Bayesian Rate Estimation

Will M. Farr[*]
Center for Interdisciplinary Exploration and Research in Astrophysics
Department of Physics and Astronomy
Northwestern University, 2145 Sheridan Road, Evanston, IL 60208

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

We show how to obtain a Bayesian estimate of the rate of signal events from a set of signal and background 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. We give examples of determining the rates of gravitatonal-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 and location parameters from an observation of a stellar field that contains a non-uniform background density of stars superimposed on the cluster stars.

[*] w-farr@northwestern.edu; http://faculty.wcas.northwestern.edu/will-farr/
[†] jrg23@cam.ac.uk
[‡] imandel@star.sr.bham.ac.uk; http://www.sr.bham.ac.uk/~imandel

I. INTRODUCTION

FIXME: Introduce the necessity of estimating rates, prior work (like [1]), Bayesian inference.
Difficulty of measuring rate in presence of background
Want to add more than just the loudest event (cf. loudest event statistic)
Unknown background distribution, Λ
Cf. number of gold-plated detections in sensitive volume – frequentist approach OK if tons of events, otherwise need this.

II. MODEL

We assume that we are presented with a data set of N events that exceed a pre-specified threshold in ranking statistic, x_{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:

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

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. We will use d to denote ranking statistic-ordered events, and d_{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

$$\frac{dN_f}{dx} = f(x, \theta) \tag{2}$$

and

$$\frac{dN_b}{dx} = b(x, \theta), \tag{3}$$

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

$$F(x, \theta) \equiv \int_{-\infty}^{x} ds\, f(s, \theta) \tag{4}$$

and

$$B(x, \theta) \equiv \int_{-\infty}^{x} ds\, b(s, \theta). \tag{5}$$

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

$$P(n = 1 \in [x, x + h]) = f(x, \theta) h + \mathcal{O}\left(h^2\right). \tag{6}$$

   and similarly for background events.

4. The probability of two or more events in a small range of ranking statistic is negligible

$$P(n = 2 \in [x, x + h]) = \mathcal{O}\left(h^2\right). \tag{7}$$

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 amplitude of the foreground and background rates. Let

$$f(x,\theta) = R_f \hat{f}(x,\theta'), \tag{8}$$

and

$$b(x,\theta) = R_b \hat{b}(x,\theta'), \tag{9}$$

where $\hat{F}(\infty,\theta') = \hat{B}(\infty,\theta') = 1$, and $\theta' = \theta \setminus \{R_f, R_b\}$. 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, f_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 $f_i = \{0, 1\}$.

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

$$p\left(\{f_i\}, R_f, R_b, \theta | d_{\mathrm{to}}, N\right) = \frac{p\left(d_{\mathrm{to}} | \{f_i\}, N, R_f, R_b, \theta\right) p\left(\{f_i\}, N, R_f, R_b, \theta\right)}{p(d_{\mathrm{to}}, N)}. \tag{10}$$

The normalization constant, called the evidence, $p(d_{\mathrm{to}}, N)$, 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

$$p\left(d_{\mathrm{to}} | \{f_i\}, N, R_f, R_b, \theta\right) = \left[\prod_{\{i|f_i=1\}} \hat{f}\left(x_i, \theta\right)\right] \left[\prod_{\{i|f_i=0\}} \hat{b}\left(x_i, \theta\right)\right]. \tag{11}$$

This is the probability that the first observed event is a fore/background event (if $f_i = 1, 0$) with ranking statistic x_1 and the second observed event is a fore/background event (if $f_i = 1, 0$) 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

$$p\left(\{f_i\}, N, R_f, R_b, \theta\right) = p\left(\{f_i\} | N, R_f, R_b\right) p\left(N | R_f, R_b\right) p\left(R_f, R_b, \theta\right) = p\left(\{f_i\}, N | R_f, R_b\right) p\left(R_f, R_b, \theta\right). \tag{12}$$

The probability that the i'th state flag is $f_i = 1$ 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

$$p\left(\{f_i\} | N, R_f, R_b\right) = \prod_{\{i|f_i=1\}} \left(\frac{R_f}{R_f + R_b}\right) \prod_{\{i|f_i=0\}} \left(\frac{R_b}{R_f + R_b}\right) = \left(\frac{R_f}{R_f + R_b}\right)^{N_f} \left(\frac{R_b}{R_f + R_b}\right)^{N_b}, \tag{13}$$

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

$$p\left(N | R_f, R_b\right) = \frac{\left(R_f + R_b\right)^N}{N!} e^{-(R_f + R_b)}, \tag{14}$$

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:

$$p\left(\{f_i\}, N | R_f, R_b\right) = \frac{R_f^{N_f} R_b^{N_b}}{N!} \exp\left[-\left(R_f + R_b\right)\right]. \tag{15}$$

The second term in Eq. (12) 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 [2], independent of the shape parameters

$$p\left(R_f, R_b, \theta\right) = \alpha \frac{1}{\sqrt{R_f R_b}} p(\theta), \tag{16}$$

where α is a normalisation 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. (15) regularize the posterior as $R_f, R_b \to \infty$.

Putting everything together, the posterior is

$$p\left(\{f_i\}, R_f, R_b, \theta | d_{\mathrm{to}}, N\right) = \frac{\alpha}{p(d_{\mathrm{to}}, N)\, N!} \left[\prod_{\{i|f_i=1\}} R_f \hat{f}\left(x_i, \theta\right)\right] \left[\prod_{\{i|f_i=0\}} R_b \hat{b}\left(x_i, \theta\right)\right] \exp\left[-\left(R_f + R_b\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}} \tag{17}$$

When sampling the posterior, the first term can be omitted and the equals sign replaced by proportionality, since it is independent of the parameters of interest, but 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 $p\left(\{f_i\}, R_f, R_b, \theta | d_{\mathrm{to}}, N\right)$ and the corresponding statistic-ordered posterior $p\left(\{f_i\}, R_f, R_b, \theta | d, N\right)$, which means $p\left(\{f_i\}, R_f, R_b, \theta | d, N\right) = p\left(\{f_i\}, R_f, R_b, \theta | d_{\mathrm{to}}, N\right)$. However, the evidence $p(d, N) = N!\, p(d_{\mathrm{to}}, N)$, since there are $N!$ ways in which N events can be ordered in time and would have the same set of ranking statistics.

