Source Detection in Astronomical Images by Bayesian Model Comparison

Marcus Frean[*,†], Anna Friedlander[*], Melanie Johnston-Hollitt[**] and Christopher Hollitt[*]

[*]*School of Engineering and Computer Science, Victoria University of Wellington*
[†]*Email: marcus.frean@ecs.vuw.ac.nz*
[**]*School of Chemical and Physical Sciences, Victoria University of Wellington*

Abstract. The next generation of radio telescopes will generate exabytes of data on hundreds of millions of objects, making automated methods for the detection of astronomical objects ("sources") essential. Of particular importance are faint, diffuse objects embedded in noise. There is a pressing need for source finding software that identifies these sources, involves little manual tuning, yet is tractable to calculate. We first give a novel image discretisation method that incorporates uncertainty about how an image should be discretised. We then propose a hierarchical prior for astronomical images, which leads to a Bayes factor indicating how well a given region conforms to a model of source that is exceptionally unconstrained, compared to a model of background. This enables the efficient localisation of regions that are "suspiciously different" from the background distribution, so our method looks not for brightness but for *anomolous distributions of intensity*, which is much more general. The model of background can be iteratively improved by removing the influence on it of sources as they are discovered. The approach is evaluated by identifying sources in real and simulated data, and performs well on these measures: the Bayes factor is maximized at most real objects, while returning only a moderate number of false positives. In comparison to a catalogue constructed by widely-used source detection software with manual post-processing by an astronomer, our method found a number of dim sources that were missing from the "ground truth" catalogue.

Keywords: Source detection, astronomy, Bayesian model comparison
PACS: 02.70.-c, 42.30.Tz

The sheer scale of data generated by next-generation radio telescopes makes automated methods for finding astronomical objects essential. Existing approaches require time-intensive manual parameter tuning, and manual post-processing by an astronomer, and are not fully adequate to find all objects of interest [1, 2, 3].

Most automated source detection algorithms can be described as flood-filling or region-growing, driven by (possibly transformed) pixel intensities [4]. These often require restricting the sources to those that are at least as bright as some threshold, typically set at 3σ or 5σ above root mean square (rms) noise. However, some of the most scientifically important objects in astronomy are dim, low surface brightness sources, with intensities in the range of background noise [2].

There are a small number of Bayesian source detection algorithms (reviewed in [4], see also [5], and in the current volume), most of which use Markov-chain Monte Carlo sampling (MCMC) to estimate the relative probability that each pixel (or pixel grouping) in an image arises from background or is a source pixel [6, 4]. For examples, see [7], [8], and [9].

The work most similar to that in this paper is that of Hobson and McLachlan [10], who performed Bayesian model selection using MCMC to explore the parameters of background and source models, with a source model of circularly symmetric Gaussian-profile objects and a background model of Gaussian noise. They present two versions of the algorithm: one in which all sources are found simultaneously, and one in which sources are found iteratively. Brewer *et. al.* [11] have successfully scaled a model similar to Hobson and McLachlan's [10] up to ~ 1000 sources.

In contrast to the existing Bayesian source detection methods, in our procedure source and background are modelled as distributions over ranges of pixel intensities. We first show how to avoid the worst problems caused by naive binning procedures. In common with other approaches we then assume a parametrised form (shape) for astronomical sources and find the most plausible shape parameters per source, however sources are not assumed to be circularly symmetric or of a particular size, and we make only weak assumptions about the form of the noise distribution.

By exploiting the relative consistency of image background, we suggest a tractable way to infer the log posterior ratio (Bayes Factor) that a given region is source *vs* background. This avoids the difficulties inherent in approaches that are based on comparing to "typical" sources and yeilds an objective function which, when maximized, finds sources.

Instead of MCMC, we make use the analytic integral over histograms that is possible via the Dirichlet-multinomial distribution.

1. DISCRETISATION BY BINNING

Radio astronomy images have continuous-valued pixel intensities, which can be converted to a smaller set of discrete values or "bins" [12]. A histogram of pixel intensities can be constructed by assigning each pixel in an image to a bin on the basis of its intensity $z_{x,y}$, where bins are ranges of intensity values. The number of bins must also be set (either manually, or by the binning method itself). The simplest way to partition data into bins is to create K equal width partitions in the data range. Alternatively, bins may be defined according to their occupancy, by ranking the data and assigning partitions based on this ranking. Equal occupancy partitions are constructed by dividing the N ranked pixels evenly into K bins, and setting bin ranges accordingly.

