PTEP

Prog. Theor. Exp. Phys. **2021** 123F01(18 pages)
DOI: 10.1093/ptep/ptab137

A nonparametric method to assess the significance of events in the search for gravitational waves with false discovery rate

Hirotaka Yuzurihara[1,*], Shuhei Mano[2], and Hideyuki Tagoshi[3]

[1]*Institute for Cosmic Ray Research, The University of Tokyo, Higashi-Mozumi 238, Kamioka-cho, Hida-shi, Gifu 506-1205 Japan*
[2]*The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan*
[3]*Institute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8582, Japan*
*E-mail: yuzu@icrr.u-tokyo.ac.jp

Received May 6, 2021; Revised October 11, 2021; Accepted October 18, 2021; Published October 23, 2021

We present a consistent procedure to assess the significance of gravitational wave events observed by laser interferometric gravitational wave detectors based on the background distribution of the detection statistic. We propose a non-parametric method to estimate the p-value. Based on the estimated p-values, we propose a new procedure to assess the significance of a particular event with a q-value which is the minimum false discovery rate that can be attained when calling the event significant. The q-value gives us a criterion on the significance of events which is different from $P_{\rm astro}$ as used in the LIGO–Virgo analysis and others. The proposed procedure is applied to the 1-OGC and 2-OGC catalogs. For most of the events which were claimed significant in these catalogs, we obtain the same results. However, there are differences in the significance for several marginal events. Since the proposed procedure does not require any assumptions on signal and noise, it is very simple and straightforward. The procedure is also applicable to other searches for gravitational waves whose background distribution of the detection statistic is difficult to know.

Subject Index F33, F34

1. Introduction

The first gravitational wave event from binary black hole (BBH) coalescence, GW150914, was observed by the advanced LIGO detectors in their first observing run (O1) [1]. After the first detection, tens of gravitational wave events were reported [2]. During the second observing run (O2), the first gravitational waves from a binary neutron star coalescence, GW170817 [3], were observed by LIGO [4] and Virgo [5]. The follow-up observations by electromagnetic telescopes identified the host galaxy in NGC 4993. The event strongly suggests the existence of radioactive decay from a rapid neutron-capture process [6]. The discovery of these events has opened up gravitational wave astronomy. During the third observing run (O3), many candidate events were reported [7], and four events have been published individually [8–11]. Very recently, the GWTC-2 catalog, which reports the gravitational wave signals from compact binary coalescences during the first half of the O3 observation, was released [12]. In the coming years the network of gravitational wave detectors consisting of two LIGO detectors, Virgo, and

KAGRA [13] plans to perform coincident observation runs. As the detectors' sensitivities improve and the observation time becomes longer, we expect to observe more and more gravitational wave events.

In compact binary coalescence searches, we search for gravitational wave signals by maximizing the detection statistic over the template bank in a short time window. When the value of the detection statistic exceeds a given threshold, we record it as a trigger. Accordingly, for a given threshold, as the observing time and the template bank becomes larger, the probability of false triggers produced by noise (false alarm probability) becomes larger. This is called the multiple comparisons problem. Several methods have been proposed to control the false alarm probability. The Bonferroni correction is one of these [14, Chapter 9]. However, these methods generally reduce the detection probability while controlling the false alarm probability.

Recently, the false discovery rate (FDR) was proposed to treat these problems (see Sect. 3 for the formal definition of the FDR). To the author's best knowledge, the first introduction of FDR to the gravitational wave community was in Ref. [15], but the paper did not discuss any actual problems. Recently, $P_{\rm astro}$ was introduced as a measure of true discovery of a particular event [16]. In the recent catalog of gravitational waves from compact binary mergers [2], a candidate event is considered to have a gravitational wave origin if the false alarm rate is less than one per 30 days and $P_{\rm astro}$ is larger than 0.5.

In this paper we propose the use of the *q-value*, which is a measure of FDR. We present a consistent procedure to assess the significance of candidate events by using the q-value. We first introduce a definition of the p-value by using the background distribution of the detection statistic. Then, we propose a new procedure to evaluate the q-value of each event by extending the procedure proposed in Ref. [17]. Their original procedure is not applicable for a search for gravitational waves from compact binary coalescences, because it requires a complete list of p-values. However, in gravitational wave searches a complete list of p-values is usually not available because we store only triggers whose detection statistic is larger than a certain threshold. We apply these procedures to the publicly available analysis results, the analysis, the 1-OGC and 2-OGC catalogs [18,19], and evaluate the q-value of each candidate event. We compare the significance of each candidate event evaluated by using $P_{\rm astro}$. We find that we obtain almost consistent results on the significance of each candidate event. However, we also find that, although the conclusion on the significance may change depending on the threshold for the q-value and $P_{\rm astro}$, the conclusion on the significance of events can be different for marginally significant events. We find one such event in the 2-OGC catalog.

The main advantage of our procedure is that it is completely nonparametric, i.e. we do not assume any parametric model behind the data. Our procedure can be applied to other gravitational wave searches. The evaluation of the p-value in a non-parametric way, the procedure to evaluate the q-value, and estimation of q-values for the LIGO–Virgo O1 and O2 candidate events by using this procedure are all new things in this paper.

The paper is organized as follows. In Sect. 2 we discuss statistical hypothesis testing in the search for gravitational waves from compact binary coalescences. In Sect. 3 we present a procedure to assess the significance of a particular event with a false discovery rate. In Sect. 4, the proposed procedure is applied to the results of the analysis of the O1 data. Section 5 is devoted to a summary and discussion.

2. Estimation of *p*-value

We first introduce the statistical terminology used in this paper. The definitions statistical terminology can be found in a standard textbook, such as Ref. [14]. By analyzing the data from gravitational wave detectors, we obtain *events* which have larger signal-to-noise ratio than a threshold. Each event is classified as either *signal* or *noise*. If the event originates from a gravitational wave, it is called a signal. Otherwise, it is called noise. In the statistical literature, the noise model is called the *null hypothesis* (in this paper, also called *background*) and the signal model is called the *alternative hypothesis*.

In the analysis of gravitational waves from compact binary coalescences, event search is done by maximizing the detection statistic over the templates. The detection statistic is also maximized over time within a certain time length.

In statistical hypothesis testing, the *p-value* of an event is a measure of the significance of the event. It is the probability that the event or rarer events occur under the null hypothesis. If the *p*-value of the event is significantly small, the null hypothesis is rejected. Let us consider the statistical hypothesis testing of each event based on the background distribution of the detection statistic.

2.1 *A conventional p-value*

In the LIGO–Virgo O1 analysis, the following *p*-value was used [20,21] (see Appendix A for a discussion of its derivation):

$$p_{\rm conv}(\rho) = 1 - e^{-\mu(\rho)}, \qquad \mu(\rho) = \frac{n_{\rm bg}(\rho)}{t_{\rm bg}} t_{\rm obs}, \tag{1}$$

where ρ is the detection statistic of an event. In this paper we call this the conventional *p*-value. Here, $t_{\rm obs}$ and $t_{\rm bg}$ are the time lengths of the analyzed data and time length for the estimation of the background distribution, respectively. The estimation of the background data is usually generated by time-shifting data from different detectors [21]. Moreover, $n_{\rm bg}(\rho)$ is the number of noise events in the background data whose detection statistics are equal to or larger than ρ. It is

$$n_{\rm bg}(\rho) = \sum_{i=1}^{n_{\rm bg}(0)} 1_{\{r_i \geq \rho\}}, \tag{2}$$

where r_i is the detection statistic of the ith event in the background data, and $1_{\{\,\cdot\,\}} = 1$ if $\{\,\cdot\,\}$ is true and 0 otherwise. From the definition, $n_{\rm bg}(0)$ is the total number of noise events in the background data. Therefore, $\mu(\rho)$ in Eq. (1) is the mean of number of events whose detection statistics are greater than or equal to ρ. The ratio $n_{\rm bg}(\rho)/t_{\rm bg}$ is usually called the *false alarm rate* of the event whose detection statistic is ρ.