Not surprisingly, this ranking-statistic ordered posterior can be computed directly by assuming that the flags, $\{f_i\}$, are un-observed data and treating the sets $\{x_i | f_i = 1\}$ and $\{x_i | f_i = 0\}$ 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

$$\begin{aligned} p\left(\{y_i\} | r\right) \mathrm{d}^N y_i &= P\left(\text{zero events below } y_0\right) \times P\left(\text{one event between } y_0 \text{ and } y_0 + \mathrm{d}y_0\right) \\ &\quad \times P\left(\text{zero events between } y_0 + \mathrm{d}y_0 \text{ and } y_1\right) \ldots \\ p\left(\{y_i\} | r\right) &= \lim_{\delta y_i \to 0} \exp\left[-R\left(y_0\right)\right] \left[r\left(y_0\right) + \mathcal{O}\left(\delta y_0\right)\right] \times \exp\left[-\left[R\left(y_1\right) - R\left(y_0 + \delta y_0\right)\right]\right] \times \ldots \\ &= \left[\prod_i r\left(y_i\right)\right] \exp\left[-R\left(\infty\right)\right]. \end{aligned} \tag{18}$$

Applying this once to the foreground samples, once to the background samples and taking the product, we obtain $p(d, \{f_i\}, N | R_f, R_b, \theta)$ and thence $p(\{f_i\}, R_f, R_b, \theta | d, N) = p(d, \{f_i\}, N | R_f, R_b, \theta) p * R_f, R_b, \theta)/p(d, N)$. With the identification $p(d, N) = N!\, p(d_{\mathrm{to}}, N)$, as justified above, we reproduce Eq. (17).

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

$$p\left(R_f, R_b, \theta | d, N\right) = \sum_{\{f_i\} \in \{0,1\}^N} p\left(\{f_i\}, R_f, R_b, \theta | d, N\right) \propto \prod_i \left[R_f \hat{f}\left(x_i, \theta\right) + R_b \hat{b}\left(x_i, \theta\right)\right] \exp\left[-\left(R_f + R_b\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}}. \tag{19}$$

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. (17)), Eq. (19) 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 marginalisation over either R_b or R_f can be achieved analytically using

$$I_n = \int_0^\infty x^{n-\frac{1}{2}}\, \mathrm{e}^{-x} \mathrm{d}x = \frac{(2n-1)!!}{2^n} \sqrt{\pi} \tag{20}$$

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

Eq. (17) is unchanged if the ranking statistic is multi-dimensional; in this case, the rates are

$$R_f = \int d^k \vec{x}\, f(x, \theta) \tag{21}$$

and

$$R_b = \int d^k \vec{x}\, b(x, \theta), \tag{22}$$

where f and b are rate densities on the k-dimensional space of ranking statistics. We give an example of fitting for multi-dimensional rate densities in § III C.

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

1. 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. (17)) would be modified by an additional factor of $\exp[R_f \hat{F}(x_{N-k+1}) + R_b \hat{B}(x_{N-k+1})]$, where we have assumed events are ordered by loudness, so that x_{N-k+1} is the k-th loudest event. This term accounts for the data-dependent threshold that a loudest event statistic employs.

For the usual $k = 1$ case, the marginalised posterior (Eq. (19)) becomes

$$p_{\mathrm{LE}}\left(R_f, R_b, \theta | d\right) \propto \left(R_f \hat{f}(x_N, \theta) + R_b \hat{b}(x_N, \theta)\right) \exp\left[-\left(R_f(1 - \hat{F}(x_N, \theta)) + R_b(1 - \hat{B}(x_N, \theta))\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}}. \tag{23}$$

where x_N denotes the loudness of the loudest event. Marginalising over R_b we obtain

$$p_{\mathrm{LE}}\left(R_f, \theta | d\right) \propto \left(\frac{\hat{b}(x_N, \theta)}{2(1 - \hat{B}(x_N, \theta))} + R_f \hat{f}(x_N, \theta)\right) \frac{\sqrt{\pi}}{\sqrt{1 - \hat{B}(x_N, \theta)}\,\sqrt{R_f}} \exp(-R_f(1 - \hat{F}(x_N, \theta))). \tag{24}$$

This posterior has a maximum in R_f at

$$R_f = \frac{\hat{f}(x_N, \theta) - (1 - \hat{F}(x_N, \theta)\tilde{b}(x_N, \theta) \pm \sqrt{(\hat{f}(x_N, \theta) - (1 - \hat{F}(x_N, \theta)\tilde{b}(x_N, \theta))^2 - 4\tilde{b}(x_N, \theta)(1 - \hat{F}(x_N, \theta))\hat{f}(x_N, \theta)}}{4\hat{f}(x_N, \theta)(1 - \hat{F}(x_N, \theta)}. \tag{25}$$

where $\tilde{b}(x_N, \theta) = \hat{b}(x_N, \theta)/(1 - \hat{B}(x_N, \theta))$. If $\tilde{b}(x_N, \theta) \ll 1$, i.e., the loudest event is very likely to be foreground, we obtain the result $(1 - \hat{F}(x_N, \theta))\, R_f \approx 1/2$. This is 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. If $\tilde{b}(x_N, \theta) >> 1$, however, the posterior on R_f is peaked at 0.

2. 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. The posterior can then be approximated by

$$p_{\mathrm{FD}}\left(R_f, R_b, \theta | d\right) \propto \prod_i \left[\hat{f}\left(x_i, \theta\right)\right] R_f^N \exp\left[-\left(R_f + R_b\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}}. \tag{26}$$

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

$$p_{\mathrm{FD}}\left(R_f, \theta | d\right) \propto \prod_i \left[\hat{f}\left(x_i, \theta\right)\right] R_f^{N - \frac{1}{2}} \exp\left[-R_f\right] p(\theta). \tag{27}$$

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. We note that this is the rate of events occurring above threshold and so if we were to use this statistic with different choices of threshold we would have to renormalise the rates to a fixed threshold in order to compare the results.

III. 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., 3, 4]. 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 signal to noise ratio (SNR) of a template, $h(f)$, given data, $d(f)$, is

$$\rho_h \equiv \frac{\langle h, d\rangle}{\sqrt{\langle h, h\rangle}}, \tag{28}$$

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

$$\langle a, b\rangle \equiv 4\Re \int_0^\infty df\, \frac{a^*(f) b(f)}{S(f)}. \tag{29}$$

We suppose for simplicity that the templates are sufficiently distinct that

$$\langle h_i, h_j\rangle \simeq \delta_{ij}. \tag{30}$$

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

$$x \equiv \max_h \rho_h, \tag{31}$$

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

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

$$\hat{B}(x) = \left(\frac{1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)}{2}\right)^N \tag{32}$$

Imposing the threshold, $x > x_{\min}$, the cumulative distribution of the background becomes

$$\hat{B}(x) = \frac{\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right)^N - \left(1 + \operatorname{erf}\left(\frac{x_{\min}}{\sqrt{2}}\right)\right)^N}{2^N - \left(1 + \operatorname{erf}\left(\frac{x_{\min}}{\sqrt{2}}\right)\right)^N} \tag{33}$$

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

The SNR of a gravitational wave signal in an interferometric detector scales as $1/d$ [5], 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

$$\hat{F}(x) = 1 - \frac{x_{\min}^3}{x^3}. \tag{34}$$

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. (33) and (34) with $R_f^{\rm true} = 10.4$ and $R_b^{\rm true} = 95.1$, using 1000 templates. **[Specify $x_{\min}$?]** 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 Figure 1. We used a Markov Chain Monte Carlo simulation to draw samples of state flags and rates from the joint posterior (Eq. (17)).

