

Dirichlet Methods for Bayesian Source Detection in Radio Astronomy Images

by

Anna Friedlander

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Computer Science.

Victoria University of Wellington
2014

Abstract

The sheer volume of data to be produced by the next generation of radio telescopes — exabytes of data on hundreds of millions of objects — makes automated methods for the detection of astronomical objects (“sources”) essential. Of particular importance are low surface brightness objects, which are not well found by current automated methods.

This thesis explores Bayesian methods for source detection that use Dirichlet or multinomial models for pixel intensity distributions in discretised radio astronomy images. A novel image discretisation method that incorporates uncertainty about how the image should be discretised is developed. Latent Dirichlet allocation — a method originally developed for inferring latent topics in document collections — is used to estimate source and background distributions in radio astronomy images. A new Dirichlet-multinomial ratio, indicating how well a region conforms to a well-specified model of background versus a loosely-specified model of foreground, is derived. Finally, latent Dirichlet allocation and the Dirichlet-multinomial ratio are combined for source detection in astronomical images.

The methods developed in this thesis perform source detection well in comparison to two widely-used source detection packages and, importantly, find dim sources not well found by other algorithms.

Acknowledgments

Marcus, thank you for your generosity with your time and ideas, for believing in me, and always encouraging me to do better.

Thank you Melanie for the privilege of working alongside you and your team of radio astronomers.

Thank you Chris for all your practical assistance and insightful suggestions.

I gratefully acknowledge the financial assistance provided by a KAREN Capability Build Fund grant in support of radio astronomy.

To my husband Karl: thank you for your support, encouragement, and unwavering optimism.

To my son Malachi: your beautiful smile, kind heart and cheeky sense of humour always brighten my day.

Contents

1	Source detection in radio astronomy	1
1.1	Source detection in radio astronomy	2
1.1.1	Characteristics of radio astronomy images	2
1.1.2	Existing approaches to source detection	7
1.1.3	Bayesian source detection methods	11
1.2	Outline of the thesis	13
1.3	Original contributions	14
2	Discretising images	17
2.1	Histogram binning strategies	18
2.1.1	Width-based partitions	19
2.1.2	Frequency-based partitions	22
2.2	A novel method for softening bin borders	23
2.2.1	Creating sets of bins with the Dirichlet distribution	24
2.3	Discretised images: exploration	30
2.4	Extending the exploratory approach	34
3	Latent Dirichlet allocation	37
3.1	Problem framework	37
3.2	Latent Dirichlet allocation	39
3.2.1	The generative model	39
3.2.2	Inferring the latent variables	41
3.2.3	Application of LDA to images	45

3.3	Methods	47
3.4	Results	49
3.5	Discussion	56
3.5.1	Parameter and computational issues	56
3.5.2	Future work	58
3.6	Conclusions	60
4	The DMR: derivation	61
4.1	Notation	62
4.1.1	Representing an image in terms of binned values	62
4.2	Scoring a region	68
4.2.1	The Dirichlet-multinomial distribution	68
4.2.2	A Dirichlet-multinomial ratio	69
4.3	The gradient of the DMR	71
4.3.1	The gradient, given a specific parameterisation for W	73
4.4	The DMR in higher dimensions	75
4.5	Comparison with other Bayesian methods	76
4.6	Comparison with latent Dirichlet allocation	77
5	The DMR: evaluation	81
5.1	Methods	81
5.1.1	Parameters	83
5.1.2	Derivation of α^S and α^B	93
5.1.3	Dirichlet-multinomial ratio	95
5.1.4	Gradient ascent	102
5.1.5	Evaluation methods	104
5.2	Results and discussion	105
5.2.1	One dimensional data	105
5.2.2	Two dimensional data	110
5.3	Conclusions and future work	116
5.3.1	Simulated data	116
5.3.2	Real data	118

5.3.3	Binning strategies	121
5.3.4	Potential extensions to the DMR	121
6	The DMR with LDA output	123
6.1	Methods	124
6.2	Results	126
6.3	Discussion and future work	128
7	Conclusions	131
7.1	Summary	131
7.2	Future work	132
7.2.1	Image discretisation	132
7.2.2	Latent Dirichlet allocation	133
7.2.3	The Dirichlet-multinomial ratio	134
7.2.4	The Dirichlet-multinomial ratio using LDA output	136
7.2.5	From exploration to usable software	136

List of Figures

1.1	Point source	5
1.2	Radio galaxy	6
1.3	Tailed radio galaxy	7
1.4	Radio relic	8
1.5	Supernova remnant	9
2.1	A histogram of pixel intensities in an astronomical image . .	21
2.2	Dirichlet samples on the simplex	27
2.3	Dirichlet bin borders	28
2.4	Binned views of an image	29
2.5	Sources found by the exploratory multinomial model	33
3.1	Graphical model for LDA	40
3.2	Matrices used in LDA	43
3.3	Image segmented by LDA	50
3.4	The number of bins can affect results	51
3.5	Radio astronomy images segmented by LDA	52
3.6	LDA results on an image polluted with artefacts	53
3.7	A comparison between LDA and <i>Duchamp</i>	55
4.1	Two methods for defining a region	65
4.2	Variation in background across an image	79
5.1	One dimensional data from a two dimensional image	84

5.2	Skewed generalised Gaussian curves	86
5.3	Simulated one dimensional data	87
5.4	Astronomical images used for evaluation of the DMR	89
5.5	Simulated astronomical images	91
5.6	Exhaustive calculation of the DMR and its gradient (1)	97
5.7	Exhaustive calculation of the DMR and its gradient (2)	98
5.8	Exhaustive calculation of the DMR and its gradient (3)	99
5.9	Background compensation correction to the DMR	101
5.10	Hard-bordered versus soft-bordered test regions	107
5.11	Peaks found in a one dimensional “slice” of an image	108
5.12	The DMR at different thresholds (1D)	109
5.13	Exhaustive calculation of the DMR (2D)	112
5.14	The DMR at different thresholds (2D)	113
5.15	Gradient ascent on the DMR (2D)	114
5.16	Iterative source removal and rebinning	116
5.17	Comparison with the ground truth catalogue	117
5.18	False positives due to multiple peaks on one source	120
6.1	Astronomical images used for evaluation	125
6.2	Results of the DMR combined with LDA	127

List of Tables

3.1	Astronomical images used for evaluation of LDA	47
3.2	Comparison of the performance of LDA with <i>Duchamp</i> . . .	54
5.1	Astronomical images used for evaluation of the DMR	88
5.2	Performance of the DMR (simulated 1D)	106
5.3	Performance of the DMR (simulated 2D)	111
5.4	Performance of the DMR (real 2D)	111
6.1	Performance of DMR using LDA output	128

Chapter 1

Source detection in radio astronomy

The sheer volume of data to be produced by next generation radio telescopes makes automated methods for the detection of astronomical objects (“sources”) essential. Current methods require time-intensive parameter tuning and do not find all sources, particularly scientifically interesting low surface brightness objects.

This thesis explores Bayesian methods that use Dirichlet or multinomial models for pixel intensity distributions in source detection in discretised radio astronomy images.

The methods developed in this thesis exploit the consistency of image background, while minimising the problems inherent in modelling the diversity of astronomical objects. The posterior probability that a given region is source or background is inferred, avoiding the difficulties in approaches based on comparing “typical” exemplars of source and background.

1.1 Source detection in radio astronomy

Existing automated approaches to detecting scientifically important astronomical objects require time-intensive manual parameter tuning, and manual post-processing by an astronomer. The sheer volume of data to be produced by the next generation of radio telescopes — exabytes of data on hundreds of millions of objects — will make efficient and timely detection of astronomical objects by such manually intensive processing at best impracticable and at worst impossible. Fully automated approaches that do not require manual tuning and post-processing are therefore essential to finding objects of interest in astronomical images [53].

Further, current algorithms for automatic source detection are not fully adequate to find all objects of interest. Spatially extended sources, particularly those that are faint, are poorly handled by existing automated approaches, as are sources in the presence of artefacts¹, and sources in images in which the signal-to-noise ratio varies across the image [30, 53, 55].

1.1.1 Characteristics of radio astronomy images

Radio astronomy images can be thought of as primarily background (made up of unresolved sources) with an unknown number of spatially extended sources. Identifying the sources requires distinguishing them from background, a task made difficult by the diversity of pixel intensities within and between sources. In general, source pixels are brighter than background pixels, though there is considerable overlap between background and source intensity ranges.

The intensity of background pixels in radio telescope images is non-uniform,

¹“Artefacts” are patterns in radio astronomy images which do not represent actual signal but rather reflect the lack of full sampling of the Fourier plane by the radio telescope.

but follows a Gaussian distribution in local areas. The variability of background however is lower than that of sources and in that sense it is easier to identify.

The noise in radio astronomy images varies over an image. Images typically display more noise around their edges (for example, see Figure 4.2 in Chapter 4).

Sources in radio astronomy images can be divided into the following classes:

- Point sources (Figure 1.1), which are unresolved sources in an image; at or below the resolution element of a telescope. Point sources are abundant in radio astronomy images. They are relatively bright and not spatially extended, and are well-found by existing algorithms².
- Galaxy clusters and associated structures, which are of interest to astronomers. These include:
 - Radio galaxies, which consist of a small, circular, central source from which two roughly symmetrical, elliptical lobes emanate 180 degrees from each other (“bipolar” lobes). The central source is typically brighter than the lobes. Radio galaxies such as the linear source Centaurus A (Figure 1.2) may be bright and extended, or they may be diffuse, extended, low surface brightness sources. One or more of the lobes or central source may be absent or may be distorted. A galaxy’s symmetry may be distorted, and lobes may be warped and may differ in size, shape, and brightness. One lobe may appear brighter than the other. Unrelated objects (for example, point sources) may appear on or near the galaxy. Tailed radio galaxies are a sub-class of radio galaxies (Figure 1.3).

²Hancock *et. al.* [27] report that $\sim 99\%$ of point sources are found by widely used source detection packages (that is, a $\sim 1\%$ false negative rate); similarly $\sim 99\%$ of reported point sources are actual point sources (a $\sim 1\%$ false positive rate).

- Radio relics (found on the periphery of clusters) appear as dim, diffuse, extended single or double arcs (Figure 1.4).
- Radio halos (found around the centres of galaxy clusters) are diffuse, circular structures.
- Supernova remnants (SNRs; Figure 1.5) are also of interest to astronomers. SNRs generally appear as hollow circular rings, although regions of the rings are often occluded, and symmetry may be distorted. Old and near SNRs appear faint, diffuse and large in images. Young and distant SNRs may be miss-classified as point sources.

Artefacts in images appear as bright circles, rings and radial lines. Artefacts will be less prevalent in next generation radio telescope images, due to the increased number of elements in next generation arrays [53].

A number of characteristics of astronomical objects of interest to astronomers make the object-detection task non-trivial. These include distorted symmetry and other variations in the appearance of sources, occlusions, and low surface brightness, particularly in diffuse objects with intensities close to background.

Image format

Radio telescope images are typically stored in FITS (Flexible Image Transportation System) format. For the astronomical images used in this thesis, image data is stored in two dimensional arrays of pixel intensities with implicit Cartesian coordinates. Pixel intensity values are continuous, and range from small positive to small negative values in any given image. Image meta-data including information about the equatorial coordinates of the image is also stored with this format [73].

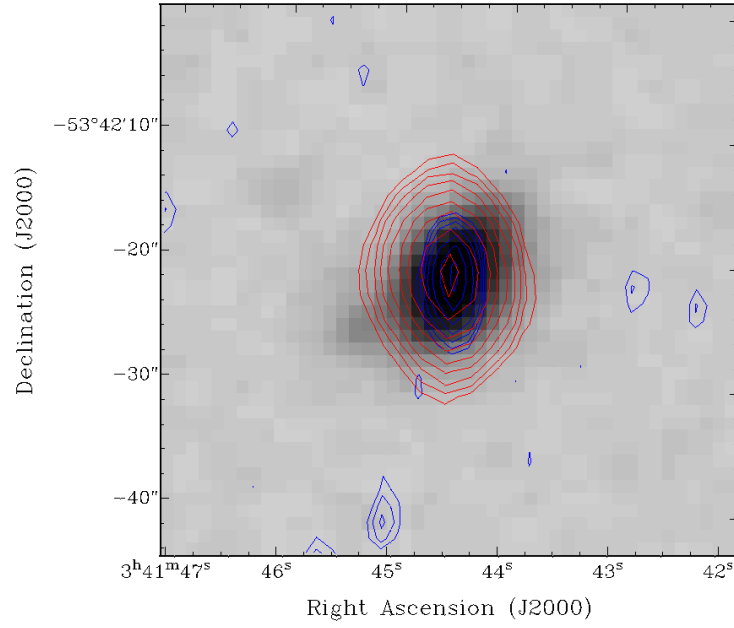


Figure 1.1: **A point source.** The red contours show the 1.4 GHz radio data and the blue show the 2.4 GHz data overlaid on the optical image from the Digitized Sky Survey. The resolution of these radio data are 8.8×4.6 arcseconds at position angle of 2.4 degrees and 5.0×2.6 arcseconds at 2.5 degrees at 1.4 and 2.4 GHz, respectively. Despite the fact that the optical host galaxy is clearly seen, the radio data are unresolved at both frequencies and so this is a point source in a radio image. The source appears elliptical due to the beam of the telescope [35].

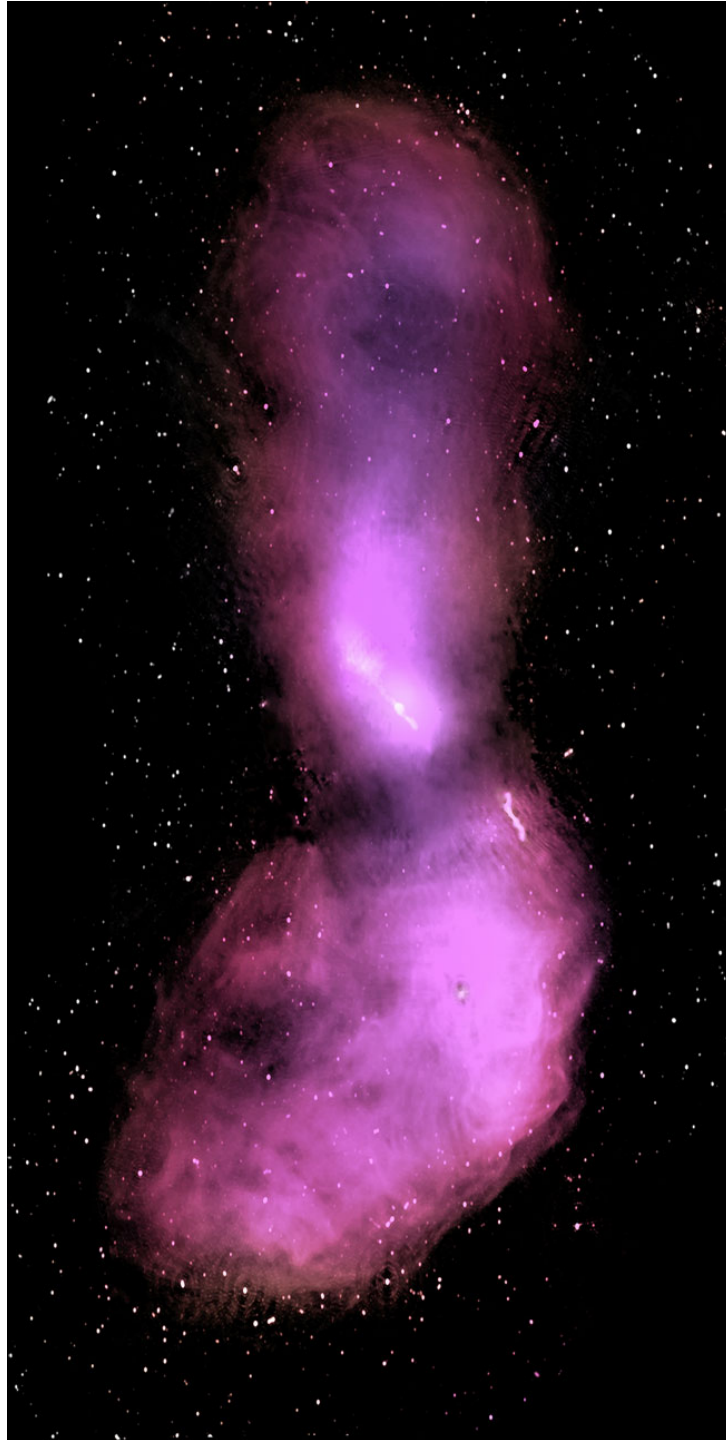


Figure 1.2: **The radio galaxy Centaurus A.** Centaurus A is the nearest powerful radio galaxy of the Fanaroff-Reilly class 1 radio galaxies, which typically have bipolar jets [17]. The image is false colour graded from white (bright pixel intensities) to purple to black (dim pixel intensities).

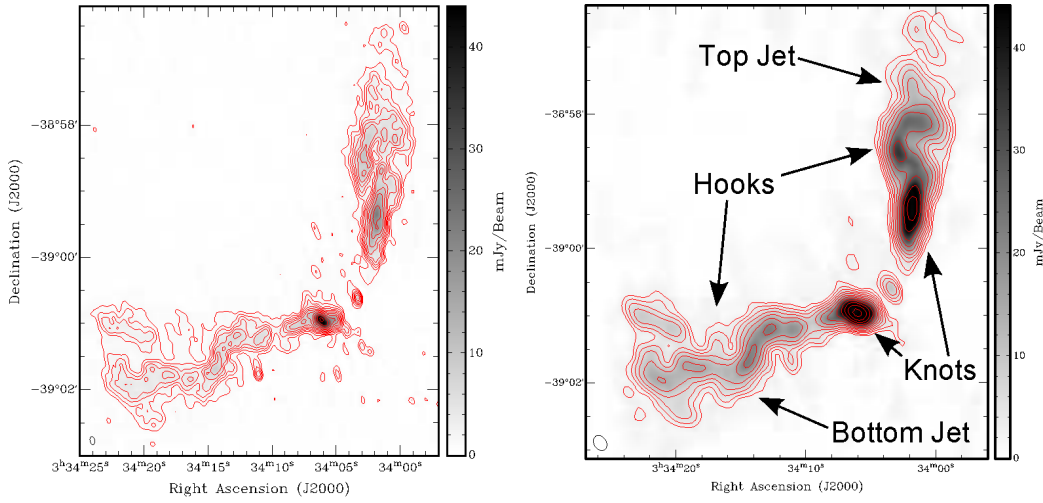


Figure 1.3: **The tailed radio galaxy PKS J0334-3900 [58].** Note the bright compact knots and the low surface brightness emission associated with the jets.

1.1.2 Existing approaches to source detection

An ideal source-finding algorithm should locate regions in an image that differ from the background and are unlikely to have arisen by statistical variation in noise. False positives are more acceptable than false negatives (as astronomers can do post-processing to confirm and characterise sources), however, a large number of false positives would make output difficult to interpret at best and unusable at worst.

In a 2012 review, Masias *et. al.* [46] characterised current approaches to automated and semi-automated source detection in astronomy as a two-step process: an image transformation step followed by a detection step.

