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

Ilya Mandel†
School of Physics and Astronomy
University of Birmingham
Edgbaston B15 2TT Birmingham
United Kingdom

Jon Gair‡

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/
† imandel@star.sr.bham.ac.uk; http://www.sr.bham.ac.uk/~imandel
‡ jrg23@cam.ac.uk

I. INTRODUCTION

FIXME: Introduce the necessity of estimating rates, prior work (like [1]), Bayesian inference.

II. MODEL

We assume that we are presented with a data set of N events. 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:

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

We assume that both the foreground and background events are samples from an inhomogeneous Poisson process **[pardon my ignorance, but I thought an "inhomogeneous" Poisson process meant that the rate changed as a function of time – and I don't think that's what we are assuming...]** 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

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

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

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

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 $R_f \equiv \mathbf{F}(\infty,\theta)$ and $R_b \equiv \mathbf{B}(\infty,\theta)$ 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 $f_i = \{0,1\}$.

Bayes' theorem relates the posterior probability of the state flags, rates, and shape parameters, $p\left(\{f_i\}, R_f, R_b, \theta | d\right)$, the likelihood of the data, $p\left(d | \{f_i\}, R_f, R_b, \theta\right)$, and the prior probability of state flags, rates and shape parameters before any data are obtained, $p\left(\{f_i\}, R_f, R_b, \theta\right)$:

$$p\left(\{f_i\}, R_f, R_b, \theta | d\right) = \frac{p\left(d | \{f_i\}, R_f, R_b, \theta\right) p\left(\{f_i\}, R_f, R_b, \theta\right)}{\mathbf{p(d)}}. \tag{10}$$

The **normalization constant**, called the evidence, $p(d)$, 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

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

The prior distribution can be factorized as

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

The first term is not actually a "prior" in the usual sense; the conditional probability of the flags $\{f_i\}$ on the rates is given by

$$p\left(\{f_i\} | R_f, R_b, \theta\right) = \frac{R_f^{N_f} R_b^{N_b}}{N_f! N_b!} \exp\left[-\left(R_f + R_b\right)\right] N_f! N_b! = R_f^{N_f} R_b^{N_b} \exp\left[-\left(R_f + R_b\right)\right], \tag{13}$$

where N_f and N_b are the number of foreground and background events. This expression follows from Property 1 of inhomogeneous Poisson processes and the $N_f! N_b!$ different ways of assigning foreground and background flags to events for fixed N_f, N_b.

[I agree with the final result, but I find this way of deriving it to be confusing. First of all, the number of ways to assign flags to events for fixed N_f and N_b is $N!/(N_f! N_b!)$, not $N_f! N_b!$ as written. Secondly, you are computing the probability of having a particular ordered list of flags, so there's no reason to consider different permutations. Thirdly, you are not properly including the fact that the length of the flags vector must be exactly N, the total number of data points. (Well, you have the right result, so you are obviously doing these things properly, but the explanation just doesn't seem clear to me. :)) I would make that N explicit, by writing Eq. (12) as

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

Then

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

Note that there are no permutations involved here – once we assume that there are precisely $N = N_f + N_b$ events, the probability for the i'th event to have flag f_i is either $R_f/(R_f + R_b)$ or $R_b/(R_f + R_b)$ depending on whether f_i is 1 or 0. Meanwhile,

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

since the distribution of total event number is a Poisson process with rate $R_f + R_b$. Multiplying these, we get the same result as in (13), but in a way that even I can understand. :)]

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 Jeffrey's prior on rates **[citation?]**, independent of the shape parameters

$$p\left(R_f, R_b, \theta\right) \propto \frac{1}{\sqrt{R_f R_b}} p(\theta), \tag{14}$$

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. (13) regularize the posterior as $R_f, R_b \to \infty$.

Putting everything together, the posterior is

$$p\left(\{f_i\}, R_f, R_b, \theta | d\right) \propto \left[\prod_{\{i|f_i=1\}} R_f \hat{f}\left(x_i, \theta\right)\right] \left[\prod_{\{i|f_i=0\}} R_b \hat{b}\left(x_i, \theta\right)\right] \exp\left[-\left(R_f + R_b\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}} \tag{15}$$

Not surprisingly, this posterior is equivalent to one derived 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:

$$\begin{aligned} p\left(\{f_i\}, R_f, R_b, \theta | d\right) &\propto p\left(d, \{f_i\} | R_f, R_b, \theta\right) p\left(R_f, R_b, \theta\right) \\ &= \left[\prod_{\{i|f_i=1\}} f\left(x_i, \theta\right)\right] \left[\prod_{\{i|f_i=0\}} b\left(x_i, \theta\right)\right] \exp\left[-\left(R_f + R_b\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}}. \end{aligned} \tag{16}$$

[I can't derive the preceding equation in 1 line. ;) So I'd suggest either making the alternative derivation more explicit, or, if you don't think it's worth the space, just saying that one could derive the same thing in this alternative way – otherwise, just repeating Eq. (15) doesn't seem very useful.]

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

$$p\left(R_f, R_b, \theta | d\right) = \sum_{\{f_i\} \in \{0,1\}^N} p\left(\{f_i\}, R_f, R_b, \theta | d\right) \propto \prod_i \left[R_f \hat{f}\left(x_i, \theta\right) + R_b \hat{b}\left(x_i, \theta\right)\right] \exp\left[-\left(R_f + R_b\right)\right] \frac{p(\theta)}{\sqrt{R_f R_b}}. \tag{17}$$

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. (15)), Eq. (17) contains only continuous parameters.

Eq. (15) is unchanged if the ranking statistic is multi-dimensional; in this case, the rates are

$$R_f = \int d^{\mathbf{k}}\tilde{\mathbf{x}}\, f(x, \theta) \tag{18}$$

and

$$R_b = \int d^{\mathbf{k}}\tilde{\mathbf{x}}\, b(x, \theta), \tag{19}$$

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.