In Figure 2, we show the marginalized posterior densities for the foreground and background rates. (Refer to Eq. (19).) Figure 3 shows the posterior foreground probability for each event marginalized over all other events' types and the foreground and background rates.

[General comment on the text: I think we need to say a bit more about the figures and the interpretation of the results, instead of just relying on the reader to make the best of it. For example,

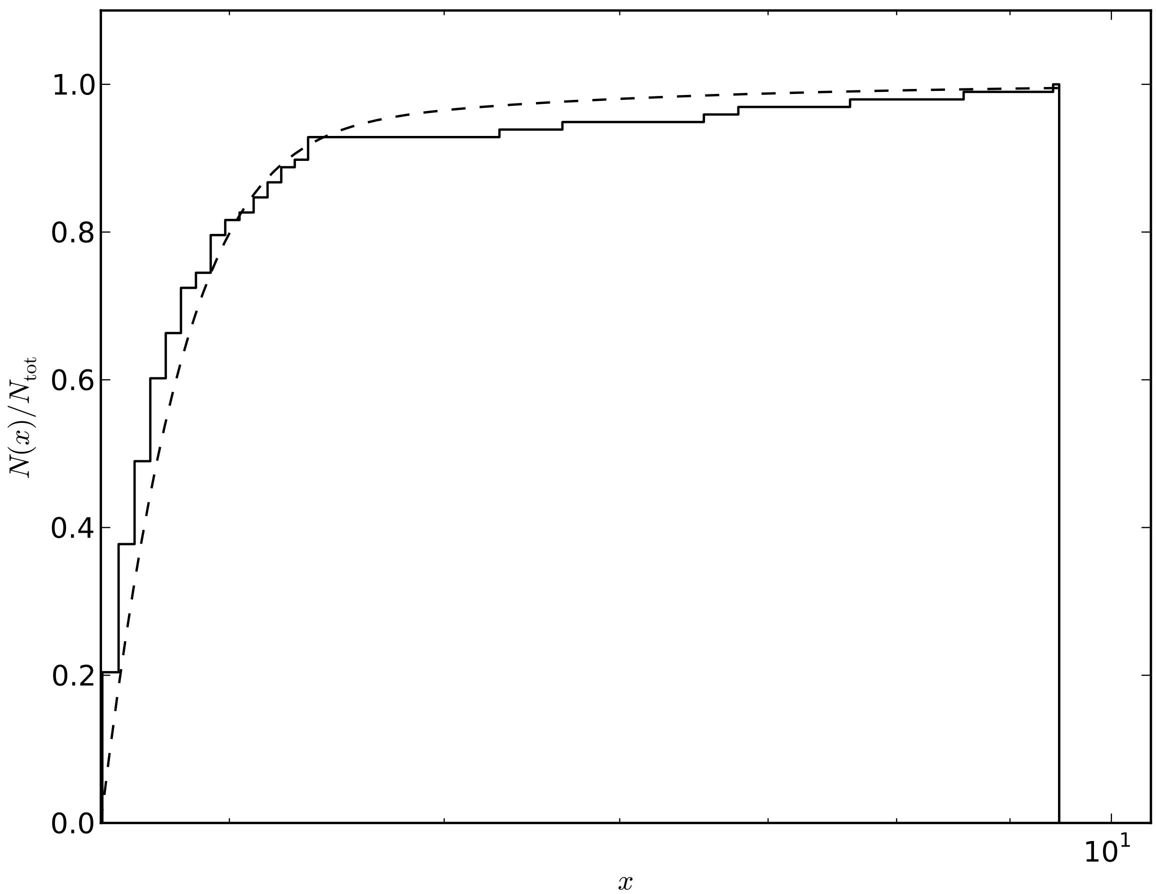

FIG. 1. The cumulative distribution of the ranking statistics for the synthetic data used to test the formalism on the model from §III A. The solid line gives the cumulative distribution of the synthetic data; the dashed line gives the theoretical cumulative distribution for the models in Eqs. (33) and (34) combined with $R_f = 10.4$ and $R_b = 95.1$.

even though it may seem obvious to us, we should explicitly say that Fig. 2 shows that we succeed in recovering rates within the expected uncertainty envelope. On Fig. 3, could say something obvious like point out that we cannot separate background from foreground for individual events, but that doesn't necessarily matter for rate measurements. Now, one interesting question that Fig. 3 brings to mind is the following: if we only used "gold-plated" events (say, those with $p(x|fore)/p(x|back) > 0.99$, how much worse / more uncertain would the rate estimation be?]

B. Gravitational Waves With Overlapping Templates

In §III A 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., 6, 7], 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