Masias *et. al.* divided approaches in the transformation step into basic transformation methods, Bayesian approaches, matched filtering, and multi-scale approaches. Basic transformation methods predominate, and include median filtering, local thresholding using information such as pixel intensity mean and standard deviation, Gaussian image convolution, di-

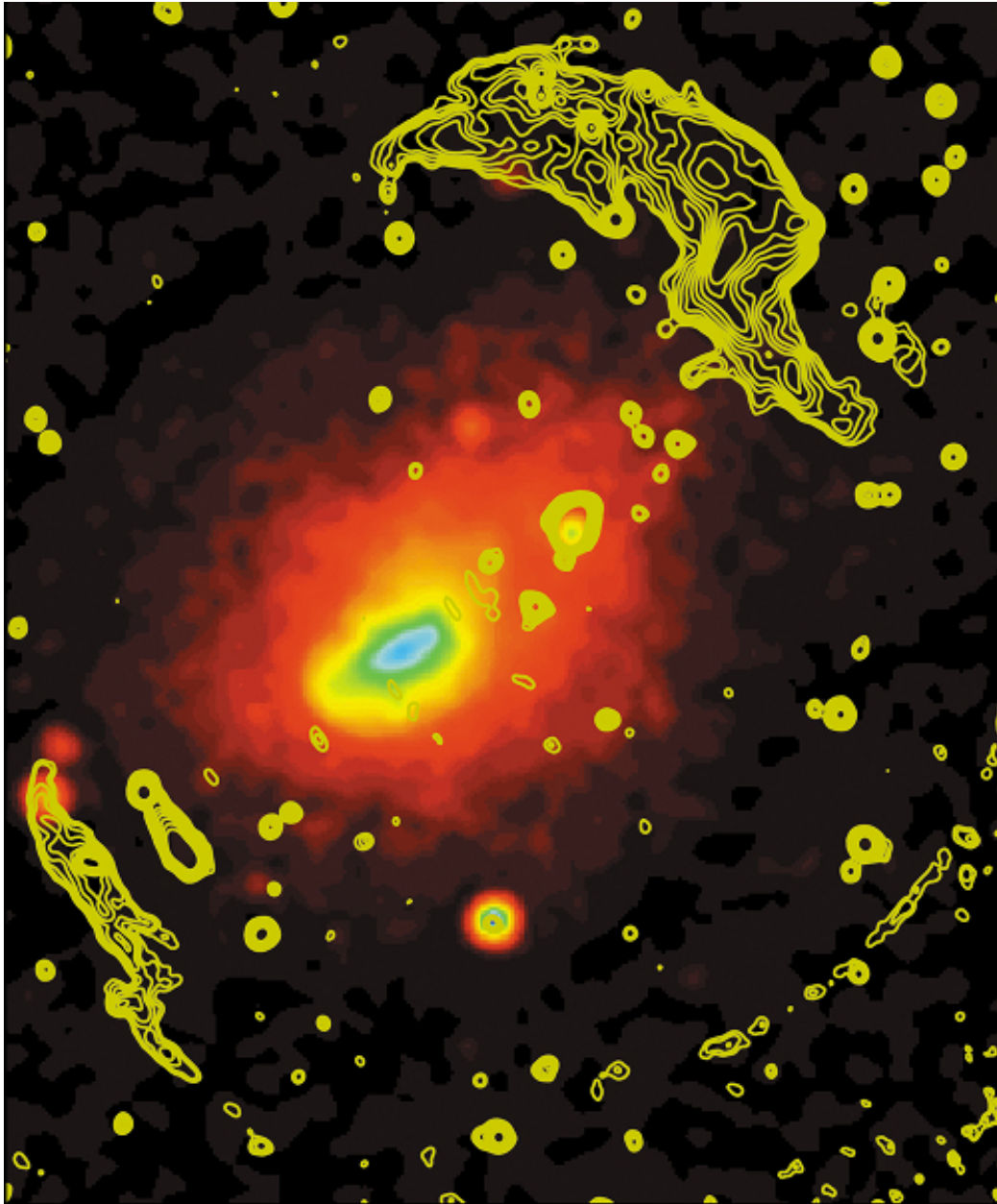


Figure 1.4: **Radio relic ABELL 3667.** The yellow contours are the radio continuum data, shown overlaid on a coloured X-ray image [34].

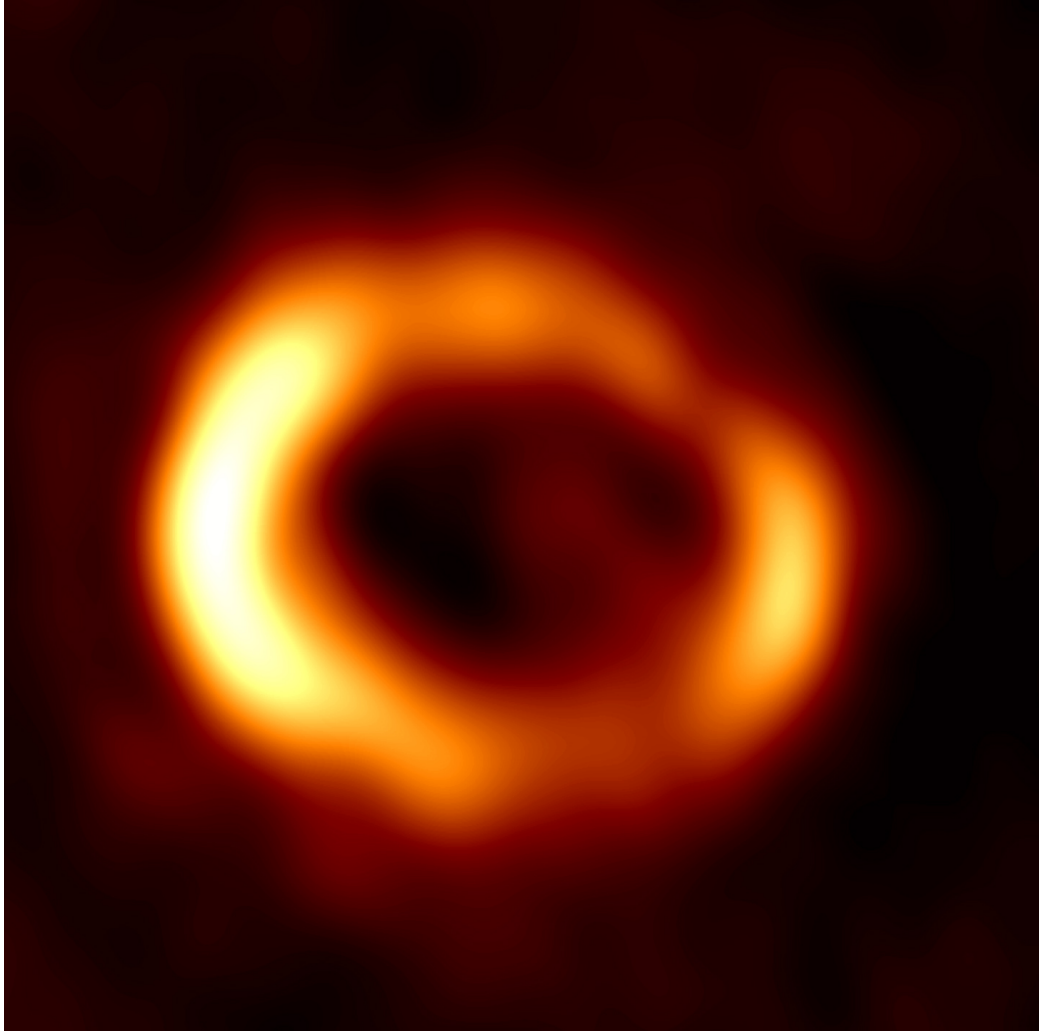


Figure 1.5: **Supernova remnant SN 1987A.** A high frequency (44 GHz) radio image of SN 1987A [77]. The image is false colour graded from white (bright) to red to black (dim). Note that SNRs are dynamic structures, expanding outwards. The SNR is expanding out into the surrounding molecular gas, which is not uniform. The bright arcs (limb-brightened structure) are due to the SNR expanding into denser material, which is hotter, at the locations of the bright arcs.

lation and erosion, and template matching.

Bayesian approaches transform pixel intensity data to a probability map indicating regions where sources are likely to be located (for example: [20]; also see Section 1.1.3 below). Matched filtering involves convolving an image using an expected source profile as a filter (for example: [70]), while multi-scale approaches work by applying a transform to decompose images into different frequencies (for example: [57]).

In the detection step Masias *et. al.* categorised the majority of methods into thresholding (local or adaptive) and local peak search (region growing from peaks in, for example, pixel intensities). The authors also list a small number of “other” techniques (neural networks, watershed transform, contrast radial function, and connected component trees).

The majority of automated source detection algorithms can therefore be described as flood-filling or region-growing driven by (possibly transformed) pixel intensities [46].

Such intensity based thresholding algorithms often require a parameter to be set (either manually or automatically as part of the algorithm) that restricts the sources to be found as those that are at least as bright as some threshold: typically 3σ or 5σ above root mean square (rms) noise. However, as noted earlier in this chapter, some of the most scientifically important objects in astronomy are dim, low surface brightness sources, with intensities in the range of background noise [53]. Restricting the search for sources that contain pixels above a particular threshold restricts the set of sources that may be found. This is due to the nature of thresholding algorithms: if the search is not restricted in this way, far too many background regions will be identified as “sources”, and the output will be unusable.

Threshold-based algorithms are often reported to have very high rates of true detections with very low rates of false detections and missed sources.

However, these rates must be taken with caution, as they are often restricted to the sub-set of sources in an image above a particular brightness. For example, Hancock *et. al.* [27] report perfect detection rates with no false detections or missed sources — in the case of simulated point sources 10σ above rms noise.

The source detection algorithms developed in this thesis are evaluated against two source detection software packages commonly used in radio astronomy: Duchamp [74] and BLOBCAT [26]; both of which perform pixel-intensity based thresholding in the transformation step and flood-filling in the detection step.

1.1.3 Bayesian source detection methods

Beside the many thresholding and flood-filling / region-growing algorithms, there are a small number of Bayesian source detection algorithms [46].

Most of these Bayesian source detection algorithms use Markov-chain Monte Carlo sampling (MCMC) to estimate the relative probability that each pixel (or pixel grouping) in an image is a background pixel or a source pixel [20, 46].

Savage and Oliver [63] used MCMC for source detection in infrared astronomical images. Each pixel was labelled with the relative probability of being a background pixel or a source pixel, with a flat and uniform background model, and a source model consisting of background plus a point source modeled as a circularly symmetric Gaussian of known size. This labelling yields a probability map which can be used to find local maxima (subject to some threshold) corresponding to locations of point sources. Once the location of putative sources is found by this method, the nature of the source is identified: MCMC sampling is used to find the best-fit

model of uniform background, point source, or extended source models (where point sources are circularly symmetric and of known size, and extended sources are circularly symmetric with variable size).

Similarly, Hobson and McLachlan [29] performed Bayesian model selection using MCMC to explore the parameters of background and source models, with a source model of circularly symmetric Gaussian-shaped 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. A significant speed up to this technique is presented in Carvalho *et. al.* [9], in which multiple local maximisation of the posterior distribution replaces sampling, and a Gaussian approximation to the posterior is used for the evaluation of Bayesian evidence in model selection. Brewer *et. al.* [7] have successfully scaled a model similar to Hobson and McLachlan's [29] up to ~ 1000 sources.

Guglielmetti *et. al.* [24] developed a Bayesian source detection method for X-ray data. They used a two component mixture model where an astronomical image is assumed to consist of a smooth background with additive source signal; therefore the two models are 1. background plus noise and 2. background plus source plus noise.

Feroz and Hobson [20] developed one of the few non-MCMC based Bayesian source detection algorithms. They used the Bayesian model selection algorithm nested sampling [66], which they argue outperforms MCMC based methods, and is computationally less expensive. Similarly to other Bayesian source detection methods, circularly symmetric Gaussian-shaped sources are assumed.

The Bayesian source detection methods developed in this thesis have some similarities with the existing algorithms in astronomy. For example, in Chapter 3, in latent Dirichlet allocation (LDA), Gibbs sampling is performed on individual pixels which are probabilistically assigned background or source labels, with flood-filling performed on the transformed

image to identify islands of source pixels. In Chapters 4 and 5, a Bayesian Dirichlet-multinomial ratio (“DMR”) is derived to iteratively find Gaussian elliptical shaped regions that conform to a source model (or differ from a background model).

In contrast to the existing Bayesian source detection methods however, source and background are modelled as distributions over ranges of pixel intensities. As with the existing Bayesian algorithms, the DMR assumes a parametrised form for astronomical sources and estimates these parameters from a posterior distribution, however sources are not assumed to be circularly symmetric or of a particular size. LDA does not require a rigid model for the size or shape of sources.

It is worth noting that in several cases, these Bayesian source detection methods are tested on simulated data in which the sources are generated with the same parameters as the model of source used in the detection methods. For example [20, 29, 63] generate circularly symmetric Gaussian sources of a particular size, and use a circularly symmetric Gaussian of that size as a source model. Using the same parameters for simulated sources and the method’s source model is likely to artificially inflate the success of the model, and is unjustified in the case of radio astronomy data, where sources can take a great variety of shapes and sizes.

1.2 Outline of the thesis

Within the framework provided by Masias *et. al.* [46], the source detection algorithms in this thesis can be characterised as follows:

- **Image transformation step:** the continuous valued pixel intensities in astronomical images are converted to discrete values by histogram binning or discretisation. This process is described in **Chapter 2**.

- **Detection step:** two approaches are taken:
 - Background and source models are inferred using the topic modelling technique latent Dirichlet allocation, these models are used to classify pixels according to which model they were most likely generated by. Flood-filling is then performed to find contiguous regions of “likely to be source” pixels. The use of LDA for source detection is described in **Chapter 3**. Results of this technique as compared to results obtained by a standard source-detection package are presented.
 - A Dirichlet-multinomial ratio (DMR) is derived, and, given a background model and a loosely specified source model, gradient ascent is performed to find peaks in the ratio indicating regions that are unlikely to be background. The derivation of the DMR is outlined in **Chapter 4**. Results of the DMR on real and simulated data are presented in **Chapter 5**.

A combined LDA-DMR technique is described and evaluated in **Chapter 6**. The work in the thesis and its limitations are summarised and possible future work suggested in **Chapter 7**.

1.3 Original contributions

The contributions of this thesis are as follows:

- A novel method for incorporating uncertainty about where bin borders should be located in discretised data is introduced in **Chapter 2**.
- The first use of the topic modelling technique latent Dirichlet allocation (developed to extract latent topics from collections of documents) in source detection in astronomical images is described in

Chapter 3. This application of LDA to source detection differs from that taken in previous approaches to the segmentation of non-astronomical images in that it relies only on pixel intensity and location, whereas previous approaches employ techniques to extract image interest points and pre-segment images before applying LDA.

- A new Dirichlet-multinomial ratio, indicating how well a region conforms to a model of background versus a model of foreground is derived in **Chapter 4**.
- LDA output is utilised in the DMR in **Chapter 6**.

Chapter 2

Discretising images

Astronomical images have continuous-valued pixel intensities. These continuous values can be converted to discrete values — “discretised” — by histogram binning.

Histogram binning is a method of mapping a large number of continuous values (in this case, pixel intensities) to a smaller set of discrete values (or “bins”), where all values within a range are assigned a discrete label representing that range [76].

Discretisation of continuous pixel intensity values makes it possible to model the data according to a multinomial distribution, and to use the Dirichlet distribution, which is the conjugate prior for the multinomial distribution [12, 21, 33, 40]. This allows the use of the topic modelling technique latent Dirichlet allocation [6] (LDA; Chapter 3) and the derivation of a Dirichlet-multinomial ratio (DMR; Chapters 4 and 5). LDA and the DMR are combined in Chapter 6¹.

¹Note that a compound Dirichlet-multinomial distribution is used in both LDA and the DMR. In both cases the multinomial parameter is integrated out, leaving only the data and the Dirichlet hyperparameter. LDA estimates a conditional distribution while the DMR calculates a marginal joint distribution [6, 51].

Discretisation is a simplification and reduction in the complexity of the information in the data. It is therefore important that the process of discretisation does not destroy information essential to the task at hand. In the case of source finding in radio astronomy images, it is important that information that can be used to discriminate source pixels from background pixels is not lost. Though discriminative ability may be decreased by discretisation, an ideal discretisation retains or improves discriminative accuracy [39, 42].

In this chapter, simple data partitioning methods for defining histogram bins for discretisation are described in Section 2.1. A novel way of “softening” bin borders to incorporate uncertainty about the location of partition-points in the data is described in Section 2.2. An exploratory approach to source finding in discretised images is described in Section 2.3.

2.1 Histogram binning strategies

Histogram binning is the simplest of a range of discretisation methods [39, 42, 76].

The majority of existing discretisation methods are supervised (using a training set of instances with known class labels) [39, 42, 76]; these methods are not appropriate for radio astronomy images as the class (source or background) of an image’s pixels is not known *a priori*.

Of the few unsupervised algorithms, cluster-based [10] and dynamic-qualitative [43] — both of which seek to reduce within-interval variation and maximise between-interval variation — would be appropriate, but, given the time-consuming nature of these algorithms (particularly relative to histogram binning), neither is practical [76].

A histogram of pixel intensities can be constructed by assigning each pixel

in an image to an interval or “bin” on the basis of its intensity, where bins are ranges of intensity values defined² as $[min_k, max_k)$. A pixel x falls into bin k if $min_k \leq x < max_k$ (except in the case of the last/top bin K , which is defined as $[min_K, max_K]$: a pixel x is assigned bin K if $min_K \leq x \leq max_K$). For adjacent ranges $[min_i, max_i)$ and $[min_j, max_j)$, $max_i \leq min_j$ and so each data instance falls into exactly one bin; coverage is ensured by setting $min_j = max_i$.

In order to discretise intensity values by histogram binning, bin borders (values that define min_k and max_k for each bin $k \in K$) must be defined. The number of bins must also be set (either manually, or by the binning method itself).

A simple way to create bins in data is to partition the data on the basis of either interval-width or bin-occupancy. These methods are described below.

2.1.1 Width-based partitions

Width-based methods divide the number-line between x_{min} and x_{max} where x_{min} is the minimum pixel intensity value in the data, and x_{max} is the maximum value [76].

Equal width partitions

One of the simplest ways to partition data into K bins is to create K equal width partitions in the data range $[x_{min}, x_{max}]$. The width of intervals is defined as $w = (x_{max} - x_{min})/K$, and the bins are defined at $[x_{min}, x_{min} + w)$, $[x_{min} + w, x_{min} + 2w)$, ..., $[x_{min} + (K - 1)w, x_{max}]$. This is an equal partitioning of the number-line between x_{min} and x_{max} [76].

²Note that some authors use the definition $(min_k, max_k]$, i.e., $x \in k$ if $min_k < x \leq max_k$ [39].

Equal width discretisation of non-uniformly distributed data may result in an inaccurate representation of the data; for example, outliers may skew the histogram binning [14, 39, 42].

In the case of radio astronomy data, the vast majority of an image's pixels are typically within a small range at the bottom of end of the data range, with a few, relatively small, peaks at the top of the range (see Figure 2.1). This means that an equal width binning scheme will produce many empty bins (which means wasted representation power) and, more importantly, a lack of resolution at the low end of the intensity range. Because many scientifically interesting objects have pixel intensities close to background intensities, equal width binning strategies risk losing information that differentiates background from source pixels. In the case of radio astronomy data, some of these effects may be alleviated with an iterative binning process that removes bright source pixels (outliers) from the histogram binning process once the sources containing these pixels have been identified (for an example, see Chapter 5).

Exponentially increasing width partitions

This problem of including source pixels and background pixels in the same bin(s) may be addressed by exponentially increasing the interval-widths in a partition.

In this method, intervals are defined recursively with interval k having width $0 < \gamma < 1$ times less wide than interval $k + 1$. The last (top) interval wid_k is defined as $\gamma(x_{\max} - x_{\min})$, with bin $K - 1$ having width γwid_k . For this partitioning method, the number of bins K will be defined by the parameter γ and the range of the data.

