|\lambda) P_k(k) = \sum_{k \in C_k} \prod_{i=1}^{N_{\text{bin}}} e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} P_k(k).$$ (3)

The likelihood of data $\{d_i\}$ given $\lambda$ can be expressed as

$$\mathcal{L}((d_i)|\lambda) = \sum_{k \in C_k} \mathcal{L}(k|\lambda) \mathcal{L}((d_i)|k),$$ (4)

where $\mathcal{L}((d_i)|k)$ is the likelihood of data $\{d_i\}$ given $k$, which can be reasonably related to $P_k$ ($P_k$ can be directly determined by the data of the observed events $\{d_i\}$) by $P(k) \propto \mathcal{L}((d_i)|k)$. To evaluate $\mathcal{L}((d_i)|\lambda)$, we use 10,000 sets of samples $\{S_1, \ldots, S_{10,000}\}$, and each sample $S_m = (M_1, \ldots, M_{N_{\text{bin}}})$ is randomly chosen from the posterior samples of $N_{\text{obs}}$ observed events, and for each posterior sample $M_t$, the probability of being chosen is $\propto \frac{1}{\pi(\theta)}$ ($\pi(\theta)$ is its default prior). By counting the numbers of samples in each chirp mass bin for a given $S_m$, we can get a corresponding $k_m$, i.e., $\{S_1, \ldots, S_{10,000}\}$ is converted to $\{k_1, \ldots, k_{10,000}\}$. Thus, Equation (4) can be

$^4$ https://gpflow.readthedocs.io/en/master/notebooks/advanced/kernels.html

$^5$ As introduced in Rasmussen & Williams (2006), the squared-exponential kernel provides a strong smoothness that sometimes is unrealistic for modeling some physical processes, and alternatively, the Matern-3/2 kernel may be more flexible.

$^6$ If we choose a smaller one like $\sigma = 2$, the prior is too narrow, which will bias the results of inference; if we choose a larger one like $\sigma = 10$, the inference becomes more time-consuming.
Note that the mass functions are normalized.

\[ \lambda_i = \sum_{s \in \Omega(0,2),03a} \int \mathcal{P}_{\text{det}}(r_i) \, dV \frac{1}{1+z} \, dz \, d\Omega = F(r_i), \]

where \( P_{\text{det}}(z) \) is the probability that an event with parameters \( \theta \) at redshift \( z \) is detectable, \( dV_c/dz \) is the differential comoving volume per unit redshift, and \( r_i \) is the merger rate of the BBHs with chirp masses in the \( i \)th bin. Following Abbott et al. (2021b), we calculate \( F(r_i) \) with injections. Using a Monte Carlo integral over the found injections, Equation (6) is estimated as

\[ \lambda_i = \sum_{s \in \Omega(0,2),03a} \frac{r_i \langle VT \rangle_{N_{\text{inj},i}}}{N_{\text{inj},i}} \sum_{j=1}^{N_{\text{inj},i}} \frac{p_j(\theta_j)}{P_{\text{inj}}(\theta_j)}, \]

where \( p_j(\theta_j) \) is the normalized probability distribution of the assumed BBH population in the \( i \)th bin and \( P_{\text{inj}}(\theta_j) \) is for the injection campaign. Since the number of bins is sufficient (i.e., the bins are narrow enough) and the mass ratio distribution of the injection campaign (i.e., a power law with index \( \beta_q = 2 \)) is close to the inferred distribution in Abbott et al. (2021b; i.e., power law with index \( \beta_q = 1.4^{\pm 0.4}_{} \)), we can make the assumption that the mass ratio and chirp mass in each bin share the same distribution as the injection campaign. As for spin parameters, we use the distribution inferred by the default model in Abbott et al. (2021b). Thus, we have \( p_j(\theta_j) / P_{\text{inj}}(\theta_j) = p(\theta_j) / P_{\text{inj}}(\theta_j) = p(\theta_j) / P_{\text{inj}}(\theta_j) \).