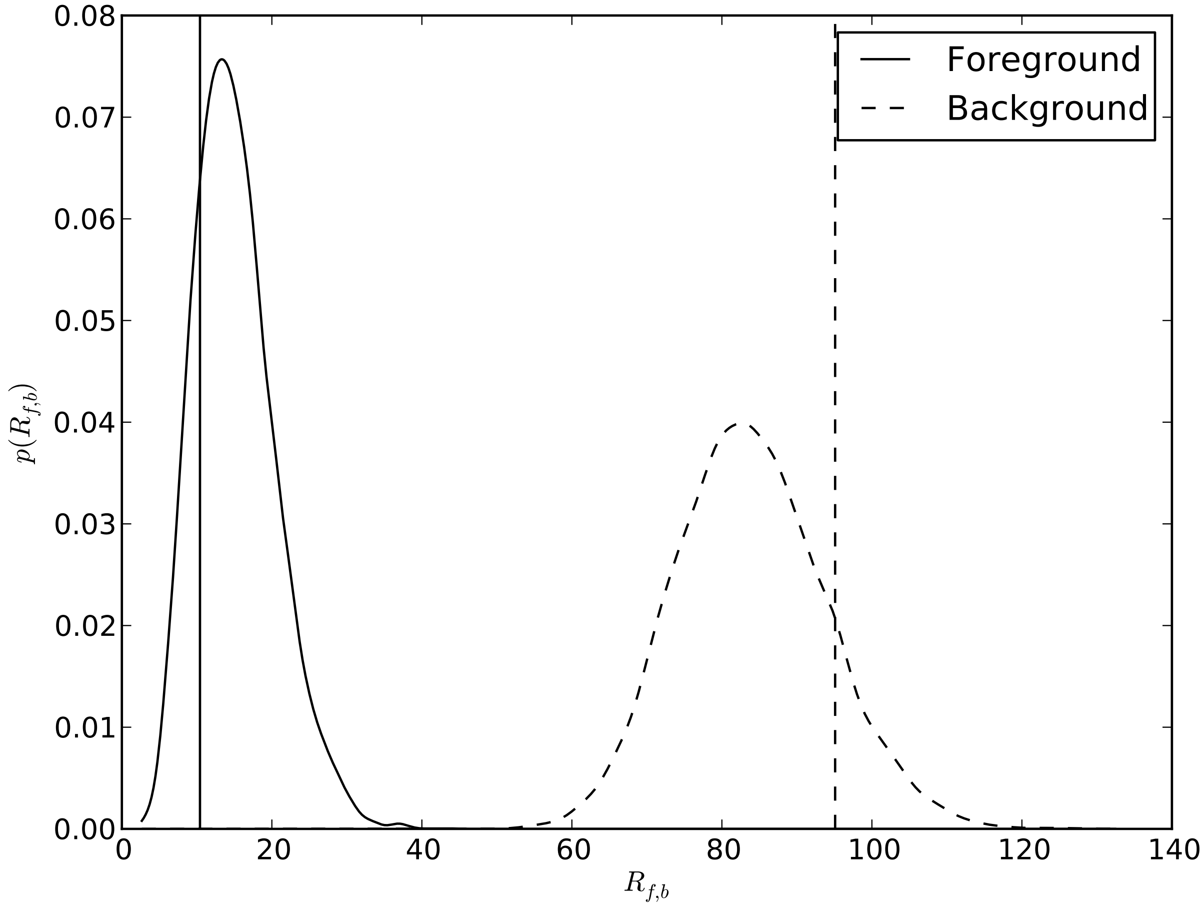

FIG. 2. The marginalized posterior densities for R_f (solid line) and R_b (dashed line) for the analytic model discussed in §III A. The vertical lines indicate the "true" values used to generate the synthetic data set.

$\theta = \{N_{\text{eff}}\}$ is a shape parameter for the background cumulative distribution:

$$\hat{B}\left(x, N_{\text{eff}}\right) = \frac{\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right)^{N_{\text{eff}}} - \left(1 + \operatorname{erf}\left(\frac{x_{\min}}{\sqrt{2}}\right)\right)^{N_{\text{eff}}}}{2^{N_{\text{eff}}} - \left(1 + \operatorname{erf}\left(\frac{x_{\min}}{\sqrt{2}}\right)\right)^{N_{\text{eff}}}}. \tag{35}$$

Results from such an analysis appear in Figures 4 and 5. We use the same parameters and data set as in §III A, with $R_f = 10.4$, $R_b = 95.1$, and $N_{\text{eff}} = 1000$, but now allow N_{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_{eff} situation in §III A.

[Again, I think we need to explicitly compare this method with two alternatives. Alternative one is the loudest event statistic. Alternative two is just counting gold-plated events, and dividing the volume in which we are sensitive to them by their number. In principle, alternative one should be suboptimal when the number of events is more than 0 or 1, while alternative two should be OK when the number of events is large, but error-prone otherwise. But if we want people to use this technique, we have to defend the point that there is a regime where it is clearly superior (and say what that regime is).]

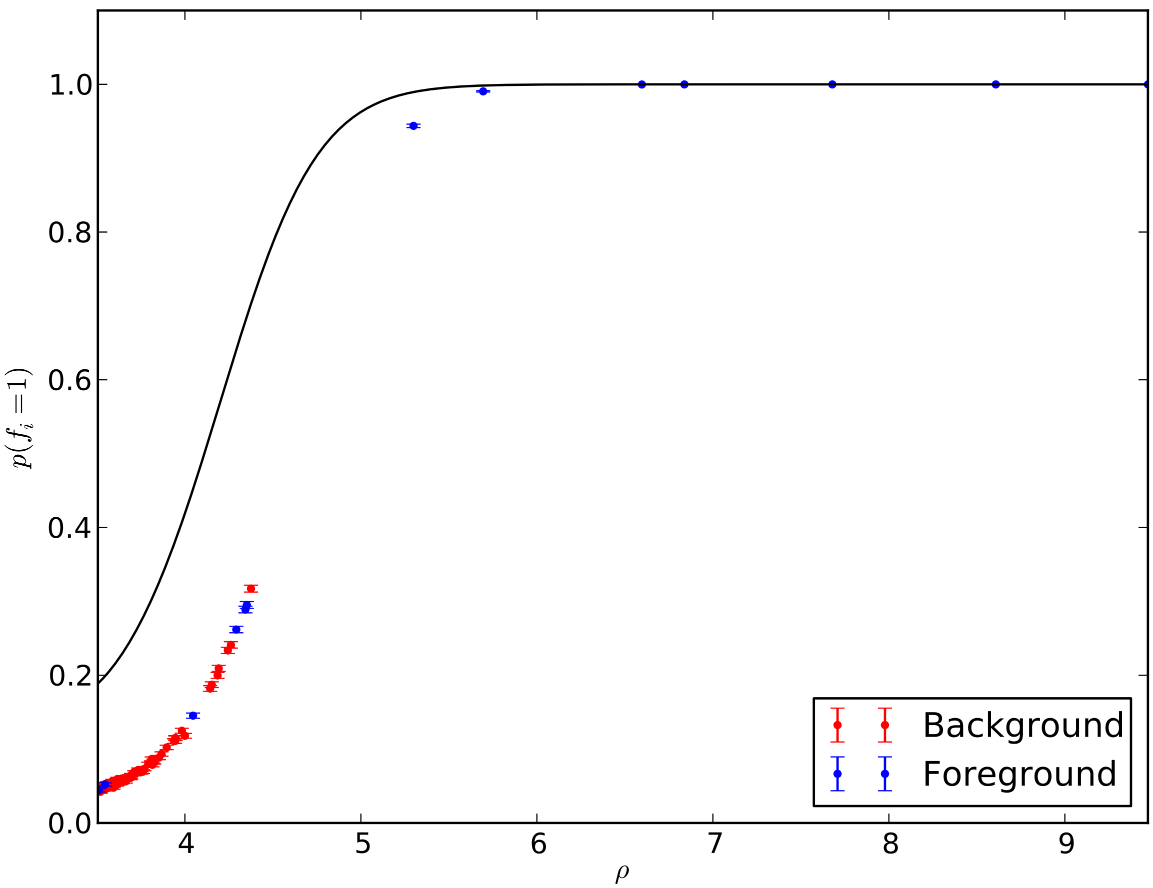

FIG. 3. Foreground probability for each event in the synthetic data set of §III A marginalized over all other parameters. True foreground events are in blue **(dark grey)**, background events in red **(light grey)**. The solid line is the likelihood ratio $p(x|\text{foreground})/p(x|\text{background}) = \hat{f}(x)/\hat{b}(x)$ for this model **evaluated at the simulated parameters (?)**; this exceeds the marginalized foreground probability for many of the events because in this data set there are approximately nine times as many background events as foreground events. **[I'm not sure what the reader is meant to take away from the solid lines If there's something worth taking away, we should describe the message in the text. Otherwise, if it only confuses things, we should drop it.]**

C. Star Cluster Parameters With Background Contamination

Look up von Hippel paper about Bayes 9, the cluster isochrone fitting software—we should cite, since our approach to cluster membership.