2.2 *Nonparametric estimation of p-value*

Now we introduce a non-parametric method to estimate the *p*-value. Let us assume the background distribution is continuous. If we know the probability density function of the detection statistic under the null hypothesis, $f(r)$, the *p*-value of an event whose detection statistic is ρ is given by

$$p(\rho) = \int_{\rho}^{\infty} f(r)\, dr = 1 - F(\rho), \qquad F(\rho) = \int_{0}^{\rho} f(r)\, dr. \tag{3}$$

In reality, the background distribution is unknown; nevertheless, it can be estimated non-parametrically (free from the assumption of a parameterized distribution) by using simulated

Table 1. Outcomes and counts.

	Significant	Not significant	Total
Noise	F	$n_0 - F$	n_0
Signal	T	$n_1 - T$	n_1
Total	S	$n_{\rm obs} - S$	$n_{\rm obs}$

background data. An estimator of the null distribution F is given by

$$\hat{F}(\rho) := \frac{1}{n_{\rm bg}(0)} \sum_{i=1}^{n_{\rm bg}(0)} 1_{\{r_i \leq \rho\}}.$$

It is important to distinguish F and $\hat{F}$. The former is the (unknown) true background distribution, while the latter is an estimator of the background distribution. By the Glivenko–Cantelli theorem, $\hat{F}$ converges to F almost surely and uniformly in ρ [14]. Therefore, an estimator of the p-value of an event whose detection statistic is ρ is given by

$$\hat{p}(\rho) = 1 - \hat{F}(\rho) := \frac{n_{\rm bg}(\rho)}{n_{\rm bg}(0)}, \tag{4}$$

where we used the fact that $\rho \neq r_i$, $i = 1, \ldots, n_{\rm bg}(0)$. Note that $\hat{p}(\rho)$ is the probability of obtaining the event whose detection statistic is larger than ρ in the background data and has been called (an estimator of) the *false alarm probability* in the gravitational wave community [22]. In addition, $\hat{p}(\rho)$ is proportional to the mean $\mu(\rho)$ in Eq. (1). The estimator in Eq. (4) is a consistent estimator of the p-value in Eq. (3), i.e. $\hat{p}(\rho)$ converges to $p(\rho)$ almost surely for each ρ by the strong law of large numbers.

For later discussion, let us recall a basic property of p-values. The p-value of a statistic ρ following any continuous null distribution $F(\rho)$ follows the uniform distribution, because

$$\begin{aligned}\mathbb{P}(p(\rho) < u) &= \mathbb{P}(\rho > F^{-1}(1-u)) = 1 - \mathbb{P}(\rho \leq F^{-1}(1-u)) \\ &= 1 - F(F^{-1}(1-u)) = u\end{aligned}$$

is the distribution function of the uniform distribution where $0 \leq u \leq 1$ and $\mathbb{P}(x)$ is the probability of x. It is worth mentioning that we cannot expect the conventional p-value given by Eq. (1) with ρ following $F(\rho)$ to follow the uniform distribution (see Appendix A). In the discussion that follows, we discuss the p-value defined by Eq. (3).

3. Assessment of significance with false discovery rate

In this section we describe statistical hypothesis testing using detection statistics and how to assess significance with the false discovery rate. When we perform the statistical test, each event can be categorized as one of four possible outcomes, which are summarized in Table 1.

There are two kinds of truth (noise or signal) and two kinds of claim (significant or not significant). F and T are the number of noise and signal events called significant, respectively, and S is the total number of events called significant; n_0 and n_1 are the number of noise and signal statistics, respectively, and $n_{\rm obs}$ is the total number of events in the observed data.

In statistical hypothesis testing, a p-value threshold is selected to keep the number of false positives F small. When we select the threshold α, the expected number of false positive is $\alpha n_{\rm obs}$. If $n_{\rm obs}$ is very large, α should be selected to be very small.

Here, the probability $\mathbb{P}(F \geq 1)$ is called a *familywise error rate*. The familywise error rate is simply called the false alarm probability in the gravitational wave community, but we call it the familywise false alarm probability in this paper to avoid confusion. *Family* means that we test a hypothesis by using $n_{\rm obs}$ tests. To control the familywise error rate such that $\mathbb{P}(F \geq 1) \leq \alpha$, that is, the rate that a noise event is classified as significant is less than α, one of the solutions is to change the threshold α to $\alpha/n_{\rm obs}$. This method is called Bonferroni's procedure [14, Chapter 9].

Unfortunately, controlling the familywise error rate is practical only when extremely few events are expected to be signal. Otherwise, controlling the familywise error rate will be too conservative and the statistical power of the test procedure will be too poor. Benjamini and Hochberg [23] introduced the *false discovery rate*, which is defined as the expected value of F/S, $\mathbb{E}(F/S, S > 0)$, where F and S were introduced in Table 1, and gave a test procedure to keep the FDR less than a threshold. A fairly recent survey of an FDR is Ref. [24]. Note that the false positive rate and the FDR are quite different measures. A false positive rate of 5% means that 5% of noise events are called significant. On the other hand, an FDR of 5% means that 5% of events called significant are noise events. Controlling the FDR should be more powerful than controlling the familywise error rate, since the FDR is less than or equal to the familywise error rate [23].

Storey and Tibshirani [17] introduced the *q-value* for a particular event, which is the expected proportion of false positives incurred if calling the event significant. Let us define FDR(u), which is the FDR when calling all events significant whose p-value is less than or equal to a threshold u with $0 < u \leq 1$: namely,

$$\mathrm{FDR}(u) = \mathbb{E}\left[\frac{F(u)}{S(u)}, S(u) > 0\right], \tag{5}$$

where $\mathbb{E}(x, y > 0)$ is the expectation of x given $y > 0$. Here, $F(u)$ is the number of noise events whose p-value is smaller than or equal to the threshold u, and $S(u)$ is the number of both noise and signal events whose p-value is smaller than or equal to the threshold u. The definition of the q-value is the minimum FDR that can be attained when calling the event significant, namely,

$$q_i := \min_{u \geq p_i} \mathrm{FDR}(u), \tag{6}$$

where $i = 1, \ldots, n_{\rm obs}$ and the p-value given by Eq. (3) of the ith event is denoted by p_i. Note that FDR(u) is not always monotonically increasing in the threshold u. Taking the minimum guarantees that the estimated q-value is increasing in the same order as the p-value.

Let us recall the procedure for estimating the q-value proposed in Ref. [17]. Their estimator of FDR(u) is

$$\widehat{\mathrm{FDR}}(u) = \frac{\hat{\pi}_0 n_{\rm obs} u}{S(u)}, \tag{7}$$

where $\hat{\pi}_0$ is an estimator of $\pi_0 = n_0/n_{\rm obs}$, which indicates the overall proportion of noise events in the data. Roughly speaking, Eq. (7) is a sample mean whose population mean is given by Eq. (5). Since the p-value of a statistic follows the uniform distribution under the null hypothesis (see Sect. 2), the numerator of Eq. (7) is an estimator of $F(u)$.

How to estimate $\hat{\pi}_0$ is the central issue. In gravitational wave searches, very few events are expected to be signal. In such a case, we can assume $\hat{\pi}_0 \simeq 1$. In Appendix B we show that this assumption is justified by using the 1-OGC and 2-OGC catalogs. We thus set $\hat{\pi}_0 = 1$.

We can construct an estimator of the q-value by plugging the estimator of the p-value in Eq. (4) and the estimator of the FDR in Eq. (7) into the expression in Eq. (6) and setting

$\hat{\pi}_0 = 1$. The result is

$$\hat{q}_i = \min_{u \geq \hat{p}_i} \frac{n_{\rm obs} u}{\#\{\hat{p}_j \leq u;\, j \in \{1, \ldots, n_{\rm obs}\}\}}, \tag{8}$$

where $\hat{p}_i = \hat{p}(r_i)$.

4. Application to the 1-OGC and 2-OGC results

In this section we evaluate the q-value of events in the 1-OGC catalog [18] and the 2-OGC catalog [19] using the data available at https://github.com/gwastro/1-ogc and https://github.com/gwastro/2-ogc. The available data set contains information on events such as time, false alarm rate in units of year^{-1}, the value of the ranking statistic, two masses, the dimensionless spin component value of each star perpendicular to the orbital plane, etc. The data set consists of the *complete* and *bbh* data sets. There are 146,214 and 12,741 events respectively in the *complete* and *bbh* data sets of 1-OGC, and 733,231 and 502,994 events in the *complete* and *bbh* data sets of 2-OGC, respectively. The complete data set contains all candidate events from the full analysis, and the bbh data set contains the candidate events from the BBH region targeted analysis [18,19].