Putting Equation (6) into Equation (5), the likelihood of data \( \{d_i\} \) given the distribution of chirp mass \( r \) can be rewritten as

\[ \mathcal{L}(\{d_i\} | r) = \sum_{k \in \{\theta_1, \ldots, \theta_{100}\}} \prod_{i=1}^{N_{\text{inj}}} e^{-F(\theta_i)} \frac{F(r_i)^{k_i}}{k_i!}, \]

### 3. Reconstruction with Simulated Data

To check the performance of our method, we carry out the reconstruction procedure with the simulated data. We assume an initial distribution for chirp mass, which we are going to recover, as

\[ P(M | m_{\text{min}}, m_{\text{max}}, \alpha, \mu, \sigma, \lambda_{\text{peak}}) = (1 - \lambda_{\text{peak}}) P \times (M - \alpha, m_{\text{min}}, m_{\text{max}}) + \lambda_{\text{peak}} G(M | \mu, \sigma), \]

where \( P(M - \alpha, m_{\text{min}}, m_{\text{max}}) \) is a normalized power-law distribution with spectral index \( -\alpha \) and low-mass (high-mass) cutoff \( m_{\text{min}} \), \( G(M | \mu, \sigma) \) is a normalized Gaussian distribution with the mean \( \mu \) and width \( \sigma \), and \( \lambda_{\text{peak}} \) is the fraction of the Gaussian component. Here, \( m_{\text{min}}, m_{\text{max}}, \alpha, \mu, \sigma, \) and \( \lambda_{\text{peak}} \) are set to \( 5 M_\odot, 80 M_\odot, 3, 25 M_\odot, 5 M_\odot, \) and \( 0.12 \), respectively. The distribution for mass ratio follows a power law with \( p(\theta) \propto q^2 \) and \( m_2 > 2 M_\odot \). For simplicity, the orbit-aligned dimensionless spin parameters are fixed to zero. In the injection, we assume that the merger rate is uniformly distributed in the source frame with a boundary of \( m_1 = 10,000 \text{ Mpc} \), and the sky locations and binary orientation are set isotropic. Besides, we use the waveform template IMRPHENOMD (Husa et al. 2016) to generate the simulated signals and inject them into the noise generated by the representative ADVANCED LIGO MID NOISE power spectral density (Abbott et al. 2018), which is appropriate for the state of LIGO in O3. We define a BBH event being “detected” when its network signal-to-noise ratio is > 12. As for the parameter estimation for each “detected” event, we adopt the same waveform template as injection (i.e., IMRPHENOMD) and use the user-friendly GW parameter inference package BAYES (Ashton et al. 2019a) and nested sampling sampler PyMultiNest (Buchner 2016).

By applying the method described in Section 2, we reconstruct the chirp mass function in the cases of 40, 300, and 1000 simulated detections. As shown in the left column of Figure 1, in each case we provide the Bayes factors of the models with different length scales, and each Bayes factor is evaluated relative to the model with the strongest Bayesian evidence. The right column of Figure 1 shows the normalized distributions of chirp mass with the most favored length scales, and we can find that when the number of detected events increases, the injected distribution is more accurately recovered. In the case of 40 detections, the distribution in high-mass range rises up, and this may be caused by the fluctuation of detections around the lower boundary, which can be avoided with more detections. Generally, our method is capable of recovering the underlying mass distribution of BBHs.

The reconstructions with other length scales are displayed in Figure 4 in Appendix B. As expected, it is harder to recover the detail of the underlying structure with the mass function with larger \( l \), while the mass function with shorter \( l \) is overfitted and

---

7 The injection campaign’s data are now publicly available in [https://dcc.ligo.org/LIGO-P2000434/public](https://dcc.ligo.org/LIGO-P2000434/public).

8 The total number of injections is \( 7.7 \times 10^7 \) for O3a and \( 7.1 \times 10^7 \) for O1 and O2 (Abbott et al. 2021b).