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). We assume that a star cluster has a Plummer surface-density profile [8, 9],

$$\hat{f}(\vec{x}, \theta) = \frac{1}{\pi r_0^2 \left(1 + \frac{|\vec{x} - \vec{x}_0|^2}{r_0^2}\right)^2}, \tag{36}$$

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[10], $\vec{x} \in [0, 1]^2$, and a background that has a density

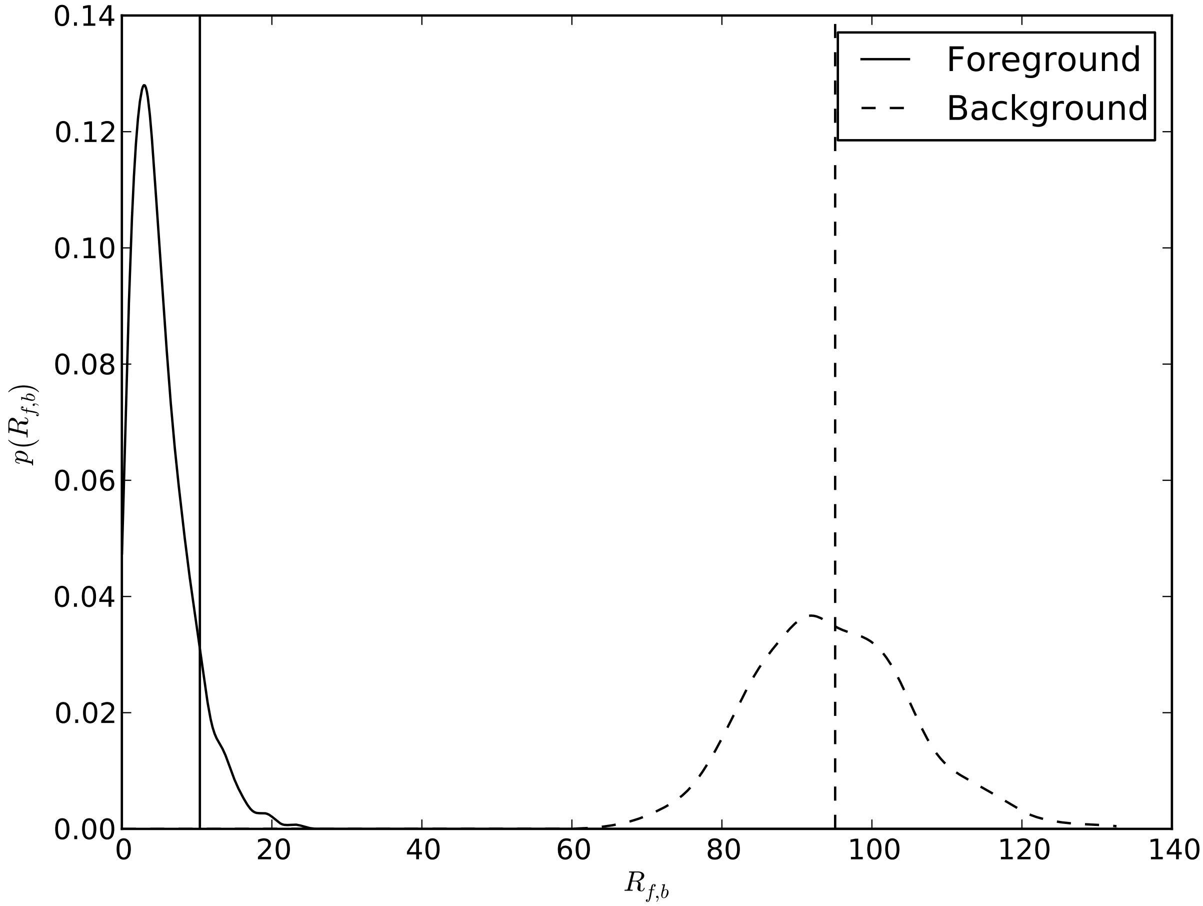

FIG. 4. The foreground (solid lines) and background (dashed lines) rate posterior, marginalized over all flags and the N_{eff} parameter, for the gravitational wave template detection scenario with overlapping templates discussed in §III B. The true values of the rates, $R_f = 10.4$ and $R_b = 95.1$, are indicated with vertical lines. The distributions are not significantly wider than those of Figure 2, in spite of the extra parameter. **[R_f seems to be more of an outlier than before; why was there a significant shift, given that N_{eff} seems fairly spot on?]**

gradient at an arbitrary orientation with respect to the observational axes:

$$\hat{b}\left(\vec{x}, \theta\right) = 1 + \vec{\gamma} \cdot \left(\vec{x} - \vec{x}_{1/2}\right), \tag{37}$$

where $\vec{\gamma}$ is the gradient, and $\vec{x}_{1/2} = [1/2, 1/2]$ is the centroid of the observational domain. **[I think that if the domain is finite, the previous equation is only normalized for all $\vec{\gamma}$ if $\vec{x}_{1/2}$ is in the center of a symmetric domain (as it happens to be in this case).]** **Yes, as you say, this function is only normalized for $\vec{x}_{1/2}$ is the center, and also for small enough $\vec{\gamma}$; both of these conditions obtain here.**

We use simulated data drawn from our model with parameters

$$\theta_0 \equiv \{x_0, y_0, r_0, \gamma_x, \gamma_y\} = \left\{\frac{1}{2}, \frac{1}{2}, 0.18, -\frac{1}{2}, \frac{1}{2}\right\}, \tag{38}$$

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 6 shows the density of stars (cluster and background) on the sky for our particular data set.

To analyze our synthetic data set, we analytically marginalized over the state flags (i.e. cluster membership), using the likelihood in Eq. (19). We did this to take advantage of the *emcee* sampler of Foreman-Mackey *et al.* [11], which