There are three major problems with the simple discretisation/binning methods for radio astronomy data: there is a lack of any informed basis for deciding on what bin boundaries to use and so any particular choice amounts to an unjustified assertion; because radio astronomy data is non-uniform, relevant groupings of values may be split across bins or combined in bins in a way that reduces discriminative power [13]; discretisation results in sudden changes at partition values, which introduces variation into the system that did not previously exist. Moreover, data-points that lie near a boundary of a bin are less well represented by that bin than data-points that lie towards the middle of the bin's interval [14].

It would be more principled to assign proportions of each pixel across a range of bins to reflect uncertainty about which bin a pixel belongs to. Instead of making a "hard" assignment of pixels to bins, many sets of bins $b \in B$ (each of size K) could be created, with bin sets differing in their partition points. Each pixel then has a number of "possible" bin assignments, so we have a *distribution* over all the bins for each pixel, reflecting our uncertainty about where the borders should lie. This solves the problem of the sudden changes at bin borders, and is a method for converting continuous values into distributions over discrete values (counts). It reduces the impact both of the uninformed choice of bins and of any within-class splitting over bins and between-class combining within bins [13]. We use the Dirichlet distribution [15, 16, 17], parametrized by $\alpha = \alpha_1, ..., \alpha_K$, to create these sets of bins. A sample from a Dirichlet distribution yeilds a K-element vector whose elements are positive and sum to one. A Dirichlet distribution with symmetric α will produce roughly equal values for all elements, and the larger the α values overall, the more consistent the distributions drawn, since as parameters α grow, the Dirichlet reverts to a categorical having probabilities p_i proportional to α_i. The cumulative sum of those elements can be interpreted as break-point levels $L_0, \ldots, L_K$ of a partition over $[0,1]$. To partition pixel values into bin borders, levels L could be scaled to the range between minimum and maximum pixel values $[z_{min}, z_{max}]$ in some set, or alternatively applied to a *ranking* of those values. Draws from a Dirichlet distribution with such an α vector can thus be used as partition points in a set of roughly equal width, or roughly equal-occupancy bins, respectively.

2. SOURCE DETECTION

For ease of description we use a single discretisation (binning) of the pixel intensities in the argument that follows, but it is straightforward to combine the resulting algorithm with the Dirichlet bins approach given above, and the results section gives performance for both variants. We give a somewhat informal description here to convey the essential ideas – further elaboration can be found in [18].

Given an astronomical image and a generative model for the distribution of background pixels, a likelihood can be assigned to a particular region of that image, reflecting how well the region conforms to the model. If, in addition, there is a model for foreground, a comparison of the likelihoods can indicate which of the two best fits the region. A simple way to do this would be to use categorical distributions (i.e. simple histograms) for both models. However, while background is relatively consistent, every source is different: given this variability, what is the best categorical distribution to use? This *ad hoc* choice will affect the subsequent comparison of likelihoods.

Consider the following generative model (hierarchical prior) for astronomical images. Elliptical regions ("sources") are randomly distributed in the image, with a very broad prior on their sizes and eccentricities. Each such region has an associated categorical distribution ("normalised histogram") over the bin indices of pixels to be found within it. Since this distribution is unknown we consider it a draw from a Dirichlet distribution. With parameters $\alpha_i = 1$, the histogram

