PHYSICAL REVIEW D **91,** 023005 (2015)

Counting and confusion: Bayesian rate estimation with multiple populations

Will M. Farr[*]

Department of Physics and Astronomy, Center for Interdisciplinary Exploration and Research in Astrophysics, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, USA and School of Physics and Astronomy, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom

Jonathan R. Gair[†]

Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom

Ilya Mandel[‡]

School of Physics and Astronomy, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom

Curt Cutler[§]

Jet Propulsion Lab, 4800 Oak Grove Drive, Pasadena, California 91109, USA and Theoretical Astrophysics, California Institute of Technology, Pasadena, California 91125, USA

(Received 21 February 2013; published 22 January 2015)

We show how to obtain a Bayesian estimate of the rates or numbers of signal and background events from a set of events when the shapes of the signal and background distributions are known, can be estimated, or approximated; our method works well even if the foreground and background event distributions overlap significantly and the nature of any individual event cannot be determined with any certainty. We give examples of determining the rates of gravitational-wave events in the presence of background triggers from a template bank when noise parameters are known and/or can be fit from the trigger data. We also give an example of determining globular-cluster shape, location, and density from an observation of a stellar field that contains a nonuniform background density of stars superimposed on the cluster stars.

DOI: 10.1103/PhysRevD.91.023005 PACS numbers: 95.75.Pq, 02.50.Tt, 02.70.Rr, 04.30.Tv

I. INTRODUCTION

The task of estimating rates of events when a mixture of foreground and background events is present in data is a common one in physical and astrophysical applications. This problem comes up, among others, in gravitational-wave data analysis [e.g., [1–6]] and in astronomical observations of a field of objects of mixed provenance [7,8]. In this paper, we introduce a robust formalism for estimating event rates from the data when the shape of foreground and background distributions are known (or parametrized), but the provenance of individual events as either background or foreground is unknown.

We use a Bayesian approach and consider all available data to ensure that the inferred rates are both unbiased and maximally constrained in the presence of limited observations. Bayes' theorem yields the *posterior* probability density function on a set of parameters, $\vec{\theta}$, given the observed data, d, under a model M:

$$p(\vec{\theta}|d,M) = \frac{p(\vec{\theta}|M)p(d|\vec{\theta},M)}{p(d|M)}, \tag{1}$$

where $p(\vec{\theta}|M)$ are the *prior* probabilities of the model parameters, $p(d|\vec{\theta},M)$ is the *likelihood* of obtaining the data given a particular choice of parameters, and the normalizing factor $p(d|M)$ is known as the *evidence*.

Two alternative approaches to rate estimation have been suggested and are commonly used. One, known as the *loudest-event statistic* [9–11], uses only the information from the highest-ranked event in the data to infer the rate distribution. This approach has been used successfully [1–6] when the number of loud foreground events is small (typically zero or one) to obtain upper limits on foreground rates. However, the loudest-event statistic ignores all events except the loudest one, and so suffers from an unnecessary loss of information; therefore, we expect it to yield a much larger variance than strictly necessary when multiple events are present in the data. In practice, the loudest-event statistic is typically applied repeatedly to multiple "chunks"

[*]wfarr@star.sr.bham.ac.uk; http://faculty.wcas.northwestern.edu/will-farr/.
[†]jrg23@cam.ac.uk
[‡]imandel@star.sr.bham.ac.uk; http://www.sr.bham.ac.uk/imandel.
[§]Curt.J.Cutler@jpl.nasa.gov

of data, using the estimated rate posterior from each chunk as a rate prior for the next chunk's analysis [2–4]. Even when used in this mode, the method discards information, with the amount of information loss depending on the (arbitrary) division of the data into chunks.

Another possible approach is based on the use of only loud, "gold-plated" events, ones which are certain (or nearly certain) to come from the foreground, to derive rates. We refer to this approach as the *foreground-dominated statistic*. The foreground-dominated statistic may yield accurate results when the foreground and background are cleanly separated, at least for the loudest events, and the number of such loud events is sufficiently large. However, it cannot properly account for marginal events. In addition, the results of the method are very sensitive to contamination by the background events, and therefore the method requires a careful choice of *threshold* or reliable *membership* information to distinguish foregrounds and backgrounds for individual events. While either the loudest-event statistic or the foreground-dominated statistic can approach the accuracy of our proposed method in specific regimes, both are suboptimal in a general case.

Reference [12] considered the problem of determining an intrinsic rate and population parameters in the presence of missing data, either due to thresholding, poor sensitivity, or contamination from noise events. The approach is complementary to ours: we consider the problem of accurately counting the events of different classes present in a data set, while Ref. [12] deals with translating such counts into physical rates by properly accounting for the selection effects on the data set.

Our key results appear in Eq. (18), which provides the joint posterior probability distribution on the foreground and background rates and shape parameters and the provenance of individual events as either foreground or background. Equation (21) is a marginalized version of Eq. (18), useful when the provenance of individual events is not relevant. In practice, these posteriors are best sampled with stochastic techniques such as Markov chain Monte Carlo.

In order to demonstrate our method, we consider three different examples. The first two come from the field of gravitational-wave data analysis, but could equally arise in any application that employs matched filtering [13] to extract weak signals with known shapes from the data. The last example considers the case of a globular cluster on a background of field stars. Throughout, we compare the results obtained with our technique to the loudest-event and foreground-dominated statistics, which make use of a limited subset of the available information.

II. MODEL

We first consider one-dimensional data, but will generalize to the multidimensional case below. We assume that we are presented with a data set of N events that exceed a pre-specified threshold in ranking statistic, $x_{\rm min}$. Each event may be due to either a signal of interest or an uninteresting background. Each event is associated with a ranking statistic, x. Our data set therefore consists of the ranking statistics for the set of events:

$$d = \{x_i | i = 1, \ldots, N\}. \tag{2}$$

The number of events N is also part of the observed data, but we separate out N and the observed ranking statistics, d, for convenience. We can choose how to label our events. Ultimately we will label the events in order of ranking statistic, i.e., $x_1 < x_2 < \cdots < x_N$, but some of the derivations that follow are simpler if the events are ordered by time of arrival (i.e., randomly with respect to the x_i). We will use d to denote ranking-statistic-ordered events, and $d_{\rm to}$ to denote time-ordered events.

We assume that both the foreground and background events are samples from an inhomogeneous Poisson process with respective differential rates

$$\frac{dN_f}{dx} = f(x, \theta) \tag{3}$$

and

$$\frac{dN_b}{dx} = b(x, \theta), \tag{4}$$

where the θ argument represents additional "shape" parameters that may affect the distribution, and for which we will eventually fit. The cumulative rates of the two processes are therefore

$$F(x, \theta) \equiv \int_{-\infty}^{x} ds f(s, \theta) \tag{5}$$

and

$$B(x, \theta) \equiv \int_{-\infty}^{x} ds b(s, \theta). \tag{6}$$

The assumption that the foreground and background events form an inhomogeneous Poisson process implies

(1) The number of events in any range of ranking statistics, $x \in [x_1, x_2]$ is Poisson distributed with rate $F(x_2, \theta) - F(x_1, \theta)$ or $B(x_2, \theta) - B(x_1, \theta)$.
(2) The numbers of events in non-overlapping ranges of ranking statistics are independent.
(3) The probability of exactly one foreground event between x and $x + h$ is given by

$$P(n = 1 \in [x, x + h]) = f(x, \theta)h + \mathcal{O}(h^2) \tag{7}$$

and similarly for background events.
(4) The probability of two or more events in a small range of ranking statistic is negligible

$$P(n = 2 \in [x, x + h]) = \mathcal{O}(h^2). \quad (8)$$

The foreground and background rates can in general depend on several parameters; the goal of our analysis is to determine the posterior probability distributions for these parameters that are implied by the data. At the least, we will want to know the overall dimensionless amplitude of the foreground and background rates. Let

$$f(x, \theta) = R_f \hat{f}(x, \theta'), \quad (9)$$

and

$$b(x, \theta) = R_b \hat{b}(x, \theta'), \quad (10)$$

where $\hat{F}(\infty, \theta') = \hat{B}(\infty, \theta') = 1$, and $\theta' = \theta \backslash \{R_f, R_b\}$. Then $R_f \equiv F(\infty, \theta)$ and $R_b \equiv B(\infty, \theta)$ are the total number of foreground and background events expected and $\hat{f}(x, \theta')$ and $\hat{b}(x, \theta')$ are the likelihood of obtaining an event with ranking statistic x under the foreground and background distributions. In what follows, we will drop the prime, using θ to denote all parameters of the rate distributions except R_f and R_b.

We do not know *a priori* which of the events are foreground and which are background. For each event, we introduce a flag, g_i, which is either 0 (background) or 1 (foreground). These "state" flags are parameters in our model, along with R_f, R_b, and θ. We can marginalize over our uncertainty in the state of any given event by summing posteriors over $g_i = \{0, 1\}$.