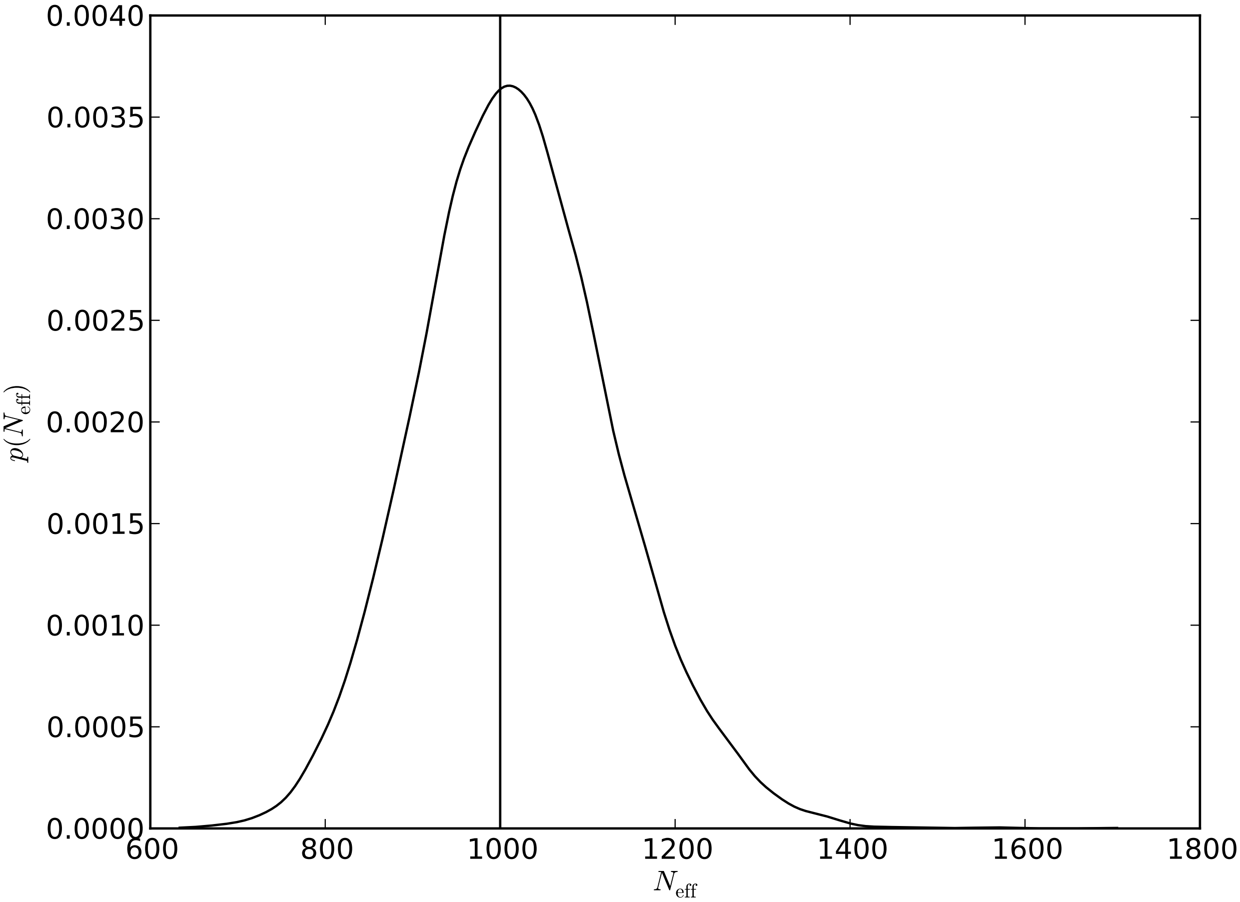

FIG. 5. The posterior on the number of effective templates, $N_{\rm eff}$, for the model and data discussed in §III B, marginalized over all state flags and rates. The true value, $N_{\rm eff} = 1000$, is indicated by the vertical line.

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,

$$p\left(r_0\right) = \frac{\sqrt{R_f}}{r_0}. \tag{39}$$

[Why is this the Jeffreys prior (BTW, it's Jeffreys, or perhaps Jeffreys', but definitely not Jeffrey's ;))? It's not obvious why the prior on r_0 should depend on R_f?] **Here is the calculation:**

$$p\left(r_0\right) = \sqrt{\left\langle\left(\frac{\partial \log \mathcal{L}}{\partial r_0}\right)^2\right\rangle} = \sqrt{R_f\left\langle\left(\frac{\partial \log \hat{f}}{\partial r_0}\right)^2\right\rangle} = \sqrt{R_f\left\langle\left(\frac{2}{r_0} - \frac{4r_0}{r^2+r_0^2}\right)^2\right\rangle} = \sqrt{\frac{4R_f}{3r_0^2}} \propto \frac{\sqrt{R_f}}{r_0}. \tag{40}$$

Note that the calculation of the prior uses the likelihood before analytical marginalization over the flags, f_i. We are ignoring the fact that R_f and r_0 are correlated. In truth, the Jeffreys prior would be the determinant of the matrix built out of terms like $\partial \log \mathcal{L}/\partial\theta_i \times \partial \log \mathcal{L}/\partial\theta_j$, but let's not get carried away.... As another aside, note that we could have predicted the $1/r_0$ dependence because r_0 has units, so the only unit-invariant prior is to be flat in $\log r_0$; I'm don't have a similar argument for the appearance of the $\sqrt{R_f}$ factor. (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 ~ 1000 foreground stars.)

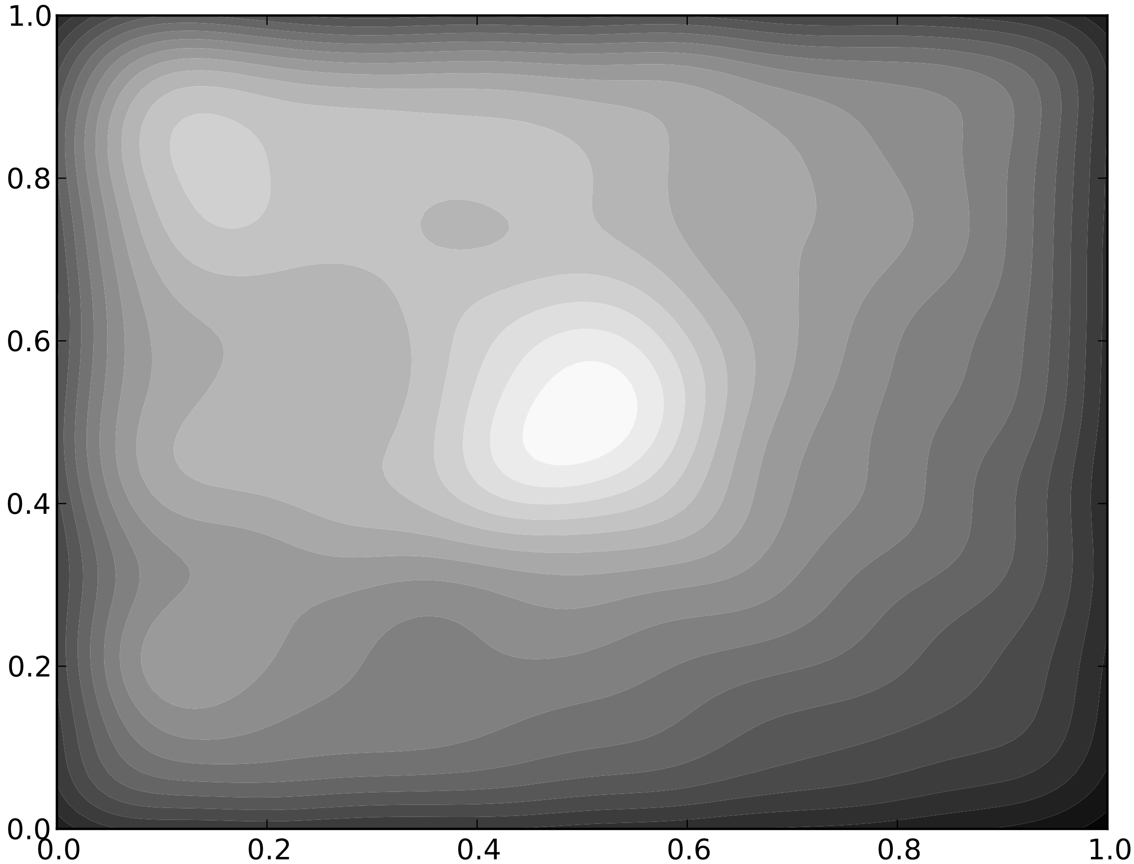

FIG. 6. The density contours on the sky of our synthetic data set of cluster and background stars for the parametrs in Eq. (38). There are a factor of 10 more stars (10000) in the background than in the cluster (1000), but the average density of cluster and background stars is similar.