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. 2]. 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

$$\rho_h \equiv \frac{\langle h, d\rangle}{\sqrt{\langle h, h\rangle}}, \tag{20}$$

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

$$\langle a, b\rangle \equiv \int_0^\infty df\, \frac{a^*(f)b(f)}{S(f)}. \tag{21}$$

We suppose for simplicity that the templates are sufficiently distinct that

$$\langle h_i, h_j\rangle \simeq \delta_{ij}. \tag{22}$$

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,

$$x \equiv \max_h \rho_h, \tag{23}$$

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 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

$$\hat{B}(x) = \left(\frac{1 + \mathrm{erf}\left(\frac{x}{2}\right)}{2}\right)^N \tag{24}$$

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

$$\hat{B}(x) = \frac{\left(1 + \mathrm{erf}\left(\frac{x}{2}\right)\right)^N - \left(1 + \mathrm{erf}\left(\frac{x_{\rm min}}{2}\right)\right)^N}{2^N - \left(1 + \mathrm{erf}\left(\frac{x_{\rm min}}{2}\right)\right)^N} \tag{25}$$

The SNR of a gravitational wave signal in an interferometric detector scales as $1/d$ [3], 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

$$\hat{F}(x) = 1 - \frac{x_{\rm min}^3}{x^3}. \tag{26}$$

Note that this scenario has no shape parameters 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. (25) and (26) with $R_f^{\rm true} = 10.4$ and $R_b^{\rm true} = 95.1$, using 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 Figure 1. We used a Markov Chain Monte Carlo simulation to draw samples of state flags and rates from the joint posterior (Eq. (15)).

In Figure 2, we show the marginalized posterior densities for the foreground and background rates. (Refer to Eq. (17).) Figure 3 shows the posterior foreground probability for each event marginalized over all other events' types and the foreground and background rates.

B. Gravitational Waves With Overlapping Templates

In §III A we assumed that the overlap between templates in the template bank vanished, so the SNR's in different templates are independent random variables. In fact, template banks are not constructed in this way [4, 5], because signals can fall in the gaps between the non-overlapping templates. We can model this effect by assuming that

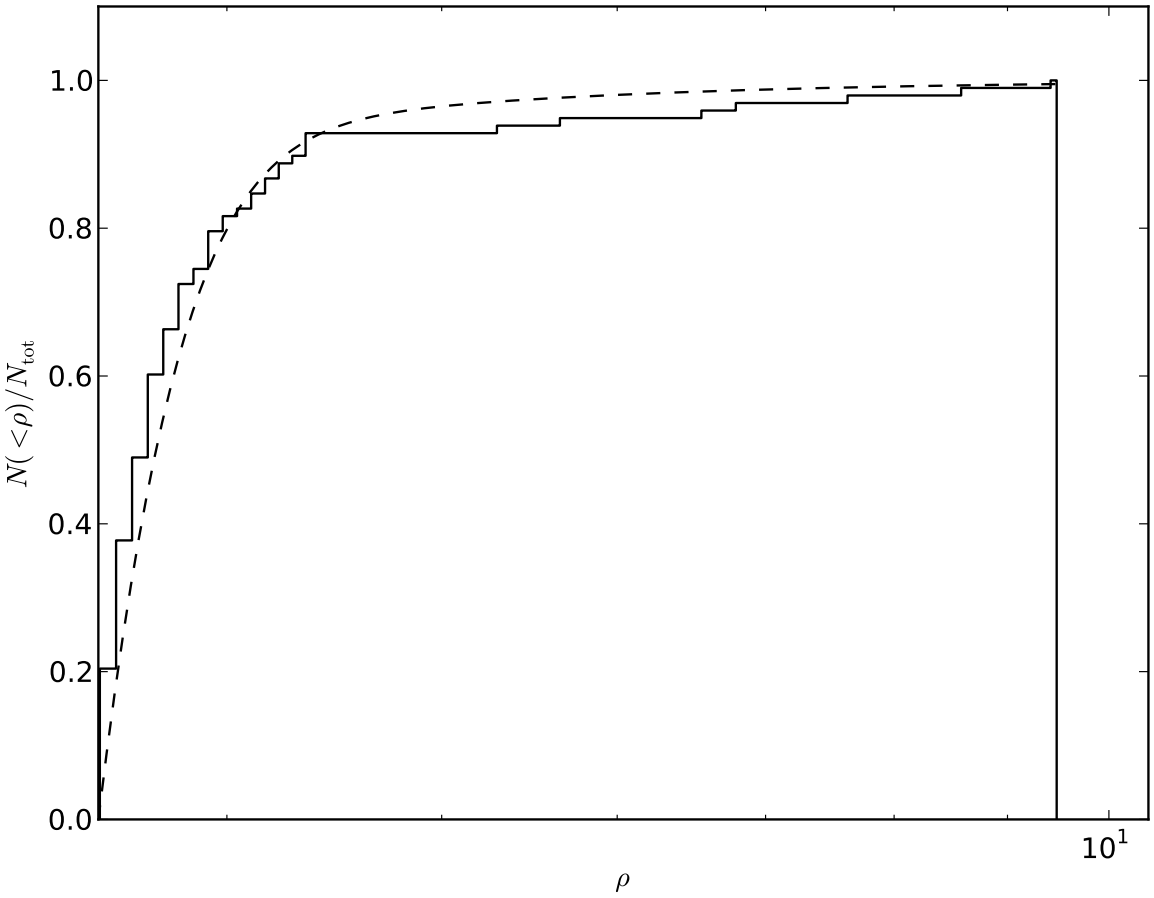

FIG. 1. The cumulative distribution of the ranking statistics for the synthetic data used to test the the formalism on the model from §III A. The solid line gives the cumulative distribution of the synthetic data; the dashed line gives the theoretical cumulative distribution for the models in Eqs. (25) and (26) combined with $R_f = 10.4$ and $R_b = 95.1$.