Depending on the γ value used, there may still not be sufficient resolution in the low values to differentiate source from background in radio astronomy images.

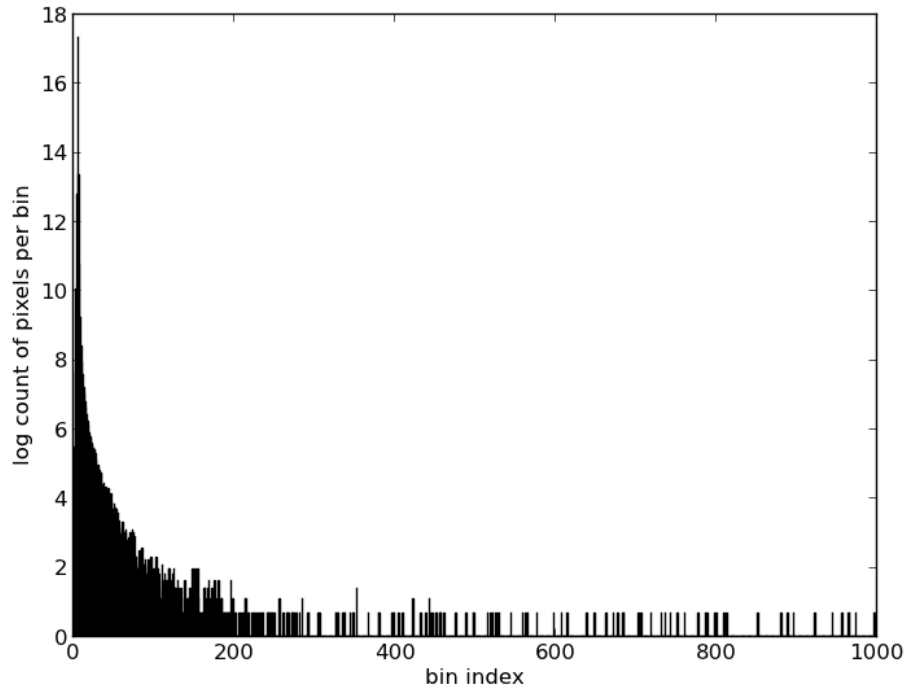


Figure 2.1: **A histogram of pixel intensities in an astronomical image.** Pixel intensities from the Australia Telescope Large Area Survey Chandra Deep Field-South (ATLAS CDFS) [54], with the **log** count in each of 1000 equal width bins. Note the large peak in the low intensities and the outliers in the high intensities.

2.1.2 Frequency-based partitions

Alternatively, bins may be defined according to their occupancy. In contrast to width-based methods that partition the number line, frequency-based partitions rank the data and make partitions on this ranking [76].

Equal occupancy partitions

Equal occupancy partitions are constructed by dividing the count of pixels N by K bins, to determine the number of pixels n that should fall into each bin k . Pixels are allocated to bins by ranking pixels by intensity and partitioning this ranking into K partitions of (approximately, depending on whether N is divisible by K) n data points each. The minimum and maximum values in the partition defining the bin boundaries $[min_k, max_k)$ are those of the first and last ranked data points within the partition.

In contrast to width-based partitions, there are no empty bins in equal occupancy partitions (no “wasted” bins). With respect to astronomical data, low intensity source pixels are less likely to be subsumed into a bin with mostly background pixels (though this is not guaranteed). However, information with respect to the number line, such as bright “peaks” in the data, is lost, as the resulting distribution of data into bins is not representative of the distribution of the original, non-uniform, data [11].

Exponentially decreasing occupancies

An exponentially decreasing partitioning of data with respect to occupancy divides the pixel ranking such that each bin k has $0 < \gamma < 1$ times as many pixels as bin $k - 1$.

The first (bottom) bin occupancy occ_1 is created by assigning it the first γN -many pixels (where N is the total number of pixels in the data) and

defining the minimum and maximum values for bin 1 as the intensities of the first ranked and last ranked pixels within $[min_1, max_1)$ respectively. The rest of the bins are defined recursively (bin 2 has the next γocc_1 -many pixels in the ranking, and so on).

2.2 A novel method for softening bin borders

There are three major problems with the simple discretisation/binning methods described in Section 2.1 for radio astronomy data:

1. 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;
2. 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 [11]; and
3. 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 [75].

For the source-detection algorithms used in this thesis, bin identities are used to convert pixels in an image or a region in that image into counts in bins. Given that there is a lack of any informed basis for the choice of bin boundaries used, and the problems inherent in binning non-uniform data, 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 where each pixel is allocated to one bin only, it is possible to create several sets of bins $b \in B$,

each of size K , with bin sets b_i and b_j having differing partition points. Each pixel would then have a number of “possible” bin assignments across sets of bins (each pixel x will fall into one bin k for each set of bins b).

Given a number of plausible sets of bins B , it is possible to count how often a given pixel x appears in each bin $k \in B$. Once normalised (by dividing the counts in each bin by the number of $b \in B$, so that the sum of counts for each pixel x across bins is equal to 1) this gives a *distribution* over all the bins for x , which reflects 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) in a way that is smooth. It reduces the impact both of the uninformed choice of bins and of any within-class splitting over bins and between-class combining in bins [11]³.

2.2.1 Creating sets of bins with the Dirichlet distribution

The Dirichlet distribution is a natural way to create such sets of bins. In this section, the Dirichlet distribution is described, followed by an explanation of how it may be used to create sets of bins.

³Note that there are existing techniques for transforming continuous attributes to distributions over intervals/bins. Ishibuchi *et. al.* [31] use labelled data and fuzzy logic to create linguistic association rules; Yang and Webb [75] create overlapping, rather than disjoint, intervals constructed such that a given data-point falls near the middle of its interval, away from interval boundaries. In contrast, the approach developed in this chapter directly addresses uncertainty about where those boundaries lie.

The Dirichlet distribution

Informally, a Dirichlet distribution can be thought of as a distribution over categorical or multinomial probability mass functions (pmfs) of size K ; or, in other words, a distribution over the $(K - 1)$ dimensional probability simplex [21, 51].

Firgyik *et. al.* [21] illustrate the intuition behind the Dirichlet distribution using dice. A die can be thought of as a pmf with $K = 6$. A given die may be fair (with equal probability of rolling each number) or it may have imperfections that result in a biased or weighted die. A Dirichlet distribution can be thought of as a bag of dice, from which dice may be drawn. The bag may contain dice that are all reasonably fair, biased in some particular manner, or some mixture.

A Dirichlet distribution has a parameter $\alpha = \alpha_1, \dots, \alpha_K$ and probability density function on the $(K - 1)$ dimensional simplex:

$$f(p_1, \dots, p_{K-1}; \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K p_i^{\alpha_i-1} \quad (2.1)$$

where $\Gamma(\cdot)$ denotes the gamma function; and with $p_i \in [0, 1]$, $\sum_{i=1}^{K-1} p_i < 1$, and $p_K = 1 - p_1 - \dots - p_{K-1}$ (this parametrisation of the Dirichlet probability density function is due to the fact that the probability density is positive only on the $(K - 1)$ dimensional simplex and zero elsewhere; the simplex exists in K dimensional space) [21, 28, 33, 40, 48, 51].

The α parameter controls where the density on the probability simplex lies. In the special case when $\alpha_i = \alpha_j = 1 \ \forall i, j \in K$, the density is uniformly distributed over the simplex. Every multinomial distribution over K is equally likely to be drawn from the Dirichlet distribution with this α parameter. The Dirichlet with uniform α parameter $\alpha_i = \alpha_j < 1$ pushes the density to the corners of the simplex; multinomial distributions with one p_k much greater than all the others (with $(1 - p_k) \ll 1$ and

$p_i \ll 1 \forall i \neq k$) are typical of the Dirichlet with this α parameter. When $\alpha_i = \alpha_j > 1$, the density lies towards the middle of the simplex. Distributions drawn from the Dirichlet with $\alpha_i = \alpha_j \gg 1$ approximate uniform distributions over K . When α parameters are not uniform, the density on the simplex is not symmetric. For non-uniform α parameters with values > 1 , if $\alpha_i > \alpha_j$, distributions are likely to be drawn with $p_i > p_j$ [21]. Figure 2.2 illustrates these cases.

Interpreting Dirichlet samples as bin borders

A single sample from a Dirichlet distribution with an α -vector of K components yields a K -element vector \mathbf{p} for which $\sum_{k=1}^K p_k = 1$ and all $p_k > 0$ (such as $\mathbf{p} = [0.1, 0.7, 0.2]$ with $K = 3$). The cumulative sum of the elements of \mathbf{p} , denote \mathbf{p}_d , can be interpreted as partition-points over $[0, 1]$ (such as $\mathbf{p}_d = [0.1, 0.8, 1.0]$).

To translate \mathbf{p}_d into bins, these partitions can either be interpreted over the number-line between minimum and maximum pixel values $[x_{min}, x_{max}]$ or over a ranking of pixels, as described in Sections 2.1.1 and 2.1.2.

A Dirichlet distribution with symmetric α -vector with $\alpha_i = \alpha_j > 1 \forall i, j$ will produce roughly even values for all $p_k \in \mathbf{p}$ (and the larger the α values, the more uniform the distributions drawn). Draws from a Dirichlet distribution with such an α -vector can be used as partition points in a set of roughly equal width or roughly equal-occupancy bins.

To produce increasing or decreasing bins (width or occupancy), a “stretch factor” β can be applied to the α values, such that $\alpha_j = \beta\alpha_i$, similar to the derivation of exponentially increasing width bin borders (Section 2.1.1) and exponentially decreasing occupancy bin borders (Section 2.1.2).

Figure 2.3 shows examples of bin borders produced by Dirichlet distributions with symmetric and asymmetric α -vectors.

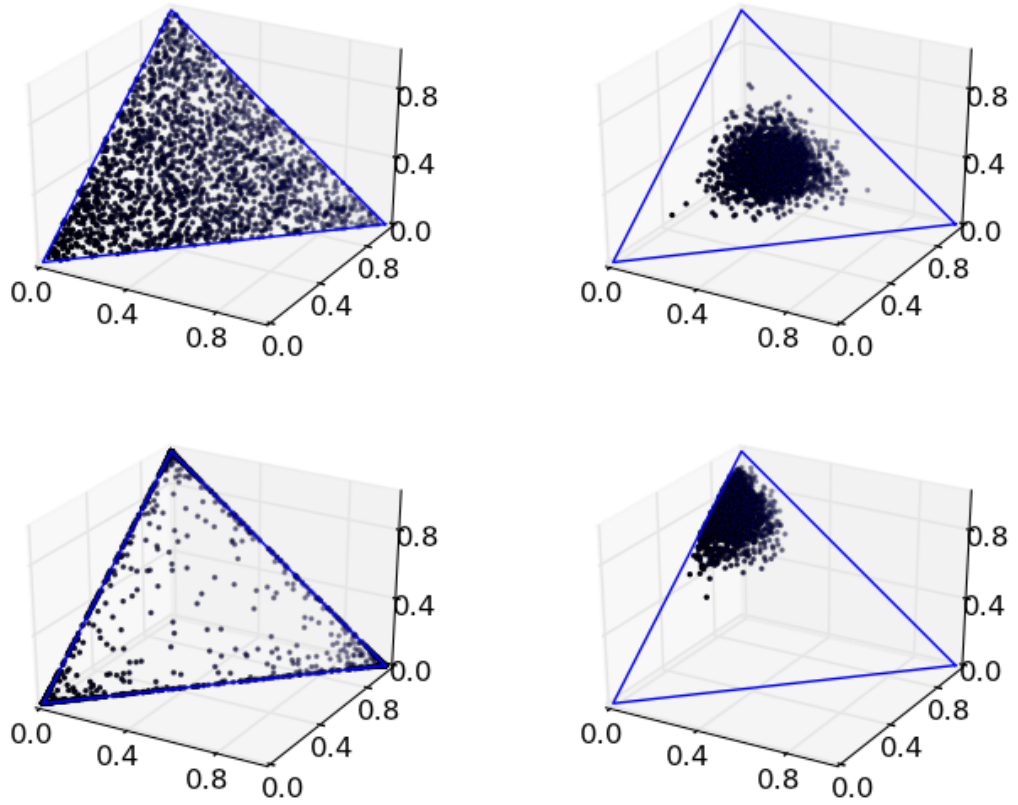


Figure 2.2: **Dirichlet samples on the simplex.** α -vectors: $[1, 1, 1]$ (top left), $[10, 10, 10]$ (top right), $[0.1, 0.1, 0.1]$ (bottom left), $[2, 5, 20]$ (bottom right).

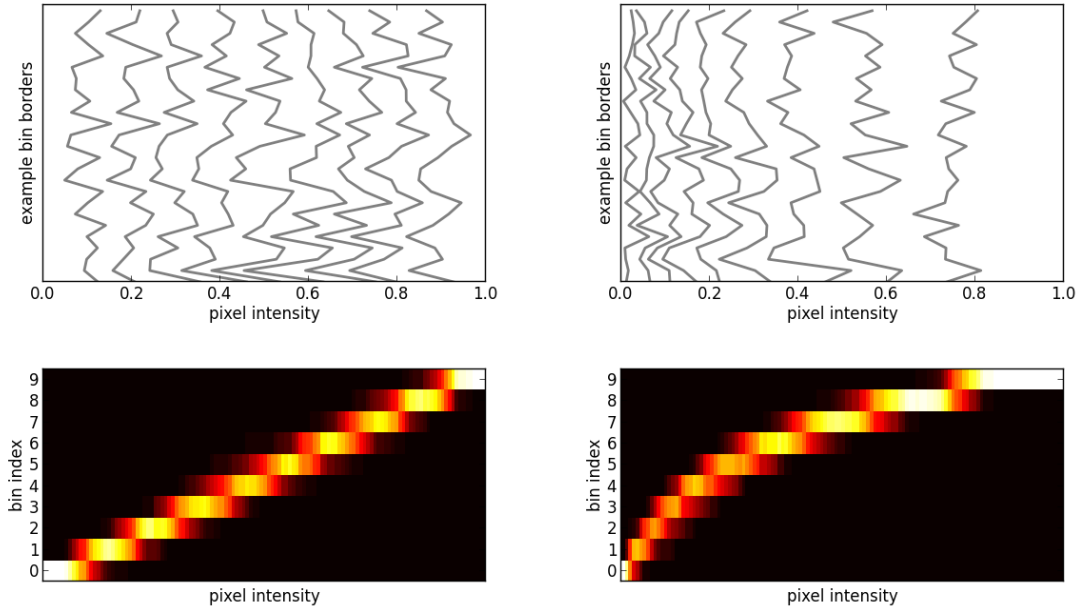


Figure 2.3: **Dirichlet bin borders.** Sets of bin borders created by sampling from a Dirichlet distribution. The borders on the left are generated from a Dirichlet with symmetric α -vector with $\alpha_i = 10 \forall i$. On the right, $\alpha_{i+1} = \beta \times \alpha_i$, with the stretch factor $\beta = 1.3$. Their sum $A = \sum_i \alpha_i$ is set to 100 in both cases. In the top-most plots, each row shows an independent draw from the Dirichlet with this α -vector. The bottom plots show the distribution of pixel intensities across bins for a set of Dirichlet bins, using a colour map from black (zero) to red (low) to white (high). For example in the plots at right, a pixel intensity of 0.8 falls into bin 9 in most cases, but sometimes falls into bin 8.

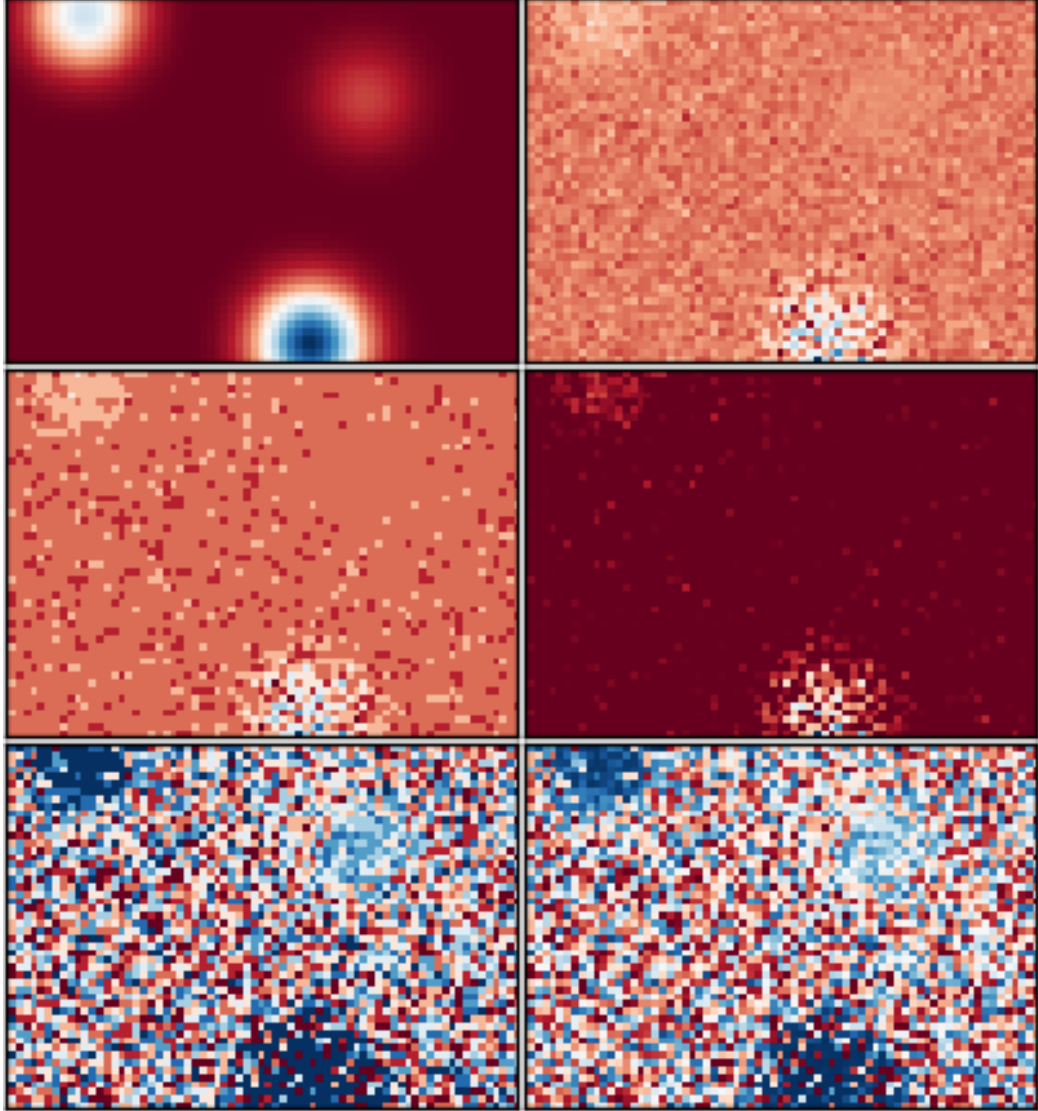


Figure 2.4: **Different views of an image with different binning strategies.** The image is shown at top, created by generating three sources (at left), and then adding zero-mean Gaussian noise (right). The next four images show the image discretised by four different histogram binning methods: equal width (middle left), Dirichlet equal width (middle right), equal occupancy (bottom left), and Dirichlet equal occupancy (bottom right). $K = 10$ bins for all four binning strategies. For the Dirichlet bins, $\alpha_i = \alpha_j = 10 \forall i, j \in K$.