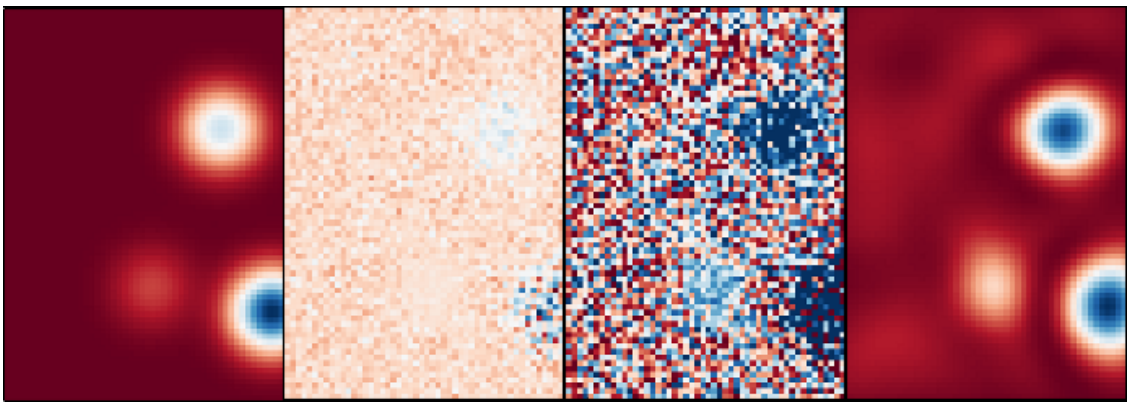

FIGURE 1. Exhaustive calculation of DMR on simulated data. From left to right: three simulated sources; the image overlaid with Gaussian noise; the noisy image binned with Dirichlet equal occupancy bins; the value of DMR where the kernel is centered on (x,y) calculated using these bins, shown using a colour gradient from red (background) to blue (source). The other parameters of W (namely σ_x, σ_y and rotation ϕ) were held constant. All three sources are recovered. The blue peaks in function-space for two of the three sources in the image at right indicate high values for these sources, and roughly equal values despite the bottom source being brighter than the top one. The objective function's value for the dim source is relatively lower than that for the other two sources, however it still exceeds a value that would indicate "background", relative to the true background regions.

is a draw uniformly from the simplex of possible categorical distributions. Each source then generates its pixel's bin indices *i.i.d.* from its unique histogram.

Similarly, all the remaining pixels (the "background") have their bin indices drawn from a categorical which is a taken to be a single draw from a Dirichlet, except that in this case the α parameters are large. There are various ways one might set the background α, but the simplest is to set them equal to the bin counts of the overall image plus one, which (via conjugacy of the Dirichlet with Categorical) yeilds the *maximum a posteriori* distribution one would obtain were there no sources present.

This prior has a strong preference for a particular distribution when it comes to background pixels, but makes almost no assumptions about "source characteristics" other than their elliptical shape. For example, it does not assume that pixels from source are brighter than those from background, only that they are likely to be different. Finally, to avoid forcing source boundaries to have hard borders and be precisely elliptical, we assume the bin indices actually arise from a mixture of the source and background categoricals with the mixing coefficient following a Gaussian-shaped kernel.

Taking a Bayesian view, source detection "inverts" the above to assess the likely origin of the histogram leading to the pixels in a certain region, without committing to the unknown histogram itself. Given a region, the relative plausibilities for the two possible origins are captured by the Bayes factor $\log \frac{P(S|\mathbf{n})}{P(B|\mathbf{n})}$, where S and B are source and background models respectively. Each of these probabilities is a Dirichlet-Multinomial compound distribution, and involves an integral over the unknown histogram involved. A very convenient property of this distribution is that the integral is analytic and easy to evaluate. We denote the Bayes factor "DMR" (Dirichlet Multinomial Ratio) for brevity. The task of locating sources becomes one of optimization of the parameters in DMR (other than hyperparameters α), namely those controlling source location and shape. Since DMR is analytic it can be optimized efficiently by gradient ascent.

DMR can be calculated as follows. Let $b_{(x,y)}$ be the index of the bin that results from the pixel intensity at location (x,y) in the image. For a region R with a hard border the aggregate bin counts would be

$$n_k = \sum_{(x,y)\in R} \delta_{(b_{(x,y)}=k)} \tag{1}$$

where $\delta_{()}$ is the delta function. For "soft" borders, all the counts should be weighted by a Gaussian-shaped kernel function $W^{\theta}_{x,y}$ over locations (x,y). Parameters θ specify the position and shape of the region, namely its center coordinates (m_x, m_y) and full covariance (parameterised by approximate half-widths σ_x and σ_y together with a rotation ϕ). Note that the Gaussian kernel is not normalised and at its peak reaches 1. Other shapes could also be used. We can

now generate a vector $\mathbf{n}^\theta$ containing the aggregate of weighted bin counts, with k^{th} component

$$n_k^{(\theta)} = \sum_{(x,y)} W_{x,y}^{\theta} \ \delta_{(b_{(x,y)}=k)} \tag{2}$$