a template bank of N actual templates will behave as if it had $N_{\rm eff}$ *independent* templates, and fit for the shape parameter $N_{\rm eff}$. That is, we assume that $N_{\rm eff}$ is a shape parameter for the background cumulative distribution:

$$\hat{B}\left(x, N_{\rm eff}\right) = \frac{\left(1+\mathrm{erf}\left(\frac{x}{2}\right)\right)^{N_{\rm eff}} - \left(1+\mathrm{erf}\left(\frac{x_{\rm min}}{2}\right)\right)^{N_{\rm eff}}}{2^{N_{\rm eff}} - \left(1+\mathrm{erf}\left(\frac{x_{\rm min}}{2}\right)\right)^{N_{\rm eff}}}. \tag{27}$$

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 $N_{\rm eff} = 1000$, 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 §III A.

C. 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. We assume that a star cluster has a Plummer surface-density profile

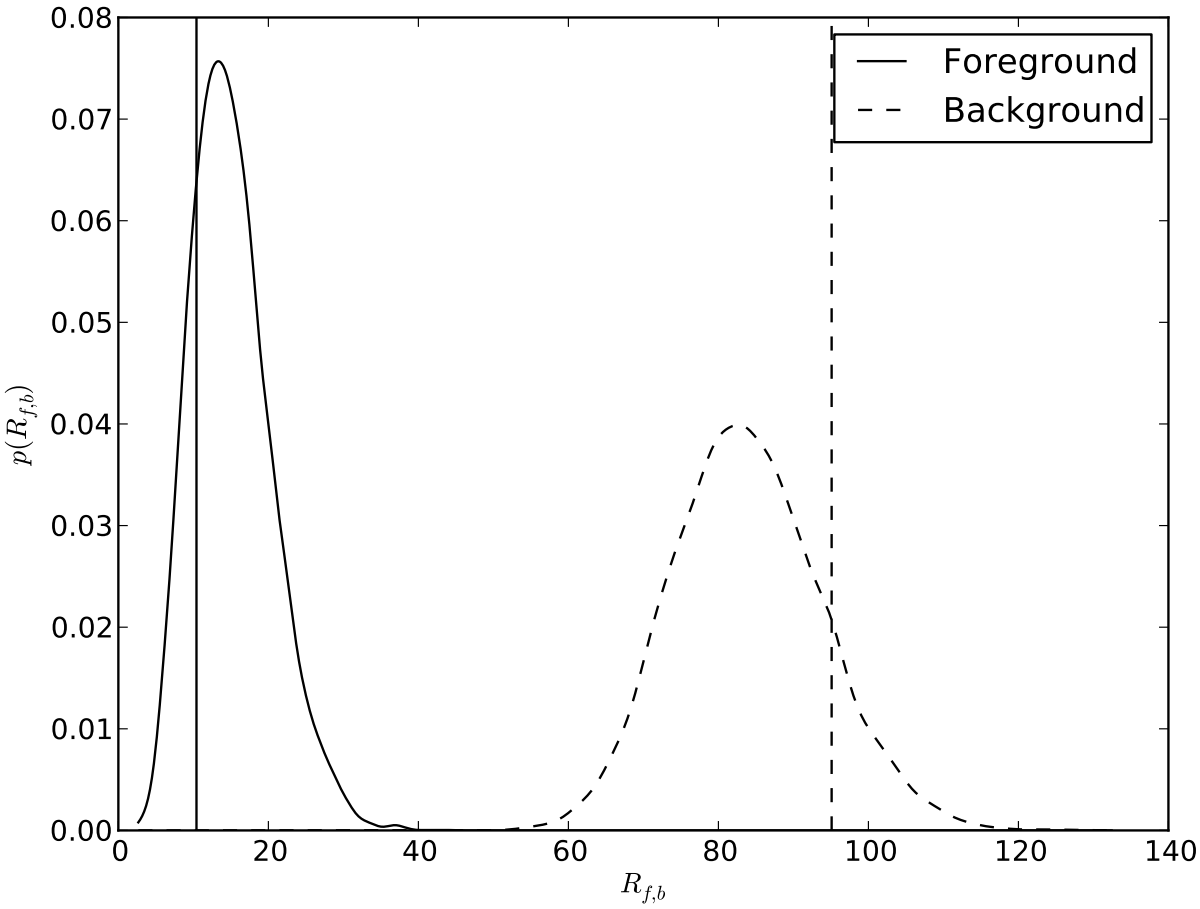

FIG. 2. The marginalized posterior densities for R_f (solid line) and R_b (dashed line) for the analytic model discussed in §III A. The vertical lines indicate the "true" values used to generate the synthetic data set.

[6, 7],

$$\hat{f}(\vec{x},\theta) = \frac{1}{\pi r_0^2 \left(1 + \frac{|\vec{x}-\vec{x}_0|^2}{r_0^2}\right)^2}, \tag{28}$$

where $\vec{x}_0$ is the location on the sky of the center of the cluster, and r_0 is a radial scale parameter. We assume a square observational domain[8], $\vec{x} \in [0,1]^2$, and a background that has a density gradient at an arbitrary orientation with respect to the observational axes:

$$\hat{b}\left(\vec{x},\theta\right) = 1 + \vec{\gamma}\cdot\left(\vec{x} - \vec{x}_{1/2}\right), \tag{29}$$

where $\vec{\gamma}$ is the gradient, and $\vec{x}_{1/2} = [1/2, 1/2]$ is the centroid of the observational domain. We use assume a gradient for the background of $\vec{\gamma} = [-0.5, 0.5]$. We assume that the total number of background stars, $R_b = 10000$, exceeds by a factor of ten the total number of cluster stars, $R_f = 1000$. We choose a true cluster center $\vec{x}_0 = \vec{x}_{1/2}$, and a radial scale parameter of $r_0 = 0.1$. In all, the shape parameters for the true distributions are

$$\theta_0 \equiv \{x_0, y_0, r_0, \gamma_x, \gamma_y\} = \left\{\frac{1}{2}, \frac{1}{2}, \frac{1}{10}, -\frac{1}{2}, \frac{1}{2}\right\}. \tag{30}$$

Figure 6 shows the density on the sky due to the cluster foreground and the background with these parameters. We drew a synthetic data set from this density according to the corresponding inhomogeneous Poisson process.