9 Note that the mass functions are normalized.
Jeffreys (1961) pointed out that $\log_{10} B > 1$ can be interpreted as a strong (very strong) preference for one model over another, and $\log_{10} B > 2$ as decisive evidence. Thus, considering the mass distributions (with different length scales) and their corresponding Bayes factors, we can exclude the fake structures in the low-mass range (i.e., $<20 M_\odot$) and determine whether the distribution that ignores the peak at $25 M_\odot$ is disfavored or not, especially for the case with 300 and 1000 detections.

4. Reconstruction with Data of GWTC-2

In this section, we apply our method to study the BBH population properties using the real data from GWTC-2, which includes the BBH observations during O1, O2, and O3a. Following Abbott et al. (2021b), we choose the false-alarm rate of 1 yr$^{-1}$ as the threshold to select the events. To increase the purity of the sample and avoid misidentification of BBH (Tang et al. 2020), the GW190426, GW190719, and GW190909 are not included in this analysis. In addition, we further exclude the unusual asymmetry system GW190814 (Abbott et al. 2020a), which is a significant outlier (Abbott et al. 2021b). For BBH events in O1 and O2, we use the "overall" posterior samples released in Abbott et al. (2019a),10 while for BBH events in O3a, the “publication” posterior samples are acquired from Abbott et al. (2020a).11

Our main results are reported in Figure 2, where we provide the Bayes factors (relative to the most favored model) of the models with different length scales (as shown in the left panels) and the most favored chirp mass distribution (with best length scale) reconstructed from the observed data (as shown in the left panels). Interestingly, we find that there is an additional feature located in the range of $(20, 30) M_\odot$, and there might be a peak located in $(5, 10) M_\odot$. Meanwhile, abrupt cutoff in the high-mass range is significantly disfavored. The reconstructions with different length scales are presented in Figure 5 of

---

10 https://dcc.ligo.org/LIGO-P1800307/public
11 https://dcc.ligo.org/LIGO-P2000061/public
Appendix B. For the results with small length scales, like $l = 0.3$ and $l = 0.5$, there are three peaks located at $(5, 10) M_\odot$, $(10, 20) M_\odot$, and $(20, 30) M_\odot$, respectively. When $1 < l < 1.5$, the two peaks in the low-mass range disappear, while the one in $(20, 30) M_\odot$ still exists. As for the results with large length scales, like $l = 5$, all additional features disappear, and the distribution becomes single-power-law-like. Based on the Bayes factors corresponding to the models with different length scales, we conclude that there must be some additional features in the chirp mass function of BBHs rather than a single-power-law-like distribution. The peak in $(20, 30) M_\odot$, or maybe a break at $\sim 30 M_\odot$, is supported by current observations, while the additional features in the range of $(5, 10) M_\odot$ and $(10, 20) M_\odot$ cannot be confidently identified yet, which will be clear with future observations.

The reconstruction of component mass function from the inferred chirp mass distribution depends on the distribution of mass ratio. As pointed out by Tiwari & Fairhurst (2021), there is no significant evidence for the mass ratio having a mass-dependent distribution. Therefore, we simply assume the distribution of mass ratio $p(q) \propto q^{1.4}$ as obtained in The Abbott et al. (2021b), and additionally, the secondary mass is constrained to be $> 2 M_\odot$. Then, the distribution of primary mass can be obtained (more details are presented in Appendix A). As expected, it has a feature similar to the distribution of chirp mass. Figure 3 shows the distribution of primary mass that is converted from the chirp mass distribution (reconstructed with $l = 0.9$). We notice that the astrophysical BBHs’ primary mass function is not single-power-law-like either, and there must be some features presenting in the range of $30 – 40 M_\odot$. Four distributions reconstructed by Abbott et al. (2021b) are also presented in Figure 3 for comparison. The peak in our model that is located in $(30, 40) M_\odot$ is lower than the peak of the POWER-LAW PEAK model (Abbott et al. 2021b), while if comparing to the BROKEN POWER-LAW model, there is a depression in the range of $(20, 30) M_\odot$ for our model. In the high-mass range, there is no significant peak or cutoff; this may be explained by the fact that we obtain the primary mass distribution from the reconstructed chirp mass function with a fixed mass ratio distribution.