Since the p-values of events are not available in these catalogs, we need to evaluate them from the false alarm rates (FARs). An estimate of the FAR is given by $n_{\rm bg}(\rho)/t_{\rm bg}$, where $t_{\rm bg}$ is the length of data used for background estimation, and $n_{\rm bg}(\rho)$ is defined by Eq. (2). The events in the catalog are defined by taking an event which gives a maximum detection statistic within a certain time window Δt and which is in the template bank used in the analysis. Thus, the total number of background events, $n_{\rm bg}(0)$, is given as $t_{\rm bg}/\Delta t$. In both 1-OGC and 2-OGC, $\Delta t = 10\,{\rm s}$ is used. Then, from Eq. (4), we obtain an estimate of the p-value of an event as

$$\hat{p}(\rho) = \frac{n_{\rm bg}(\rho)}{n_{\rm bg}(0)} = {\rm FAR} \times \Delta t. \tag{9}$$

We note that the candidate events in these data sets are not all events in the sense that only events with relatively low false alarm rates are recorded. This is for practical reasons, to reduce the computation time of the analysis. This is a typical situation in gravitational wave analysis.

Since all the candidate events are not available, we cannot use the algorithm originally proposed in Ref. [17], which is explained as Algorithm 2 in Appendix C. Instead, we propose an alternative procedure for estimating the q-value, which is a modified version of Algorithm 2. Appendix C explains why Algorithm 1 yields estimates of the q-value defined in Eq. (8).

Algorithm 1. We compute estimates of the q-value defined in Eq. (8). Let m to be the number of false alarm rates which are less than some value. Assume p-values in the region around and larger than $\hat{p}_{(m)}$ are noise.

1. Compute estimates of the p-value: $\hat{p}_i =$ (false alarm rate of ith event) $\times\, \Delta t$, where $i = 1$, …, m.
2. Let $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(m)}$ be the ordered p-values.
3. Set $\hat{q}_{(m)} = n_{\rm obs}\hat{p}_{(m)}/m$.
4. For $i = m-1, m-2, \ldots, 1$, compute
$$\hat{q}_{(i)} = \min\left(\frac{n_{\rm obs}\hat{p}_{(i)}}{i}, \hat{q}_{(i+1)}\right).$$
5. The estimated q-value for the ith most significant event is $\hat{q}_{(i)}$.

□

4.1 *1-OGC results*

In the 1-OGC catalog [18], the true discovery rate (TDR) and $P_{\rm astro}$ are given to evaluate the significance of events. A true discovery is the complement of a false discovery, FDR $= 1 -$ TDR. Note, however, that the evaluation of TDR in Ref. [18] is a very conservative estimate, defined as

$$\widehat{\rm TDR}(\tilde{\rho}_c) = \frac{\mathcal{T}(\tilde{\rho}_c)}{\mathcal{T}(\tilde{\rho}_c) + \mathcal{F}(\tilde{\rho}_c)}, \tag{10}$$

where $\mathcal{T}(\tilde{\rho}_c)$ is the rate that signals of astrophysical origin are observed with a ranking statistic $\geq \tilde{\rho}_c$, and $\mathcal{F}(\tilde{\rho}_c)$ is the FAR. In Ref. [18], to estimate $\mathcal{T}(\tilde{\rho}_c)$ the two significant events GW150914 and GW151229 were assumed to be real astrophysical signals, and $\mathcal{T} \sim 15\,{\rm yr}^{-1}$ was obtained. In order to take account of the uncertainty in the estimate based on only two events, a Poisson distribution was assumed for the observed number, and as a lower 95% bound, $\mathcal{T} \sim 2.7\,{\rm yr}^{-1}$ was obtained. In Ref. [18], this value is used in the equivalent of Eq. (10) for all events other than GW150914 and GW151226.

On the other hand, $P_{\rm astro}$ is the posterior probability given that a particular event has astrophysical origin. In the 1-OGC catalog [18], it is estimated as

$$P_{\rm astro}(\tilde{\rho}_c) = \frac{\Lambda_{\rm S} P_{\rm S}(\tilde{\rho}_c)}{\Lambda_{\rm N} P_{\rm N}(\tilde{\rho}_c) + \Lambda_{\rm S} P_{\rm S}(\tilde{\rho}_c)}, \tag{11}$$

where $P_{\rm S}(\tilde{\rho}_c)$ and $P_{\rm N}(\tilde{\rho}_c)$ are the probability densities of an event having ranking statistic $\tilde{\rho}_c$ given the event is signal or noise, respectively, and $\Lambda_{\rm S}$ and $\Lambda_{\rm N}$ are the rates of signal and noise events.[1] In order to estimate $\Lambda_{\rm S} P_{\rm S}(\tilde{\rho}_c)$, an analytic model of the signal distribution and a fixed conservative rate of mergers are used by assuming two events (GW150914 and GW151226) are of astrophysical origin.[2]

Figure 1 shows the q-value computed using Algorithm 1 from the p-values of events in the *complete* data set. Table 2 summarizes the results of the estimated p-value and q-value for the 10 most significant events.

Figure 2 shows the q-values computed using Algorithm 1 from p-values of events in the *bbh* data set. Table 3 summarizes the results of the estimated p-values and q-values for the 10 most significant events, together with the inverse of the false alarm rate, $1 - \widehat{\rm TDR}$, and $P_{\rm astro}$ as given in the 1-OGC catalog. For the first two events, since only the upper limit of the false alarm rate was evaluated in Ref. [18], the estimated p-value of these events should be considered an upper limit to the p-value. $1 - \widehat{\rm TDR}$ and $1 - P_{\rm astro}$ are not given for the top two events in Ref. [18], since these events are used to estimate these values for the other events.

Following Ref. [18], we discuss the significance of events in the *bbh* case. In Table 3, if we call the events whose q-value is smaller than 0.05 significant, the top three events are significant. The expected proportion of false discoveries incurred in the three events is less than 0.05. Since the q-value of GW151012 (151012+09:54:43) is 9.83×10^{-5}, this is significant enough as to be a true signal. In Ref. [18], since $P_{\rm astro}$ for GW151012 is 9.76×10^{-1}, which is larger than 0.5, GW151012 is called significant. Thus, the results for the q-value and $P_{\rm astro}$ are consistent for this event.

In Table 3 we find two marginally not significant events, 160103+05:48:36 and 151213+00:12:20, whose q-values are 8.31×10^{-2} and 8.53×10^{-2} respectively. On the other

[1] $P_{\rm astro}$ is also called *purity* in other fields of physics [25].

[2] Note that the method used to estimate $P_{\rm astro}$ in Ref. [18] is different from that used in the GWTC-1 catalog by the LIGO–Virgo collaboration [2] and in the 2-OGC paper [19].

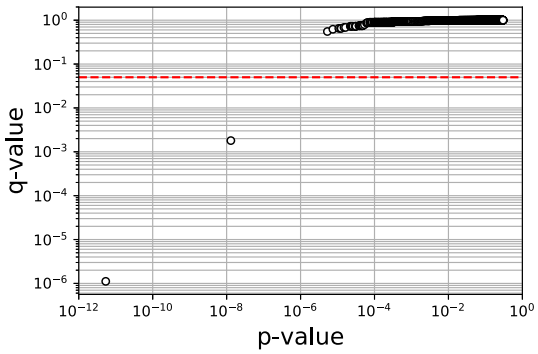

Fig. 1. The q-values computed using Algorithm 1 from p-values of events in the *complete* data set of 1-OGC. The red dashed line indicates a q-value of 0.05.

Table 2. Estimated p-values and q-values of the events of the *complete* data set. Events are sorted by false alarm rate, and the 10 most significant events are shown. The inverse false alarm rates (FAR^{-1}) are obtained from the 1-OGC catalog. The p-values are computed by Eq. (4), and the q-values are computed by Algorithm 1.