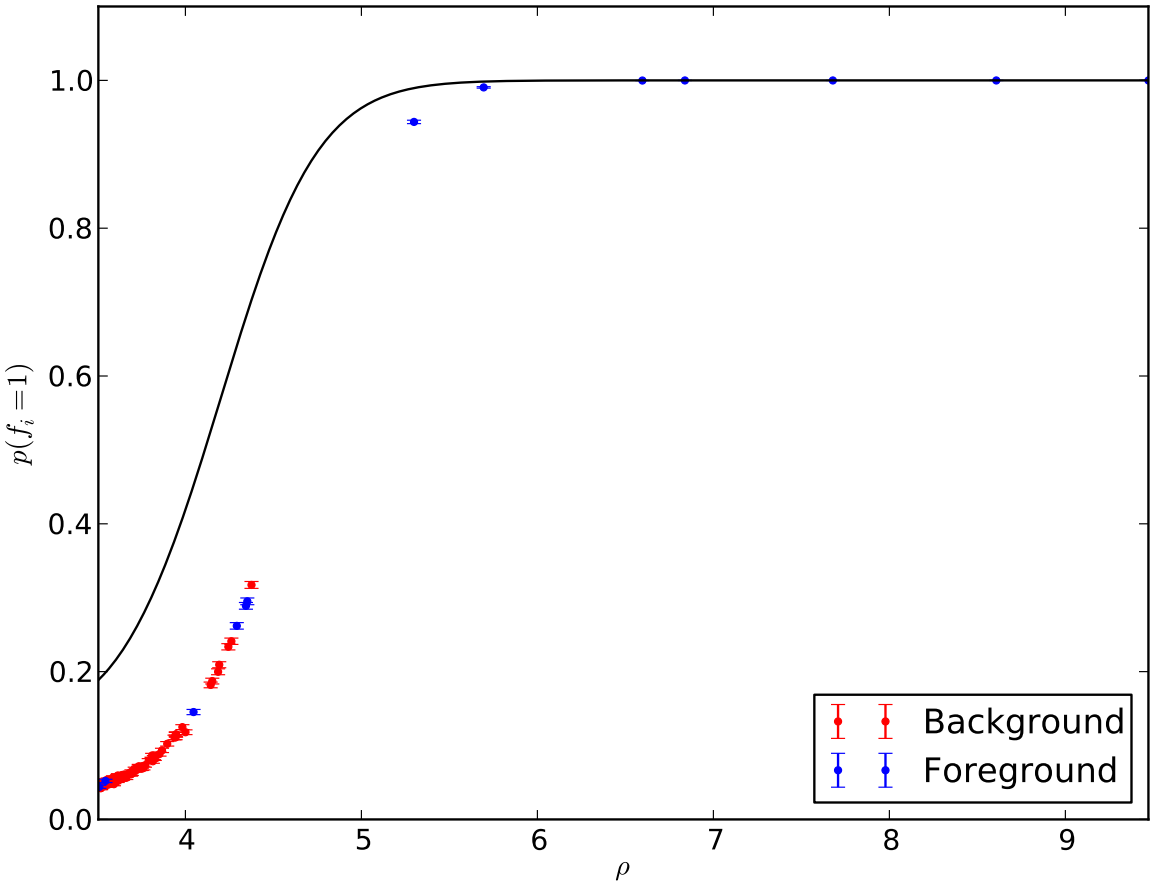

FIG. 3. Foreground probability for each event in the synthetic data set of §III A marginalized over all other parameters. True foreground events are in blue, background events in red. The solid line is the likelihood ratio $p(x|\text{foreground})/p(x|\text{background}) = \hat{f}(x)/\hat{b}(x)$ for this model; 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.

To analyze our synthetic data set, we analytically marginalized over the state flags, using the likelihood in Eq. (17). We did this to take advantage of the *emcee* sampler of Foreman-Mackey *et al.* [9], 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 the Jeffrey's prior on r_0,

$$p\left(r_0\right) = \frac{\sqrt{R_f}}{r_0}. \tag{31}$$

(Note that this factor of $\sqrt{R_f}$ cancels with the Jeffrey's 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.)

Figure 7 shows the posterior for the location parameters, $\vec{x}_0$; the center of the cluster is localized to within a few percent of its 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 recoverd by our analysis.

IV. CONCLUSION

We conclude.