Given any vector of counts $\mathbf{n}$ in K bins of a histogram, integrating out the multinomial distribution gives the following marginal joint likelihood [17]:

$$P(\mathbf{n}|\alpha) = \frac{\Gamma(A)}{\Gamma(N+A)} \prod_k \frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)} \tag{3}$$

where $\Gamma(.)$ denotes the gamma function, $A = \sum_k \alpha_k$ and $N = \sum_k n_k$. The ratio of posterior probabilities [19] comparing the two possible origins for the counts is:

$$\text{DMR}(\theta) = \log \frac{P(\mathbf{n}^{(\theta)}|S)}{P(\mathbf{n}^{(\theta)}|B)} \ + \log \frac{P(S)}{P(B)} \tag{4}$$

where S denotes use of hyperparameter vector α^S, and similarly for B. To reflect the prior belief that radio astronomy images are dominated by background, a heuristic value for the the ratio of source to background pixels in the image may be used in place of the fraction in the second term: we used 0.05. Denoting the derivative of the log of the gamma function by ψ, the gradient of DMR with respect to the region parameters θ is [18]:

$$\frac{\partial}{\partial\theta}\text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \ \frac{\partial n_k^{(\theta)}}{\partial\theta} \tag{5}$$

$$\text{where} \qquad Q_k = \psi(n_k^{(\theta)} + \alpha_k^S) - \psi(n_k^{(\theta)} + \alpha_k^B)$$
$$Q_{\text{base}} = \psi(N^{(\theta)} + A^S) - \psi(N^{(\theta)} + A^B)$$

The final term $\frac{\partial n_k^{(\theta)}}{\partial\theta}$ can be calculated [18] on a per-parameter basis from Equation 2.

3. METHODS

The performance of DMR in identifying astronomical sources was evaluated against simulated and real data.

Real data comprised 25 randomly selected 500×500 pixel windows of two large astronomical images[1]. Each window was treated as a whole image for purposes of evaluation of DMR; windows had an average of 36 sources each.“Ground truth" sources and their parameters (central position and coordinates describing a bounding box around each source) were identified using the source-finding package BLOBCAT [21] with manual postprocessing by an astronomer. Noisy borders of the images were excluded from source detection in construction of the ground truth catalogue.[2] BLOBCAT was restricted to finding sources at least 4σ above rms noise, and the final catalogue was restricted to sources at least 5σ above rms noise. For pixel-intensity based thresholding methods such as BLOBCAT, search must be restricted in this way to avoid identifying background regions as sources. However, some of the most scientifically important objects in astronomy are those that are dim, with intensities in the range of background noise [2]. The DMR does not restrict search in this way.

Simulated images were created by defining three “source" regions using a Gaussian kernel function times a random multiplier for the mean and overlaying them with zero-mean Gaussian noise. The noise had fixed-variance (with the same variance as background pixels) for one source, variance that increased proportionally with the intensity of the source for another, and a variance below that of the background for the third source.

---

[1] Australia Telescope Large Area Survey European Large Area ISO Survey S1 and Australia Telescope Large Area Survey Chandra Deep Field-South [20].

[2] The area of the two images for source extraction was defined by: rms noise $\leq 100\mu$ Jy beam^1; bandwidth smearing $\geq 80\%$; and mosaicked primary beam response $\geq 40\%$. The defined area covers 3.566 square degrees of the CDFS image 2.697 square degrees of the ELAIS image [22].