UTC time	FAR^{-1} (year)	p-value	q-value
150914+09:50:45	$>6.55 \times 10^{4}$	$<4.84 \times 10^{-12}$	$<1.11 \times 10^{-6}$
151226+03:38:53	$>5.91 \times 10^{4}$	$<5.36 \times 10^{-12}$	$<1.11 \times 10^{-6}$
151012+09:54:43	2.44×10^{1}	1.30×10^{-8}	1.80×10^{-3}
151019+00:23:16	5.96×10^{-2}	5.32×10^{-6}	5.52×10^{-1}
150928+10:49:00	4.24×10^{-2}	7.48×10^{-6}	6.22×10^{-1}
151218+18:30:58	2.93×10^{-2}	1.08×10^{-5}	6.51×10^{-1}
160103+05:48:36	2.63×10^{-2}	1.21×10^{-5}	6.51×10^{-1}
151202+01:18:13	2.53×10^{-2}	1.25×10^{-5}	6.51×10^{-1}
160104+03:51:51	2.12×10^{-2}	1.49×10^{-5}	6.84×10^{-1}
151213+00:12:20	1.93×10^{-2}	1.64×10^{-5}	6.84×10^{-1}

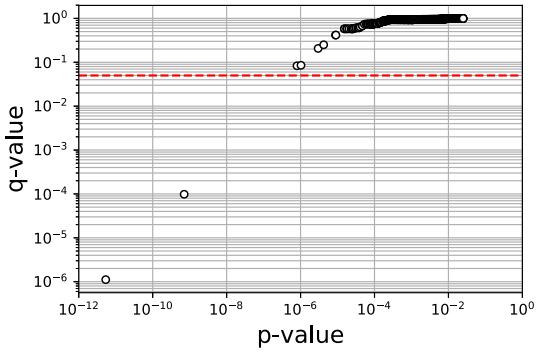

Fig. 2. The q-values computed using Algorithm 1 from p-values of events in the *bbh* data set of 1-OGC. The red dashed line indicates a q-value of 0.05.

hand, $P_{\rm astro}$ for these events is small, at 6.07×10^{-2} and 4.66×10^{-2} respectively. So, in Ref. [18] these two events are called not significant. Although the conclusions are the same, the significance is slightly different between q-value and $P_{\rm astro}$ in Ref. [18], and this difference might be interesting. However, since these two events do not appear in the 2-OGC catalog (see the next subsection), we do not investigate these events further.

Table 3. As Table 2, but obtained from the *bbh* data set. FAR, $1-\widehat{\mathrm{TDR}}$, and $1-P_{\mathrm{astro}}$ are obtained from the 1-OGC catalog.

UTC time	FAR^{-1} (year)	p-value	q-value	$1-\widehat{\mathrm{TDR}}$	P_{astro}
150914+09:50:45	$>6.55\times10^{4}$	$<4.84\times10^{-12}$	$<1.11\times10^{-6}$	—	—
151226+03:38:53	$>5.91\times10^{4}$	$<5.36\times10^{-12}$	$<1.11\times10^{-6}$	—	—
151012+09:54:43	4.46×10^{2}	7.10×10^{-10}	9.83×10^{-5}	8.29×10^{-4}	9.76×10^{-1}
160103+05:48:36	3.96×10^{-1}	8.00×10^{-7}	8.31×10^{-2}	4.83×10^{-1}	6.07×10^{-2}
151213+00:12:20	3.09×10^{-1}	1.03×10^{-6}	8.53×10^{-2}	5.45×10^{-1}	4.66×10^{-2}
151216+18:49:30	1.06×10^{-1}	2.98×10^{-6}	2.07×10^{-1}	7.77×10^{-1}	1.72×10^{-2}
151222+05:28:26	7.51×10^{-2}	4.22×10^{-6}	2.50×10^{-1}	8.31×10^{-1}	1.20×10^{-2}
151217+03:47:49	3.59×10^{-2}	8.82×10^{-6}	4.17×10^{-1}	9.12×10^{-1}	5.99×10^{-3}
151009+05:06:12	3.51×10^{-2}	9.02×10^{-6}	4.17×10^{-1}	9.13×10^{-1}	5.20×10^{-3}
151220+07:45:36	2.07×10^{-2}	1.53×10^{-5}	5.78×10^{-1}	9.47×10^{-1}	3.20×10^{-3}

The value of $1-\widehat{\mathrm{TDR}}$ is about one of magnitude larger than the q-value for all events. Since $\widehat{\mathrm{TDR}}$ in Ref. [18] is a very conservative estimate, this difference is not surprising. Even in this case, $1-\widehat{\mathrm{TDR}}$ for GW151012 is 8.29×10^{-4}. Thus, this can be called significant. But, $\widehat{\mathrm{TDR}}$ for 160103+05:48:36 and 151213+00:12:20 is 0.483 and 0.545. Thus, these cannot be called marginal events.

In the LIGO–Virgo GWTC-1 catalog of gravitational waves from compact binary mergers during O1 and O2 [2], a necessary condition that an event is considered to be a gravitational wave signal is that the FAR of the event is less than one per 30 days, which corresponds to a p-value of $10\,\mathrm{s}/30\,\mathrm{days}=3.9\times10^{-6}$. By linearly fitting the data in Figs. 1 and 2, we can evaluate that this p-value corresponds to q-values of 0.411 and 0.240, respectively. A q-value of 0.05 corresponds to one per 271 days and one per 246 days of FAR, respectively. The q-value threshold of 0.05 is more stringent than the FAR of one per 30 days.

When we compare q-values of the same event, the q-value in Table 3 is smaller than that in Table 2. The reason for this difference is that the events in the data sets are computed from different numbers of templates. A smaller number of templates decreases the false alarm rate and the p-value. Accordingly, it produces a different q-value.

4.2 *2-OGC results*

Figure 3 shows the q-values as a function of the p-values in the *complete* data set. Table 4 summarizes the results of the estimated q-values of the events for the 30 most significant events.

Figure 4 shows the q-values as a function of p-values in the *bbh* data set. Table 5 summarizes the results of the estimated q-values for the top 30 events. P_{astro} computed in 2-OGC paper Ref. [19] is also shown in this table.[3]

We now discuss the significance of events for the *bbh* case. In Table 5, if we call the events whose q-value is smaller than 0.05 significant, the top 13 events are significant. In Ref. [19], these 13 events are called significant since P_{astro} is larger than 0.5. Thus, the results for q-value and P_{astro} are consistent with each other. On the other hand, we obtain a different result for 151205+19:55:25. The q-value of this event is 0.07, while P_{astro} is 0.525. Thus, this is definitely

[3]The method used to estimate P_{astro} in Ref. [19] is based on a mixture model developed in Ref. [16] and employed in the GWTC-1 catalog by the LIGO–Virgo collaboration [2].

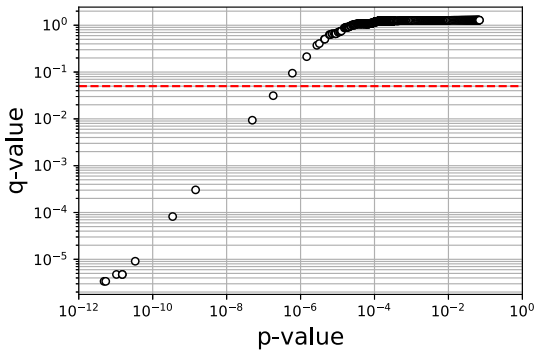

Fig. 3. The q-values computed using Algorithm 1 from p-values of events in the *complete* data set of 2-OGC. The red dashed line indicates a q-value of 0.05.

Table 4. Estimated p-values and q-values of the events from the *complete* data set of 2-OGC. Events are sorted by false alarm rate, and the top 30 events are shown. The inverse false alarm rates (FAR^{-1}) are obtained from the data set. The p-values are computed by Eq. (4), and the q-values are computed by Algorithm 1.