2.3 Source detection in discretised images: exploration

Within the framework of discretisation of continuous-valued images by histogram binning,⁴ an exploratory approach to classifying sub-windows of images as source or background was implemented.

In this approach, the distribution of pixel intensities in an image or window (sub-image) was modelled as a multinomial distribution [40, 72] across bins.

The task of finding sources in a radio astronomy image was framed as finding and identifying background, with sources defined by proxy as non-background regions. This approach exploits the fact that the variability of image background is much lower than for sources: while there is a great deal of diversity between astronomical objects — every source is different — the background in astronomical images is much more consistent. This approach avoids the difficulties inherent in approaches based on comparing “typical” exemplars of source and background⁵.

The following exploratory algorithm was implemented:

- An equal width binning of the overall image was created, that is, equal width bins were created by dividing the pixel intensity range into a set of equally spaced intervals b_K (as described in Section 2.1.1). Given that radio astronomy images are dominated by back-

⁴The quantisation of intensities in radio astronomy images was first described by Gull and Daniell; see: [25].

⁵Note that previous approaches have framed the source detection problem as one of identifying and removing background. For example, the software package SExtractor [2] approximates an image’s background by a low-order polynomial, as a function of the mean and median pixel intensity. The background is then subtracted from the image and flood-filling performed to find islands of connected pixels. This background subtraction is likely however to remove dim sources along with background [9].

ground pixels, the counts in b_K for the whole image (denoted M^B) was used as a **model for background**.

- Using a square window of size N pixels, the image was iterated across by moving the window from the top-left to the bottom-right corner in steps of y pixels.
- For each such window, the distribution of pixel intensities in the bins b_K (denoted X) was compared to the overall distribution M^B , returning the likelihood of the window's pixel intensities X , under the background model M^B .
- The image was output with the 10 percent least likely windows under the background model highlighted.

With the equal width binning of the overall image as the background model, the probability p_k of a pixel's intensity falling within bin $k \in [1..K]$ (where $\sum_{k=1}^K p_k = 1$) is the normalised counts in the bins of the histogram of pixel intensities in the overall image.

Given this model M^B , the likelihood of pixel intensities for each window (where x_k is the count of pixels in the window that fall into bin k ; with a total of N pixels in the window) is calculated using the probability mass function of the multinomial distribution. That is,

$$\begin{aligned} f(x_1, \dots, x_K; N, p_1, \dots, p_K) &= Pr(X_1 = x_1, \dots, X_K = x_K) \\ &= \frac{N!}{x_1! \dots x_K!} p_1^{x_1} \dots p_K^{x_K} \end{aligned} \quad (2.2)$$

where

$$\sum_{i=1}^K x_i = N$$

The log likelihood was calculated to avoid working with extremely small

numbers:

$$\begin{aligned} \log Pr(X_1 = x_1, \dots, X_K = x_K) &= \log(N!) - \log\left(\prod_{k=1}^K x_k!\right) + \log\prod_{k=1}^K p_k^{x_k} \\ &= \sum_{n=1}^N \log(n) - \sum_{k=1}^K \sum_{m=1}^{x_k} \log(m) + \sum_{k=1}^K x_k \log(p_k) \end{aligned} \quad (2.3)$$

where x_k is the count in the $k^{th} \in [1..K]$ bin for a given window, and p_k is the background model probability of a pixel falling into the window (the normalised count in that bin for whole image).

In practice, a look-up table of cumulative sums of logs of integers $[1..N]$ can be constructed; then the first term in Equation 2.3 is the N^{th} entry in the table; the second term is a sum of entries in the table (one per bin-count). Because the window-size does not vary as it moves across the image, the first term in Equation 2.3 is a constant and so can be dropped from the calculation.

The third term in Equation 2.3 is the dot product of two vectors: the counts in each bin x_k for the window and the log probabilities for each bin p_k (the normalised counts in each bin for the overall image).

It is important to note that one of the assumptions of the multinomial distribution is violated; the counts in each bin are not independent because the intensity of a particular pixel is not independent of its neighbours [40, 72]. Pixels are not randomly distributed across an image, but are found together as “sources” and “background” in an image.

Figure 2.5 shows some sample output for this method, illustrating that the simple multinomial model explored in this chapter correctly identifies regions containing sources as unlikely to be background regions. This indicates that:

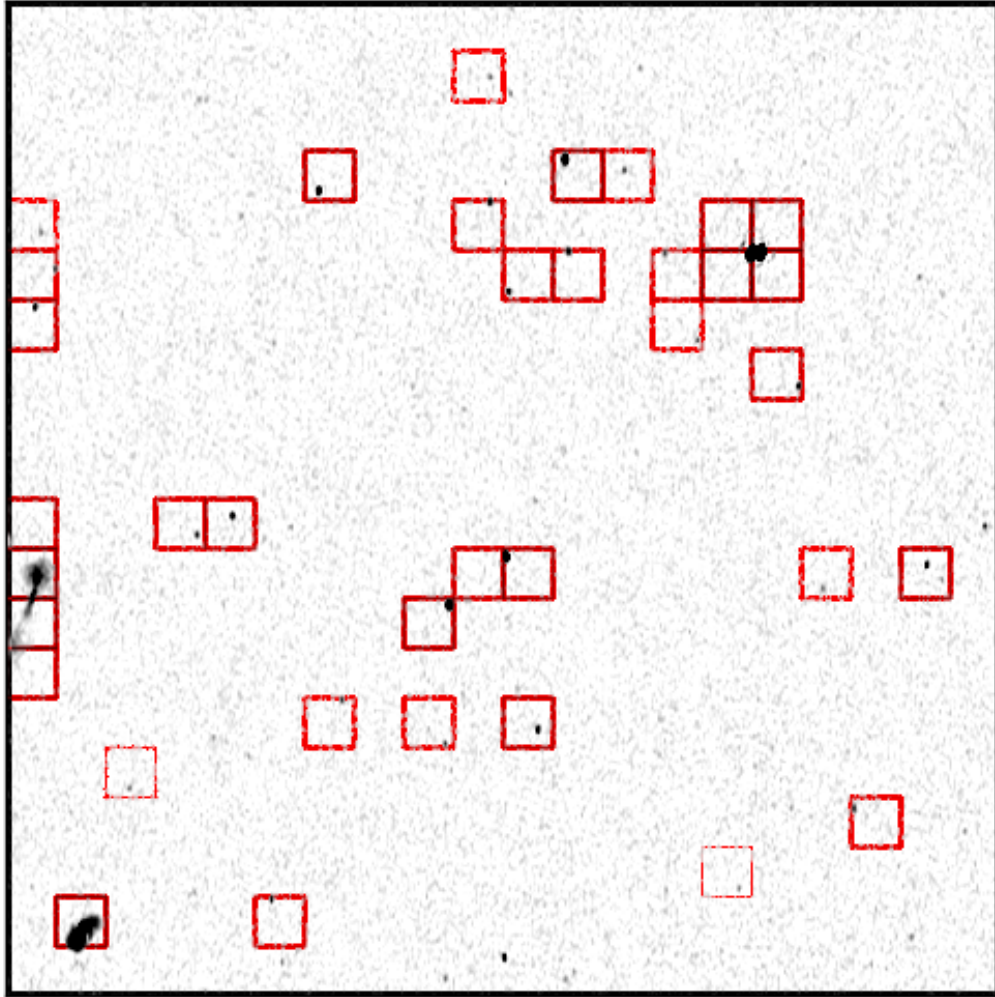


Figure 2.5: **Sources found by the exploratory multinomial model.** A small section of a radio astronomy image (from the Extended Chandra Deep Field South [47], contrast-adjusted to show sources) showing the 10 per cent least likely windows under the multinomial model with the equal width histogram binning of the overall image as the background model M^B . The window size is 50 and the step size is 50 (that is, there is no overlap between windows). The windows unlikely to be background under the model M^B are shown by red bounding boxes. The transparency of the each window's borders is scaled according to how unlikely the window is: more unlikely windows are given less transparent borders and vice-versa.

1. discretisation via histogram binning preserves enough information to locate sources in astronomical images (at least in this example);
2. a multinomial distribution is a reasonable model for the distribution of pixels in an image; and
3. using an entire astronomical image — including both source and background pixels — as a model for background does not prevent regions in which sources are located from being correctly found.

For these reasons, the exploratory approach was extended, as described in the next section.

2.4 Extending the exploratory approach

The ideas developed in this chapter — discretisation via histogram binning and the exploratory approach to source detection using a multinomial model of background — are extended in the remainder of this thesis.

In Chapter 3, multinomial models for background and source are inferred via Gibbs sampling on a conditional distribution of a compound Dirichlet-multinomial generative model described by latent Dirichlet allocation [6]. These models are used to segment images into background and source regions.

In the exploratory multinomial approach in this chapter, the observation that images are primarily made up of background pixels motivated the use of the histogram binning of a whole image as a proxy for background. In this framework, the task is primarily that of identifying and removing background; sources are defined as “non-background” (or regions that are unlikely to be background).

In Chapters 4 and 5, whole images are again used as a model of background. A likelihood of how well a region conforms to a model of background is derived, and peaks in the likelihood are found using gradient ascent. In contrast to the exploratory approach in this chapter however, the binned image is used as pseudocounts in an α -vector of the Dirichlet hyperparameter of the marginal joint of a compound Dirichlet-multinomial distribution, rather than directly as a multinomial distribution.

Both latent Dirichlet allocation (Chapter 3) and the Dirichlet-multinomial ratio (Chapters 4 and 5) are performed on discretised images. Histogram binning methods, both “hard” and “softened” as described in this chapter, are used for this discretisation.

Chapter 3

Latent Dirichlet allocation

In this chapter, the topic modelling technique latent Dirichlet allocation [6] is used to infer multinomial models for background and source in discretised images. These models are used to segment the images, in order to identify sources.

3.1 Problem framework

As described in Chapter 1, radio astronomy images can be thought of as primarily background with an unknown number of point and spatially extended sources. Identifying the sources requires distinguishing them from background, a task made difficult by the diversity within and between sources. The variability of background is lower than that of sources and in that sense it is easier to identify.

Source detection may be approached as a problem of identifying and excluding regions of background and merging what remains into a modest number of sources. This requires specification of a method for labelling a region as likely or unlikely to be background.

One plausible approach is to assume background is equivalent to “non-signal” with some noise, and this is the basis for many existing algorithms [46]. However, this assumption can fail. Background pixels may not be restricted to a single narrow range of pixel intensities, or to just one such band, and may lie in source intensity ranges [53].

A second possible approach is to use a human domain expert to identify “valid” background regions, and use this to build a probabilistic model for segmentation. However, such manual intervention is problematic given the volumes of data to be produced by next generation telescopes [53], and is vulnerable to biases of the human perceptual system.

In this chapter, a method for producing models of background and source without manual intervention is proposed, where the models are extracted from image data containing both background and sources. This task is nontrivial: individual pixels in an astronomical image are not spatially independent (source pixels are more likely to be found with other source pixels, and similarly, background pixels are more likely to be found together than with source pixels), but regions of the image may contain only background pixels, only source pixels, or an unknown mixture. This motivates the use of the “mixed-membership model” latent Dirichlet allocation (LDA) [6].

Though the problem of finding sources in astronomical images is framed as one of building a good model of background (so as to identify non-background — i.e., source — regions), LDA allows background and source models to be built concurrently [6].

Source detection in radio astronomy images is performed via flood-filling, based on a probabilistic model of pixel intensities inferred by LDA. An additional application of this technique in segmenting grayscale images is presented.

3.2 Latent Dirichlet allocation

3.2.1 The generative model

A “topic model” is a generative model for documents¹ based on latent topics, where topics are modelled as distributions over a vocabulary of words and documents are modelled as mixtures of topics [67].

Informally, a topic can be thought of as a “semantic theme”. The words that have high probability under a particular topic would be expected to convey some sense of meaning [5]. For example, words like “environment”, “conservation”, “sustainable”, and “ecology” found together with high probability in a particular topic might convey a theme of “environmental sustainability”.

Topics are discovered by fitting the generative model to data and finding the best set of latent variables to explain the observed data; for example, the best mixture of topics in a document and distributions of words in a topic [67].

LDA [6] is one such generative probabilistic model for sets of discrete data such as collections of documents, where a document is a multinomial distribution over topics, and topics are multinomial distributions over the vocabulary of words in the collection².

Each document in a collection of documents is represented as a “bag of

¹A “generative model” is a description of a probabilistic procedure for generating documents, used to form a conditional probability density function and infer the latent topics (rather than actually generate documents) [67].

²Note that categorical distributions would be most appropriate for distributions over words in topics and distributions over topics in documents; however the multinomial distribution is used for its convenient properties, including its conjugacy with the Dirichlet distribution (and in fact a K dimensional categorical distribution is equivalent to one trial of a K dimensional multinomial distribution [6, 12, 41]).

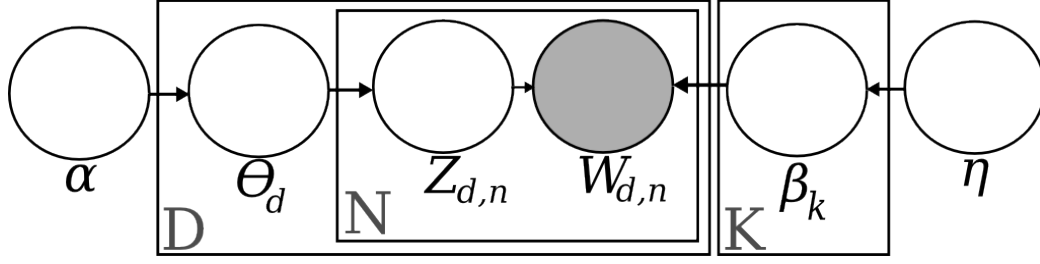


Figure 3.1: **Graphical model for LDA.** Nodes are random variables (shaded are observed; unshaded are latent), directed edges show dependence. Boxes denote replication. Image adapted from [4].

words” in LDA. Under the generative model described by LDA, a document is generated by:

1. drawing topic proportions for that document; and then
2. given the document’s topic proportions, for each word that will be in the document:
 - (a) drawing the topic for that word;
 - (b) generating the actual word by drawing it from the distribution corresponding with its assigned topic [3].

More formally, and with reference to the graphical model in Figure 3.1, the generative model assumes there are K topics, each β_k of which is a multinomial distribution over all of the words in the collection of documents (the collection’s vocabulary). The topic distributions are drawn from a Dirichlet distribution with parameter vector η (a Dirichlet distribution can informally be thought of as a distribution of multinomial distributions [21]).

There are D documents in a collection, each with topic proportions θ_d , a multinomial distribution over topics drawn from a Dirichlet distribution with parameter vector α (where $\theta_{d,k}$ is the topic proportion for the k^{th} topic in the d^{th} document).

The n^{th} word in the d^{th} document is assigned topic $z_{d,n}$ (drawn from the multinomial distribution θ_d), with the *observed* word $w_{d,n}$ drawn from the multinomial topic distribution for topic $\beta_{z_{d,n}} \in (\beta_1, \dots, \beta_K)$ [6].

The generative model corresponds to the joint distribution of latent and observed variables [3, 6]:

$$P(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^K P(\beta_k | \eta) \prod_{d=1}^D P(\theta_d | \alpha) \left(\prod_{n=1}^N P(z_{d,n} | \theta_d) P(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \quad (3.1)$$

To summarise, in the generative model, K topics $\beta_1 \dots \beta_K$ are each drawn from $Dir(\eta)$ (the Dirichlet distribution with parameter η). Each document d is then generated by:

1. drawing $\theta_d \sim Dir(\alpha)$
2. for each word $W_{d,n}$ drawing:
 - (a) topic assignment $Z_{d,n} \sim Mult(\theta_d)$; and
 - (b) generating $W_{d,n} \sim \beta_{Z_{d,n}}$ [3, 6, 67].

3.2.2 Inferring the latent variables

The following few paragraphs outline the derivation of the conditional distribution for inferring the latent variables in LDA, as given by Blei *et al.* (2003) and Steyvers and Griffiths (2007) [6, 67]. The full derivation can be found in these papers, particularly, [6].

The posterior distribution of the latent variables given the observed data and assuming topics $\beta_{1:K}$ are fixed is [3, 6]:

$$P(\theta_{1:D}, z_{1:D} | w_{1:D}, \beta_{1:K}) = \frac{P(\theta | \alpha) \prod_{n=1}^N P(z_n | \theta) P(w_n | z_n, \beta_{1:K})}{\int_{\theta} P(\theta | \alpha) \prod_{n=1}^N \sum_{z=1}^K P(z_n | \theta) P(w_n | z_n, \beta_{1:K})} \quad (3.2)$$

This is intractable to calculate (a “multiple hypergeometric function” [6, 13]). However, the latent topic assignments $z_{d,n}$ can be estimated via Gibbs sampling³ [22] on a conditional distribution over a compound Dirichlet-multinomial distribution with the multinomial parameters θ_d (the per-document topic distributions) and β_k (the per-topic distributions over words) integrated out [6, 67]:

$$P(z_i | z_{-i}, w_{1:N}) \propto P(w_i | \beta_{1:K}) \prod_{k=1}^K \Gamma(n_k(z_{-i})) \quad (3.3)$$

where $n_k(z_{-i})$ is the number of times topic k has been seen in the collection of topic assignments z_{-i} (i.e., all the topic assignments except for assignment z_i).

Integrating out the multinomial parameters in the model leaves the observed words $w_{d,n}$, and the α and η hyperparameter vectors. These hyperparameters may be inferred or simply set with empirically-derived values [67].

The distribution in Equation 3.4 can be iteratively sampled from to infer each latent topic assignment z_i given the observed words w_i in each document d_i in the collection, and all other topic assignments z_{-i} .

$$p(z_i = j | z_{-i}, w_i, d_i) \propto \frac{C_{w_{ij}}^{WT} + \eta}{\sum_{w=1}^W C_{wj}^{WT} + W\eta} \frac{C_{d_{ij}}^{DT} + \alpha}{\sum_{k=1}^K C_{d_{ik}}^{DT} + K\alpha} \quad (3.4)$$

To perform Gibbs sampling, each word in each document in the collection is initially randomly assigned a topic, and two count matrices are created: C^{WT} of topic assignments to each word in the vocabulary (with C_{wj}^{WT} the number of times topic j is assigned to word w in the collection), and C^{DT} of topic assignments per document (with $C_{d_{ik}}^{DT}$ the number of times topic k is assigned in document d_i) [67] (see Figure 3.2).

³An algorithm for sampling from a multivariate probability distribution by iteratively generating an instance of each variable conditioned on all others.