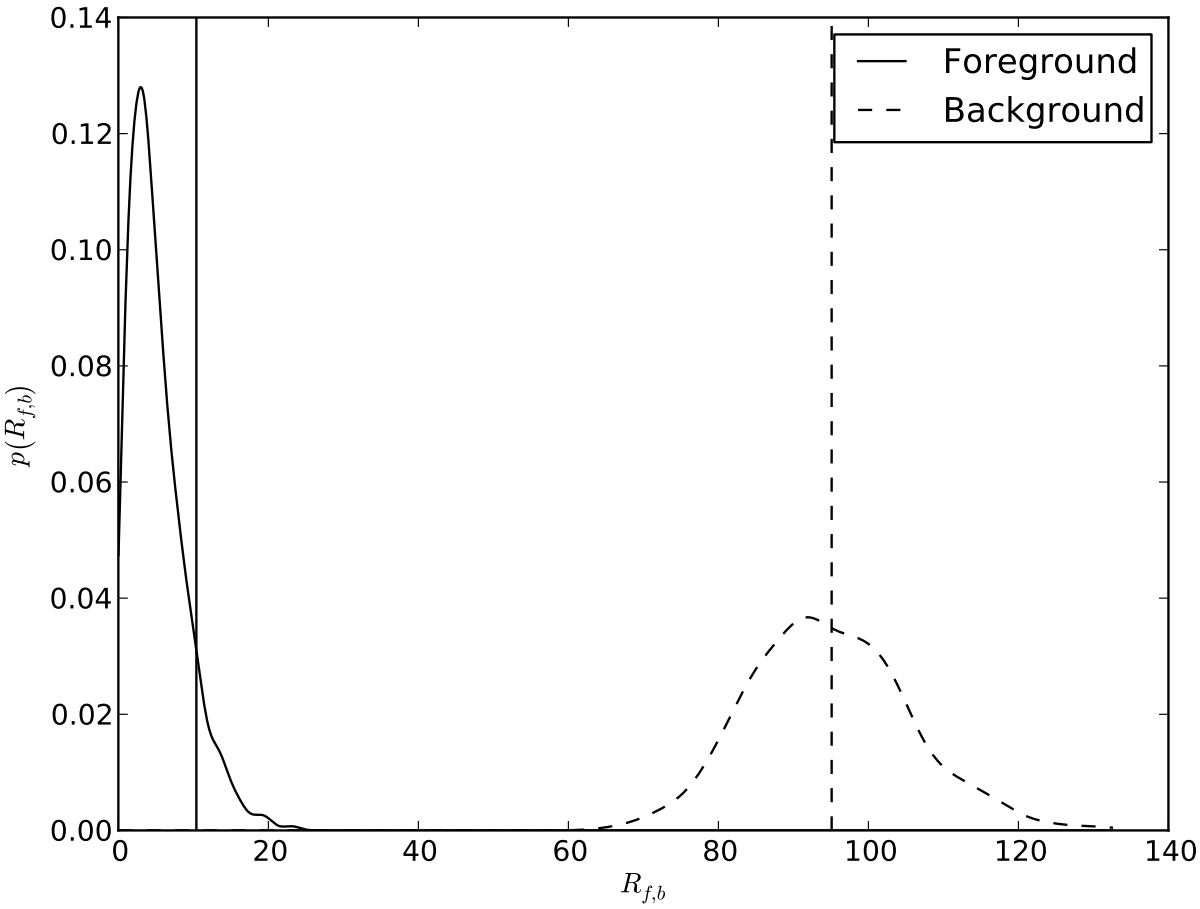

FIG. 4. The foreground (solid lines) and background (dashed lines) rate posterior, marginalized over all flags and the $N_{\rm eff}$ parameter, for the gravitational wave template detection scenario with overlapping templates discussed in §III B. The true values of the rates, $R_f = 10.4$ and $R_b = 95.1$, are indicated with vertical lines. The distributions are not significantly wider than those of Figure 2, in spite of the extra parameter.

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] L. S. Colaboration (Virgo Collaboration), Phys.Rev. **D85**, 082002 (2012), arXiv:1111.7314 [gr-qc].
[3] L. S. Finn and D. F. Chernoff, Phys.Rev. **D47**, 2198 (1993), arXiv:gr-qc/9301003 [gr-qc].
[4] S. Caudill, S. E. Field, C. R. Galley, F. Herrmann, and M. Tiglio, Class.Quant.Grav. **29**, 095016 (2012), arXiv:1109.5642 [gr-qc].
[5] K. Cannon, C. Hanna, and D. Keppel, Phys.Rev. **D84**, 084003 (2011), arXiv:1101.4939 [gr-qc].
[6] H. C. Plummer, Mon.Not.Roy.Astron.Soc. **71**, 460 (1911).
[7] S. J. Aarseth, M. Henon, and R. Weilen, Astron. and Astrophys. **37**, 183 (1974).
[8] The observational domain is not infinite, so the normalization of the cluster density in Eq. (28) is not quite correct. In our modeling we properly take this into account, but for simplicity here we ignore it.