Figure 7 shows the posterior for the location parameters, $\vec{x}_0$; the center of the cluster is localized to within about 10% of the cluster scale. Figure 8 shows the posteriors inferred on the cluster and background numbers, R_f and R_b, and Figure 9 shows the posterior for the cluster's scale parameter. In spite of the significant background, the cluster scale and total number are accurately recovered by our analysis.

IV. CONCLUSION

Separating classes of events

Comments on background probability estimation – Cannon et al. in prep.

Relevance (vs. loudest statistic, or frequentist approach using only gold-plated events)

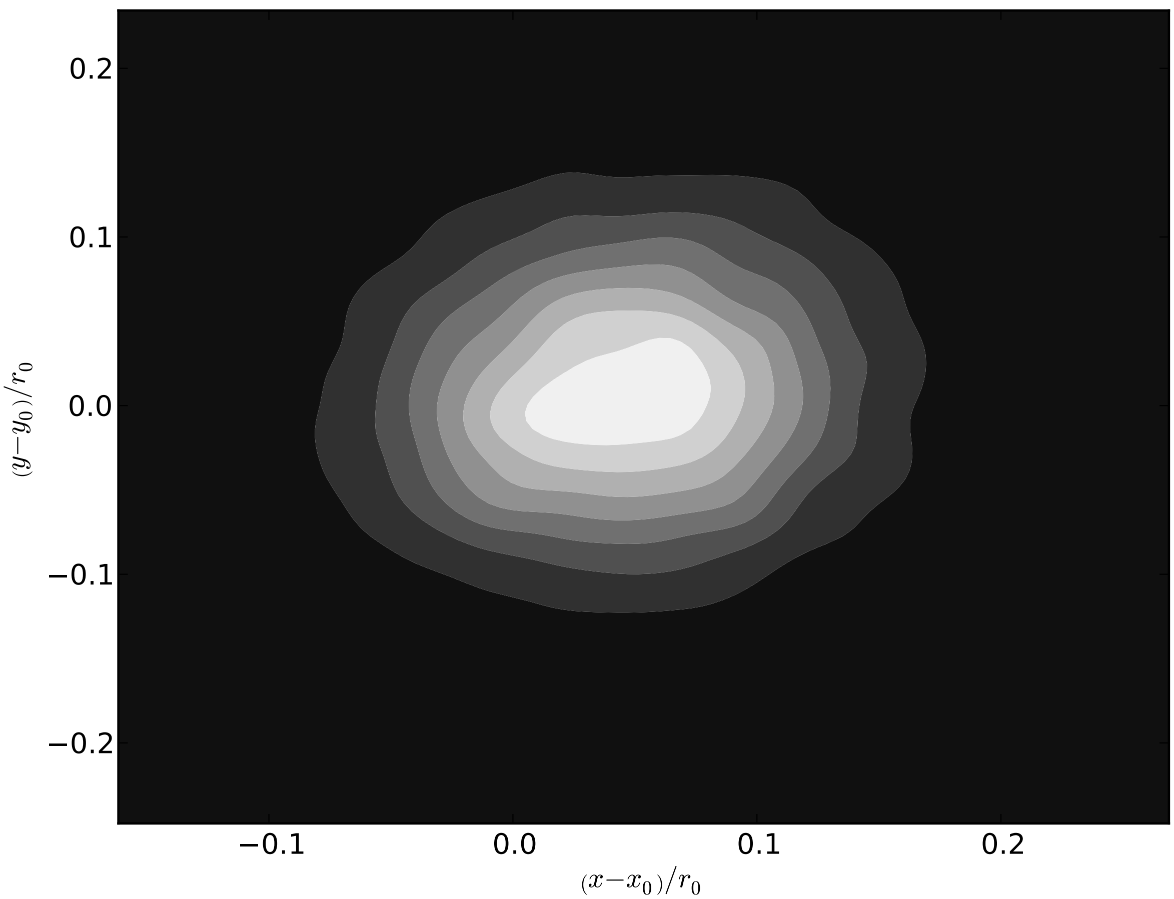

FIG. 7. Contours of the posterior probability distribution for the center of the cluster, $\vec{x}_0$, in the example from § III C. The center $(x, y) = (x_0, y_0)$ is determined to within a few percent of the structural radius of the cluster, r_0 (see Eq. (38)). **[Given the cluster size of 0.1, it seems a bit surprising that the center is off by ~ 0.05...]**

ACKNOWLEDGMENTS

We thank Kipp Cannon, Chad Hanna, Drew Keppel, and Richard O'Shaughnessy for discussions and suggestions about this manuscript.

[1] R. Biswas, P. R. Brady, J. D. E. Creighton, and S. Fairhurst, Classical and Quantum Gravity **26**, 175009 (2009), arXiv:0710.0465 [gr-qc].
[2] H. Jeffreys, Royal Society of London Proceedings Series A **186**, 453 (1946).
[3] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, Phys. Rev. D **85**, 122006 (2012).
[4] L. S. Colaboration (Virgo Collaboration), Phys.Rev. **D85**, 082002 (2012), arXiv:1111.7314 [gr-qc].
[5] L. S. Finn and D. F. Chernoff, Phys.Rev. **D47**, 2198 (1993), arXiv:gr-qc/9301003 [gr-qc].
[6] B. J. Owen and B. S. Sathyaprakash, Phys. Rev. D **60**, 022002 (1999).
[7] 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), arXiv:0710.2335.
[8] H. C. Plummer, Mon.Not.Roy.Astron.Soc. **71**, 460 (1911).
[9] S. J. Aarseth, M. Henon, and R. Weilen, Astron. and Astrophys. **37**, 183 (1974).