UTC time	FAR^{-1} (year)	p-value	q-value
170104+10:11:58	$>7.27 \times 10^{4}$	$<4.36 \times 10^{-12}$	$<3.38 \times 10^{-6}$
150914+09:50:45	$>6.55 \times 10^{4}$	$<4.84 \times 10^{-12}$	$<3.38 \times 10^{-6}$
151226+03:38:53	$>5.91 \times 10^{4}$	$<5.36 \times 10^{-12}$	$<3.38 \times 10^{-6}$
170823+13:13:58	$>3.03 \times 10^{4}$	$<1.05 \times 10^{-11}$	$<4.75 \times 10^{-6}$
170817+12:41:04	$>2.11 \times 10^{4}$	$<1.51 \times 10^{-11}$	$<4.75 \times 10^{-6}$
170814+10:30:43	$>2.11 \times 10^{4}$	$<1.51 \times 10^{-11}$	$<4.75 \times 10^{-6}$
170809+08:28:21	9.42×10^{3}	3.36×10^{-11}	9.10×10^{-6}
170608+02:01:16	$>9.15 \times 10^{2}$	$<3.46 \times 10^{-10}$	$<8.19 \times 10^{-5}$
151012+09:54:43	2.19×10^{2}	1.45×10^{-9}	3.04×10^{-4}
170729+18:56:29	6.41	4.95×10^{-8}	9.36×10^{-3}
170121+21:25:36	1.74	1.82×10^{-7}	3.13×10^{-2}
170727+01:04:30	5.28×10^{-1}	6.00×10^{-7}	9.46×10^{-2}
170818+02:25:09	2.16×10^{-1}	1.46×10^{-6}	2.13×10^{-1}
170722+08:45:14	1.15×10^{-1}	2.76×10^{-6}	3.73×10^{-1}
170321+03:13:21	9.84×10^{-2}	3.22×10^{-6}	4.06×10^{-1}
170310+09:30:52	7.17×10^{-2}	4.42×10^{-6}	5.04×10^{-1}
170809+03:55:52	7.00×10^{-2}	4.53×10^{-6}	5.04×10^{-1}
170819+07:30:53	5.26×10^{-2}	6.02×10^{-6}	6.29×10^{-1}
170618+20:00:39	5.02×10^{-2}	6.32×10^{-6}	6.29×10^{-1}
170416+18:38:48	4.47×10^{-2}	7.09×10^{-6}	6.53×10^{-1}
170331+07:08:18	4.37×10^{-2}	7.25×10^{-6}	6.53×10^{-1}
151216+18:49:30	3.88×10^{-2}	8.17×10^{-6}	6.59×10^{-1}
170306+04:45:50	3.64×10^{-2}	8.71×10^{-6}	6.59×10^{-1}
151227+16:52:22	3.62×10^{-2}	8.76×10^{-6}	6.59×10^{-1}
170126+23:56:22	3.54×10^{-2}	8.95×10^{-6}	6.59×10^{-1}
151202+01:18:13	3.50×10^{-2}	9.06×10^{-6}	6.59×10^{-1}
170208+20:23:00	3.02×10^{-2}	1.05×10^{-5}	7.11×10^{-1}
170327+17:07:35	3.01×10^{-2}	1.05×10^{-5}	7.11×10^{-1}
170823+13:40:55	2.75×10^{-2}	1.15×10^{-5}	7.26×10^{-1}
150928+10:49:00	2.75×10^{-2}	1.15×10^{-5}	7.26×10^{-1}

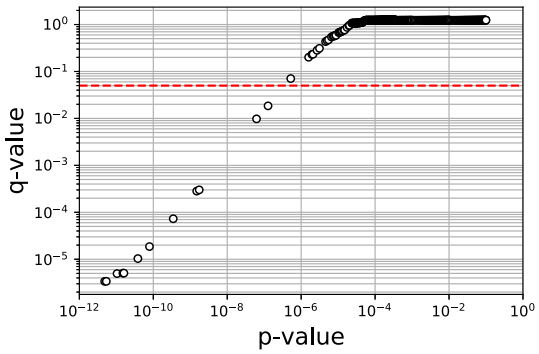

Fig. 4. The q-values computed using Algorithm 1 from p-values of events in the *bbh* data set of 2-OGC. The red dashed line indicates a q-value of 0.05.

Table 5. Estimated p-values and q-values of the events of the *bbh* data set of 2-OGC. Events are sorted by the inverse false alarm rate (FAR^{-1}), and the top 30 events are shown. The inverse false alarm rates are obtained from the data set. The p-values are computed by Eq. (4), and the q-values are computed by Algorithm 1.

UTC time	FAR^{-1} (year)	p-value	q-value	P_{astro}
170104+10:11:58	$>7.27 \times 10^{4}$	$<4.36 \times 10^{-12}$	$<3.38 \times 10^{-6}$	>0.999
150914+09:50:45	$>6.55 \times 10^{4}$	$<4.84 \times 10^{-12}$	$<3.38 \times 10^{-6}$	>0.999
151226+03:38:53	$>5.91 \times 10^{4}$	$<5.36 \times 10^{-12}$	$<3.38 \times 10^{-6}$	>0.999
170823+13:13:58	$>3.03 \times 10^{4}$	$<1.05 \times 10^{-11}$	$<4.95 \times 10^{-6}$	>0.999
170814+10:30:43	$>2.11 \times 10^{4}$	$<1.51 \times 10^{-11}$	$<5.06 \times 10^{-6}$	>0.999
151012+09:54:43	$>1.98 \times 10^{4}$	$<1.60 \times 10^{-11}$	$<5.06 \times 10^{-6}$	>0.999
170809+08:28:21	8.28×10^{3}	3.83×10^{-11}	1.03×10^{-5}	>0.999
170729+18:56:29	4.02×10^{3}	7.88×10^{-11}	1.86×10^{-5}	>0.999
170608+02:01:16	$>9.15 \times 10^{2}$	$<3.46 \times 10^{-10}$	$<7.28 \times 10^{-5}$	>0.999
170121+21:25:36	2.12×10^{2}	1.49×10^{-9}	2.83×10^{-4}	>0.999
170727+01:04:30	1.81×10^{2}	1.75×10^{-9}	3.01×10^{-4}	9.94×10^{-1}
170818+02:25:09	5.11×10^{0}	6.21×10^{-8}	9.79×10^{-3}	>0.999
170304+16:37:53	2.49×10^{0}	1.27×10^{-7}	1.85×10^{-2}	6.97×10^{-1}
151205+19:55:25	6.07×10^{-1}	5.22×10^{-7}	7.06×10^{-2}	5.25×10^{-1}
170425+05:53:34	1.99×10^{-1}	1.59×10^{-6}	2.01×10^{-1}	2.05×10^{-1}
170201+11:03:12	1.63×10^{-1}	1.95×10^{-6}	2.30×10^{-1}	2.39×10^{-1}
151217+03:47:49	1.52×10^{-1}	2.09×10^{-6}	2.33×10^{-1}	2.57×10^{-1}
151011+19:27:49	1.18×10^{-1}	2.69×10^{-6}	2.82×10^{-1}	7.95×10^{-2}
151216+09:24:16	1.01×10^{-1}	3.12×10^{-6}	3.11×10^{-1}	1.81×10^{-1}
170403+23:06:11	6.93×10^{-2}	4.57×10^{-6}	4.33×10^{-1}	3.26×10^{-2}
170202+13:56:57	6.29×10^{-2}	5.04×10^{-6}	4.54×10^{-1}	1.28×10^{-1}
170629+04:13:55	5.76×10^{-2}	5.50×10^{-6}	4.73×10^{-1}	1.90×10^{-2}
170220+11:36:24	4.81×10^{-2}	6.59×10^{-6}	5.42×10^{-1}	1.03×10^{-1}
170721+05:55:13	4.40×10^{-2}	7.20×10^{-6}	5.67×10^{-1}	5.99×10^{-2}
170123+20:16:42	4.07×10^{-2}	7.78×10^{-6}	5.70×10^{-1}	8.41×10^{-2}
170801+23:28:19	4.05×10^{-2}	7.83×10^{-6}	5.70×10^{-1}	—
170818+09:34:45	3.69×10^{-2}	8.58×10^{-6}	5.90×10^{-1}	—
170620+01:14:02	3.63×10^{-2}	8.73×10^{-6}	5.90×10^{-1}	1.51×10^{-2}
151216+18:49:30	3.07×10^{-2}	1.03×10^{-5}	6.73×10^{-1}	6.93×10^{-2}
170104+21:58:40	2.87×10^{-2}	1.10×10^{-5}	6.83×10^{-1}	1.19×10^{-1}

a marginal event. If we call events with q-value less than 0.05 significant, this event cannot be called significant. On the other hand, in Ref. [19] this event is called significant, since $P_{\rm astro}$ is larger than 0.5, and it is identified as a new marginal binary black hole merger, GW151205.