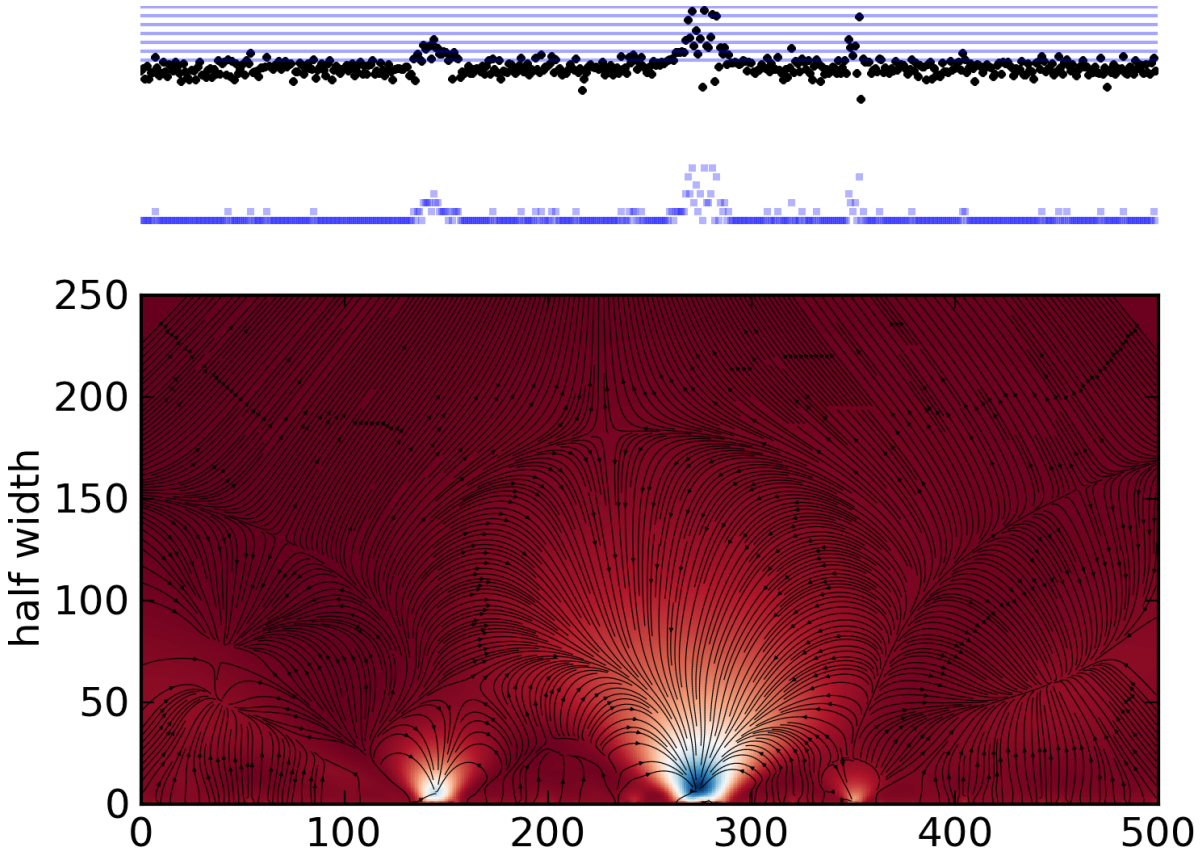

FIGURE 2. Exhaustive calculation of DMR and its gradient over locations $m \in [0..500]$ and half-widths $\sigma \in [0..250]$. Three sources are present, at m =145, 275, and 354. The three subplots correspond to: a plot of the data (black dots) with blue lines at the location of bin borders; a plot of the binned data (blue square dots); the function space with parameter m on the x-axis and σ on the y-axis. DMR is shown via the background colour (red < 0, white $= 0$, blue > 0), and its gradient as black lines.

The images were discretised using equal occupancy and equal width bin border strategies with $K = 10$. The Dirichlet-border "softening" of both types of bin borders was also used with a symmetric α vector with $\alpha_i = 10 \; \forall i$, and generating 50 sets of bins.

We set the source model α^S to the symmetric α vector with $\alpha_i = \alpha_j = 1 \; \forall i, j \in 1, ..., K$, which is non-committal about what a source distribution might be, and α^B to be the binned-counts of the whole image. Given these parameterisations for α^B and α^S, background regions are expected to have a similar distribution over bins to that of the whole image, while a source region may have any distribution over K bins. A high DMR value using Equation 4 means that the counts $\mathbf{n}$ in a region are highly dissimilar to those expected under multinomial distributions drawn from a Dirichlet distribution having parameter vector α^B. So, for example, a region in which there is a higher proportion of bright pixels than is typical in the image would receive a high DMR value. More generally, regions in the image with *any* "unusual" distributions across bins will correspond to peaks in the function.

A bounded Newton Conjugate-Gradient algorithm [23] was used to implement gradient ascent. One "round" of source-finding consisted of 50 trials of gradient ascent for simulated data, and just one trial of gradient ascent for real data (to speed up execution). The components of θ corresonding to the center of the region W were initialised to coordinates of the brightest pixel in the image, while other parameters were initialised to random values.

The peak with the highest value in each round was recorded as a found source, and removed from the data. After each source removal, histogram bin borders and the α^B vector were re-calculated on the remaining data.

For simulated data, three rounds were performed per image. For real data, rounds of gradient ascent continued until the maximum DMR found was negative or after 53 rounds (the maximum number of sources in any window), whichever came first.