Assuming time-ordered data, $d_{\rm to}$, in the following, Bayes' theorem relates the posterior probability of the state flags, rates, and shape parameters, $p(\{g_i\}, R_f, R_b, \theta | d_{\rm to}, N)$, the likelihood of the data, $p(d_{\rm to}|\{g_i\}, N, R_f, R_b, \theta)$, and the prior probability of state flags, rates and shape parameters before any data are obtained, $p(\{g_i\}, N, R_f, R_b, \theta)$:

$$p(\{g_i\}, R_f, R_b, \theta | d_{\rm to}, N) = \frac{p(d_{\rm to}|\{g_i\}, N, R_f, R_b, \theta) p(\{g_i\}, N, R_f, R_b, \theta)}{p(d_{\rm to}, N)}. \quad (11)$$

The normalization constant, called the evidence, $p(d_{\rm to}, N)$, is independent of the state flags, rates, and shape parameters.

Each foreground event is drawn from the probability distribution $\hat{f}$ and each background event is drawn from the probability distribution $\hat{b}$. The events are independent of each other. Therefore, the likelihood of the data is

$$p(d_{\rm to}|\{g_i\}, N, R_f, R_b, \theta) = \left[\prod_{\{i|g_i=1\}} \hat{f}(x_i, \theta)\right]\left[\prod_{\{i|g_i=0\}} \hat{b}(x_i, \theta)\right]. \quad (12)$$

This is the probability that the first observed event is a fore/background event (if $g_1 = 1, 0$) with ranking statistic x_1 and the second observed event is a fore/background event (if $g_2 = 1, 0$) with ranking statistic x_2, etc. If the events are ordered by ranking statistic the corresponding expression is more complicated, since x_1 is now the event from foreground or background with the smallest ranking statistic, etc. We will return to the statistic-ordered case later.

The prior distribution can be factorized as

$$\begin{aligned} p(\{g_i\}, N, R_f, R_b, \theta) &= p(\{g_i\}|N, R_f, R_b) p(N|R_f, R_b) p(R_f, R_b, \theta) \\ &= p(\{g_i\}, N|R_f, R_b) p(R_f, R_b, \theta). \end{aligned} \quad (13)$$

The probability that the ith state flag is $g_i = 1$ is given by $R_f/(R_f + R_b)$, while the probability that it is zero is $R_b/(R_f + R_b)$, provided the data are time ordered as we have assumed. Then

$$\begin{aligned} p(\{g_i\}|N, R_f, R_b) &= \prod_{\{i|g_i=1\}} \left(\frac{R_f}{R_f + R_b}\right) \prod_{\{i|g_i=0\}} \left(\frac{R_b}{R_f + R_b}\right) \\ &= \left(\frac{R_f}{R_f + R_b}\right)^{N_f} \left(\frac{R_b}{R_f + R_b}\right)^{N_b}, \end{aligned} \quad (14)$$

where N_f and N_b are the numbers of foreground and background flags, $N_f + N_b = N$. Meanwhile,

$$p(N|R_f, R_b) = \frac{(R_f + R_b)^N}{N!} e^{-(R_f + R_b)}, \quad (15)$$

since the distribution of total event number is a Poisson process with rate $R_f + R_b$. Combining these yields the conditional probability of the flags on the rates:

$$p(\{g_i\}, N|R_f, R_b) = \frac{R_f^{N_f} R_b^{N_b}}{N!} \exp\left[-(R_f + R_b)\right]. \quad (16)$$

The last term in Eq. (13) is a traditional prior. Because the rate parameters enter the posterior in the same form as Poisson rates, we choose here the Poisson Jeffreys prior on rates [14], independent of the shape parameters

$$p(R_f, R_b, \theta) = \alpha \frac{1}{\sqrt{R_f R_b}} p(\theta), \quad (17)$$

where α is a normalization constant; but of course other choices are possible. This choice has the advantage that the prior is normalizable as $R_f, R_b \to 0$, and the exponentials in Eq. (16) regularize the posterior as $R_f, R_b \to \infty$.

Putting everything together, the posterior is

$$p(\{g_i\}, R_f, R_b, \theta|d_{\rm to}, N) = \frac{\alpha}{p(d_{\rm to}, N)N!}\left[\prod_{\{i|g_i=1\}} R_f \hat{f}(x_i,\theta)\right]\left[\prod_{\{i|g_i=0\}} R_b \hat{b}(x_i,\theta)\right]\exp\left[-(R_f+R_b)\right]\frac{p(\theta)}{\sqrt{R_f R_b}}. \quad (18)$$

When sampling the posterior, the first term, which is independent of the parameters of interest, can be omitted and the equals sign replaced by proportionality; however, we have kept this term explicitly so that we can see the equivalence to ranking-statistic ordered data. Once data have been observed, there is a unique loudness ordering and time ordering of those events, and so there is a one to one correspondence between a time-ordered posterior $p(\{g_i\}, R_f, R_b, \theta|d_{\rm to}, N)$ and the corresponding statistic-ordered posterior $p(\{g_i\}, R_f, R_b, \theta|d, N)$, which means $p(\{g_i\}, R_f, R_b, \theta|d, N) = p(\{g_i\}, R_f, R_b, \theta|d_{\rm to}, N)$. However, the evidence $p(d, N) = N!p(d_{\rm to}, N)$, since there are $N!$ ways in which N events with a given set of ranking statistics can be ordered in time.

The ranking-statistic-ordered posterior can be computed directly by assuming that the flags, $\{g_i\}$, are unobserved data and treating the sets $\{x_i|g_i = 1\}$ and $\{x_i|g_i = 0\}$ as samples from an inhomogeneous Poisson process. For an inhomogeneous Poisson process with rate function $r(y)$ [cumulative rate $R(y)$], the likelihood of a set of samples $\{y_i\}$ is given by

$$\begin{aligned} p(\{y_i\}|r)\mathrm{d}^N y_i = {} & P(\text{zero events below } y_1) \\ & \times P(\text{one event between } y_1 \text{ and } y_1 + \mathrm{d}y_1) \\ & \times P(\text{zero events between } y_1 + \mathrm{d}y_1 \text{ and } y_2)\ldots, \end{aligned} \quad (19)$$

so

$$\begin{aligned} p(\{y_i\}|r) &= \lim_{\delta y_i \to 0} \exp\left[-R(y_1)\right]\left[r(y_1) + \mathcal{O}(\delta y_1)\right] \\ &\quad \times \exp\left[-\left[R(y_2) - R(y_1 + \delta y_1)\right]\right] \times \ldots \\ &= \left[\prod_i r(y_i)\right]\exp\left[-R(\infty)\right]. \end{aligned} \quad (20)$$

Applying this once to the foreground samples, once to the background samples and taking the product, we obtain $p(d, \{g_i\}, N|R_f, R_b, \theta)$ and thus $p(\{g_i\}, R_f, R_b, \theta|d, N) = p(d, \{g_i\}, N|R_f, R_b, \theta)p(R_f, R_b, \theta)/p(d, N)$. With the identification $p(d, N) = N!p(d_{\rm to}, N)$, as justified above, we reproduce Eq. (18).

We can marginalize the posterior over the flags, g_i, obtaining

$$\begin{aligned} p(R_f, R_b, \theta|d, N) &= \sum_{\{g_i\}\in\{0,1\}^N} p(\{g_i\}, R_f, R_b, \theta|d, N) \\ &\propto \prod_i \left[R_f \hat{f}(x_i,\theta) + R_b \hat{b}(x_i,\theta)\right] \\ &\quad \times \exp\left[-(R_f + R_b)\right]\frac{p(\theta)}{\sqrt{R_f R_b}}. \end{aligned} \quad (21)$$

This expression is useful if we are only interested in rates and not the probability that any particular event is foreground or background. Unlike the full posterior [Eq. (18)], Eq. (21) contains only continuous parameters. We note that the terms that depend on the overall rate parameters, R_b or R_f, are of the form $R_b^{n-1/2}\exp(-R_b)$ and so marginalization over either R_b or R_f can be achieved analytically using

$$I_n = \int_0^\infty x^{n-\frac{1}{2}}\mathrm{e}^{-x}\mathrm{d}x = \frac{(2n-1)!!}{2^n}\sqrt{\pi} \quad (22)$$

using the usual notation $(2n-1)!! \equiv (2n-1)(2n-3)\ldots 1$.

Equation (18) is unchanged if the ranking statistic is multidimensional; in this case, the rates are

$$R_f = \int d^k\vec{x} f(x,\theta) \quad (23)$$

and

$$R_b = \int d^k\vec{x} b(x,\theta), \quad (24)$$

where f and b are rate densities on the k-dimensional space of ranking statistics. We give an example of fitting for multidimensional rate densities in Sec. V D.