Assuming that the merger rate density does not evolve with redshift, we integrate the BBH merger rates over the whole surveyed chirp mass range. This merger rate is then estimated to be $26.29_{-8.96}^{+14.21} \text{Gpc}^{-3} \text{yr}^{-1}$ for the model of $l = 0.9$. The merger rates of BBHs estimated with $l$ in the range of $(0.2, 2.0)$ are found to be rather similar.

5. Summary and Discussion

We introduced a new method that relies on GPs for reconstruction of the mass function of BBHs and found that our method can well recover the injected distribution of...
simulated events. Comparing to the existing parameterization methods, our method is more faithful to the basic observation data, and it is able to distinguish the underlying features of mass distribution of BBHs when we know little about their formation processes. We applied it to the real observed data and obtained the chirp mass distributions reconstructed by models with different length scale \( l \). By comparing the Bayes factors among models (with different \( l \)), we found that the chirp mass of a single-power-law-like distribution is disfavored. On the contrary, there must be a peak located in \((20, 30) \, M_\odot\); this may be contributed by the Population III BHs (Kinugawa et al. 2021). Meanwhile, there may be one or two peaks in the range of \(<20 \, M_\odot\), which is consistent with the results of Belczynski et al. (2020). Though there is no abrupt mass cutoff in our surveyed chirp mass range, the distribution function drops rather rapidly in the range of \(30 \, M_\odot < \mathcal{M} < 60 \, M_\odot\). Such a quick drop may be due to the superposition of the mass cutoff induced by the pair pulsational instability supernovae and the emergence of a more massive but subdominant population from, for example, the hierarchical mergers (see Wang et al. 2021, and the references therein).

In order to reconstruct the distribution of component masses, we assume a mass ratio distribution \( p(q) \propto q^{-1.4}\) as obtained in Abbott et al. (2021b). The distribution of primary mass turns out to the same feature as that of chirp mass, i.e., there is a peak located in \((30, 40) \, M_\odot\); this feature echos with those of POWER LAW + PEAK and BROKEN POWER LAW in Abbott et al. (2021b). Finally, we also calculate the merger rate of \(26.29^{+14.31}_{-8.96} \, \text{Gpc}^{-3} \, \text{yr}^{-1}\), which is consistent with that found in Abbott et al. (2021b).

Our current method is not able to model the mass ratio and spin properties of the BBH population together with chirp mass, and dedicated efforts to model these parameters are left to future work. Nevertheless, in the approach of a mixture of weighted Gaussians to reconstruct the chirp mass distribution together with spin parameters and mass ratio, Tiwari & Fairhurst (2021) found that neither the mass ratio nor the aligned spin distributions show significant dependence on the mass. These authors also found two peaks located at \(7.1\) and \(13.6 \, M_\odot\) in the chirp mass function, which is consistent with some of our inferences (with small length scales, like \( l = 0.5\)). However, whether these features are real or caused by the statistical fluctuation is still being clarified with more observations in the upcoming O4/O5 runs of the LIGO/Virgo/KAGRA detectors (Abbott et al. 2018; see also https://dcc.ligo.org/public/0161/P1900218/002/SummaryForObservers.pdf).

We thank the referee for the very helpful comments and suggestions. This work was supported in part by NSFC under grant Nos. 11921003, 11933010, and 12073080; the Chinese Academy of Sciences via the Strategic Priority Research Program (grant No. XDB23040000); and the Key Research Program of Frontier Sciences (No. QYZDJ-SSW-SYS024). This research has made use of data and software obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration, and the Virgo Collaboration. LIGO is funded by the U.S. National Science Foundation. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN), and the Dutch Nikhef, with contributions by Polish and Hungarian institutes.