Because the test regions are continuous in the image, for the purposes of evaluation the region of a found source is

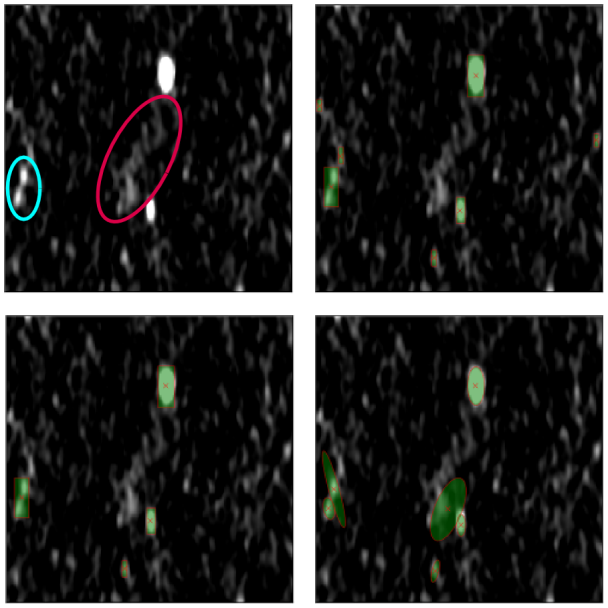

FIGURE 3. Comparison with the ground truth catalogue. *Top left:* A section of one of the CDFS test-windows with a low brightness source (radio galaxy tail; pink ring). Note also the two separate sources in the blue ring. *Top right:* sources identified by BLOBCAT (raw, unprocessed output). *Lower left:* sources in the ground-truth catalogue, and (*lower right)* found by the DMR algorithm. DMR has identified all the sources in the ground truth catalogue, and, in addition has found part of the radio galaxy tail missing from the ground truth catalogue. It has also correctly identified two separate sources that the ground truth catalogue conflated (blue ring).

TABLE 1. Performance of the DMR: precision, recall.

Binning method	Simulated data*	Real data
Equal width	0.83, 0.83	0.51, 0.74
Dirichlet equal width	0.57, 0.57	0.59, 0.75
Equal occupancy	0.79, 0.79	0.30, 0.30
Dirichlet equal occupancy	0.84, 0.84	0.32, 0.34

* For simulated data, there are three sources per image, and three rounds of gradient ascent per image, so precision = recall.

taken to be the ellipse defined by a doubling of the optimized parameters corresponding to the spread of the source region, which is 95% of the area under the curve of the associated Gaussian kernel function [24].

In the case of real data, “true sources" were those identified by the source-finding package BLOBCAT [21] with manual postprocessing by an astronomer. These sources have parameters describing a “bounding box" around the source. Precision and recall [25] were calculated in order to evaluate performance. A true positive occurs when the location of at least half of a found source’s pixels overlap with at least half of a real source’s pixels. A false positive is a found source that does not meet these criteria, a false negative is a true source that does not meet these criteria.

4. RESULTS AND DISCUSSION

Figure 1 illustrates an exhaustive calculation of DMR on simulated data over x and y positions $\theta = (m_x, m_y)$ with other parameters kept constant. The positions of the original sources are recovered almost perfectly, despite the fact that they are obscured with noise.

Figure 2 shows an exhaustive calculation of DMR and its gradient on a one dimensional $1 \times X$ “slice" through an astronomical image (ATLSB survey region A at 50” resolution [26, 27]), using Dirichlet equal width bins and a one dimensional Gaussian region, with centre and half-width parameters $\theta = (m, \sigma)$. High DMR values correspond to the

location of sources, while the rest of the image has values that tend strongly towards the background model. Gradient lines converge at the peaks of the regions of values indicating sources.

Performance of DMR is given in Table 1. Low precision of the DMR on real data is partly due to the fact that DMR found more sources than appear in the ground truth catalogue. While some of these additional found sources are spurious, some appear to be sources that were missed by BLOBCAT with manual postprocessing. This is not surprising given that the ground truth catalogue only contained sources at least 5σ above rms noise, while DMR was not restricted in this way. In fact, even the raw BLOBCAT output (without manual postprocessing) had precision of only 0.69 when compared to the ground-truth catalogue. Examples of found sources that were not found by BLOBCAT and are not in the ground truth catalogue can be seen in Figure 3.