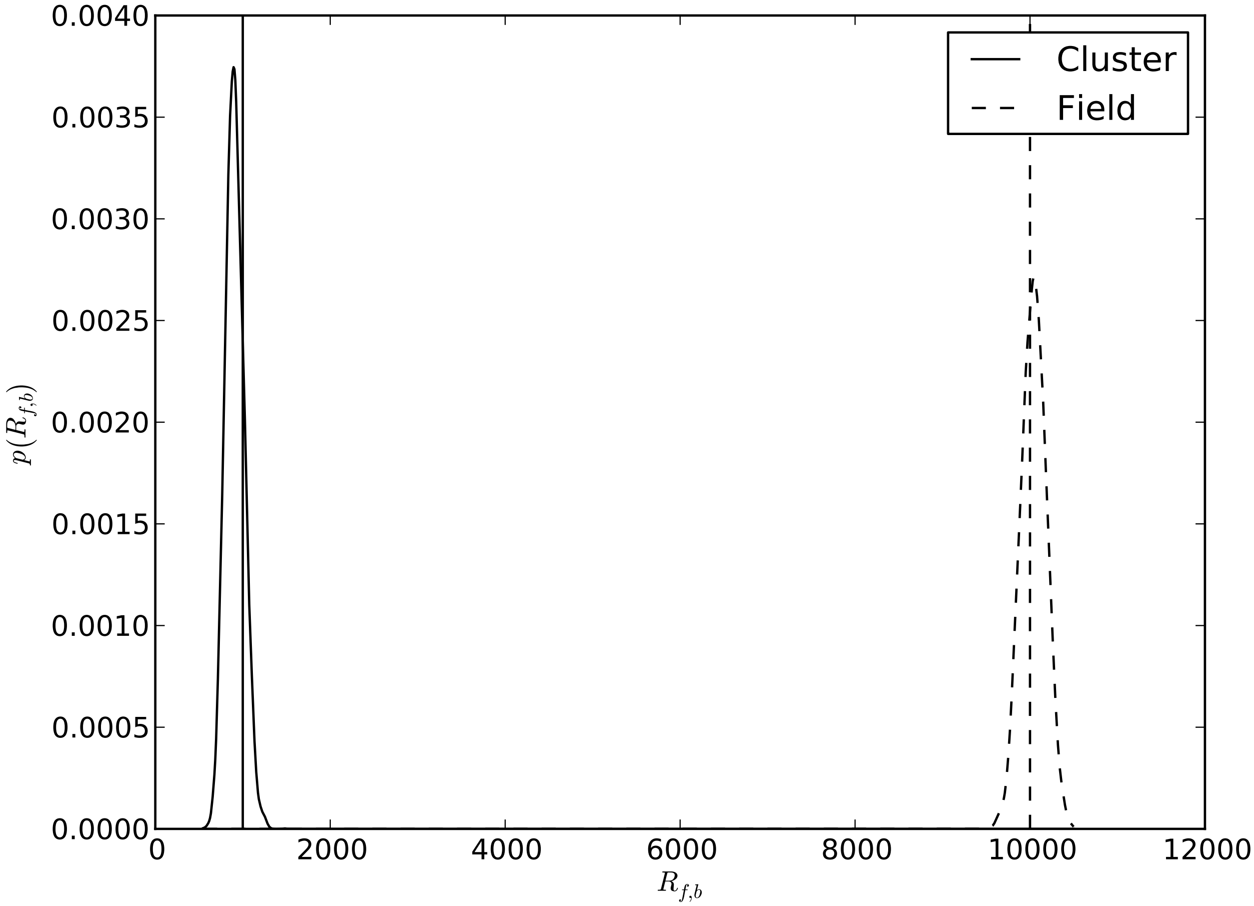

FIG. 8. Posterior densities for the number of stars in the cluster (R_f) and in the field (R_b) in the example from § III C. Vertical lines indicate the true values (see Eq. (38)).

[10] The observational domain is not infinite, so the normalization of the cluster density in Eq. (36) is not quite correct. In our modeling we properly take this into account, but for simplicity here we ignore it.

[11] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, (2012), arXiv:1202.3665 [astro-ph.IM].

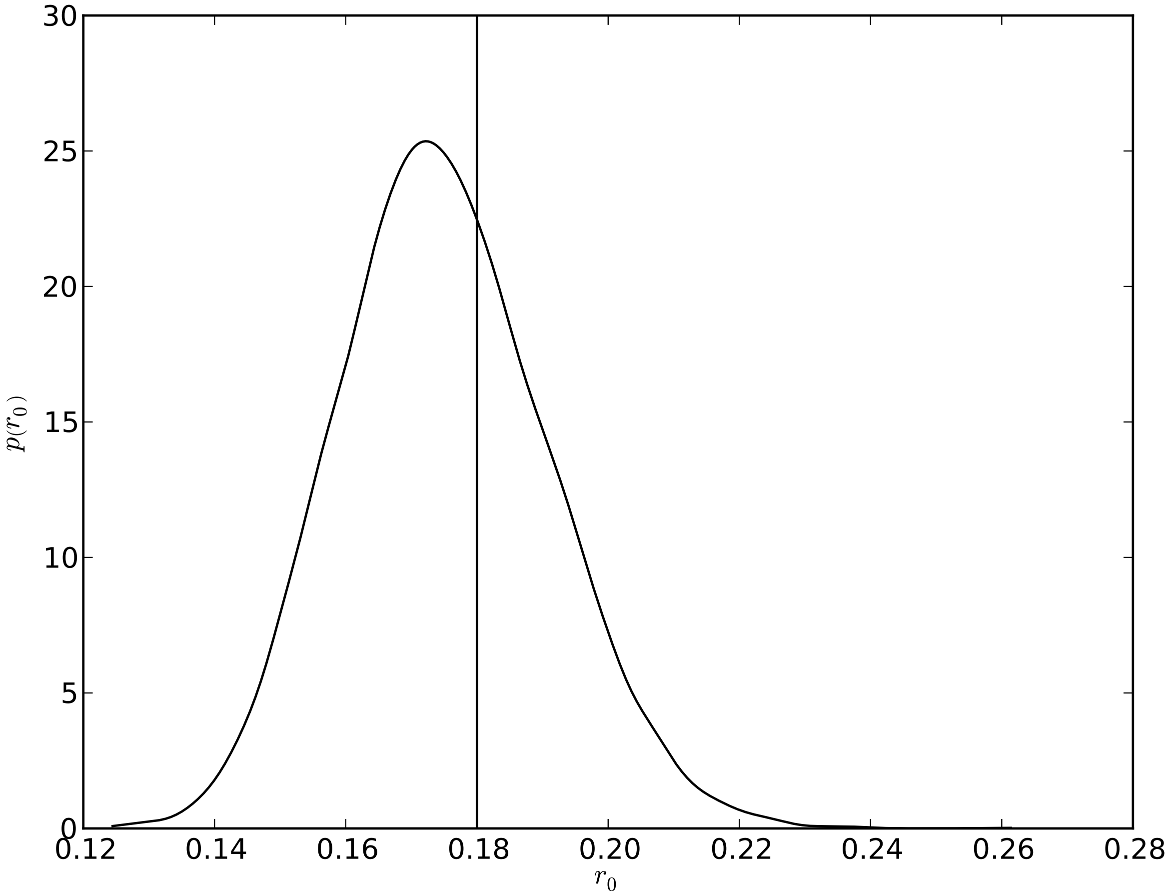

FIG. 9. Posterior density for the scale parameter for the cluster, r_0, from the example in § III C. The true value is indicated by the vertical line (see Eq. 38). **[Perhaps combine Figs. 8 and 9 (put an extra scale for r_0 at the top of 8?]**