III. COMPARISON TO OTHER RATE ESTIMATION METHODS

It is informative to relate these results to two other methods for estimating the foreground rate parameter—the loudest-event statistic and the foreground-dominated statistic.

A. Loudest-event statistic

If we were to include only the k loudest events in the posterior distribution, rather than all observed events, the posterior [Eq. (18)] would be modified by an additional factor of $\exp[R_f\hat{F}(x_{N-k+1},\theta) + R_b\hat{B}(x_{N-k+1},\theta)]$, where we have assumed events are ordered by loudness, so that x_{N-k+1} is the kth loudest event. This term accounts for the

data-dependent threshold that a loudest-event statistic employs.

For the usual $k = 1$ case [9], the marginalized posterior [Eq. (21)] becomes

$$p_{\rm LE}(R_f, R_b, \theta|d) \propto (R_f \hat{f}(x_N, \theta) + R_b \hat{b}(x_N, \theta)) \times \exp[-(R_f(1 - \hat{F}(x_N, \theta)) + R_b(1 - \hat{B}(x_N, \theta)))] \frac{p(\theta)}{\sqrt{R_f R_b}}, \quad (25)$$

where x_N denotes the loudness of the loudest event, and R_f and R_b are the number of events expected *above our original threshold* [so, for example, $R_f(1 - \hat{F}(x_N, \theta))$ is the number of foreground events expected above loudness x_N]. In the loudest-event-statistic paper [9], the authors assume the background distribution and rate are known, which corresponds to using a narrow prior on R_b. They further assume a flat prior (in the absence of other experimental data) on R_f and that the foreground and background distributions do not depend on any unknown free parameters. With these assumptions, the posterior on R_f, Eq. (25), is modified to

$$p_{\rm LE}(R_f|d) \propto (R_f \hat{f}(x_N) + R_b \hat{b}(x_N)) \times \exp[-(R_f(1 - \hat{F}(x_N)) + R_b(1 - \hat{B}(x_N)))]. \quad (26)$$

Integrating over R_f gives

$$\int_0^\infty p_{\rm LE}(R_f|d) {\rm d}R_f = \frac{R_b \hat{b}(x_N)}{(1 - \hat{F}(x_N))} {\rm e}^{-(1-\hat{B}(x_N))R_b} \times \left(\frac{\hat{f}(x_N)}{(1 - \hat{F}(x_N)) R_b \hat{b}(x_N)} + 1 \right) \quad (27)$$

and so the normalized posterior is

$$p_{\rm LE}(R_f|d) = \frac{(1 - \hat{F}(x_N))}{1 + \Lambda} (1 + R_f(1 - \hat{F}(x_N))\Lambda) \times \exp[-R_f(1 - \hat{F}(x_N))] \quad (28)$$

in which we have defined

$$\Lambda \equiv \frac{\hat{f}(x_N)}{(1 - \hat{F}(x_N)) R_b \hat{b}(x_N)}. \quad (29)$$

With the further identification $\mu \equiv R_f$ and $\hat{\epsilon} \equiv 1 - \hat{F}(x_N)$, this is Eq. (14) of [9] and we have shown how their parameter Λ is related to the foreground and background distributions used here.

Returning now to Eq. (25) and marginalizing over R_b, we obtain

$$p_{\rm LE}(R_f, \theta|d) \propto \left(\frac{\hat{b}(x_N, \theta)}{2(1 - \hat{B}(x_N, \theta))} + R_f \hat{f}(x_N, \theta) \right) \times \frac{\sqrt{\pi} p(\theta)}{\sqrt{1 - \hat{B}(x_N, \theta)} \sqrt{R_f}} \times \exp\left(-R_f(1 - \hat{F}(x_N, \theta))\right). \quad (30)$$

This posterior has a maximum in R_f at

$$R_f = \frac{\hat{f}(x_N, \theta) - (1 - \hat{F}(x_N, \theta))\tilde{b}(x_N, \theta) + \sqrt{g(x_N, \theta)}}{4\hat{f}(x_N, \theta)(1 - \hat{F}(x_N, \theta))}$$

where

$$g(x_N, \theta) = (\hat{f}(x_N, \theta) - (1 - \hat{F}(x_N, \theta))\tilde{b}(x_N, \theta))^2 - 4\tilde{b}(x_N, \theta)(1 - \hat{F}(x_N, \theta))\hat{f}(x_N, \theta) \quad (31)$$

and $\tilde{b}(x_N, \theta) = \hat{b}(x_N, \theta)/(1 - \hat{B}(x_N, \theta))$ and similarly for $\tilde{f}(x_N, \theta)$.

If $\tilde{b}(x_N, \theta) \ll \tilde{f}(x_N, \theta)$, we obtain the result $(1 - \hat{F}(x_N, \theta))R_f \approx 1/2$. This can be understood as the statement that the rate of foreground events with ranking statistic greater than x_N, $(1 - \hat{F}(x_N, \theta))R_f$, is of order 1, as expected. However, $\tilde{b}(x_N, \theta) = -{\rm d}[\ln(1 - \hat{B}(x, \theta))]/{\rm d}x$ and $(1 - \hat{B}(x, \theta)) \to 0$ as $x \to \infty$, so this term may be divergent and for many reasonable examples, we will find $\tilde{b}(x_N, \theta) \gg \tilde{f}(x_N, \theta)$, in which case the posterior on R_f is peaked at 0. This issue highlights the problem with using a loudest-event statistic with an improper prior on the background rate R_b. No matter how improbable an event with $x = x_N$ is under the background distribution, it can become likely that the event at x_N is from the background distribution by taking the background rate to be sufficiently large. Although this predicts many more events with $x < x_N$, by using only the loudest event we do not incorporate the information that no such events are seen. This problem is avoided in the new framework described here, since we use all events detected above threshold and combined rates, $R_f + R_b$, significantly greater than the total number of observed events are strongly disfavored.

This problem can also be avoided in the context of the loudest-event framework by even very weak prior information on the background rate, R_b, of the kind present in nearly all experiments. For example, we can include an upper limit on the rate, $R_{\rm max}$, in the prior for R_b.

The marginalized distribution for the foreground rate then becomes

$$p_{\mathrm{LE}}(R_f,\theta|d) \propto \left(\frac{\hat{b}(x_N,\theta)}{(1-\hat{B}(x_N,\theta))^{\frac{3}{2}}} \left[\frac{\sqrt{\pi}}{2} \mathrm{erf}\left(\sqrt{(1-\hat{B}(x_N,\theta))R_{\max}} \right) - \sqrt{(1-\hat{B}(x_N,\theta))R_{\max}}\, \mathrm{e}^{-(1-\hat{B}(x_N,\theta))R_{\max}} \right] \right.$$
$$\left. + R_f \hat{f}(x_N,\theta) \frac{\sqrt{\pi}\mathrm{erf}\left(\sqrt{(1-\hat{B}(x_N,\theta))R_{\max}} \right)}{\sqrt{(1-\hat{B}(x_N,\theta))}} \right) \frac{p(\theta)}{\sqrt{R_f}} \exp\left(-R_f(1-\hat{F}(x_N,\theta))\right), \quad (32)$$

where $\mathrm{erf}(x)$ is the error function, defined in the usual way $\mathrm{erf}(x) = (2/\sqrt{\pi})\int_0^x \exp(-u^2)\mathrm{d}u$. If $(1-\hat{B}(x_N,\theta))R_{\max} \ll 1$, Eq. (32) can be approximated by

$$p_{\mathrm{LE}}(R_f,\theta|d) \propto \left(\frac{R_{\max}}{3}\hat{b}(x_N,\theta) + R_f\hat{f}(x_N,\theta) \right) \times \frac{p(\theta)}{\sqrt{R_f}} \exp\left(-R_f(1-\hat{F}(x_N,\theta))\right) \quad (33)$$

and if $\hat{f}(x_N,\theta) \gg R_{\max}\hat{b}(x_N,\theta)$ we find the same result as before, $(1-\hat{F}(x_N,\theta))R_f \approx 1/2$.

B. Foreground-dominated statistic

If we set the threshold for including an event, $x_{\min}$, sufficiently high, we can ensure that $\hat{f}(x_i,\theta) \gg \hat{b}(x_i,\theta)$ for all ranking statistics x_i in the data set. If we can further be confident that $R_f\hat{f}(x_i,\theta) \gg R_b\hat{b}(x_i,\theta)$ for all events, then the posterior can be approximated by

$$p_{\mathrm{FD}}(R_f,R_b,\theta|d) \propto \prod_i [\hat{f}(x_i,\theta)] R_f^N \exp\left[-(R_f+R_b)\right] \frac{p(\theta)}{\sqrt{R_f R_b}}. \quad (34)$$

Note that these are posteriors on the number of events expected *above the threshold* $x_{\min}$. The threshold choice for the foreground-dominated statistic could be different from the threshold choice applied elsewhere. If the rates are estimated accurately, then a rate estimate $R_{f,1}$ above threshold $x_{\min} = x_1$ can be converted into a rate estimate $R_{f,2}$ above threshold $x_{\min} = x_2$ via $R_{f,1}(1-\hat{F}(x_2,\theta)) = R_{f,2}(1-\hat{F}(x_1,\theta))$; however, rate point estimates based on thresholding can have significant fluctuations, as discussed in the following section.