Some of the most scientifically important objects in astronomy are dim, with intensities in the range of background noise [2]. The nature of pixel-intensity based thresholding algorithms such as BLOBCAT restricts their ability to find such dim sources without also "finding" a very large number of noise regions. The implementation described here is a preliminary exploration, but encouraging. The DMR algorithm does not restrict the objects found on the basis of some threshold above rms noise, and produces raw results containing most real objects (including dim sources that are not well found by other other algorithms), while returning only a moderate number of false positives in regions of noise.

REFERENCES

1. C. Hollitt, and M. Johnston-Hollitt, *Publications of the Astronomical Society of Australia* **29**, 309–317 (2012).
2. R. Norris, A. Hopkins, J. Afonso, S. Brown, J. Condon, et al., *Publications of the Astronomical Society of Australia* **28**, 215–248 (2011).
3. R. P. Norris, J. Afonso, D. Bacon, R. Beck, M. Bell, et al., *Publications of the Astronomical Society of Australia* **30**, 20 (2013).
4. M. Masias, J. Freixenet, X. Lladó, and M. Peracaula, *Monthly Notices of the Royal Astronomical Society* (2012).
5. M. Selig, and T. A. Enßlin, *CoRR* **abs/1311.1888** (2013).
6. F. Feroz, and M. Hobson, *Monthly Notices of the Royal Astronomical Society* **384**, 449–463 (2008).
7. R. S. Savage, and S. Oliver, *The Astrophysical Journal* **661**, 1339 (2007).
8. P. Carvalho, G. Rocha, and M. Hobson, *Monthly Notices of the Royal Astronomical Society* **393**, 681–702 (2009).
9. F. Guglielmetti, R. Fischer, and V. Dose, *Monthly Notices of the Royal Astronomical Society* **396**, 165–190 (2009).
10. M. Hobson, and C. McLachlan, *Monthly Notices of the Royal Astronomical Society* **338**, 765–784 (2003).
11. B. J. Brewer, D. Foreman-Mackey, and D. W. Hogg, *The Astronomical Journal* **146**, 7 (2013).
12. Y. Yang, G. I. Webb, and X. Wu, "Discretization methods," in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 101–116.
13. E. J. Clarke, and B. A. Barton, *International Journal of Intelligent Systems* **15**, 61–92 (2000).
14. Y. Yang, and G. I. Webb, "Non-disjoint discretization for naive-Bayes classifiers," in *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002, pp. 666–673.
15. B. A. Frigyik, A. Kapila, and M. R. Gupta, *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006* (2010).
16. B. Hoadley, *Journal of the American Statistical Association* **64**, 216–229 (1969).
17. K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and Related Distributions: Theory, Methods and Applications*, vol. 888, Wiley, 2011.
18. A. M. Friedlander, *Dirichlet methods for Bayesian source detection in radio astronomy images*, Master's thesis, Victoria University of Wellington (2014).
19. R. E. Kass, and A. E. Raftery, *Journal of the American Statistical Association* **90**, 773–795 (1995).
20. R. P. Norris, J. Afonso, and P. N. Appleton, *The Astronomical Journal* **132**, 2409–2423 (2006).
21. C. A. Hales, T. Murphy, J. R. Curran, E. Middelberg, B. M. Gaensler, and R. P. Norris, *Monthly Notices of the Royal Astronomical Society* **425**, 979–996 (2012).
22. J. K. Banfield, T. M. O. Franzen, A. Hopkins, R. P. Norris, N. Seymour, K. E. Chow, C. Hales, M. T. Huynh, E. Lenc, M. Mao, and E. Middleberg, The Australia Telescope Large Area Survey I: 1.4 GHz source catalogue and counts (in prep.).
23. E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open source scientific tools for Python, `http://www.scipy.org/` (2001–).
24. L. Wasserman, *All of statistics: a concise course in statistical inference*, Springer Verlag, 2004.
25. D. Olson, and D. Delen, *Advanced data mining techniques*, Springer Verlag, 2008.
26. L. Saripalli, R. Subrahmanyan, K. Thorat, R. Ekers, R. Hunstead, H. Johnston, and E. Sadler, *The Astrophysical Journal Supplement Series* **199** (2012).
27. R. Subrahmanyan, R. Ekers, L. Saripalli, and E. Sadler, *Monthly Notices of the Royal Astronomical Society* **402**, 2792–2806 (2010).