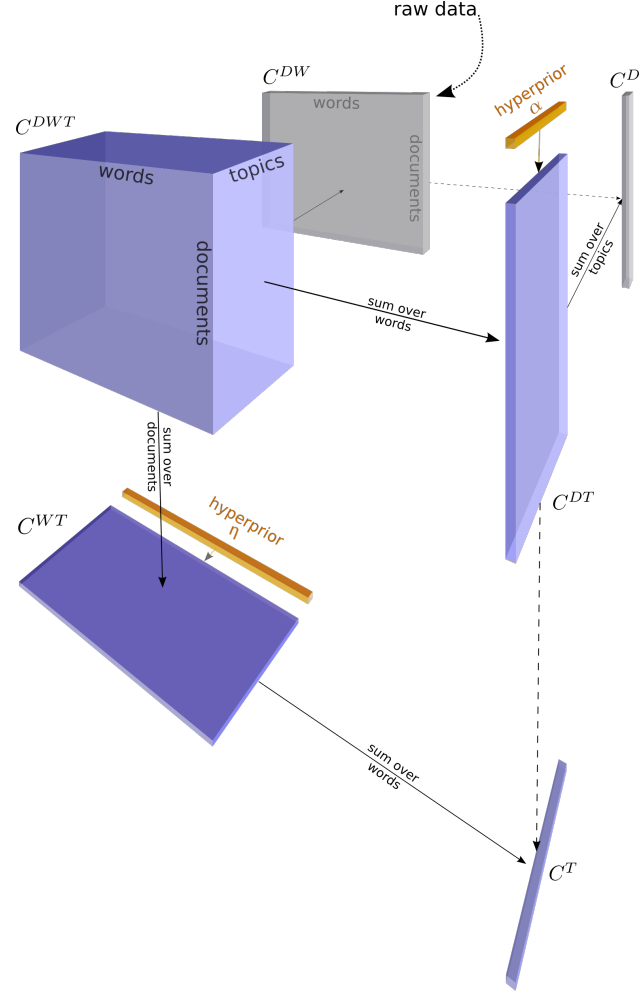


Figure 3.2: **Matrices used in LDA.** Direct counts from the data are grey, inferred ones involving topics are blue, and hyperpriors are yellow. The arrows represent summation if a dimension is being lost, and repeated addition if a dimension is being gained. The C^{DT} and C^{WT} matrices are used in Gibbs sampling to infer the latent topic assignments $z_{d,n}$. The multinomial distributions θ_d (per-document topic proportions) and β_k (distributions of words per topic) can be derived by normalising these matrices respectively [6, 67].

One iteration of Gibbs sampling involves decrementing the matrices at the entry corresponding to each word in the collection in turn, allocating that word a topic from the distribution in Equation 3.4, and incrementing the matrices accordingly [67]. Sampling is run until equilibrium is reached.

The first term in Equation 3.4:

$$\frac{C_{w_{ij}}^{WT} + \eta}{\sum_{w=1}^W C_{wj}^{WT} + W\eta}$$

describes the probability of word w_i under topic j (the number of times word w_i is assigned topic j as a proportion of the number of times any word is assigned topic j). The second term:

$$\frac{C_{d_{ij}}^{DT} + \alpha}{\sum_{k=1}^K C_{d_{ik}}^{DT} + K\alpha}$$

describes the probability of topic j under the current topic distribution in document d_i (the number of times topic j is found in document d_i as a proportion of all topic assignments in document d_i).

The distributions of words per topic, β , and topics per document, θ , can be calculated using the first and second terms respectively [67].

In essence, LDA uncovers latent topics in a document collection, where words that are likely to co-occur in documents in the collection are found together with high probability within a particular topic or topics (weighted by their overall representation in the document collection).

As an example, a collection of documents might have a vocabulary of words ["ball", "game", "win", "film", "actor", "scene"]. The first three words in the vocabulary might be found to occur together in documents with high frequency, but rarely with the last three words (and vice versa). Two topics might be extracted accordingly, a "sports" topic (under which the first three words are highly likely and the latter three unlikely) and,

similarly, a “movie” topic. A document might be primarily made up of words from one topic, or some mixture of both topics (for example, a review of a sports movie).

3.2.3 Application of LDA to images

The following analogy can be made between document collections and images: a single greyscale image is a document collection, which comprises d non-overlapping subimages (the documents). The image “vocabulary” is constructed by taking a histogram of pixel intensities in the entire image, where each of w bins (pixel intensity intervals) is a word in the vocabulary. Any of the binning strategies described in Chapter 2 can be used to construct the histogram binning. The number of occurrences of a word w_i for document d_j is the count of pixels in subimage d_j that fall into bin w_i of the overall image histogram. Topics are normalised distributions over bins.

Using this model, Gibbs sampling (as described in Section 3.2) can be run on greyscale images to uncover latent “topics”: distributions of pixel intensities that commonly co-occur in the image, for example a “background topic”.

These topics can then be used to segment the image on a pixel-by-pixel or region-by-region basis. This can be done by assigning a pixel/region a topic based on the most likely topic to have generated the pixel/region. This can be calculated using the probability mass function of the multinomial distribution (Equation 3.5, where x_i is the count of pixels in the i^{th} bin, $\sum_{i=1}^k x_i = n$, and p_i is the probability of the i^{th} bin under a particular topic).

$$\Pr(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \quad (3.5)$$

For source detection in radio astronomy images, flood-filling⁴ can then be performed on the segmented image to identify the location and size of sources in the image.

Related work in image segmentation

This application of LDA to source-detection in images differs from the approach taken in [8, 18, 19, 59, 61, 64, 65, 69, 71], in which derivations of LDA and other topic models are applied to image segmentation and object and scene classification tasks. The approach in this chapter relies only on pixel intensity and location, whereas previous approaches employ techniques to extract image interest points and pre-segment images before applying LDA.

A document is a single image in a collection in [18], represented by counts of visual words or “textons” extracted from the image collection. This model is extended in [8], in which each image is segmented into homogeneous regions, and each region comprises several interest point patches. Similarly, Russell *et. al.* [61] compute multiple segmentations on each image in a collection, and compute visual words (interest points) for each candidate segmentation. Interest points are similarly employed in “bag of words” models in [19, 59, 64, 65, 69]. Wang and Grimson [71] added spatial information to their LDA model by grouping visual words close to each other in an image.

In comparison to other flood-filling based astronomical source detection algorithms [46], the use of LDA-derived probabilities as a precursor to flood-filling is more powerful than many thresholding algorithms (such as those discussed in [23]); in LDA, commonly co-occurring bins needn’t be adjacent intensity ranges. Consider an image in which the background is made up of medium-intensity pixels while foreground objects comprise

⁴Labelling contiguous regions of pixels that have same topic label as a single region.

Table 3.1: Astronomical images used for evaluation

Image	Description	Source
A1	ATLSB ⁵ survey region A at 50'' resolution	[62, 68]
A2	ATLSB survey region A at 6'' resolution	[62, 68]
B1	ATLSB survey region B at 50'' resolution	[62, 68]
B2	ATLSB survey region B at 6'' resolution	[62, 68]
C	ATLAS CDFS ⁶	[54]

both dark and bright pixels. LDA allows the topics to reflect this, with medium-intensity bins found in one topic, and bright and dark bins in the other.

3.3 Methods

LDA was performed for segmentation and source detection in five radio astronomy images (Table 3.1) Non-astronomical greyscale images were segmented as a demonstration of this application of LDA (see Figures 3.3 and 3.4).

Astronomical images were in FITS format [73]; non-astronomical images were greyscale JPEG images.

For each image a histogram of pixel intensities was generated. An equal width binning strategy was employed with 100 or 1000 bins for astronomical images, and 10 or 100 bins for non-astronomical images.

Images were decomposed into subimages ("documents"), and counts of pixels in bins were calculated for each. A range of subimage sizes was trialled for each image.

⁵Australia Telescope Low Surface Brightness.

⁶Australia Telescope Large Area Survey Chandra Deep Field-South.

Gibbs sampling was run to infer per-word topic assignments z_i , on the distribution in Equation (3.4), as described in section 3.2. The α and η -vectors were set with the empirically derived values $\alpha_i = \alpha_j = 0.1 \ \forall i, j$; $\eta_i = \eta_j = 0.01 \ \forall i, j$ [67].

Gibbs sampling was run for 100 iterations (where each iteration was a sweep through every pixel in the image, allocating each pixel a topic based on the current distributions of words to topics and topics to documents). The final distributions of words for each topic (β_k for $k \in \{\beta_1 .. \beta_K\}$) were calculated from the hundredth sample. An average over samples was not taken as it was found that the sampler converged quickly after which the topic distributions changed very little if at all: sample 1000 was virtually identical to sample 500 and sample 100 (that is, the posterior distribution was very narrow, i.e., the model parameters were very well-constrained).

To segment the images using the inferred topic distribution, each pixel in the image was assigned the topic that it was most likely generated by using Equation (3.5). When considering a single pixel, this equation simplifies to just p_i for a given bin i . That is, if the pixel being considered falls into bin i in the overall pixel intensity histogram, the topic with the greatest probability for bin i is assigned⁷.

In the case of the astronomical images, the performance of LDA was compared with the thresholding and region growing astronomical source detection software *Duchamp* [74], on source catalogues generated by manual inspection by an astronomer (a “ground truth” reference) [62, 68].

Precision and recall were calculated in order to evaluate the performance of the LDA. Precision is the proportion of true sources of all found sources

⁷Note that this may be weighted by the topic’s overall proportion in the collection, however this was not done here.

(that is, all found sources that are really sources):

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad (3.6)$$

and recall is the proportion of found sources of all true sources in the image (all true sources that are found):

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (3.7)$$

where **tp** = “true positive”, **fp** = “false positive”, and **fn** = “false negative”⁸ [56].

In the case of non-astronomical images, two examples of segmented images are shown for illustrative purposes only (see Figures 3.3 and 3.4).

3.4 Results

Figure 3.5 demonstrates the segmentation of an astronomical image (Image A1), and the results of flood-filling on the segmented image to identify radio sources in the image.

For the astronomical images, the results of both LDA and *Duchamp* were compared to a source catalogue generated by manual inspection by an astronomer. Although LDA and *Duchamp* perform roughly equivalently with respect to spatially extended, multi-component sources, LDA had more false positives and false negatives than *Duchamp* (see Figure 3.7 for an example).

Table 3.2 shows the performance of LDA and *Duchamp* as compared to source catalogues generated via manual inspection by an astronomer on

⁸Precision and recall are sometimes called “completeness” and “reliability” in the astronomical source detection literature [27].



Figure 3.3: **Image segmentation.** Each pixel is assigned to the topic it was most likely generated by, as inferred by LDA (for illustrative purposes, each topic is assigned a greyscale value, and each region is coloured according to its topic). Left to right, top to bottom: the original image, followed by the segmented image with two, three, four, five, and six topics. Increasing the number of topics may increase the level of detail revealed by segmentation, but may also introduce spurious segmentations. Original image (top left) from [45].

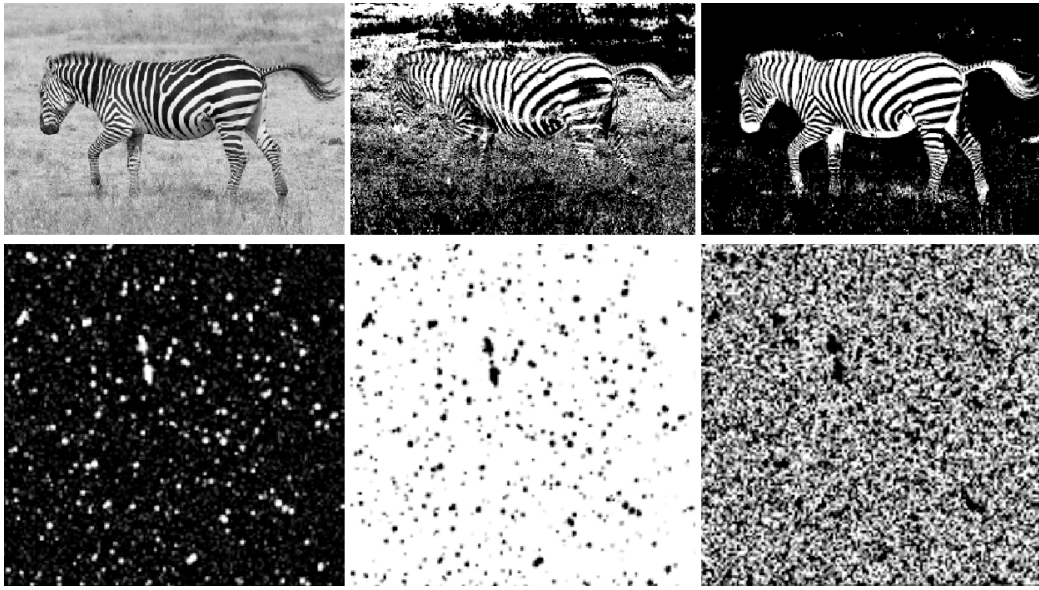


Figure 3.4: **The number of bins in the pixel intensity histogram can affect results.** Top row from left to right: a greyscale JPEG image, the segmented image using 10 bins, and 100 bins. Bottom row from left to right: an astronomical image (with contrast adjusted to see sources), the segmented image using 100 bins, and 1000 bins. Image sources: [37, 62, 68].

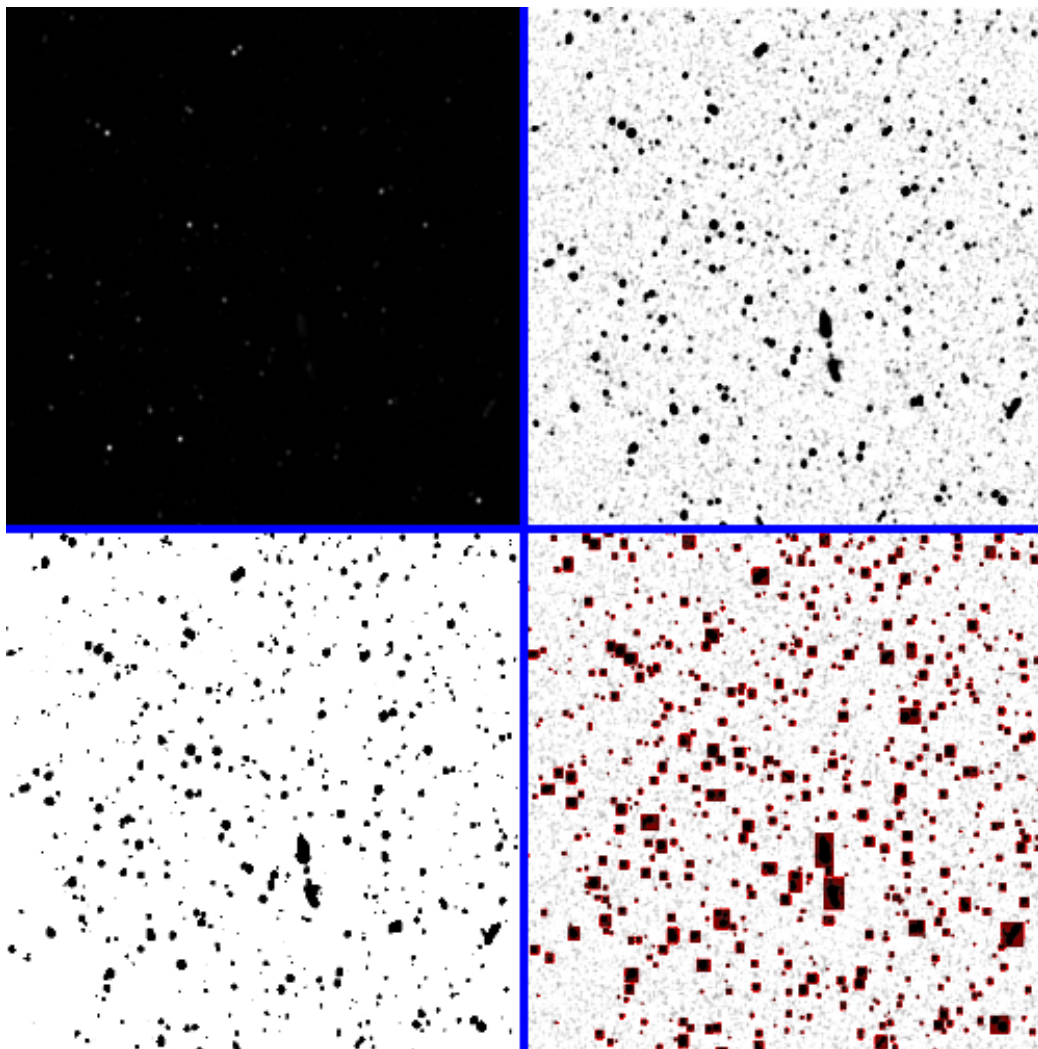


Figure 3.5: **Radio astronomy images segmented by LDA.** From left to right, top to bottom: a radio astronomy image, the image with contrast adjusted and colours inverted in order to see sources, the image segmented using topics derived by LDA (with two topics), the contrast adjusted, colour inverted image overlaid with bounding boxes showing sources identified by flood-filling on the segmented image. Blue borders have been added to delineate the images. Image source: [62, 68].

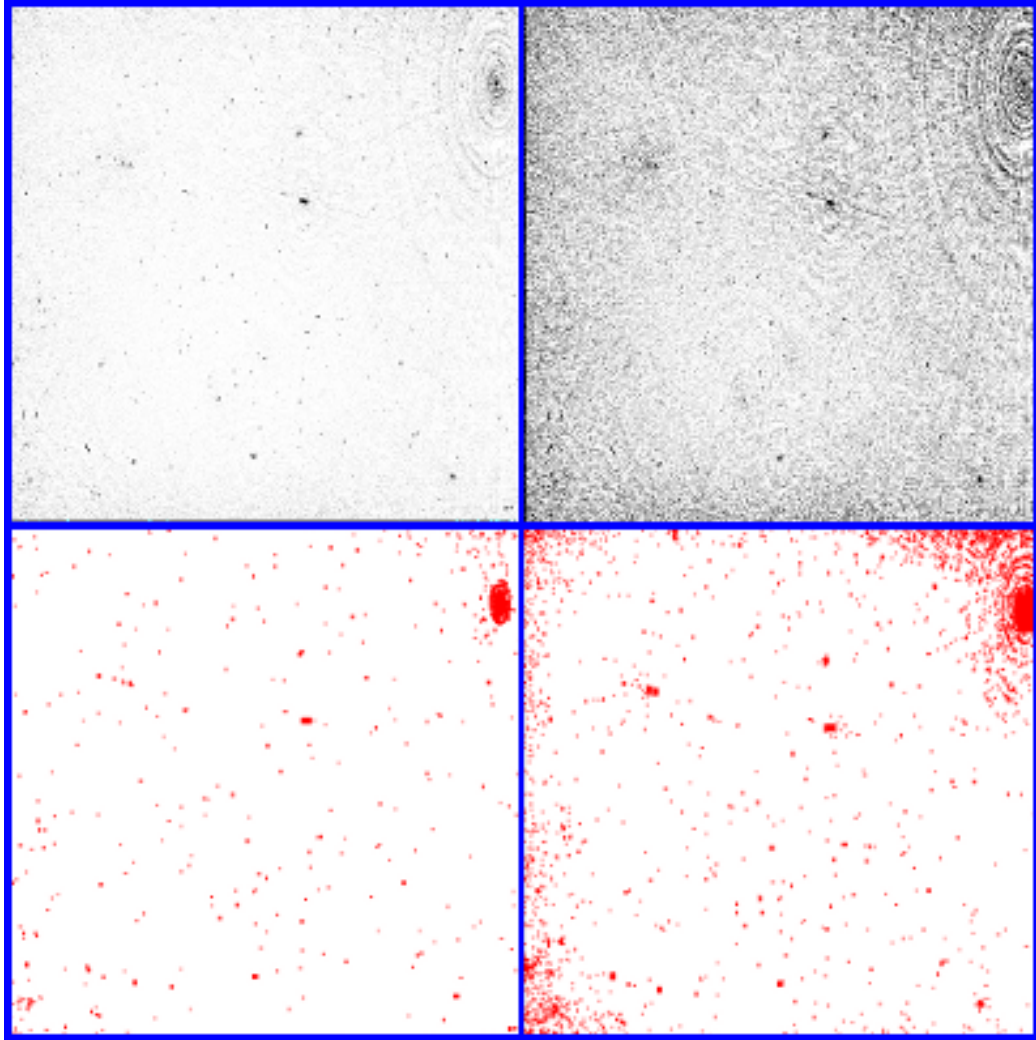


Figure 3.6: **LDA results on an image polluted with artefacts.** An astronomical image (top left) with contrast adjusted to see artefacts (top right), seen as concentric circles and radial spikes. LDA (bottom left) falsely identified fewer artefact pixels as sources than *Duchamp* (bottom right). Blue borders have been added to delineate the images. Image source: [54].