Normalization over R_b gives a constant factor and the posterior on the foreground rate becomes

$$p_{\mathrm{FD}}(R_f,\theta|d) \propto \prod_i [\hat{f}(x_i,\theta)] R_f^{N-\frac{1}{2}} \exp\left[-R_f\right] p(\theta). \quad (35)$$

Ignoring the dependence on θ, this is peaked at a rate $R_f = N - 1/2$, so we have the expected result that, in the foreground-dominated regime, the rate is approximately equal to the number of events observed (the $1/2$ comes from our use of the Jeffreys prior on the rate).

IV. THRESHOLDING

This paper is concerned with Bayesian rate estimates based on lists of events. Ideally, the lists should contain all events in the data set. However, for experimental or computational reasons one may wish to restrict the events to only those above some loudness threshold; in some cases the rate of foreground or background events, or both, is even expected to diverge at certain loudnesses. In this subsection we address the question of how the rate estimate depends on the threshold value. For a discussion of selection effects, of which thresholding is but one, on the estimate of physical rates, see Ref. [12].

To begin with, we recall the well-known fact that the Bayesian estimator is unbiased, in the following sense. For simplicity, assume that the model consists of a single rate parameter R, with prior distribution $p(R)$. Consider an ensemble of data sets whose distribution is consistent with that prior; i.e., such that $p(d)$ is given by

$$p(d) = \int p(d|R)p(R)dR. \quad (36)$$

For each data set in the ensemble, compute the Bayesian estimator for the mean of the posterior $R_B = \int Rp(R|d)dR$. Then it is immediate that

$$\int R_B(d)p(d)d(d) = \int Rp(R)dR, \quad (37)$$

i.e., the data-weighted average of the Bayesian estimator R_B equals the prior-weighted average R. Therefore all threshold values will yield, on average, the same point estimate of the rate. However this equality of averages does *not* imply that all threshold values yield the same information. In general, as the threshold is lowered to include more events, the error bar on the estimate shrinks. In this

subsection we give quantitative illustrations of how the error bar shrinks when the threshold is lowered.

Consider the following model problem. Let $p(x) = b(x) + f(x) = R_b\hat{b}(x) + R_f\hat{f}(x)$ be the rate density of events (of both foreground and background type) per unit loudness. Here we will assume that the background is normally distributed in loudness, so that b has the form

$$b(x) = \Gamma_b \exp\left(-\frac{x^2}{2}\right). \tag{38}$$

We find it useful to define x_1 as the loudness such that a data set will have on average a single noise event louder than x_1, i.e., such that

$$\int_{x_1}^{\infty} b(x)dx = R_b - B(x_1) = 1. \tag{39}$$

This condition fixes

$$\Gamma_b = \left[\sqrt{\frac{\pi}{2}}erfc\left(\frac{x_1}{\sqrt{2}}\right)\right]^{-1}, \tag{40}$$

while R_b will depend on the threshold, $x_{\rm th}$, as

$$R_b = \frac{erfc(\frac{x_{\rm th}}{\sqrt{2}})}{erfc(\frac{x_1}{\sqrt{2}})}. \tag{41}$$

Let the foreground distribution follow a power law in loudness [this is, for example, the distribution of signal-to-noise ratio (SNR) for gravitational-wave events from uniformly distributed sources in a single detector]

$$f(x) = 3\Gamma_f \frac{x_1^3}{x^4}, \tag{42}$$

where $\Gamma_f = R_f - F(x_1)$ is the mean number of foreground events with $x > x_1$. The overall foreground rate is given by

$$R_f = \Gamma_f \frac{x_1^3}{x_{\rm th}^3}. \tag{43}$$

We can write the full $p(x)$ as

$$p(x) = \left[\sqrt{\frac{\pi}{2}}erfc\left(\frac{x_1}{\sqrt{2}}\right)\right]^{-1} \exp\left[-\frac{x^2}{2}\right] + 3\Gamma_f \frac{x_1^3}{x^4}. \tag{44}$$

For any pair (x_1, Γ_f), it is straightforward to construct random event lists drawn from the corresponding $p(x)$, and straightforward to apply a threshold by "throwing away" all events with x less than the threshold value $x_{\rm th}$. If $\Gamma_f \gg 1$, then we are in the foreground-dominated regime at $x = x_1$, if $\Gamma_f \ll 1$ we are in the background-dominated regime, and if $\Gamma_f \sim 1$ the foreground and background counts above x_1 are about equal. For any thresholded event list, we use Eq. (21) to construct the probability density $p(R_f|d)$. For that event list, we define the foreground rate uncertainty, ΔR_f, by

$$(\Delta R_f)^2 \equiv \int (R_f - R_f^{\rm true})^2 p(R_f|d) dR_f, \tag{45}$$

where $R_f^{\rm true}$ is given by Eq. (43).

Figure 1 illustrates how the mean fractional foreground uncertainty, $\langle \Delta R_f \rangle / R_f$, varies with the threshold value $x_{\rm th}$ for the foreground-dominated and comparable-rate regime. In all cases we assumed that $x_1 = 8$. For large thresholds, where $R_b \ll 1$, increasing the threshold tends to increase the fractional uncertainty on the foreground rate, since fewer foreground events are included in the sample. However, as the threshold passes into the background-dominated regime, the uncertainty in the foreground rate asymptotes to

$$\frac{\Delta R_f}{R_f} \simeq \frac{1}{\sqrt{\Gamma_f}}, \tag{46}$$

which is the usual Poisson counting uncertainty on the events that stand out from the background (those with $x \gtrsim x_1$). Note that this uncertainty applies even when the total number of background events is orders of magnitude larger than the number of foreground events. When a threshold must be chosen, it is safest—in the sense of producing the minimal foreground rate uncertainty—to

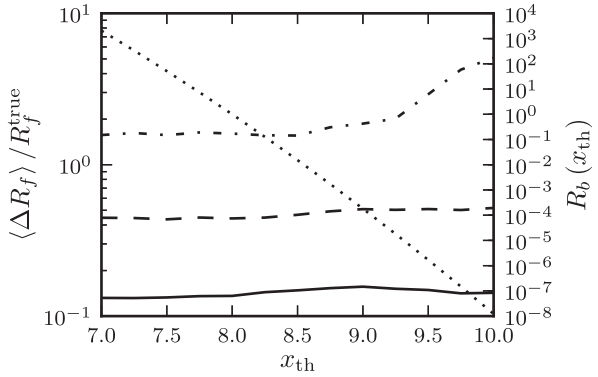

FIG. 1. The mean foreground rate uncertainty, Eq. (45), as a function of threshold for data sets with $\Gamma_f = 100$ (solid line), $\Gamma_f = 10$ (dashed line), and $\Gamma_f = 1$ (dash-dotted line). Recall that Γ_f is the mean number of foreground events above $x_{\rm th} = x_1 = 8$. The total background rate, $R_b(x_{\rm th})$, is shown by the dotted line; we fix $R_b(x_{\rm th} = x_1 = 8) = 1$, so on average there is one background event above $x = 8$. For $x_{\rm th} \gtrsim x_1$, increasing the threshold tends to increase the foreground rate uncertainty because the rate is foreground dominated and fewer events are included in the data set. For $x_{\rm th} \lesssim x_1$, the background rate dominates at small loudness, and the foreground rate uncertainty asymptotes to the counting error on the events that stand out from the background, $\Delta R_f / R_f \simeq 1/\sqrt{\Gamma_f}$.

choose the threshold *well into the background-dominated loudness regime*; the extra background events in the data set do not affect the estimate of the foreground rate, and, when the background distribution is parametrized, can help to better determine these parameters (see Sec. V B).