Finally, we investigate the correspondence between q-value and FAR. By linearly fitting the data in Figs. 3 and 4, we can evaluate that the p-value of 10 s/30 days $= 3.9 \times 10^{-6}$ corresponds to q-values of 0.463 and 0.377, respectively. A q-value of 0.05 corresponds to one per 384 days and one per 319 days of FAR, respectively. Thus, as in the case of 1-OGC, the q-value threshold of 0.05 is more stringent than the FAR of one per 30 days.

5. Summary and discussion

We have presented a consistent procedure to assess the significance of each event. We proposed an estimator of the p-values, Eq. (4), of a particular event in statistical hypothesis testing by using the empirical distribution of the detection statistic without any assumption on the background distribution. Generally, the p-value should follow a uniform distribution if all events originate from noise. The p-value defined in Eq. (4) has this property. On the other hand, the $p_{\rm conv}$ defined in Eq. (1) does not have this property in general. We thus believe that the p-value in Eq. (4) is more useful for assessing the significance of each event than $p_{\rm conv}$ in Eq. (1). Moreover, we proposed a consistent procedure to evaluate the q-value, which is a measure of FDR. In this procedure we use the property that p-values follow the uniform distribution under the null hypothesis, and we need no assumptions on the distribution of signals. We applied this procedure to the 1-OGC and 2-OGC catalog data [18,19]. There is already a procedure in the literature to evaluate the q-value [17]. However, since not all events in the analysis are available in the catalogs, we proposed a new procedure to evaluate the q-value which is a modified version of the original.

The results are shown in Tables 2, 3, 4, and 5. For the *bbh* case of 1-OGC, if we call events with q-value less than 0.05 significant, we have three significant events: GW150914, GW151226, and GW151012. This is fully consistent with the conclusion of Ref. [18]. We also found two marginally not significant events, 160103+05:48:36 and 151213+00:12:20, whose q-values are 8.31×10^{-2} and 8.53×10^{-2}, respectively. Since $P_{\rm astro}$ for these events is 6.07×10^{-2} and 4.66×10^{-2}, these are not identified as marginal events in Ref. [18].

For the *bbh* case of 2-OGC we have 13 significant events. All of them are also identified as significant based on $P_{\rm astro}$ in Ref. [19]. There is one marginal event, 151205+19:55:25. The q-value of this event is 0.07, but $P_{\rm astro}$ computed in Ref. [19] is 0.525. Thus, the q-value suggests that this is marginally not significant, while $P_{\rm astro}$ suggests this is marginally significant. It is not easy to conclude whether this signal is from an astrophysical origin or not just from these results.

The method for estimating q-value presented in this paper is very simple because we need no assumptions on the distributions of noise and signal. Note that the q-value and $P_{\rm astro}$ are based on fundamentally distinct statistical disciplines. The q-value is a frequentist measure, which is devised to estimate the FDR of events over some threshold of significance without any assumptions on signals. In contrast, $P_{\rm astro}$ is a Bayesian measure, which is devised to estimate the posterior probability of the astrophysical origin of a particular event relying on prior assumptions on signals. Nevertheless, from the results discussed above, we found that both approaches provide almost the same conclusions. The coincidence is not at all trivial, and would suggest that the prior assumptions on signals used in the computation of $P_{\rm astro}$ are close to reality. It would be useful to estimate the q-value as well as $P_{\rm astro}$ in gravitational wave searches.

This should be true especially for marginal events like 151205+19:55:25 here. We can obtain additional information on the significance of an event from different criteria.

We also note that the procedure for estimating the q-value presented in this paper can be applicable to other searches for gravitational waves. Our procedure for estimating the q-value is not restricted to the specific searches for gravitational waves whose true background distribution of the detection statistic is difficult to know, because our procedure is based on the empirical distribution, which is always available by time-shifting of time-series data from different detectors.

Acknowledgments

H.Y. and H.T. would like to thank Jishnu Suresh for fruitful discussions. We thank the authors of Refs. [18,19] for making the data sets of the catalogs public. This work was supported by MEXT, JSPS Leading-edge Research Infrastructure Program, JSPS Grant-in-Aid for Specially Promoted Research 26000005, JSPS Grant-in-Aid for Scientific Research on Innovative Areas 2905: JP17H06358, JP17H06361, JP16H02183, and JP17H06364, JSPS Core-to-Core Program A, Advanced Research Networks, JSPS Grant-in-Aid for Scientific Research (S) 17H06133 and 15H00787, the joint research program of the Institute for Cosmic Ray Research, the cooperative research program of the Institute of Statistical Mathematics, National Research Foundation (NRF) and Computing Infrastructure Project of KISTI-GSDC in Korea, Academia Sinica (AS), AS Grid Center (ASGC), and the Ministry of Science and Technology (MoST) in Taiwan under grants including AS-CDA-105-M06, Advanced Technology Center (ATC) of NAOJ, Mechanical Engineering Center of KEK, the LIGO project, and the Virgo project.

Appendix A. Derivation and meaning of $p_{\rm conv}$

As in various scientific research fields [26], there might be some confusion in the use of p-value in the gravitational wave community. In the recent American Statistical Association statement on p-value [26], the first principle is "p-values can indicate how incompatible the data are with a specified statistical model." Therefore, if we are talking about a p-value, we always have to make clear what statistical model we are considering. In this appendix we discuss the derivation and meaning of the conventional p-value $p_{\rm conv}$ defined by Eq. (1), which is the probability of observing one or more noise events as strong as a signal whose detection statistic is ρ under the noise model. In the analysis paper of event GW150914 [20], Abbott et al. the authors called $p_{\rm conv}$ a p-value; however, we have not called it a p-value here to avoid possible confusion with the p-value defined by Eq. (3).

Let us see more details of the probability in Eq. (1) proposed in the appendix of Ref. [21]. The total number of noise events in the observed data, N, is modeled parametrically with a Poisson process of mean μ:

$$\mathbb{P}(N=n)=\frac{\mu^n}{n!}e^{-\mu}, \qquad n\in\{0,1,2,\ldots\}, \tag{A1}$$

where $\mu=\mu(\rho)$. The slight difference between the expression of $\mu(\rho)$ in Eq. (1) and the expression $(1+n_{\rm bg}(\rho)t_{\rm obs})/t_{\rm bg}$ in [21, Eq. 17] (the unity in the numerator) comes from the fact that the model used in Ref. [21] involves observed events. In contrast, Eq. (1) is based only on the noise events in simulated background data, because the authors of the present paper believe that the noise model is better constructed by noise events only. In addition, Ref. [21] considered randomness in the number of candidate events and then marginalized them out. However, these steps have no influence on the final expression if $n_{\rm bg}(\rho)\ll n_{\rm bg}(0)$ (compare Eq. (A.4) and (A.12) in Ref. [21]). Then, the probability of observing one or more noise events as strong as a signal whose detection statistic is ρ under the noise model during the observation time, $\mathbb{P}(N\geq 1)$, is

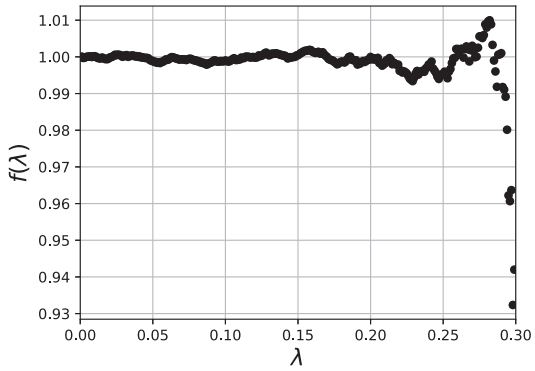

Fig. B1. Plot of $f(\lambda)$ defined in Eq. (B2) for the *complete* data set of 1-OGC.

given by Eq. (1). In the same manner, if we consider the probability of observing n_0 or more noise events as strong as a signal whose detection statistic is ρ under the noise model during the observation time, the p-value is

$$p_{\rm conv}(\rho; n_0) := \mathbb{P}(N \geq n_0) = \sum_{n \geq n_0} \frac{\mu^n}{n!} e^{-\mu}.$$

Appendix B. Discussion on $\hat{\pi}_0$

In this appendix we show that $\hat{\pi}_0$ in Eq. (7) can be approximated to be $\hat{\pi}_0 \simeq 1$. $\hat{\pi}_0$ is an estimator of $\pi_0 = n_0/n_{\rm obs}$, which indicates the overall proportion of noise events in the data. Setting $\hat{\pi}_0 = 1$ is reasonable when very few events are expected to be signal, such as in a gravitational wave search. In fact, the proposal in Ref. [23] was to set $\hat{\pi}_0 = 1$. On the other hand, for data in which some portion of the events are expected to be signal, such as in genome-wide studies, Ref. [17] proposed $\hat{\pi}_0 = \hat{f}(1)$, where $\hat{f}(\lambda)$, $\lambda \in (0, 1)$, is an estimate of $f(\lambda)$ discussed in Eq. (B2).