Table 3.2: Performance of LDA versus *Duchamp*

Image	LDA	<i>Duchamp</i>		
	Precision	Recall	Precision	Recall
A1	0.83	0.93	0.98	1.0
A2	0.98	0.99	1.0	1.0
B1	1.0	0.99	1.0	0.96
B2	1.0	0.89	1.0	1.0

sources for which the total flux (intensity) is less than 1.63 mJy⁹. LDA performed similarly to *Duchamp*.

LDA sometimes reported a single source where *Duchamp* correctly separated several; however this is due to the post-processing flood-filling, rather than the LDA algorithm itself. In other cases LDA correctly identified sources that *Duchamp* mistakenly merged.

Bright peak pixels seem key to detection by LDA. For example, in image A2 LDA detects several sources below 1.63 mJy, all of which have peak pixels at least 6σ above the rms noise¹⁰; in contrast LDA's false negatives above 1.63 mJy all have peak pixels less than 6σ above rms noise.

LDA identified fewer artefact pixels as sources than *Duchamp* in the artefact polluted Image C (Figure 3.6). This is a clear demonstration of the strength of using a probabilistic model of background. To avoid the effects of such artefacts using *Duchamp* or similar software, an astronomer would need to manually decompose the image into a number of smaller regions and manually adjust region thresholds. Our implementation of LDA avoids such manual interventions.

⁹1Jy $\equiv 1 \times 10^{-26}$ W/Hz/M²

¹⁰Noise was manually determined by an astronomer.

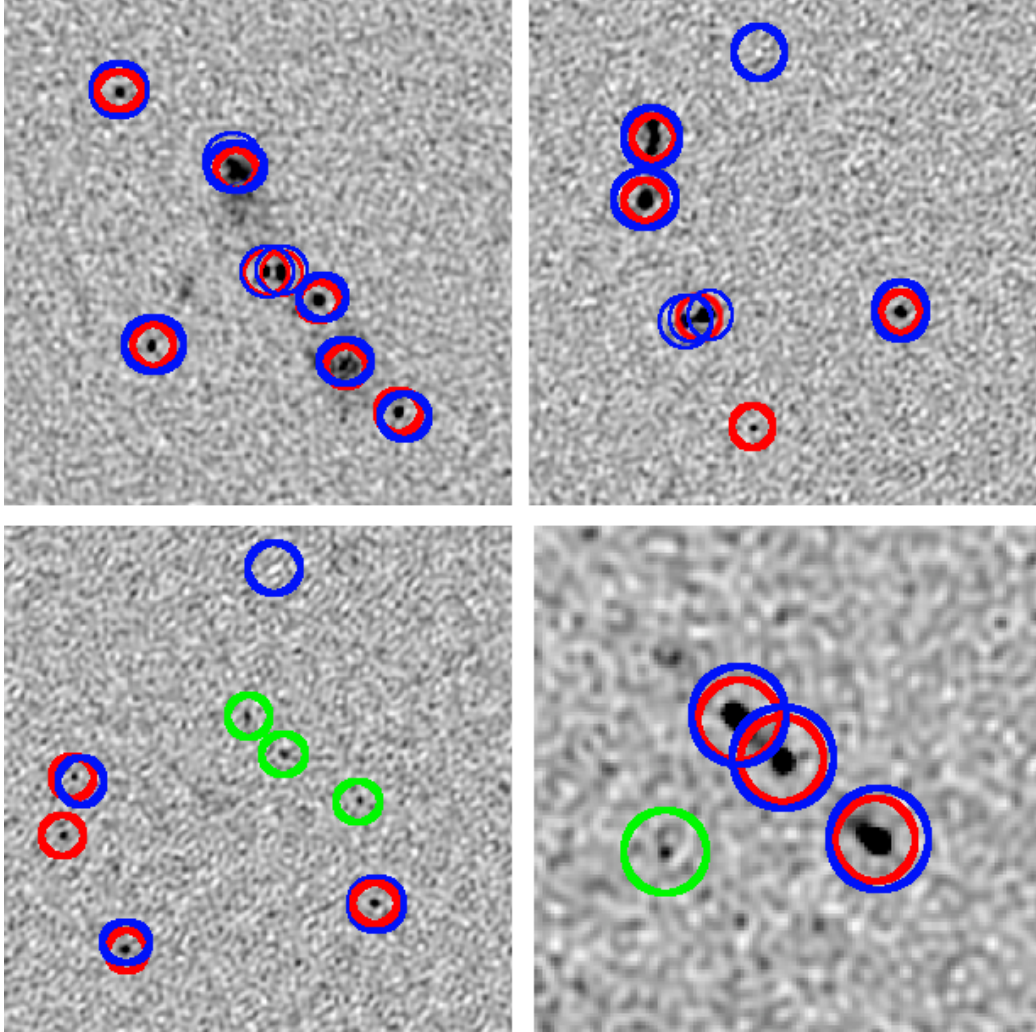


Figure 3.7: **Performance of LDA (blue) and *Duchamp* [74] (red; green) on four representative regions from Image A2.** Top left: a region containing a radio galaxy with two large jets (and no detected core) seen in projection with other point sources. Both algorithms identify multiple components. Top right: A false positive and a false negative for LDA. Bottom left: A false positive and a false negative for LDA; *Duchamp* detects three sources less than 2.5σ above the rms, missed by LDA (green). Bottom right: A radio galaxy with three components, and a point source. Both algorithms split the radio galaxy into three components. Only *Duchamp* detects the point source less than 2.5σ above the rms (green). Image source: [62, 68].

3.5 Discussion

With regards to the real astronomical images, LDA performed similarly to the standard source-detection software *Duchamp* [74] on a representative sample of radio astronomy images, particularly for sources with integrated source flux 2.5σ above the rms noise.

The two algorithms performed similarly with respect to extended, multi-component sources, but LDA had more false positive detections and non-detections than *Duchamp*.

Bright peak pixels seem to be essential for a source to be detected by the current implementation of LDA. The current implementation of LDA is therefore unlikely to detect any diffuse sources (spatially extended sources with low brightness overall and no bright peak pixels). Possible solutions to this problem are discussed in Section 3.5.2.

LDA outperformed *Duchamp* on the image polluted with artefacts — an image that would require labourious manual intervention by an astronomer to detect sources using software such as *Duchamp*. This is a clear demonstration of the utility of the probabilistic model employed.

3.5.1 Parameter and computational issues

In document collections there is a natural segregation of words and documents; LDA was developed for such discrete data [6]. However, there is no natural segregation of pixels into intensity ranges or images into regions. In this chapter, histogram binning of the image and decomposition of the image into subimages was used to discretise the images to resemble document collections. This introduces new parameters to be set. In practice, it was found that the final topics extracted did not vary over a wide range of subimage sizes chosen, however, results did vary based on the number

of bins in the histogram (see Figure 3.4). The approach taken to histogram binning in the current chapter is likely to be responsible for this implementation of LDA's poor performance on faint sources in radio astronomy images; this should be addressed in future implementations.

The analogy of pixels to words may be inaccurate for radio astronomy images. The resolution element of a radio telescope is generally several pixels, and so individual pixels are better thought of as having sub-word size (perhaps "syllables"). The performance of LDA may change if the smallest unit in the model (an individual "word") is adjusted to take this into account.

The number of topics must be set manually. For source detection two may be sufficient (one each for source and background topics); however for image segmentation in general a different number of topics may give different results. Figure 3.3 shows a grayscale image segmented with two to six topics. With two topics, the object in the image is clearly segmented from the background; increasing the number of topics reveals more details of the image; in general terms increasing the number of topics might be expected to increase the level of detail shown, but may introduce irrelevant detail, for example the segmentation of the sky in Figure 3.3. For astronomical images, more than two topics could possibly improve results: for example, three topics may result in a background topic and two source topics with one each for bright and dim sources. This may improve performance on diffuse sources without bright peak pixels.

LDA is a "bag of words" model [3, 6, 67] and so ignores the natural ordering of pixel intensities. However this may be a benefit, rather than a drawback, as this allows objects made up of non-neighbouring pixel intensity ranges to be correctly segmented from images.

Gibbs sampling to infer the latent topics in LDA is expensive both in terms of computation and time. One Gibbs sample involves iterating through each pixel in the image, allocating each a topic based on the current distri-

butions of words to topics and topics to documents, and so is linear in the number of pixels multiplied by the number of topics. As it is not unusual for astronomical images to be 8000×8000 pixels, this can be computationally difficult. Additionally, at least one large three dimensional array (C^{DWT} indexing words by documents by topic) must be kept in memory. However, LDA need not necessarily be run for every image, nor on the whole image. LDA could be run on a small representative section of one image in a collection of similar images in order to extract topics for the whole collection. This would reduce the computational expense of the approach.

3.5.2 Future work

The approach described in this chapter shows how LDA can be used for image segmentation and source detection.

The use of the final topic distributions — segmentation by assigning each pixel a hard topic label and source detection by flood-filling on the segmented image — is crude. A more nuanced approach would eliminate this hard assignment and take a more probabilistic approach to region labelling. For example, given the multinomial models for background and source, gradient ascent could be performed to find regions that have high likelihood under a particular model.

Given the reliance on bright peak pixels for source detection by LDA, more work needs to be done to improve LDA's performance on faint sources. In the particular cases analysed, the addition of more bins in the low range of pixel intensities would likely improve performance; in the general case, the optimal use of bins should be investigated.

In many cases, final background and source topics derived by LDA were simple thresholds in bins, with the background distribution over bins plac-

ing almost all weight on bins from 0 to some bin i , and the source distribution placing almost all weight from bin i to bin K . This is not ideal for discovery of all sources in images in which source and background pixel ranges overlap, and may in fact prevent the discovery of faint sources. In this way, the implementation of LDA in this chapter performs no better than pixel-intensity based thresholding algorithms that restrict search to those sources that contain pixels above a particular threshold. Exploration into why this thresholding in the distribution of pixels happens and how to prevent it is an important avenue for future work.

The implementation of LDA in this chapter does not allow for the use of “soft” (Dirichlet) bin borders, as described in Chapter 2. This is due to the fact that Gibbs sampling is done at the individual pixel (“word”) level — estimating the topic assignment of each pixel conditioned on the topic assignments of all others by:

- decrementing the C^{DWT} matrix at the given pixel’s entry,
- estimating a topic assignment for that pixel from the posterior distribution, and
- incrementing the matrix accordingly.

In the implementation of LDA in this chapter, individual pixels are described by the bin in which they belong, and topics are distributions over bins. The Dirichlet binning strategies described in Chapter 2 result in distributions over bins, with partial counts in a number of bins for each pixel: and so each pixel becomes a distribution over words, rather than a single word. Future work should investigate how latent topic assignments can be estimated given distributions over words, rather than discrete counts of words.

With regards to the additional application of LDA in segmenting non-astronomical grayscale images, the performance of this could be assessed by comparing the obtained segmentation with human segmentation of the

same images, using a large public database of images such as [45]. This would also allow comparison to the results obtained by other image segmentation algorithms.

3.6 Conclusions

This chapter presents a preliminary investigation into use of the topic model latent Dirichlet allocation for image segmentation in grayscale images and source detection in astronomical images. The method builds a probabilistic model of “non-source” pixel distributions.

LDA performed similarly to the standard source-detection software *Duchamp* [74] on a representative sample of radio astronomy images, however, for fainter sources and in particular diffuse sources, there is still some work to be done to determine whether the LDA method will be an improvement over existing algorithms.

A particular success of the approach is the superior result obtained in Image C, which is polluted with artefacts, as compared to the relatively poor performance of *Duchamp*.

The algorithm could be refined to take a more probabilistic approach to region labelling rather than the hard assignment described in the current chapter, along with further exploration of the optimal pixel binning strategy.

Chapter 4

A Dirichlet-multinomial ratio: derivation

Given an astronomical image and a model for background, it is possible to assign a likelihood to a particular region of that image on the basis of how well the region conforms to the model. If, in addition to the model for background, there is a model for foreground, a likelihood can be constructed to indicate which model best fits the region.

This chapter describes how a Dirichlet-multinomial distribution may be used to calculate such a likelihood on discretised images, in the form of an optimisable ratio, where images are discretised by methods such as those discussed in Chapter 2.

To find peaks in ratio-space — that is, regions which poorly conform to a background model, and/or regions which conform well to a foreground model — one approach is to exhaustively calculate the ratio's value for all possible regions of the data. This is inefficient and time-consuming, particularly as the size of the data and the number of parameters defining a region grows. Alternatively, the gradient of the ratio may be calculated, and gradient-ascent performed to find peaks in the ratio.

To illustrate the development of the ratio, one dimensional data is used as well as two dimensional data. In the context of two-dimensional astronomical images, one-dimensional data is a $1 \times X$ “slice” through the image at some y -position; each pixel location $x \in [0..X]$ has an intensity value $\in \mathbb{R}$. The methods described for one dimensional data are extended to two-dimensional data in each section of this chapter.

Two parameters are used to describe regions in one-dimensional data: an x -position m and an approximate half-width σ , which are denoted collectively as θ .

Gaussian elliptical regions are used for two-dimensional data, where the region’s parameters θ are center coordinates m_x and m_y , approximate half-widths in orthogonal direction σ_x and σ_y , and rotation parameter ϕ .

The gradient of the Dirichlet-multinomial ratio (DMR) with respect to θ for one and two dimensional data is derived.

Similarities and differences between the Dirichlet-multinomial ratio and latent Dirichlet allocation (Chapter 3) are discussed in Section 4.6; the two methods are combined in Chapter 6.

In Chapter 5 the results of source detection via gradient ascent on the Dirichlet-multinomial ratio are presented.

4.1 Notation

4.1.1 Representing an image in terms of binned values

Each pixel in an astronomical image has an intensity value $\in \mathbb{R}$; these values can be discretised by “binning” them under some predefined scheme, as described in Chapter 2. Within any given region in an image, a histogram of counts \mathbf{n} in K bins can be formed.

Wherever possible x is used to refer to a pixel in an image in the case of one dimensional data, and k to refer to a bin index. The raw but binned image is denoted by \mathbf{b} , and so b_x is the index of the bin that results from the pixel intensity at point x in the image.

It will be useful to define the variable $C_k^x = \delta_{b_x, k}$ where $\delta_{i,j} = 1$ iff $i = j$, and is zero otherwise. Here C is an indicator variable: in effect, simply a verbose way of representing the integer b_x as an entire vector consisting of a single 1 (corresponding to b_k , the index of the bin k that contains pixel x) and $K - 1$ many 0's. This notation appears cumbersome, but it is introduced in order to be able to extend it to normalised vectors, thus allowing C to represent “partial counts” across more than one bin¹.

The aggregated bin counts for some region \mathbf{R} can then be written:

$$n_k^{(\theta)} = \sum_{x \in R} C_k^x \quad (4.1)$$

The variable $C_k^x = \delta_{b_x, k}$ described in Equation 4.1 in may be extended to $C_k^{x,y} = \delta_{b_{(x,y)}, k}$ for two-dimensional data by denoting the aggregated bin counts for some region \mathbf{R} :

$$n_k^{(\theta)} = \sum_{(x,y) \in R} C_k^{x,y} \quad (4.2)$$

where $\delta_{i_{x,y}, j} = 1$ iff $i = j$ (that is, if pixel i at position (x, y) falls into bin j), and is zero otherwise.

Definition of a hard region: one dimensional data

A hard region has hard (“all or nothing”) borders, which can be thought of as a rectangular bounding-box around some region of the data (see Figure

¹Specifically, it is introduced to facilitate the use of Dirichlet borders (a novel way to derive partial counts as a direct consequence of the uncertainty about where bin borders should be placed, as described in Section 2.2 in Chapter 2).

4.1). For such a region, \mathbf{R} is defined as:

$$\mathbf{R}^{(\theta)} \in [x_{m-\sigma}, x_{m+\sigma}] \quad (4.3)$$

Where the parameters θ are m denoting x -position and σ denoting half-width.

A hard region can also be described by a weighting function $W_x^{(\theta)}$ over points x in the region:

$$W_x^{(\theta)} = \begin{cases} 1, & \text{if } x \in \mathbf{R}. \\ 0, & \text{otherwise.} \end{cases} \quad (4.4)$$

Alternatively, a graded weighting may be applied to a region to “soften” the borders of that region.

Definition of a soft region: one dimensional data

A region with “soft” borders can similarly be defined by a weighting function $W_x^{(\theta)}$. The function’s parameters θ could, for example, involve a center point m and approximate half-width σ , as in the Gaussian function [72]:

$$W_x^{(\theta)} = \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \quad (4.5)$$

Such a region with “soft” borders can be thought of as a weighted curve across the image, rather than a rectangular bounding box, as depicted in Figure 4.1. Soft bordered test regions may be expected to better reflect the nature of astronomical data: astronomical objects do not have hard borders but rather their intensity fades toward their edges.

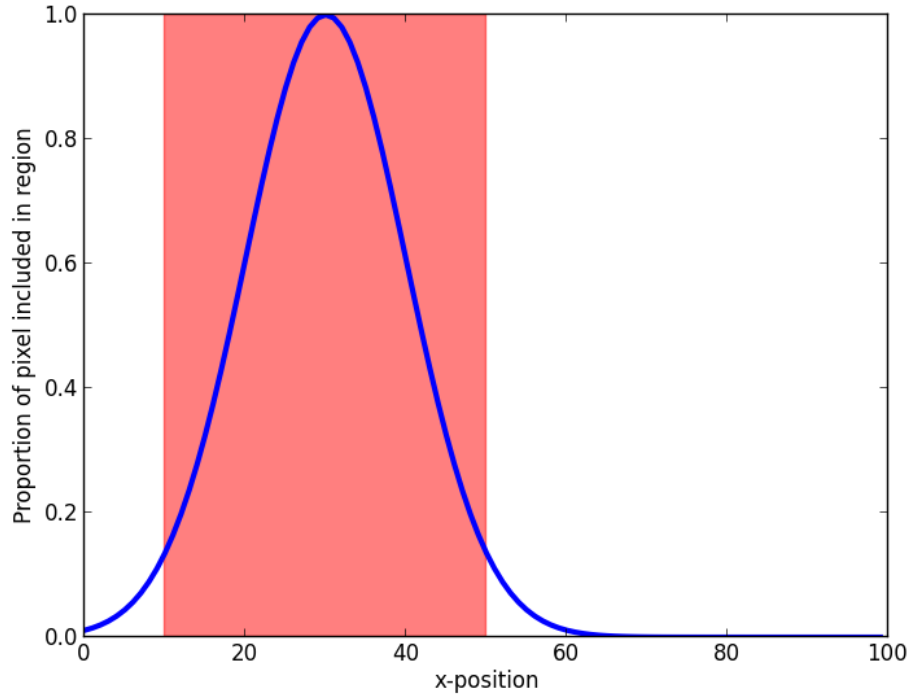


Figure 4.1: **Two methods for defining a region of data** with $m = 30$ and $\sigma = 10$. The pink rectangular region illustrates a “hard-bordered” region as defined in Equation 4.3: each pixel within the pink region contributes a count of 1 to its corresponding bin. The region under the blue curve illustrates a “soft-bordered” region as defined in Equation 4.5: a pixel at position x contributes a count corresponding with the blue curve’s position on the y -axis: for example, a pixel at position 20 on the x -axis increments the count in bin j by approximately 0.4.