Though we have only illustrated the behavior of the rate estimate quantitatively for this specific example of foreground and background rates, the conclusions hold in general. Consider the Fisher information matrix for the posterior distribution in Eq. (21). For a model with parameters $\{\theta_i\}$, the Fisher information matrix has components

$$F_{ij} \equiv \left\langle \frac{\partial \log p(\theta|d)}{\partial \theta_i} \frac{\partial \log p(\theta|d)}{\partial \theta_j} \right\rangle, \tag{47}$$

where the average is taken over the data distribution at fixed θ, $p(d|\theta)$. The components of the Fisher information matrix describe the maximum amount of information about the corresponding parameters available in a given data set; the inverse of the Fisher information matrix gives the Cramer-Rao bound on the covariance matrix of unbiased estimators of θ. Though our Bayesian analysis is not necessarily limited by the Cramer-Rao bound (since estimators constructed from it need not be unbiased and can also be affected by the prior), the Fisher information is indicative of the influence of each measurement on the posterior. For the likelihood that enters Eq. (21), the Fisher information matrix is

$$\mathbf{F} = (R_f + R_b) \times \begin{pmatrix} \left\langle \left(\frac{\hat{f}}{R_f \hat{f} + R_b \hat{b}} \right)^2 \right\rangle & \left\langle \frac{\hat{f}\,\hat{b}}{(R_f \hat{f} + R_b \hat{b})^2} \right\rangle \\ \left\langle \frac{\hat{f}\,\hat{b}}{(R_f \hat{f} + R_b \hat{b})^2} \right\rangle & \left\langle \left(\frac{\hat{b}}{R_f \hat{f} + R_b \hat{b}} \right)^2 \right\rangle \end{pmatrix}, \tag{48}$$

where the expectation values are taken over the distributions $\hat{f}$ and $\hat{b}$ (i.e., they are expectations for one event from the combined rate distribution). If the cross terms are small, then the Cramer-Rao bound on the uncertainty of R_f will be given by

$$\sigma_{R_f} \simeq \frac{1}{\sqrt{R_f + R_b}} \left[\left\langle \left(\frac{\hat{f}}{R_f \hat{f} + R_b \hat{b}} \right)^2 \right\rangle \right]^{-1/2}. \tag{49}$$

Extending a threshold into regions where the factor

$$\left(\frac{\hat{f}}{R_f \hat{f} + R_b \hat{b}} \right)^2 \tag{50}$$

becomes small—that is, into background-dominated regions—contributes little to reducing the overall uncertainty in the foreground rate. Thus, when the background distribution itself is of no interest and computational costs are high, the threshold does not need to be pushed into background-dominated regions in order to obtain an accurate foreground estimate. This is consistent with the behavior of the specific example in Fig. 1.

A. Extreme sensitivity of the LE rate estimate to a single, unusually loud event

Here we discuss a very unattractive feature of the Bayesian loudest-event estimate of R [9]: a small percentage of the time it will yield a very large overestimate.

To explain this, we will use the same model as described in the previous subsection, and we will begin with a very specific example. Let $\Gamma_f = 1$, meaning that the expected number of actual events with $x > x_1$ is one. Then there is a $1/64$ chance $[1 - \hat{F}(x_{\rm LE}) \approx 1.6\%]$ that the loudest event will have $x_{\rm LE} > 4x_1$. Consider this case, and let us also assume that there are no events (noise or actual) with $x_1 < x < x_{\rm LE}$.

The loudest-event estimate basically "throws away" the information that there are no events in this interval. The maximum of the loudest-event-statistic posterior on R_f, Eq. (28), is at $R_f = \frac{\Lambda - 1}{\Lambda(1 - \hat{F}(x_{\rm LE}))}$. If the value of Λ is sufficiently high at x_1 (and Λ will be even greater at $x_{\rm LE}$), then, for this data set, we would estimate $R_f \approx \frac{1}{1 - \hat{F}(x_{\rm LE})} \gtrsim 64$. Thus, for our assumed shape of the foreground distribution, we will estimate the rate of events above x_1 to be 64 times the true rate.

Now, if the true rate really were $\Gamma_f = 64$, then the expected number of events with $x > x_1$ would be 64. So in this case, the loudest-event estimate ignores the fact that there are ~56–72 "missing" events. However a Bayesian estimate with $x_{\rm th}$ set to x_1 incorporates this information quite naturally, and so (correctly) yields an estimated Γ_f of order one.

V. EXAMPLES

In this section we present several examples of the application of our framework to various rate estimation problems in the presence of background.

A. Gravitational waves with non-overlapping templates

Suppose we attempt to detect gravitational-wave signals in a data stream by matched filtering in the frequency domain against a set of N template waveforms [e.g., [5,13]]. We use an extremely simplified model of such a search and the ensuing analysis to demonstrate how our framework could be used in practice.

In our simplistic model, we suppose the data stream consists of stationary Gaussian noise with a power spectral density $S(f)$ combined additively with some number of gravitational-wave signals. We assume that the signals are

sufficiently rare that they do not overlap in the data stream. The SNR of a template, $h(f)$, given data, $d(f)$, is

$$\rho_h \equiv \frac{\langle h, d\rangle}{\sqrt{\langle h, h\rangle}}, \tag{51}$$

where $\langle\cdot\rangle$ denotes the noise-weighted inner product:

$$\langle a, b\rangle \equiv 4\Re \int_0^\infty df \frac{a^*(f)b(f)}{S(f)}. \tag{52}$$

We suppose for simplicity that the templates are sufficiently distinct that

$$\langle h_i, h_j\rangle \simeq \delta_{ij}. \tag{53}$$

In the following subsection, we will generalize the model to overlapping templates. We rank candidate events by their maximum SNR over the entire template bank,

$$x \equiv \max_h \rho_h, \tag{54}$$

and consider only events that have a maximum SNR above some threshold, $x > x_{\rm min}$.

For a data stream of pure noise, $d(f) = n(f)$, the SNRs of the templates are independent $N(0, 1)$ random variables. The background ranking statistic (i.e., the maximum SNR over the template bank) then has a cumulative distribution without thresholding of

$$\hat{B}(x) = \left(\frac{1 + erf(\frac{x}{\sqrt{2}})}{2}\right)^N \tag{55}$$

where $\mathrm{erf}(x)$ is the error function as before. Imposing the threshold, $x > x_{\rm min}$, the cumulative distribution of the background becomes

$$\hat{B}(x) = \frac{(1 + erf(\frac{x}{\sqrt{2}}))^N - (1 + erf(\frac{x_{\rm min}}{\sqrt{2}}))^N}{2^N - (1 + erf(\frac{x_{\rm min}}{\sqrt{2}}))^N} \tag{56}$$

for $x > x_{\rm min}$, 0 otherwise.

The SNR of a gravitational-wave signal in an interferometric detector scales as $1/d$ [15], where d is the distance to the source. Ignoring cosmological effects, the number of sources scales as d^3. Thus, we expect that the foreground cumulative distribution of events will follow

$$\hat{F}(x) = 1 - \frac{x_{\rm min}^3}{x^3}. \tag{57}$$

Note that this scenario has no shape parameters θ for the foreground and background distributions.

To demonstrate the effectiveness of our formalism, we applied it to a synthetic data set with foreground and background distributions drawn from Eqs. (56) and (57) using $x_{\rm min} = 3.5$, with $R_f^{\rm true} = 10.4$ and $R_b^{\rm true} = 95.1$ and 1000 templates. The synthetic data consisted of 13 foreground events and 85 background events; the cumulative distribution for the ranking statistic of the synthetic data appears in Fig. 2. We used a Markov chain Monte Carlo simulation to draw samples of state flags and rates from the joint posterior [Eq. (18)].

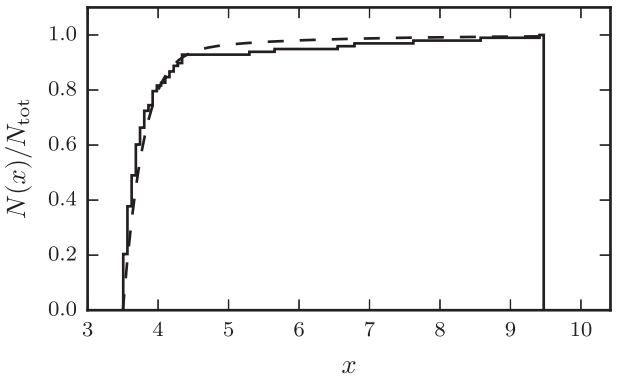

FIG. 2. The cumulative distribution of the ranking statistics for the synthetic data used to test the formalism on the model from Sec. V A. The solid line gives the cumulative distribution of the synthetic data; the dashed line gives the theoretical cumulative distribution for the models in Eqs. (56) and (57) combined with $R_f = 10.4$ and $R_b = 95.1$.

In Fig. 3, we show the marginalized posterior densities for the foreground and background rates [see Eq. (21)]. Figure 4 shows the posterior foreground probability for each event marginalized over all other events' types and the foreground and background rates.