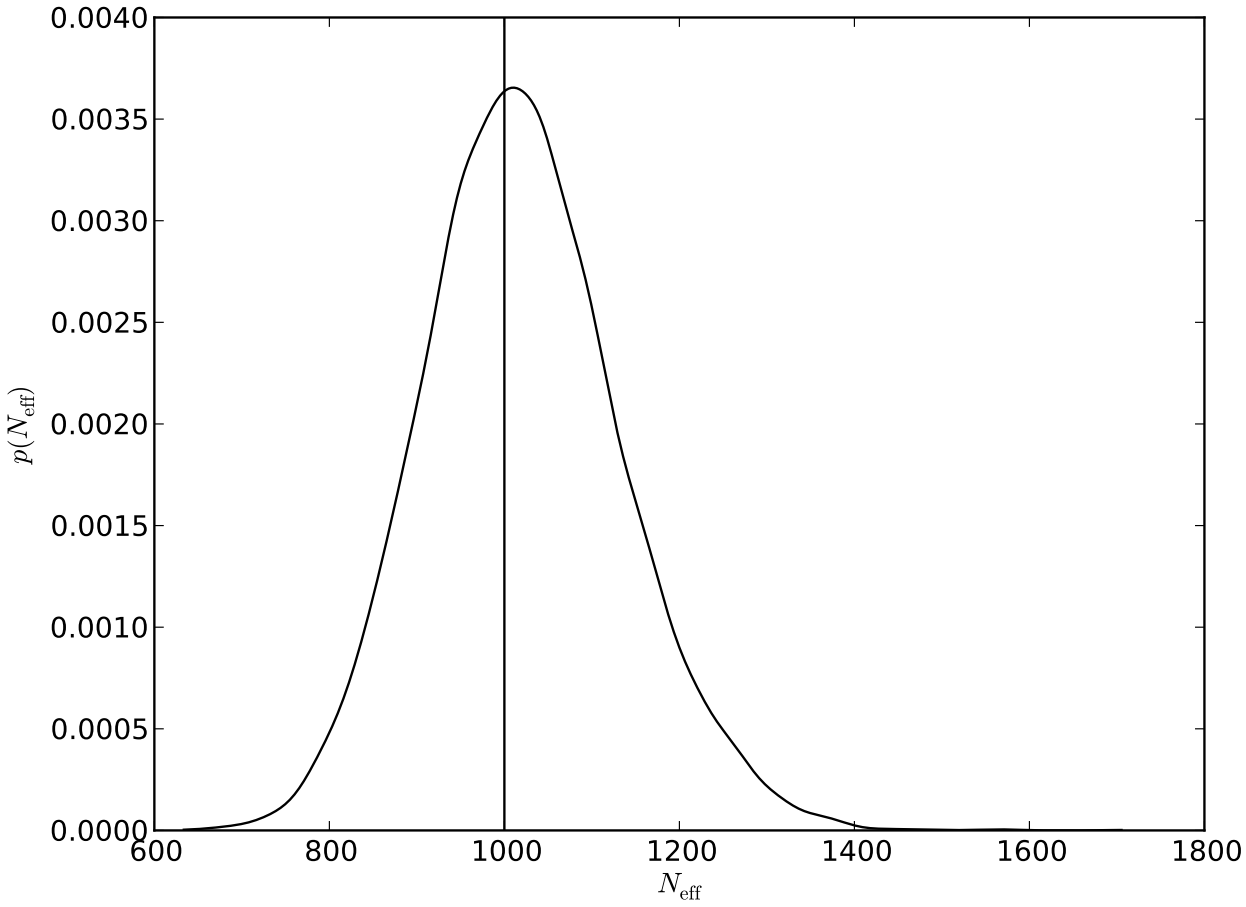

FIG. 5. The posterior on the number of effective templates, $N_{\rm eff}$, for the model and data discussed in §III B, marginalized over all state flags and rates. The true value, $N_{\rm eff} = 1000$, is indicated by the vertical line.

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

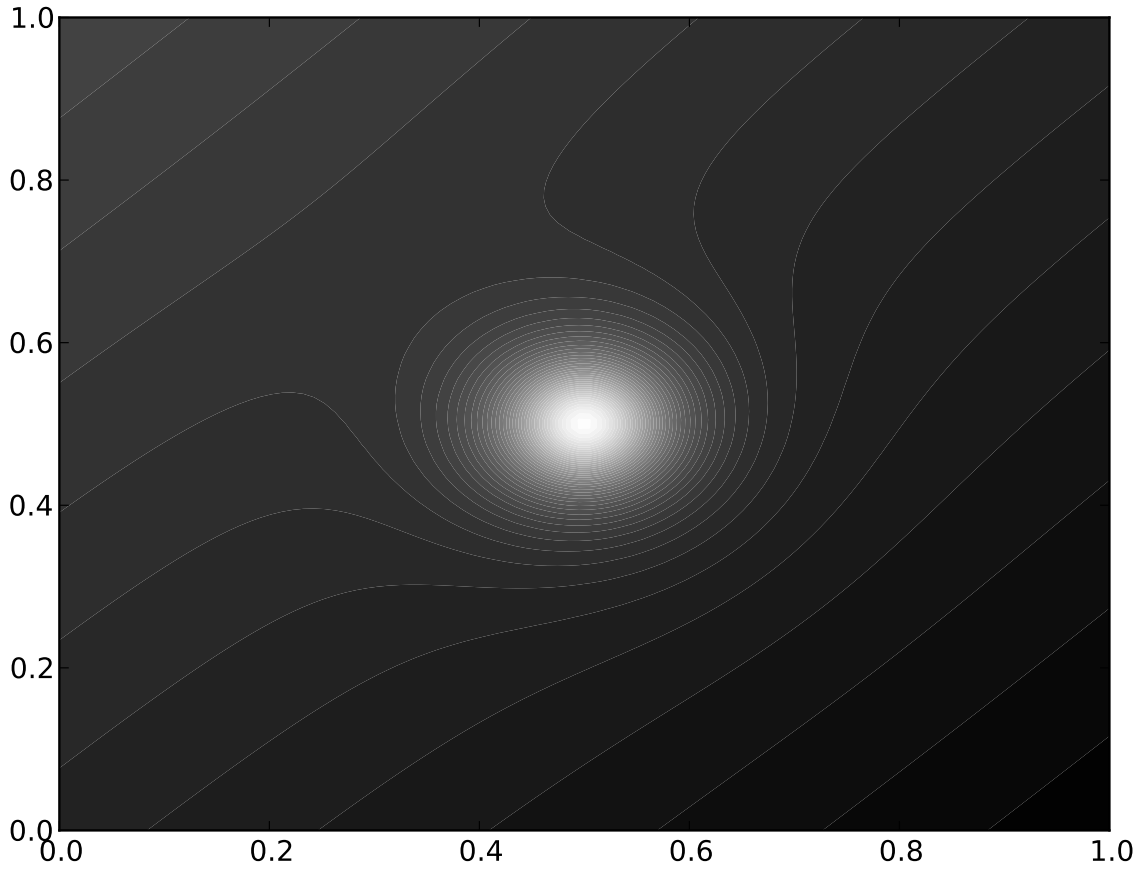

FIG. 6. The density contours on the sky of the forergound cluster and background discussed in §III C, with true parameters given in Eq. (30). There are a factor of 10 more stars (10000) in the background than in the cluster (1000), but the peak density for the cluster is much higher.