Definition of a soft region: two dimensional data

For two dimensional data, a region with “soft” borders can be defined by a weighting function $W_{x,y}^\theta$ over points (x, y) , with parameters θ specifying the position and shape of the region.

For example, to specify a Gaussian elliptical region, the function’s parameters θ could involve center coordinates m_x and m_y , approximate half-widths in orthogonal direction σ_x and σ_y , and rotation parameter ϕ .

Without the rotation parameter, the weighting function could be defined as in [60]:

$$W_{x,y}^\theta = e^f, \quad \text{with} \quad f = -\frac{\Delta_x^2}{2\sigma_x^2} - \frac{\Delta_y^2}{2\sigma_y^2} \quad (4.6)$$

where Δ signifies a displacement from the central position,

$$\Delta_x = x - m_x, \quad \text{and} \quad \Delta_y = y - m_y \quad (4.7)$$

This elliptical region can be rotated by applying a rotation matrix (incorporating the rotation parameter ϕ into Δ_x and Δ_y) [60]:

$$\begin{bmatrix} \Delta'_x \\ \Delta'_y \end{bmatrix} = \begin{bmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix} \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} = \begin{bmatrix} \Delta_x \cos(\phi) - \Delta_y \sin(\phi) \\ \Delta_x \sin(\phi) + \Delta_y \cos(\phi) \end{bmatrix} \quad (4.8)$$

The weighting function with the rotation matrix applied is therefore:

$$W_{x,y}^\theta = \exp \left(-\frac{\Delta_x'^2}{2\sigma_x^2} - \frac{\Delta_y'^2}{2\sigma_y^2} \right) \quad (4.9)$$

Expanding the terms $\Delta_x'^2$ and $\Delta_y'^2$ gives:

$$\Delta_x'^2 = (\Delta_x \cos(\phi) - \Delta_y \sin(\phi))^2 \quad (4.10)$$

$$= \Delta_x^2 \cos^2(\phi) - 2(\Delta_x \cos(\phi)\Delta_y \sin(\phi)) + \Delta_y^2 \sin^2(\phi) \quad (4.11)$$

and:

$$\Delta_x'^2 = (\Delta_x \sin(\phi) + \Delta_y \cos(\phi))^2 \quad (4.12)$$

$$= \Delta_x^2 \sin^2(\phi) + 2(\Delta_x \sin(\phi) \Delta_y \cos(\phi)) + \Delta_y^2 \cos^2(\phi) \quad (4.13)$$

Putting equations 4.11 and 4.13 back into 4.9 and simplifying yields:

$$f = \exp \left(- \left(a \Delta_x^2 + 2b \Delta_x \Delta_y + c \Delta_y^2 \right) \right) \quad (4.14)$$

with:

$$a = \frac{\cos^2(\phi)}{2\sigma_x^2} + \frac{\sin^2(\phi)}{2\sigma_y^2} \quad (4.15)$$

$$b = \frac{-\sin(2\phi)}{4\sigma_x^2} + \frac{\sin(2\phi)}{4\sigma_y^2} \quad (4.16)$$

$$c = \frac{\sin^2(\phi)}{2\sigma_x^2} + \frac{\cos^2(\phi)}{2\sigma_y^2} \quad (4.17)$$

Bin counts in a region

For one dimensional data, the aggregated bin counts in a region (using any of the histogram binning strategies described in Chapter 2) can be multiplied by the weighting function (such as those given in Equations 4.4 and 4.5) to give “weighted counts”, \hat{C} :

$$\hat{C}_k^x = C_k^x W_x^\theta \quad (4.18)$$

Similarly, for two dimensional data:

$$\hat{C}_k^{x,y} = C_k^{x,y} W_{x,y}^\theta \quad (4.19)$$

(Note: to avoid clutter, \hat{C} 's dependence on θ is omitted in this notation).

The aggregated bin counts for the region defined by θ in one dimensional data can then be denoted:

$$n_k^{(\theta)} = \sum_x \hat{C}_k^x \quad (4.20)$$

and similarly for two dimensional data:

$$n_k^{(\theta)} = \sum_{x,y} \hat{C}_k^{x,y} \quad (4.21)$$

Note that if a region is defined with “soft” borders as in Equation 4.5 the sum is extended to potentially all pixels in the image. However in practice, for one dimensional data, the sum over x can be truncated to a local mask for which $W_x^{(\theta)} \geq \epsilon$, for some suitably small value of ϵ , or alternatively, by applying the weighting function to the data as in Equation 4.20 and truncating the resulting vector to the region defined by $\mathbf{R}^{(\theta)} \in [x_{m-a\sigma}, x_{m+a\sigma}]$ for some value of a .

For two dimensional data, the Equation in 4.21 could be similarly truncated to a local mask or by applying Equation 4.21 and truncating the resulting array, for example to a bounding box centered on (m_x, m_y) .

4.2 Scoring a region

4.2.1 The Dirichlet-multinomial distribution

The Dirichlet-multinomial distribution is a compound probability distribution, where the parameter vector $\mathbf{p} = p_1, \dots, p_K$ of a multinomial distribution (with the probability that value k is drawn from \mathbf{p} given by p_k), is drawn from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = \alpha_1, \dots, \alpha_K$ [40, 51].

For some vector of counts $\mathbf{n}^{(\theta)}$ in K bins of a histogram (where the counts are taken within a region defined by $W^{(\theta)}$), integrating out the multinomial distribution gives the following marginal joint likelihood in terms of hyperparameter $\boldsymbol{\alpha}$ [40, 51]:

$$P(\mathbf{n}^{(\theta)}|\boldsymbol{\alpha}) = \frac{\Gamma(A)}{\Gamma(N+A)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \quad (4.22)$$

where

$$A = \sum_k \alpha_k \quad (4.23)$$

$$N = \sum_k n_k \quad (4.24)$$

The logarithm of this is:

$$\log P(\mathbf{n}^{(\theta)}|\boldsymbol{\alpha}) = \log \Gamma(A) - \log \Gamma(N+A) + \sum_k \log \Gamma(n_k + \alpha_k) - \log \Gamma(\alpha_k) \quad (4.25)$$

4.2.2 A Dirichlet-multinomial ratio

A good objective function should give an indication of how poorly a histogram of pixel intensities in a particular region (that is, binned counts) conforms to a model of background, where the model is histograms that might be expected in regions of pure “background” signal. (Or alternatively, how poorly a region’s histogram conforms to a model of background as compared to how well it conforms to a model of foreground).

Such a function can be derived from the ratio of posterior probabilities under the two models²:

²This is a form of a Bayes factor, which is a Bayesian model selection method for data D and M_1 and M_2 the two competing models, with $\frac{P(D|M_1)}{P(D|M_2)}$ [38]. Evaluating terms involves integrating over the parameters of the model. For the Dirichlet-multinomial ratio, conveniently the integral is analytic (Equation 4.22).

$$\text{DMR} = \log \frac{P(S|\mathbf{n}^{(\theta)})}{P(B|\mathbf{n}^{(\theta)})} \quad (4.26)$$

$$= \log \frac{P(\mathbf{n}^{(\theta)}|S)}{P(\mathbf{n}^{(\theta)}|B)} + \log \frac{P(S)}{P(B)} \quad (4.27)$$

where S is the Dirichlet-multinomial distribution with hyperparameter vector α^S and similarly for B with α^B .

If it is assumed that there is an equal mixture of source and background pixels in an image, the second term in Equation 4.27 is equal to zero and may be dropped. This may be a fair assumption if nothing is known about the actual mixture of background and source in images or regions within an image. However, because it is known that background pixels far outweigh source pixels in radio astronomy images, this information can be incorporated into the DMR.

For the gradient calculation, the second term is a constant and so may be dropped, leaving:

$$\log \frac{P(\mathbf{n}^{(\theta)}|S)}{P(\mathbf{n}^{(\theta)}|B)} = \log P(\mathbf{n}^{(\theta)}|S) - \log P(\mathbf{n}^{(\theta)}|B) \quad (4.28)$$

Writing out Equation 4.28 in full gives:

$$\begin{aligned} \log P(\mathbf{n}^{(\theta)}|S) - \log P(\mathbf{n}^{(\theta)}|B) &= \sum_k \log \Gamma(n_k + \alpha_k^S) - \log \Gamma(N + A^S) \\ &\quad - \sum_k \log \Gamma(n_k + \alpha_k^B) + \log \Gamma(N + A^B) \end{aligned} \quad (4.29)$$

Note that no reparameterisation of Equation 4.29 is necessary for extension to two-dimensional data, as $\mathbf{n}^{(\theta)}$, α^S , and α^B are all K -length vectors (where K is the number of bins in a histogram binning of the data).

4.3 The gradient of the DMR

It is possible to find peaks in the DMR by exhaustively iterating through all combinations of values of θ and calculating the value of the ratio for each. However this is obviously time consuming, and potentially intractable with increasing sizes of images and number of parameters θ . A more efficient approach is to perform gradient ascent to find maxima.

Taking the derivative of the DMR (equation 4.29) with respect to region parameters θ (denoting the derivative of the log of the Γ function by ψ), yields:

$$\begin{aligned} \frac{\partial}{\partial \theta} \text{DMR}(\theta) &= \sum_k \underbrace{[\psi(n_k + \alpha_k^S) - \psi(n_k + \alpha_k^B)]}_{\text{denote } Q_k} \frac{\partial n_k^{(\theta)}}{\partial \theta} \\ &\quad - [\psi(N + A^S) - \psi(N + A^B)] \sum_k \frac{\partial n_k^{(\theta)}}{\partial \theta} \end{aligned} \quad (4.30)$$

$$= \sum_k Q_k \frac{\partial n_k^{(\theta)}}{\partial \theta} - \sum_k \underbrace{[\psi(N + A^S) - \psi(N + A^B)]}_{\text{denote } Q_{\text{base}}} \frac{\partial n_k^{(\theta)}}{\partial \theta} \quad (4.31)$$

$$= \sum_k (Q_k - Q_{\text{base}}) \frac{\partial n_k^{(\theta)}}{\partial \theta} \quad (4.32)$$

with the above definitions for Q_k and Q_{base} .

The remaining gradient term can then be calculated from Equation 4.20 for one dimensional data:

$$\frac{\partial n_k^\theta}{\partial \theta} = \sum_x C_k^x \frac{\partial W_x^\theta}{\partial \theta} \quad (4.33)$$

$$= \sum_x C_k^x W_x^\theta \frac{\partial f}{\partial \theta} \quad (4.34)$$

$$= \sum_x \hat{C}_k^x \frac{\partial f}{\partial \theta} \quad (4.35)$$

where $f = \log W_x^\theta$

And similarly for two dimensional data from Equation 4.21:

$$\frac{\partial n_k^\theta}{\partial \theta} = \sum_{x,y} C_k^{x,y} \frac{\partial W_{x,y}^\theta}{\partial \theta} \quad (4.36)$$

$$= \sum_{x,y} C_k^{x,y} W_{x,y}^\theta \frac{\partial f}{\partial \theta} \quad (4.37)$$

$$= \sum_{x,y} \hat{C}_k^{x,y} \frac{\partial f}{\partial \theta} \quad (4.38)$$

where $f = \log W_{x,y}^\theta$

Putting these equations back together, in general the gradient of the DMR for one dimensional data is:

$$\frac{\partial}{\partial \theta} \text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \sum_x \hat{C}_k^x \frac{\partial f}{\partial \theta} \quad (4.39)$$

and similarly for two dimensional data:

$$\frac{\partial}{\partial \theta} \text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \sum_{x,y} \hat{C}_k^{x,y} \frac{\partial f}{\partial \theta} \quad (4.40)$$

Notice the data are involved in (up to) four places:

- Q_k , which depends on n_k , the k^{th} bin in $\mathbf{n}^{(\theta)}$, the aggregated weighted counts in K bins in the region defined by θ ,
- Q_{base} , which depends on $N = \sum_k n_k$ (note that this does not vary with the position of the region, except for edge effects at the edges of the image, but does vary with the region's size),
- \hat{C}_k^x or $\hat{C}_k^{x,y}$; some simplification is possible in the case that this is a strict "indicator function" (zero everywhere except one bin), but not in general, and
- $\frac{\partial f}{\partial \theta}$

but always via the weighted counts \hat{C} .

4.3.1 The gradient, given a specific parameterisation for W

Any window function W may be used in Equation 4.39 (one dimensional) or 4.40 (two dimensional) to arrive at the full DMR and its gradient calculation.

The gradients calculated can be substituted into Equation 4.39 or 4.40 giving the full gradient, which can then be provided to any gradient-based optimisation routine.

One dimensional data

Using the function for a soft window given in Equation 4.5 as an example, whose parameters θ are (m, σ) , the gradient is as follows.

The derivative with respect to m (the x -position of the mid-point of the window) is:

$$\frac{\partial f}{\partial m} = W_x^{(\theta)} \frac{(x - m)}{\sigma^2} \quad (4.41)$$

and so

$$\frac{\partial}{\partial m} \text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \sum_x \hat{C}_k^x \frac{(x - m)}{\sigma^2} \quad (4.42)$$

With respect to σ , it is

$$\frac{\partial f}{\partial \sigma} = W_x^{(\theta)} \frac{(x - m)^2}{\sigma^2} \frac{1}{\sigma} \quad (4.43)$$

and so

$$\frac{\partial}{\partial \sigma} \text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \sum_x \hat{C}_k^x \frac{(x - m)^2}{\sigma^3} \quad (4.44)$$

These two numbers can be substituted into Equation 4.39 to calculate the gradient of the DMR with respect to the complete parameter set.

Two dimensional data

Similarly for two dimensional data, using the function for a two-dimensional ellipse given in Equation 4.14 as an example, whose parameters θ are $(m_x, m_y, \sigma_x, \sigma_y, \phi)$ the gradient can be calculated as follows.

The derivative with respect to m_x is:

$$\frac{\partial}{\partial m_x} \text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \sum_{x,y} \hat{C}_k^{x,y} \frac{\partial f}{\partial m_x} \quad (4.45)$$

differentiating Equation 4.14, with the definition for Δ_x and Δ_y as per Equation 4.7, gives

$$\frac{\partial f}{\partial m_x} = 2a\Delta_x + 2b\Delta_y \quad (4.46)$$

so the gradient calculation for this parameter is:

$$\frac{\partial}{\partial m_x} \text{DMR}(\theta) = \sum_k (Q_k - Q_{\text{base}}) \sum_{x,y} \hat{C}_k^{x,y} \left[2a\Delta_x + 2b\Delta_y \right] \quad (4.47)$$

More generally, differentiating equation 4.14 with respect to each of the

five parameters $m_x, m_y, \sigma_x, \sigma_y$ and ϕ , gives the following:

$$\frac{\partial f}{\partial m_x} = 2a\Delta_x + 2b\Delta_x \quad (4.48)$$

$$\frac{\partial f}{\partial m_y} = 2b\Delta_x + 2c\Delta_y \quad (4.49)$$

$$\frac{\partial f}{\partial \sigma_x} = \frac{1}{\sigma_x^3} \left(\Delta_x^2 \cos^2(\phi) - \Delta_x \Delta_y \sin(2\phi) + \Delta_y^2 \sin^2(\phi) \right) \quad (4.50)$$

$$\frac{\partial f}{\partial \sigma_y} = \frac{1}{\sigma_y^3} \left(\Delta_x^2 \sin^2(\phi) + \Delta_x \Delta_y \sin(2\phi) + \Delta_y^2 \cos^2(\phi) \right) \quad (4.51)$$

$$\frac{\partial f}{\partial \phi} = \frac{1}{2} \left(\frac{1}{\sigma_x^2} - \frac{1}{\sigma_y^2} \right) \left((\Delta_x^2 - \Delta_y^2) \sin(2\phi) + 2\Delta_x \Delta_y \cos(2\phi) \right) \quad (4.52)$$

4.4 The DMR in higher dimensions

It would be relatively straightforward to extend the DMR to three (or higher) dimensional data, following the steps outlined in this chapter. As for one (equation 4.5) and two (equation 4.9) dimensional data, a Gaussian weighting function could be used to describe a soft region in three dimensional data. The general form of the multidimensional Gaussian function [72] is:

$$f(\mathbf{x}) = \exp \left(- \left(\mathbf{x}^T A \mathbf{x} \right) \right) \quad (4.53)$$

where $\mathbf{x} = \{x_1, \dots, x_n\}$ are the n dimensions of the data, \mathbf{x}^T is the column \mathbf{x} transposed and A is a positive-definite $n \times n$ matrix. Weighted counts \hat{C} of binned pixels could be derived as for one and two dimensional data with

aggregated bin counts for the region denoted:

$$n_k = \sum_{\mathbf{x}} \hat{C}_k^{\mathbf{x}} \quad (4.54)$$

The DMR (equation 4.29) could be used for multidimensional data with no reparameterisation necessary, as this function is given in terms of K -length vectors (where K is the number of bins in a histogram binning of the data). To find peaks in the function's value its derivative could be derived as in section 4.3.

4.5 Comparison with other Bayesian methods

Of the existing Bayesian source detection methods described in Chapter 1, the DMR is most closely related to the work of Hobson and McLachlan [29] and Feroz and Hobson [20]. Hobson and McLachlan performed Bayesian model selection using MCMC to explore the parameters of background and source models, with a source model of circularly symmetric Gaussian-shaped objects and a background model of Gaussian noise. In one version of the algorithm sources are found iteratively. Feroz and Hobson used the Bayesian model selection algorithm nested sampling with circularly symmetric Gaussian-shaped sources. In this Chapter, a Dirichlet-multinomial function was derived; in the following Chapter the DMR is used to iteratively find Gaussian elliptical shaped regions that differ from a background model.

Source and background are modelled as distributions over ranges of pixel intensities in the DMR. A parametrised form for astronomical sources is assumed and these parameters estimated from a posterior distribution, however sources are not assumed to be circularly symmetric or of a particular size.

4.6 Comparison with latent Dirichlet allocation

There are many similarities between latent Dirichlet allocation (as used in Chapter 3) and the Dirichlet-multinomial ratio.

Both techniques use multinomial distributions over bins as models for source and background; and in both methods these distributions are part of a compound Dirichlet-multinomial distribution with the multinomial parameter integrated out, leaving only the data and the Dirichlet hyperparameter(s).

However, LDA estimates a conditional distribution while the DMR calculates a marginal joint distribution [6, 51]. In addition, the multinomial models for source and background are easily derived from the LDA process, but remain latent in the DMR.

These differences are relatively minor; however there are two differences between the models that are more fundamental. In LDA, there is one α hyperparameter over all topic distributions $\beta_k \in K$ (where $K = 2$ for source and background in radio astronomy images). In the case of the DMR, there are two α hyperparameters: one each for source and background distributions. Under LDA's generative model the multinomial distributions over bins for source and background are fixed across the whole image; while under the generative model for the DMR, they may vary between sub-images.

With regards to these two points, there is a conceptual similarity between the DMR described in this chapter and the DCMLDA model described in [15, 16, 44] (DCMLDA is: Dirichlet compound multinomial latent Dirichlet allocation).

DCMLDA is largely the same as LDA as described in Chapter 3, except for one key difference: LDA has one multinomial distribution β_k for each of K topics, all of which have the same Dirichlet hyperparameter η . In the case

of DCMLDA however, each topic k has its own Dirichlet hyperparameter η_k , from which a multinomial topic distribution over words $\beta_{k,d}$ is drawn for each topic k and each document d in the document collection [15].

The DCMLDA model was developed to model the phenomenon of “burstiness”, a term which describes the behaviour of words appearing in “bursts” — that is, if a word appears once in a document, it is more likely to appear again in that same document, even if that word is relatively rare in the corpus and does not appear in most documents. This phenomenon is not well modelled by LDA, where a document is a mixture of topics, and each topic describes how likely each word in the vocabulary is to appear under that topic. Under the LDA model, each word in a topic has a particular probability of occurring, which is static from document to document [15, 16, 44].