We can compare these results to results obtained using the two approximations described earlier, the loudest-event statistic and the foreground-dominated statistic. The marginalized distribution for the foreground rate using these alternatives are shown in Fig. 5. In this case, the loudest

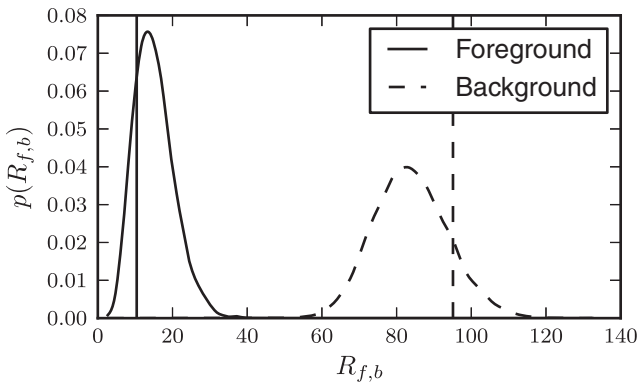

FIG. 3. The marginalized posterior densities for R_f (solid line) and R_b (dashed line) for the analytic model discussed in Sec. V A. The vertical lines indicate the "true" values used to generate the synthetic data set. Both the true foreground and background rates lie well within the probability envelope for R_f and R_b.

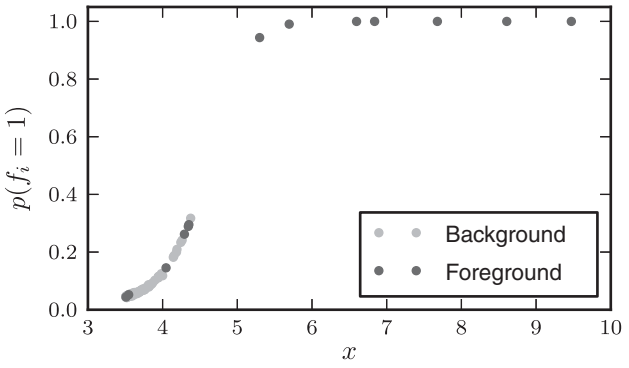

FIG. 4. Foreground probability for each event in the synthetic data set of Sec. VA marginalized over all other parameters. True foreground events are in dark grey, background events in light grey. Even though our method cannot identify the status of most events with confidence, it can still correctly estimate the rates (Fig. 3).

event had $x_N \simeq 9.47$. The loudest-event statistic depends on a specification of the maximum, $R_{\max}$, for the background rate. We show results for $R_{\max} = \infty$, i.e., the improper prior, and $R_{\max} = 10000$. The results for other reasonable choices of $R_{\max} = 100, 1000, 100000$ etc. gave exactly the same posterior, since $\hat{b}(x_N)R_{\max} \ll \hat{f}(x_N)$ for all these choices and we are therefore in the regime where the posterior is insensitive to $R_{\max}$. To apply the foreground-dominated statistic we must specify a threshold above which we assume all events are foreground. It is reasonable to do this based on a specification for the relative probability of an event being fore/background, $\hat{f}(x)/\hat{b}(x) = p_{\text{thresh}}$. Setting $p_{\text{thresh}} = 0.99$ gives $x_{\min} = 4.07$ and there are $N = 18$ (11 foreground and 7 background) events exceeding that threshold. Setting $p_{\text{thresh}} = 0.5$ gives $x_{\min} = 3.82$ and there are $N = 30$ (11 foreground and 19 background) events exceeding that threshold. Each of these thresholds gives a biased estimate of the rate because there are background events still above threshold. The "omniscient" threshold of $x_{\min} = 4.38$ produces $N = 7$ (7 foreground and 0 background) events in this data set, and therefore an unbiased estimate, but of course this threshold can only be determined because we can examine the synthetic foreground and background data samples. The threshold may seem obvious from a visual examination of Fig. 4; however, the construction of this figure relies on the application of the full framework in the first place. We show results for the first two choices of $x_{\min}$ in Fig. 5; the omniscient choice produces essentially the same posterior as our full analysis.

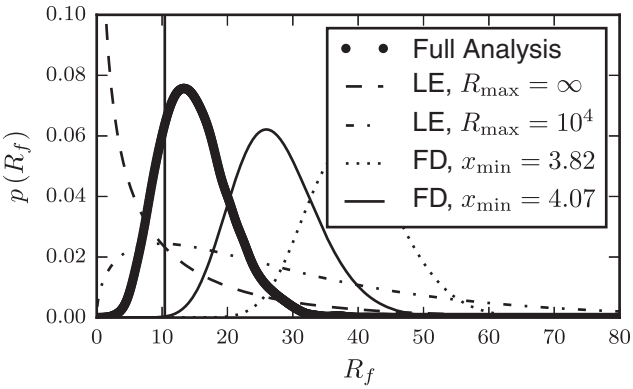

FIG. 5. Posteriors on foreground rate obtained using the method described in this paper, the loudest-event statistic and the foreground-dominated analysis for the data set from Sec. VA. For the loudest-event statistic, we present the posterior with and without an upper limit on the background rate, R_b; in both cases the rate posterior is significantly wider than the one obtained with the method described in this paper. For the foreground-dominated statistic, the limits $x_{\min} = 3.82$ and $x_{\min} = 4.07$ give likelihood ratios of $\hat{f}/\hat{b} = 0.5$ and 0.99. For this data set, the thresholds in fact include 19 and 7 background events, respectively, so the corresponding rate estimates are significantly biased. An omniscient threshold of $x_{\min} = 4.38$ would produce exactly 7 foreground and zero background events, and the resulting posterior is essentially indistinguishable from the curve for the full analysis.

The loudest-event statistic with the improper prior gives, as expected, a poor approximation to the foreground rate. The peak is more accurately located when a prior maximum rate is defined, but the distribution is much wider than using the full analysis described here in any case. This is to be expected as much of the information is being thrown away. The foreground-dominated statistic gives a reasonable approximation to the true foreground rate, and a distribution that is essentially equal to the full analysis, for the omniscient choice of threshold value that excludes all background data. For lower thresholds, even for a threshold where $p_{\text{thresh}} = 0.99$, it performs poorly since we are approximating the foreground rate by the total foreground plus background rate. This indicates that, provided the threshold is chosen appropriately, the foreground-dominated statistic can perform quite well at estimating the rate—but choosing this threshold correctly is difficult. The fact that it reproduces the posterior from the full analysis so well is indicative of the fact that most of the information about the foreground comes from the loudest events. The full analysis naturally incorporates inference about the background rate R_b along with the foreground rate and incorporates maximum information from the data set and should therefore lead to narrower posteriors in general.

B. Gravitational waves with overlapping templates

In Sec. VA we assumed that the overlap between different templates in the template bank was negligible, so the SNRs recovered by different templates are independent random variables. In fact, template banks are not constructed in this way [e.g., [16,17]], because signals could fall in the gaps between the non-overlapping templates. We can model this effect by assuming that a template bank of N actual templates will behave as if it

had $N_{\rm eff}$ *independent* templates. Rather than pre-computing $N_{\rm eff}$, we can fit for it as a shape parameter. That is, we assume that $\theta = \{N_{\rm eff}\}$ is a shape parameter for the background cumulative distribution:

$$\hat{B}(x, N_{\rm eff}) = \frac{(1 + erf(\frac{x}{\sqrt{2}}))^{N_{\rm eff}} - (1 + erf(\frac{x_{\rm min}}{\sqrt{2}}))^{N_{\rm eff}}}{2^{N_{\rm eff}} - (1 + erf(\frac{x_{\rm min}}{\sqrt{2}}))^{N_{\rm eff}}}. \tag{58}$$

Results from such an analysis appear in Figs. 6 and 7. We use the same parameters and data set as in Sec. V A, with $x_{\rm min} = 3.5$, $R_f = 10.4$, $R_b = 95.1$, and $N_{\rm eff} = 1000$, but now allow $N_{\rm eff}$ to be a parameter of the background distribution, with a flat prior. Both the rates and the number of effective templates are recovered without significant loss of accuracy relative to the fixed $N_{\rm eff}$ situation in Sec. V A.

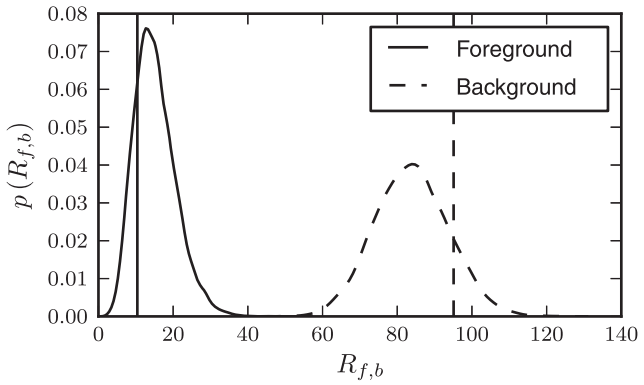

FIG. 6. The foreground (solid lines) and background (dashed lines) rate posterior, marginalized over all flags and the $N_{\rm eff}$ parameter, for the gravitational-wave template detection scenario with overlapping templates discussed in Sec. V B. The true values of the rates, $R_f = 10.4$ and $R_b = 95.1$, are indicated with vertical lines. The distributions are not significantly wider than those of Fig. 3, in spite of the extra parameter.