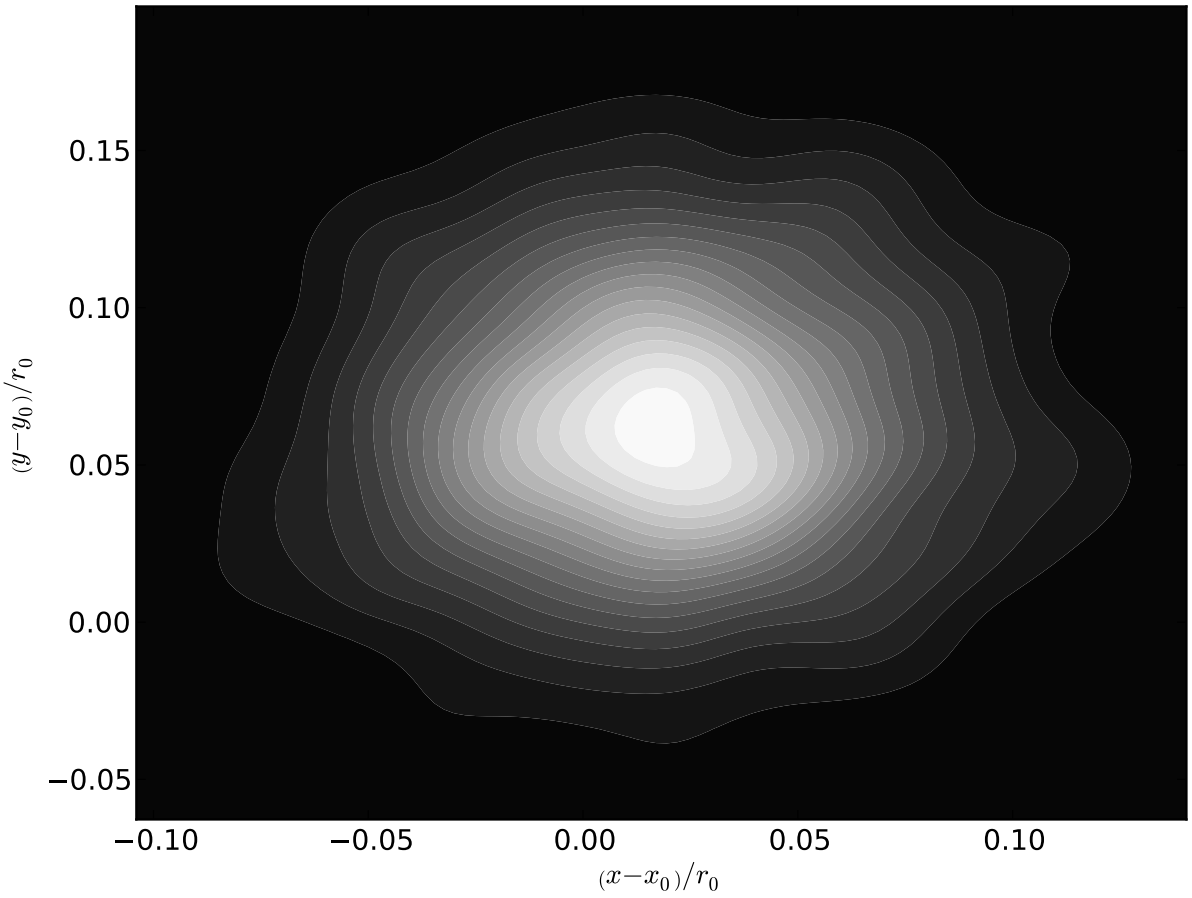

FIG. 7. Contours of the posterior probability distribution for the center of the cluster, $\vec{x}_0$, in the example from § III C. The center $(x, y) = (x_0, y_0)$ is determined to within a few percent of the structural radius of the cluster, r_0 (see Eq. (30)).