We consider a list of p-values which contains m p-values less than a certain value, and set $n_{\rm obs} = m$. We assume that the maximum p-value in this list is $\hat{p}_{(m)}$. In this case, n_0 is the number of noise events whose p-values are between 0 and $\hat{p}_{(m)}$. We consider the function

$$n(\lambda) = \frac{\#\{\hat{p}_i > \lambda; i \in \{1, \ldots, m\}\}}{1 - \lambda/\hat{p}_{(m)}}, \tag{B1}$$

where $0 < \lambda < \hat{p}_{(m)}$. If all p-values larger than λ_0 are noise, $\mathbb{E}(n(\lambda)) = n_0$ for $\lambda > \lambda_0$, since p-values follow a uniform distribution.

As an estimator of $\hat{\pi}_0$, let us consider the function

$$f(\lambda) = \frac{n(\lambda)}{m} = \frac{\#\{\hat{p}_i > \lambda; i \in \{1, \ldots, m\}\}}{m(1 - \lambda/\hat{p}_{(m)})}, \tag{B2}$$

where $0 < \lambda < \hat{p}_{(m)}$. If all p-values larger than λ_0 are noise, $\mathbb{E}(f(\lambda)) = \pi_0$ for $\lambda > \lambda_0$. In particular, if all p-values are noise, $\mathbb{E}(f(\lambda)) = 1$ for $\lambda \in (0, \hat{p}_{(m)})$.

Figure B1 shows a plot of $f(\lambda)$ for the *complete* data set of 1-OGC. In this plot, we use $m = 124,524$ events whose p-value is less than 0.3. We can see that $f(\lambda)$ in Eq. (B2) is almost unity for $0 < \lambda < 0.25$. We have $0.99 < f(\lambda) < 1.01$ in this region. This means that almost all p-values are noise except for a very few p-values around zero. The larger scatter in $0.25 < \lambda < 0.3$ is due to the statistical fluctuation caused by the smaller number in the numerator of Eq. (B2). Since we are mainly interested in events with small p-value less than 10^{-2}, we set $\hat{\pi}_0 = 1$.

The situation is similar in the *bbh* case. Figure B2 is a plot of $f(\lambda)$ for $0 < \lambda < 0.025$ for the *bbh* data set of 1-OGC. In this plot, we use $m = 10,429$ events whose p-value is less than 0.025.

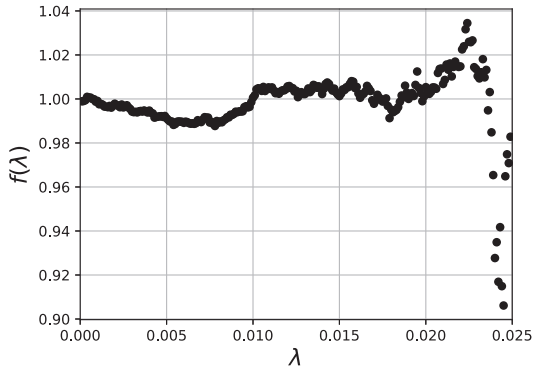

Fig. B2. Plot of $f(\lambda)$ defined in Eq. (B2) for the *bbh* data set of 1-OGC.

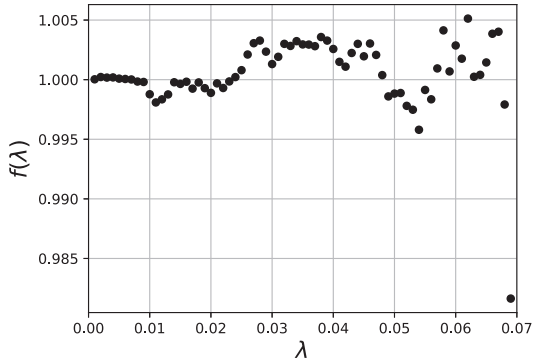

Fig. B3. Plot of $f(\lambda)$ defined in Eq. (B2) for the *complete* data set of 2-OGC.

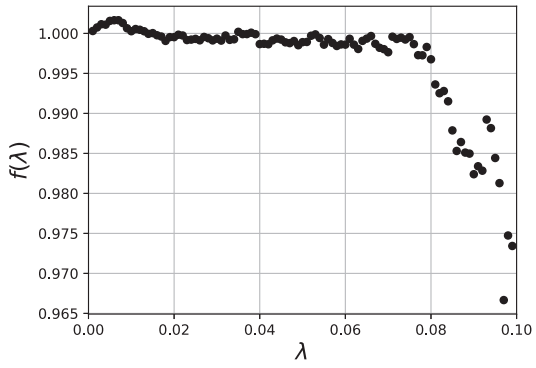

Fig. B4. Plot of $f(\lambda)$ defined in Eq. (B2) for the *bbh* data set of 2-OGC.

We have $0.98 < f(\lambda) < 1.02$ for $0 < \lambda < 0.020$. We have a larger deviation from unity for $0.020 < \lambda < 0.025$. This is due to the statistical fluctuation caused by smaller number in the numerator of Eq. (B2). From the same reason as for the *complete* data set, we set $\hat{\pi}_0 = 1$.

Figure B3 is a plot of $f(\lambda)$ for the *complete* data set of 2-OGC. In this plot, we use 103,185 events whose p-value is less than 0.07. We can see that $f(\lambda)$ in Eq. (B2) is almost unity for $0 < \lambda < 0.05$. We have $0.995 < f(\lambda) < 1.005$ in this region. The small deviation from unity near $\lambda = 0.07$ is due to the statistical fluctuation.

Figure B4 is a plot of $f(\lambda)$ for $0 < \lambda < 0.10$ for the *bbh* data set of 2-OGC. In this plot, we use 152,759 events whose p-value is less than 0.10. We have $0.996 < f(\lambda) < 1.002$ for $0 < \lambda < 0.08$.

The larger deviation from unity for $0.020 < \lambda < 0.025$ is due to the statistical fluctuation. For the same reason as for the 1-OGC data set, we set $\hat{\pi}_0 = 1$ in Step 3 of Algorithm 1.

Appendix C. Derivation of Algorithm 1 to estimate *q*-values

In this appendix we discuss the estimation procedure of q-values. We first introduce an algorithm which is a slight modification of the procedure given in Remark B of the appendix of Ref. [17]. The input is the list of detection statistics obtained from the observed data, and detection statistics in simulated background data.

Algorithm 2. Compute estimates of q-values defined in Eq. (8).