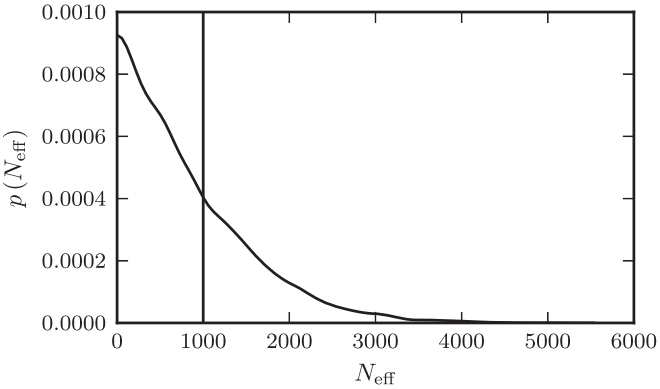

FIG. 7. The posterior on the number of effective templates, $N_{\rm eff}$, for the model and data discussed in Sec. V B, marginalized over all state flags and rates. The true value, $N_{\rm eff} = 1000$, is indicated by the vertical line.

If we consider the two alternative methods, the loudest-event and foreground-dominated statistics, and apply the same foreground-dominated thresholds as before, we will recover the same foreground distributions as are shown in Fig. 5. This is because the parameter $N_{\rm eff}$ affects only the background distribution, to which the foreground-dominated statistic is insensitive, and in the loudest-event case, after marginalization over $N_{\rm eff}$ we find $\int_0^{N_{\rm max}} \hat{b}(x_N, N_{\rm eff}) {\rm d}N_{\rm eff} \ll 3N_{\rm max}/R_{\rm max}\hat{f}(x_N, N_{\rm eff})$ and so we are still in the foreground-dominated regime in which the loudest event tells us nothing about the background. Neither of these alternative methods can inform us about the value of $N_{\rm eff}$, a property of the distribution of background events identified by filtering with this template bank. Moreover, the choice of threshold value for the foreground-dominated statistic becomes significantly more complicated in this case, since $p_{\rm thresh}$ now depends on $N_{\rm eff}$.

C. Uncertainty in the foreground and background distributions

The framework outlined above relies on the existence of models for the foreground, $\hat{f}(x, \theta)$, and background, $\hat{b}(x, \theta)$, distributions parametrized by a small number of model parameters, θ. While in many situations simple analytic functions such as power laws will provide an adequate description, this will not always be the case. In the absence of a good analytic model, the space of the ranking statistic x could be divided into bins and $\hat{f}(x)$ and $\hat{b}(x)$ are taken to be flat in each of these bins. The number of free parameters characterizing each of $\hat{f}$ and $\hat{b}$ is then the number of bins used. While such a framework is model free, the increase in model parameters will mean that more observed events will typically be required to achieve the same precision on the rates and foreground/background distributions.

In the context of gravitational-wave experiments, additional information on the ranking-statistic distributions for the foreground can be obtained using mock signal injections into the data, while distributions for the background can be estimated by analyzing time slides of data sets from different detectors relative to each other [e.g., [18]]. This information can be readily incorporated in the current framework by assuming there is another set of N_I events with ranking statistics $\{w_i\}$, known to be drawn from the foreground distribution ($g_i = 1$) and a set of N_T events with ranking statistics $\{z_i\}$ known to be drawn from the background distribution ($g_i = 0$). These events will typically not be drawn with the correct rate parameters, so they do not contribute to the estimates of R_f and R_b, but they do contribute an extra factor

$$\prod_{l=1}^{N_I} \hat{f}(w_l, \theta) \prod_{m=1}^{N_T} \hat{b}(z_i, \theta) \tag{59}$$

to the right-hand sides of Eqs (18) and (21). This approach provides a way to incorporate extra information into the analysis in order to simultaneously fit for the shape of the background and foreground as well as the rates. In the limit that there are many more events in the time slide and injection data set, this will reduce to the analysis that was described above with fixed ranking statistic distributions $\hat{f}(x)$ and $\hat{b}(x)$ given by the injection and time slide data. We note that this analysis makes the assumption that the background distribution is the same in the time slide and real data and that the foreground distribution is the same between the injection and real data. The former assumption is probably reasonable, modulo correlations of non-gravitational-wave origin between data in different detectors, but the latter relies on knowledge of the relative the astrophysical rates of different events, which is more uncertain. These astrophysical uncertainties could be handled with a hybrid approach, in which injections are used to characterize the statistic distribution for sources of a particular type, while additional rate or shape parameters are introduced to characterize the variation in the astrophysical rate of mergers as a function of source type.

D. Star cluster parameters with background contamination

Our final example concerns fitting for the location and shape parameters of a cluster of stars observed on top of a stellar background with a density gradient. In this example, stars are either members of the cluster (i.e., foreground) or background contamination, with a spatially varying density (i.e., our rate functions are two dimensional). Our method of analysis here is similar to that of De Gennaro *et al.* [8], but here we marginalize over membership flags and are simultaneously fitting foreground and background densities (i.e., rates) *and* cluster properties.

We assume that a star cluster has a Plummer surface-density profile [19,20],

$$\hat{f}(\vec{x},\theta) = \frac{1}{\pi r_0^2\left(1+\frac{|\vec{x}-\vec{x}_0|^2}{r_0^2}\right)^2}, \tag{60}$$

where $\vec{x}_0$ is the location on the sky of the center of the cluster, r_0 is a radial scale parameter, and $\vec{x} = (x, y)$ is the position on the sky. We assume a square observational domain,[1] $\vec{x} \in [0,1]^2$, and a background that has a density gradient at an arbitrary orientation with respect to the observational axes:

$$\hat{b}(\vec{x},\theta) = 1 + \vec{\gamma}\cdot(\vec{x}-\vec{x}_{1/2}), \tag{61}$$

where $\vec{\gamma}$ is the gradient, and $\vec{x}_{1/2} = [1/2, 1/2]$ is the centroid of the observational domain.

We use simulated data drawn from our model with parameters

$$\theta_0 \equiv \{x_0, y_0, r_0, \gamma_x, \gamma_y\} = \left\{\frac{1}{2}, \frac{1}{2}, 0.18, -\frac{1}{2}, \frac{1}{2}\right\}, \tag{62}$$

with $R_f = 1000$ and $R_b = 10000$. For this set of parameters, the average density of the background and the peak density of the cluster are comparable; there are an order of magnitude more background stars than cluster stars in the field. Figure 8 shows the density of stars on the sky and the particular synthetic data set used for this analysis. Because the peak density of the cluster is equal to the background density at the center of the domain, there is no single star in the domain that is more likely to be a cluster member than a background star (i.e., $\langle g_i \rangle \lesssim 0.5$ for all stars); nevertheless, we will see that our method provides good constraints on the cluster parameters.

To analyze our synthetic data set, we analytically marginalized over the state flags (i.e., cluster membership), using the likelihood in Eq. (21). We did this to take advantage of the emcee sampler of Foreman-Mackey *et al.* [21], which requires all parameters to be in $\mathbb{R}$. We applied a prior on the shape parameters that is flat in $\vec{x}_0$ and $\vec{\gamma}$, and an (approximately) Jeffreys prior on r_0,

$$p(r_0) = \frac{\sqrt{R_f}}{r_0}. \tag{63}$$

(Note that this factor of $\sqrt{R_f}$ cancels with the Jeffreys prior on the rate, $1/\sqrt{R_f}$; we have verified that the priors on these parameters are irrelevant to our results, as would be expected from the measurement of $\sim$1000 foreground stars.)

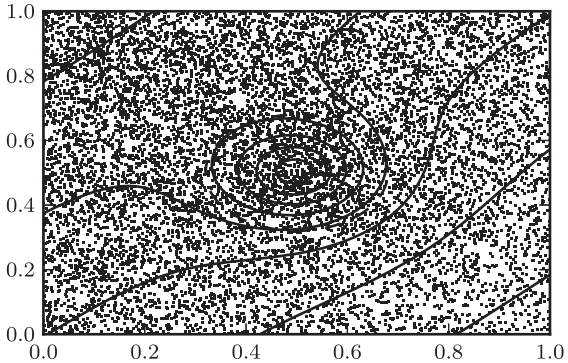

FIG. 8. Density contours and synthetic data for the example in Sec. V D. The contours describe the true density profile with the parameters in Eq. (62). The points are the realization of this density profile used as synthetic data in Sec. V D; the dashed line encloses one Plummer scale radius about the true cluster center. Because the peak cluster density is equal to the background density at the cluster center, the cluster is barely apparent to the eye.

[1]The observational domain is not infinite, so the normalization of the cluster density in Eq. (60) is not quite correct. In our modeling we properly take this into account, but for simplicity here we ignore it.