**Software:** Bilby (Ashton et al. 2019b, version 0.5.5, ascl:1901.011, https://git.ligo.org/lscsoft/bilby/, PyMultiNest (Buchner 2016, version 2.6, ascl:1606.005, https://github.com/JohannesBuchner/PyMultiNest, PyCBC (Biwer et al. 2019; Nitz et al. 2019, gwastrophy: PyCBC Release v1.14.1, https://github.com/gwastrophy/pycbc/tree/v1.14.1), gpflow (Matthews et al. 2017; van der Wilk et al. 2020, https://gpflow.readthedocs.io/en/master/intro.html).

**Appendix A**

**Reconstruction of Primary Mass**

We reconstruct the distribution of primary mass from the chirp mass distribution and a mass-independent mass ratio distribution \( p(q) \propto q^{-1.4}\), additionally, the secondary mass \( m_2 \) is constrained to be \( >2 \, M_\odot\). The details of operation are as follows.

For each posterior distribution of chirp mass \( r \) (that is obtained in the Bayes inference), we first sample a sufficient number of samples \( \{\mathcal{M}'_j\} \) from each bin; since the bins are narrow enough, it is reasonable to assume that the samples are uniformly distributed in each bin, and the number of samples of \( i^{th} \) bin \( N_i \) is proportional to the corresponding merger rate \( r_i \) and width of bin \( \Delta \mathcal{M}_i \) (i.e., \( N_i \propto r_i \times \Delta \mathcal{M}_i \)). For each chirp mass sample \( \mathcal{M}'_j \), we sample a mass ratio \( q'_j \) from power law: \( p(q|alpha = 1.4, q_{min} = 1) \), where \( q_{min} \) can be calculated from \( \mathcal{M}'_j \) and \( m_2 > 2 \, M_\odot\). Then, we can obtain the samples of primary mass \( \{m'_j\} \) from \( \{\mathcal{M}'_j, q'_j\} \), and the distribution of primary mass can be obtained by Gaussian kernel density estimation.

**Appendix B**

**Reconstruction with Different Length Scales**

For each scenario of simulation, we display three reconstructions with different length scales (see Figure 4). Referring to the corresponding Bayes factors in Figure 1, we can rule out the single-power-law-like distribution and identify the real structures (i.e., the peak at \(25 \, M_\odot\) in the injected distribution; additionally, we can rule out the fake features that are caused by the fluctuation of data as shown in the left column of Figure 4.

Figure 5 shows the different reconstructions from real data. Considering their related Bayes factors in Figure 2, we can identify the peak in \((20, 30) \, M_\odot\) and rule out the single-power-law-like distribution, while the two peaks in \((5, 10) \, M_\odot\) and \((10, 20) \, M_\odot\) cannot be fully confirmed yet.
Appendix C

Result Checking

Section 4 shows the inferred chirp mass distribution of BBHs with our method. Here we use the inferred distribution to predict an observed distribution and compare it to the empirical distribution of observed chirp mass distributions in Figure 6. Note that empirical distribution is calculated from the reweighted samples of the posteriors of the observed events (as introduced in Section 2.2.2) rather than directly calculated from their posteriors. From Figure 6 we can see that the dark-colored band overlaps with the light-colored band in general; while our model does not perform so well on the two edges of the mass range, this may be because we fix the lower and upper bounds of the chirp mass (as 4.5 and 87 $M_\odot$) when fitting.

Figure 4. Normalized distributions of chirp mass reconstructed with several different length scales upon 40 (top), 300 (middle), and 1000 (bottom) simulated detections.

Figure 5. Distributions of chirp mass reconstructed with several different length scales upon real data.
Figure 6. A posterior predictive check: the cumulative density function of the observed chirp mass distribution for the reconstruction with length scale $l = 0.9$. The observed event distribution is shown in the darker colors. The thickness of the bands indicates the 90% credibility range.