1. Compute the estimated p-values
$$\hat{p}_i \equiv \hat{p}(\rho_i) = \frac{n_{\rm bg}(\rho_i)}{n_{\rm bg}(0)},$$
where $i = 1, \ldots, n_{\rm bg}(0)$, ρ_i is the detection statistic of the ith event, and $n_{\rm bg}(\rho)$ is given in Eq. (2).
2. Let $\hat{p}_{(1)} \le \hat{p}_{(2)} \le \cdots \le \hat{p}_{(n_{\rm obs})}$ be the ordered p-values.
3. Set $\hat{q}_{(n_{\rm obs})} = \hat{\pi}_0 \hat{p}_{(n_{\rm obs})}$.
4. For $i = n_{\rm obs} - 1, n_{\rm obs} - 2, \ldots, 1$, compute
$$\hat{q}_{(i)} = \min\left(\frac{\hat{\pi}_0 n_{\rm obs} \hat{p}_{(i)}}{i}, \hat{q}_{(i+1)} \right).$$
5. The estimated q-value for the ith most significant event is $\hat{q}_{(i)}$ defined in Eq. (8). □

Since Algorithm 2 is our starting point to construct Algorithm 1, we reproduce it here. If we set $m = n_{obs}$, Algorithm 1 reduces to Algorithm 2.

Since p-values in the region around and larger than $\hat{p}_{(m)}$ are noise, if we take the threshold u in $[\hat{p}_{(m)}, 1)$, we obtain $S(u) = n_1 + n_0 u$ where n_0 and n_1 are defined in Table 1. Accordingly, Eq. (7) is

$$\widehat{\rm FDR}(u) = \frac{\hat{\pi}_0 n_{\rm obs} u}{n_1 + \hat{\pi}_0 n_{\rm obs} u},$$

which is monotonically increasing in u. Therefore, we may replace Step 3 with $\hat{q}_{(m)} = \hat{\pi}_0 n_{\rm obs} \hat{p}_{(m)}/m$. How Step 4,

$$\hat{q}_{(i)} = \min\left(\frac{\hat{\pi}_0 n_{\rm obs} \hat{p}_{(i)}}{i}, \hat{q}_{(i+1)} \right), \tag{C1}$$

still gives Eq. (8) for $i = m - 1, m - 2, \ldots, 1$ can be seen by induction. Assume Eq. (C1) gives Eq. (8) for $i = m - 1, m - 2, \ldots, k + 1$. Note that

$$\hat{q}_{(m)} \ge \hat{q}_{(m-1)} \ge \cdots \ge \hat{q}_{(k+1)} \ge \hat{q}_{(k)}. \tag{C2}$$

We show that Eq. (C1) gives Eq. (8) for $i = k$, namely,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} \ge \hat{q}_{(k)}, \qquad \forall u \ge \hat{p}_{(k)}, \tag{C3}$$

and the equality holds for some $u \ge \hat{p}_{(k)}$.

If $\hat{\pi}_0 n_{\rm obs} \hat{p}_{(k)}/k \le \hat{q}_{(k+1)}$, then $\hat{q}_{(k)} = \hat{\pi}_0 n_{\rm obs} \hat{p}_{(k)}/k$. Note that the equality of Eq. (C3) holds if $u = \hat{p}_{(k)}$. For $u \in (\hat{p}_{(k)}, \hat{p}_{(k+1)})$,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} = \frac{\hat{\pi}_0 n_{\rm obs} u}{k} > \hat{q}_{(k)}.$$

For $u = \hat{p}_{(k+1)}$,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} = \frac{\hat{\pi}_0 n_{\rm obs} \hat{p}_{(k+1)}}{k+1} \ge \hat{q}_{(k+1)} \ge \hat{q}_{(k)},$$

where the second last inequality holds from Eq. (C1) and the last inequality holds from Eq. (C2). Using a similar argument iteratively proves the assertion.

If $\hat{\pi}_0 n_{\rm obs} \hat{p}_{(k)}/k > \hat{q}_{(k+1)}$, then $\hat{q}_{(k)} = \hat{q}_{(k+1)}$. For $u \in [\hat{p}_{(k)}, \hat{p}_{(k+1)})$,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} = \frac{\hat{\pi}_0 n_{\rm obs} u}{k} > \hat{q}_{(k+1)} = \hat{q}_{(k)}.$$

For $u = \hat{p}_{(k+1)}$,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} = \frac{\hat{\pi}_0 n_{\rm obs} \hat{p}_{(k+1)}}{k+1} \ge \hat{q}_{(k+1)} = \hat{q}_{(k)}. \tag{C4}$$

Suppose the second last equality holds, namely, the equality of Eq. (C3) holds at $u = \hat{p}_{(k+1)}$. Then, for $u \in (\hat{p}_{(k+1)}, \hat{p}_{(k+2)})$,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} = \frac{\hat{\pi}_0 n_{\rm obs} u}{k+1} > \hat{q}_{(k+1)} = \hat{q}_{(k)}.$$

For $u = \hat{p}_{(k+2)}$,

$$\frac{\hat{\pi}_0 n_{\rm obs} u}{\#\{\hat{p}_{(j)} \le u;\, j \in \{1, \ldots, m\}\}} = \frac{\hat{\pi}_0 n_{\rm obs} \hat{p}_{(k+2)}}{k+2} \ge \hat{q}_{(k+2)} \ge \hat{q}_{(k+1)} = \hat{q}_{(k)}.$$

Using a similar argument iteratively proves the assertion. If the second last equality of Eq. (C4) does not hold, there exists some l such that $k+2 \le l \le m$, and

$$\frac{\hat{\pi}_0 n_{\rm obs} \hat{p}_{(l)}}{l} = \hat{q}_{(l)} = \hat{q}_{(l-1)} = \cdots = \hat{q}_{(k)},$$

because $\hat{q}_{(m)} = \hat{\pi}_0 n_{\rm obs} \hat{p}_{(m)}/m$. The assertion can be shown in a similar manner.

References

1 B. P. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Phys. Rev. Lett. **116**, 061102 (2016).
2 LIGO Scientific Collaboration, the and Virgo Collaboration, Phys. Rev. X **9**, 031040 (2019).
3 B. P. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Phys. Rev. Lett. **119**, 161101 (2017).
4 B. Abbott et al., Rept. Prog. Phys. **72**, 076901 (2009).
5 T. Accadia et al., J. Instrum. **7**, P03012 (2012).
6 R. Abbott et al. Astrophys. J. Lett. **848**, L12 (2017). [LIGO Scientific Collaboration, Virgo Collaboration]
7 LIGO Scientific Collaboration, GraceDB (available at: https://gracedb.ligo.org/, date last accessed November 7, 2021).
8 B. P. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Astrophys. J. Lett. **892**, L3 (2020).
9 R. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Phys. Rev. D **102**, 043015 (2020).
10 R. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Astrophys. J. Lett. **896**, L44 (2020).
11 R. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Phys. Rev. Lett. **125**, 101102 (2020).
12 R. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], arXiv:2010.14527 [gr-qc] [Search INSPIRE].
13 Y. Aso et al., [KAGRA Collaboration], Phys. Rev. D **88**, 043007 (2013).
14 E. L. Lehmann and J. P. Romano, Testing Statistical Hypotheses, 3rd ed. (Springer, New York, 2008).

15 L. Baggio and G. A. Prodi, Class. Quant. Grav. **22**, S1373 (2005).
16 W. M. Farr, J. R. Gair, I. Mandel, and C. Cutler, Phys. Rev. D **91**, 023005 (2015).
17 J. D. Storey and R. Tibshirani, Proc. Natul. Acad. Sci. USA **100**, 9440 (2003).
18 A. H. Nitz et al., Astrophys. J. **872**, 195 (2019).
19 A. H. Nitz et al., Astrophys. J. **891**, 123 (2020).
20 B. P. Abbott et al., [LIGO Scientific Collaboration, Virgo Collaboration], Phys. Rev. X **6**, 041015 (2016); 8, 039903 (2018) [erratum].
21 S. A. Usman et al., Class. Quant. Grav. **33**, 215004 (2016).
22 C. Capano, T. Dent, C. Hanna, M. Hendry, C. Messenger, Y.-M. Hu, and J. Veitch, Phys. Rev. D **96**, 082002 (2017).
23 Y. Benjamini and Y. Hochberg, J. Roy. Stat. Soc. B **57**, 289 (1995).
24 Y. Benjamini, J. Roy. Stat. Soc. B **72**, 405 (2010).
25 C. Bini, Data Analysis in Particle Physics (available at: https://www.roma1.infn.it/ bini/StatEPP_new.pdf, date last accessed November 7, 2021).
26 R. L. Wasserstein and N. A. Lazar, Amer. Statist. **70**, 129 (2016).