Figures 9 and 10 shows the posteriors for the cluster location and scale parameters. The center of the cluster, $\vec{x}_0$, is localized to within about 5% of the cluster scale, and the cluster radius with a relative error of about 10%. In spite of the significant background, the cluster parameters are recovered to a relative accuracy consistent with the expected uncertainty from $N_{\rm eff} \simeq R_f = 1000$ measurements. Figure 11 shows the posteriors inferred on the cluster and background numbers, R_f and R_b.

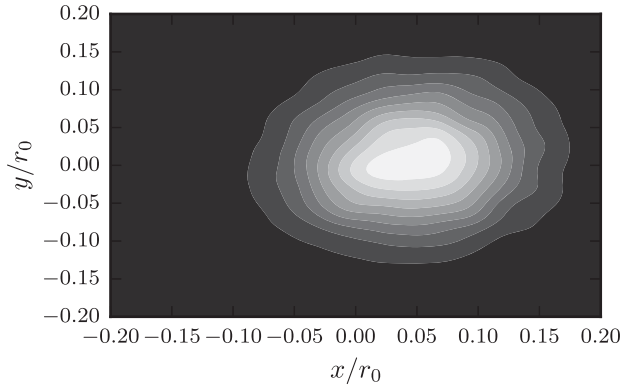

FIG. 9. Contours of the posterior probability distribution for the center of the cluster, $\vec{x}_0$, for the example from Sec. V D. The center $(x, y) = (x_0, y_0)$ is determined to within about 5% of the structural radius of the cluster, r_0 [see Eq. (62)].

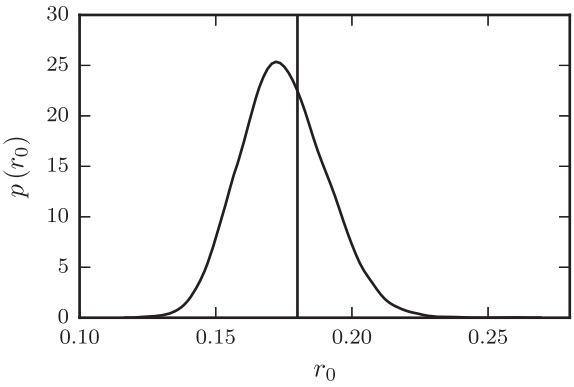

FIG. 10. Posterior density for the scale parameter for the cluster, r_0, for the example from Sec. V D. The true value is indicated by the vertical line [see Eq. (62)].

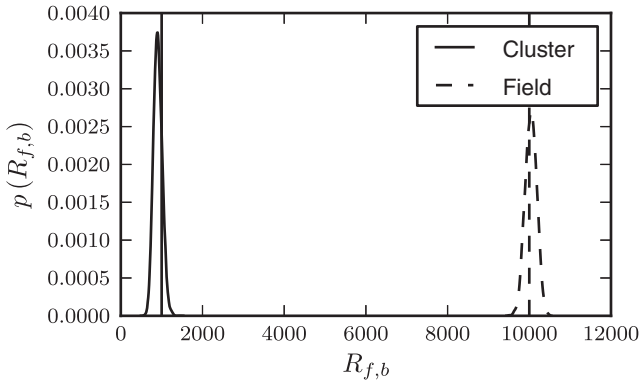

FIG. 11. Posterior densities for the number of stars in the cluster (R_f) and in the field (R_b) in the example from Sec. V D. Vertical lines indicate the true values [see Eq. (62)].

VI. DISCUSSION

In this paper, we have developed a Bayesian framework for rate estimation when the data consists of a mixture of foreground and background events. We demonstrated the application of this framework using several examples from gravitational-wave data analysis in the presence of signatures of binary mergers and noise triggers, and astronomical image analysis in the presence of several populations of stars. We showed that this framework is generally superior to both the loudest-event statistic and the foreground-dominated statistic.

Through most of this paper, we have assumed that the shape of the foreground and background distributions is known, or at least can be modeled with several additional parameters. This is not necessarily easy to do. For example, in the case of gravitational-wave data analysis, the shape of the foreground distribution of events may depend on the details of a complex data-analysis pipeline as well as the astrophysical source distribution, while the background event distribution depends on data quality and may deviate significantly from the simple Gaussian-noise behavior modeled in Sec. V. Several approaches have been developed to accurately model both distributions, e.g., through the use of injected signals [18] or other methods [22] to model the foreground distribution. However, this is a difficult problem (e.g., because of the need to estimate the background at the very tails of the distribution), and will require significant future work. In Sec. V C, we discussed some of the possible approaches when the shapes of the background and foreground distributions cannot be confidently described by models with a few adjustable parameters.

A further complication is that we have considered the rate of events in the data as products of some analysis pipeline. This rate may be different from the physical rate of interest, such as the rate of compact-binary mergers per unit time per unit volume which generate gravitational waves, or the physical numbers of stars in the cluster and field populations which produce the observed luminosities. Again, the conversion between the two will depend on the details of the data-analysis algorithm and ranking statistic, including any selection effects [12], and would need to be determined on a case-by-case basis. See Ref. [11] for an example of such conversion when the underlying framework is the loudest-event statistic.

Furthermore, in a practical application there could be multiple classes of events, not just foreground and background. For example, we are not necessarily interested in the rate of gravitational-wave signals per se, but separately in the rate of signals from mergers of binary

neutron stars and binary black holes—populations that may sometimes be difficult to distinguish. Our approach is readily extendable to this particular complication, however. Note that it is symmetric with respect to foreground and background events (as expected, since one physicist's background is another physicist's foreground). We could relabel foreground and background events into other competing event classes, and further classes could be added in a straightforward way. However, the ability to distinguish classes relies on different distributions of their statistics. In general, rankings may need to be extended to include other statistics in addition to the signal "loudness" statistic in order to indicate both event significance and the probability of event attribution to a particular class.

ACKNOWLEDGMENTS

We thank Kipp Cannon, Thomas Dent, Chad Hanna, Drew Keppel, Richard O'Shaughnessy, David Hogg, and Ted von Hippel for discussions and suggestions about this manuscript. I. M. and W. M. F. acknowledge the hospitality of KITP, supported in part by the National Science Foundation under NSF Grant No. PHY11-25915. C. C.'s work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract to the National Aeronautics and Space Administration. C. C. also gratefully acknowledges support from NSF Grant No. PHY1068881. J. G.'s work is supported by the Royal Society.

[1] B. Abbot *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **79**, 122001 (2009).
[2] J. Abadie *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **83**, 122005 (2011).
[3] J. Abadie *et al.*, Phys. Rev. D **86**, 069903(E) (2012).
[4] B. Abbot *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **80**, 047101 (2009).
[5] J. Abadie *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **85**, 082002 (2012).
[6] J. Aasi *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **87**, 022002 (2013).
[7] S. E. Koposov, H.-W. Rix, and D. W. Hogg, Astrophys. J. **712**, 260 (2010).
[8] S. De Gennaro, T. von Hippel, W. H. Jefferys, N. Stein, D. van Dyk, and E. Jeffery, Astrophys. J. **696**, 12 (2009).
[9] R. Biswas, P. R. Brady, J. D. E. Creighton, and S. Fairhurst, Classical Quantum Gravity **26**, 175009 (2009).
[10] P. R. Brady, J. D. E. Creighton, and A. G. Wiseman, Classical Quantum Gravity **21**, S1775 (2004).
[11] P. R. Brady and S. Fairhurst, Classical Quantum Gravity **25**, 105002 (2008).
[12] C. Messenger and J. Veitch, New J. Phys. **15**, 053027 (2013).
[13] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, Phys. Rev. D **85**, 122006 (2012).
[14] H. Jeffreys, Proc. R. Soc. A **186**, 453 (1946).
[15] L. S. Finn and D. F. Chernoff, Phys. Rev. D **47**, 2198 (1993).
[16] B. J. Owen and B. S. Sathyaprakash, Phys. Rev. D **60**, 022002 (1999).
[17] P. Ajith, S. Babak, Y. Chen, M. Hewitson, B. Krishnan, A. M. Sintes, J. T. Whelan, B. Brügmann, P. Diener, N. Dorband, J. Gonzalez, M. Hannam, S. Husa, D. Pollney, L. Rezzolla, L. Santamaría, U. Sperhake, and J. Thornburg, Phys. Rev. D **77**, 104017 (2008).
[18] S. Babak *et al.*, Phys. Rev. D **87**, 024033 (2013).
[19] H. C. Plummer, Mon. Not. R. Astron. Soc. **71**, 460 (1911).
[20] S. J. Aarseth, M. Henon, and R. Wielen, Astron. Astrophys. **37**, 183 (1974).
[21] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[22] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **88**, 024025 (2013).