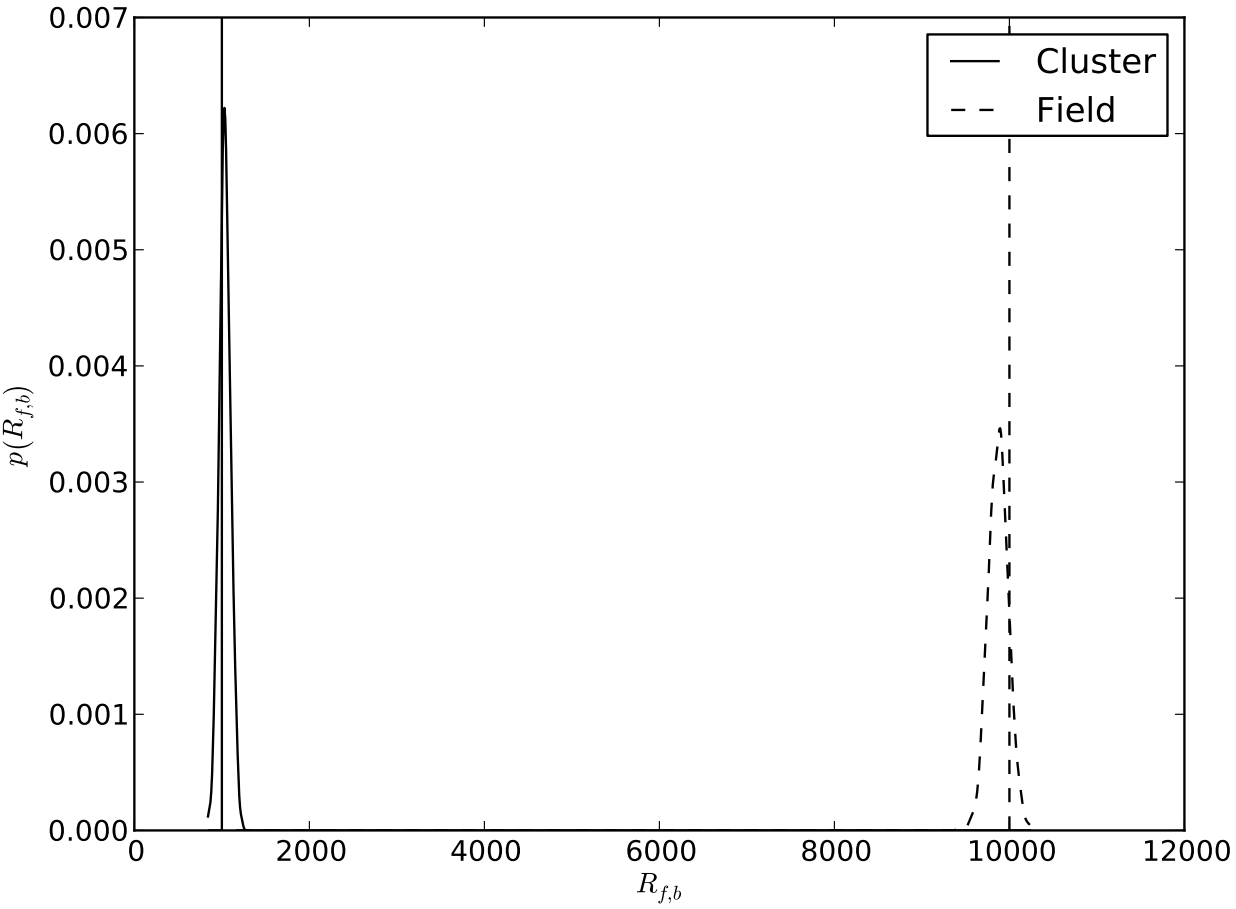

FIG. 8. Posterior densities for the number of stars in the cluster (R_f) and in the field (R_b) in the example from § III C. Vertical lines indicate the true values (see Eq. (30)).

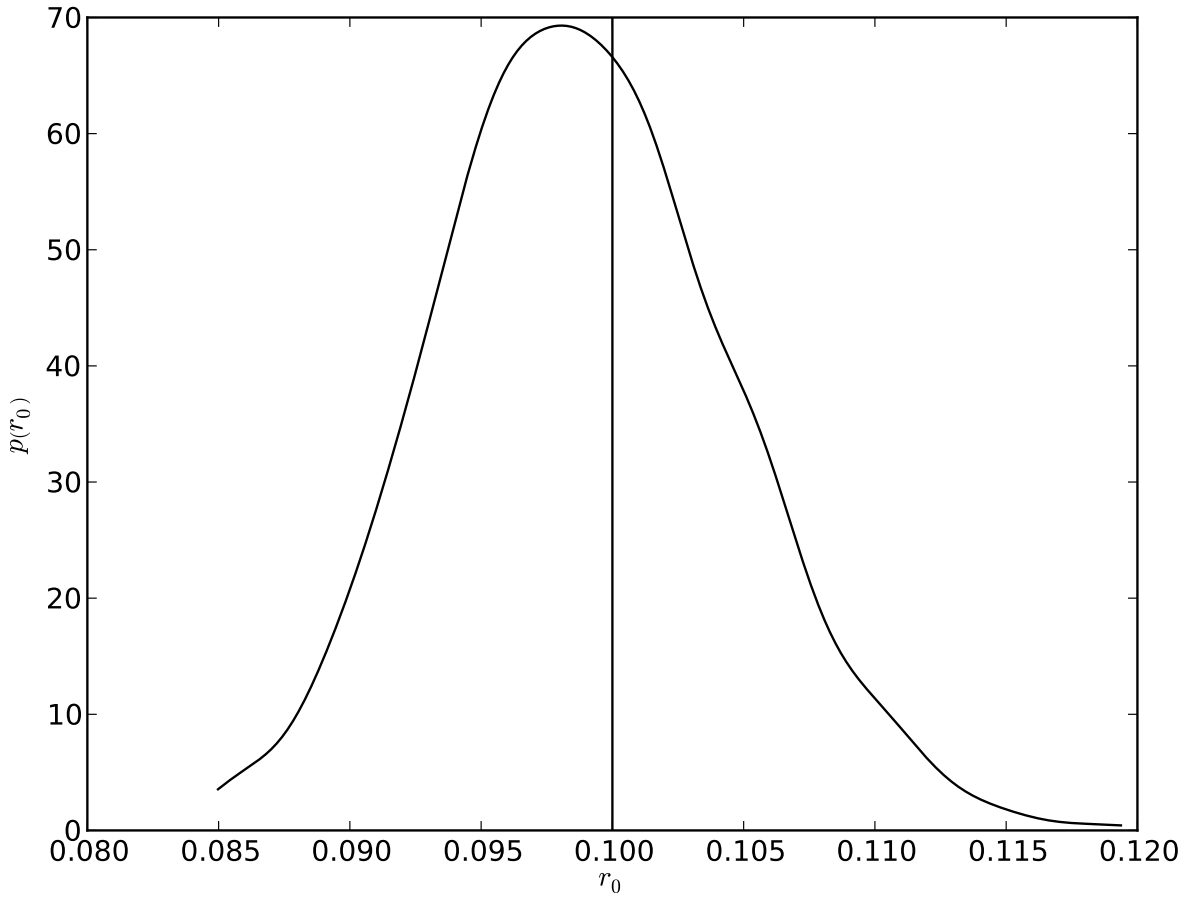

FIG. 9. Posterior density for the scale parameter for the cluster, r_0, from the example in § III C. The true value is indicated by the vertical line (see Eq. 30).