Doyle and Elkan [15] give the following example: given a “sports” topic in a document collection, in which the words “rugby” and “hockey” have equal probability of occurring, the LDA model will give both these words equal probability of occurring within a document. However, this is a poor model given that one occurrence of the word “rugby” in a document means a subsequent occurrence of the word “rugby” is more likely than a first occurrence of the word “hockey”.

In the context of the DMR for radio astronomy images, this allows the variability of source and background distributions, particularly over an image, to be accounted for. For example, the variability of an image’s background distribution, background regions at the edge of radio astronomy images are more noisy than those at the centre of an image (see Figure 4.2 for an illustration of this point).

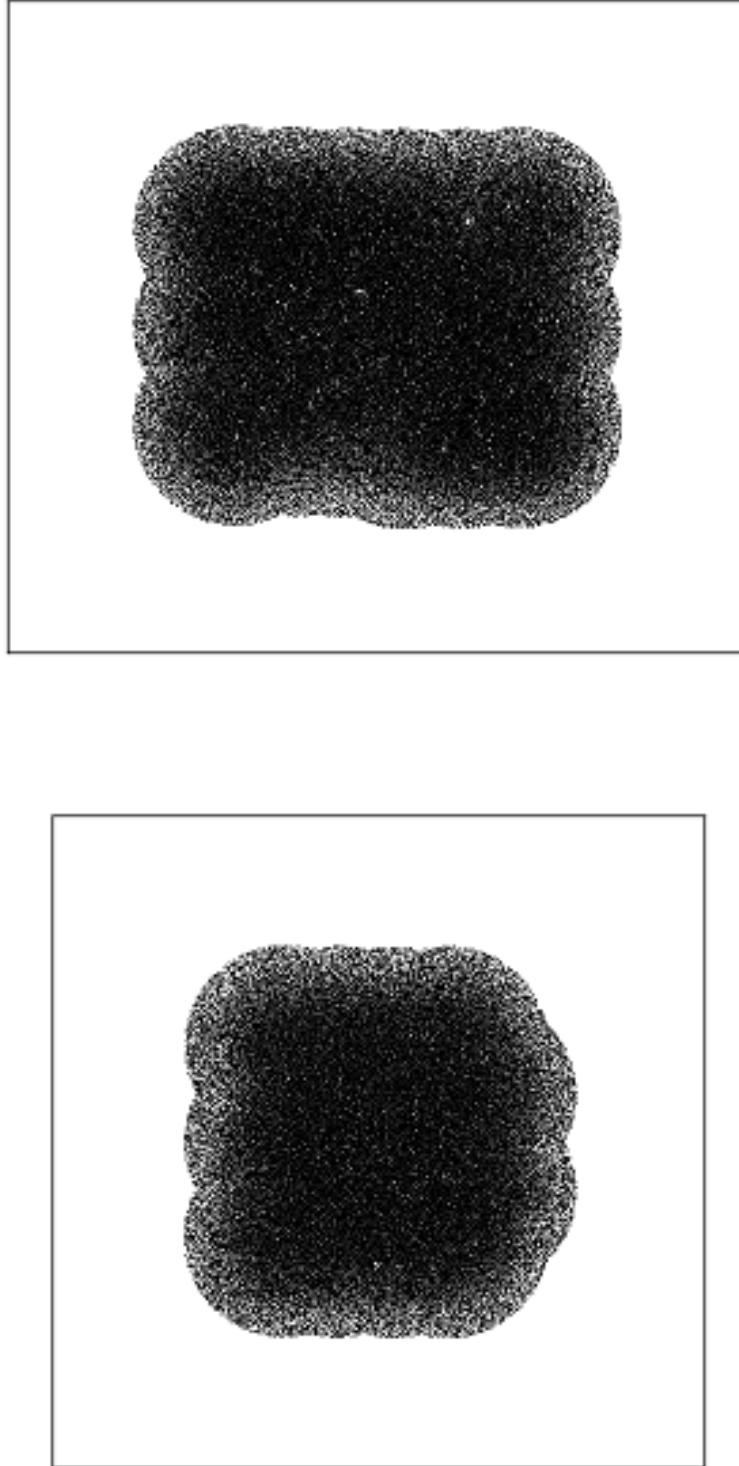


Figure 4.2: **Background variation.** Two radio astronomy images with contrast adjusted to show the variation in background across an image. Note the greater noise at the edges of the images. Images: Chandra Deep Field South (CDFS; top) and European Large Area ISO Survey S1 (ELAIS; bottom) [54]

Chapter 5

A Dirichlet-multinomial ratio: evaluation

The Dirichlet-multinomial ratio (DMR) developed in Chapter 4 is evaluated in this chapter.

The performance of the DMR in identifying astronomical sources is evaluated against simulated data as well as against subimages from two large astronomical images. Performance is compared to a ground truth catalogue constructed using the source detection software BLOBCAT [26] with manual postprocessing by an astronomer. The Dirichlet bin border-softening technique, described in Chapter 2 is employed, and results obtained using this technique are compared to “hard” binning strategies.

5.1 Methods

The Dirichlet-multinomial ratio was applied to real and simulated data, in order to test whether peaks in the function corresponded to locations and sizes of sources (i.e., non-background regions of the images). The

method was tested using hard-bordered and soft-bordered test regions defined by a Gaussian curve (one dimensional data) or Gaussian ellipse (two dimensional data). Equal occupancy and equal width histogram binning strategies (as described in Section 2.1 in Chapter 2) were employed, and the Dirichlet-border “softening” of both types of bin borders (Section 2.2, Chapter 2) was also used. Gradient ascent was performed to find peaks in the DMR. A potential improvement to the DMR (background compensation) is described.

Outline of the procedure

1. Given:
 - data (real or simulated),
 - a weighting function W^θ for test regions, and
 - a binning strategy (equal width, Dirichlet equal width, equal occupancy, or Dirichlet equal occupancy)
2. Exhaustively calculate the DMR over a range of values for parameters θ (optional);
3. Perform gradient-ascent to locate sources in the data (iteratively removing found sources from the data); and
4. Compare the parameters of found sources against those of actual sources to evaluate performance.

Each of these steps is described in more detail below.

5.1.1 Parameters

Data: one dimensional

Both real and simulated data were used in early exploration of the methods described in this chapter; simulated data was used for evaluation.

Real data comprised $1 \times X$ “slices” through astronomical images at some y -position. An example of these data is shown in Figure 5.1.

Simulated data were created by generating three “sources” and then overlaying the sources with Gaussian noise. In each simulated image, one source was generated using the Gaussian function (where m is an x -position and σ an approximate half-width; A controls the height of the peak):

$$A \exp \left(-\frac{(x - m)^2}{2\sigma^2} \right) \quad (5.1)$$

with noise added such that faint sources had lower than average variance. Two sources were generated by a skewed generalised Gaussian [49]:

$$\left[\frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} \exp \left(-\left(\frac{|x - m|}{\alpha} \right)^\beta \right) S(\epsilon(x - m)) \right] \frac{A}{\max_{x'}} \quad (5.2)$$

where m is an x -position, α is a scale parameter and β is a shape parameter ($\beta < 2$ produces tails that are heavier than Gaussian, $\beta = 2$ produces Gaussian tails, and $\beta > 2$ produces lighter than Gaussian tails [49]). The skewness is provided by S , which is the sigmoid function:

$$S(t) = \frac{1}{1 + \exp^{-t}} \quad (5.3)$$

applied to $t = \epsilon(x - m)$ with ϵ a random value $\in [0, 1]$. The multiplier $\frac{A}{\max_{x'}}$ controls the peak of the curve, with $\max_{x'}$ the maximum value returned by the weighting function. One of these sources was overlaid with zero-mean, fixed-variance Gaussian noise (with the same variance as background pixels); the other with zero-mean Gaussian noise with variance

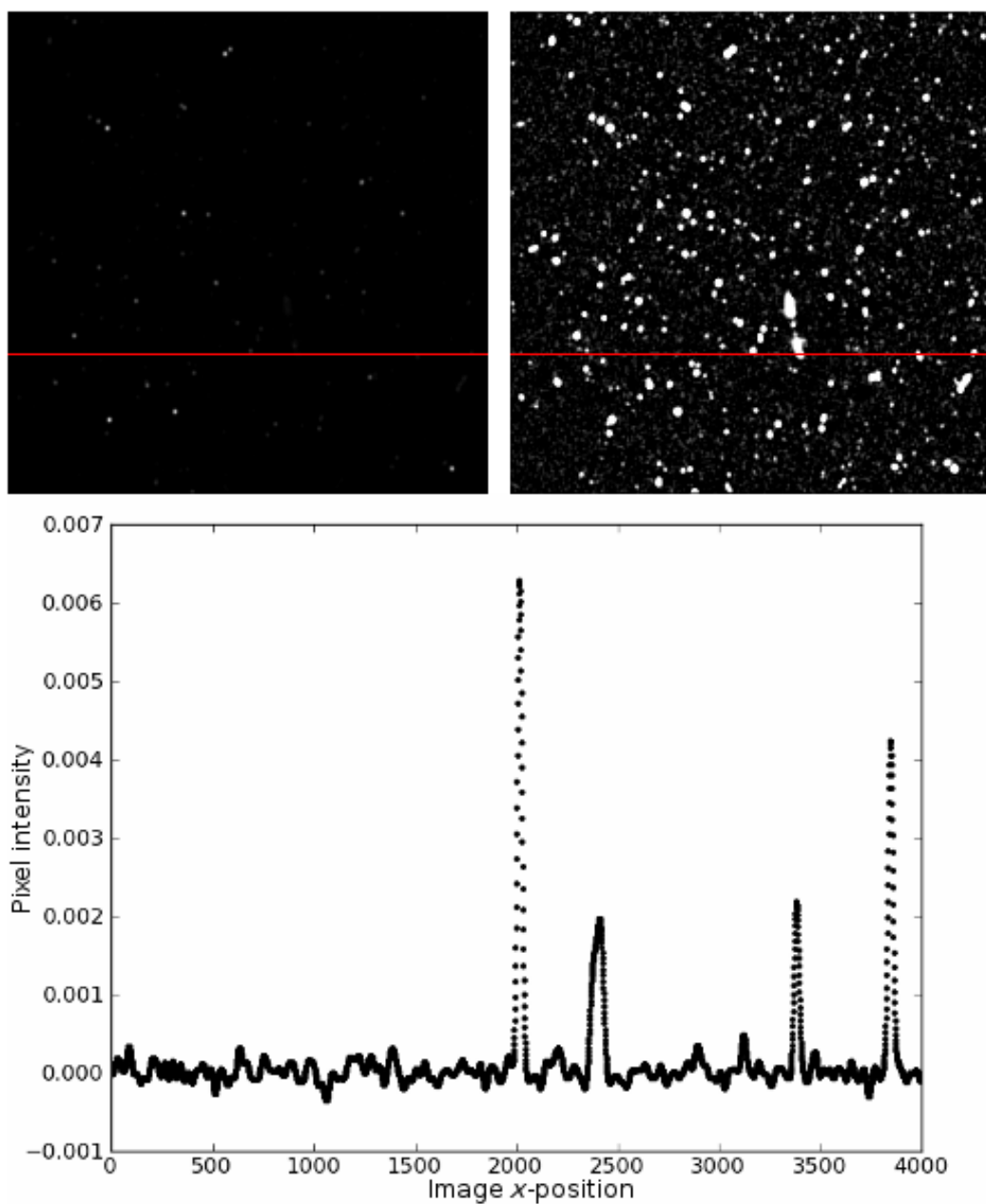


Figure 5.1: **One dimensional data.** A “slice” through an astronomical image. The image is shown at top-left, and at top-right with contrast adjusted to see sources. The slice through the data is shown as a red line in both images. At bottom, a plot of x -position by pixel intensity is shown. Image source: [62, 68].

that increases proportionally with the height of the generated source. Examples of curves created by Equation 5.2 are shown in Figure 5.2.

This skewed generalised Gaussian function (Equation 5.2) was chosen as well as the Gaussian function (Equation 5.1) to generate sources as the soft-bordered test region used for evaluation of the DMR is the Gaussian function (Equation 4.5). It is important therefore that all of the sources are not generated using the same function that generates the test region, as this would likely artificially inflate the success of the Dirichlet-multinomial ratio.

Some examples of the simulated data are shown in Figure 5.3.

Data: two dimensional

In the case of two dimensional data, simulated data were used for exploratory work, and real data were used to test the methods described in this chapter.

Real data comprised 500×500 pixel windows of two large astronomical images [54]. Each window was treated as a whole image for purposes of evaluation of the DMR; windows had an average of 35.9 sources each (see Table 5.1 and Figure 5.4).

“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 [26] with manual postprocessing by an astronomer. Note that the noisy borders of the images were excluded from source detection in construction of the ground truth catalogue¹.

¹The area of the two images for source extraction was dened by: rms noise $\leq 100\mu$ Jy beam¹; 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 [1].

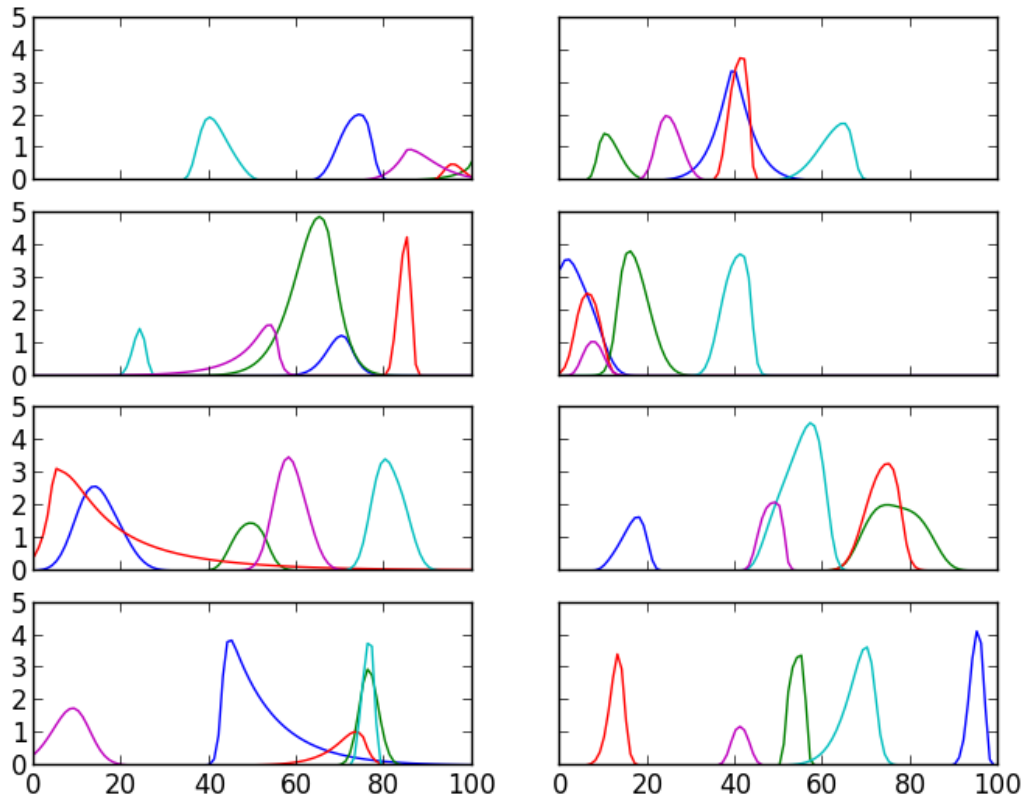


Figure 5.2: **Skewed generalised Gaussian curves.** Examples of curves generated by the skewed generalised Gaussian formula in Equation 5.2. The parameters are: $m \in [0, 100]$, $\alpha \in [2, 10]$, $\beta \in [0.5, 5]$ and $\frac{A}{\max_{x'}} \in [0.5, 5]$.

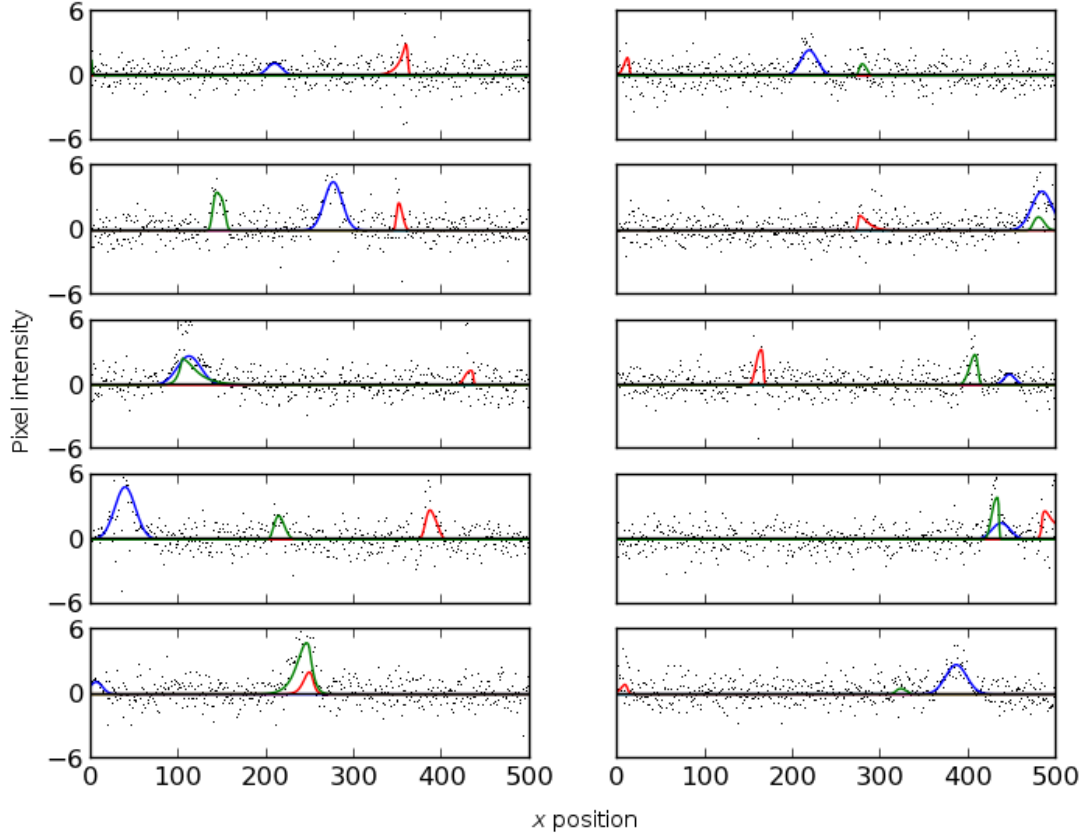


Figure 5.3: **Simulated one dimensional data.** The data (sources plus Gaussian noise) are shown as black dots with the underlying sources shown as coloured curves. The blue curve is generated by the Gaussian formula (Equation 5.1), while the red and green curves are generated by the skewed generalised Gaussian formula (Equation 5.2). The parameters for the Gaussian curve are: $m \in [0, 500]$, $\sigma \in [6, 14]$, $A \in [0.5, 5]$; and for the skewed generalised Gaussian: $m \in [0, 500]$, $\alpha \in [6, 14]$, $\beta \in [0.5, 5]$ and $\frac{A}{\max_{x'} \dots} \in [0.5, 5]$.

Table 5.1: Astronomical images used for evaluation

	ELAIS ² [54]	CDFS ³ [54]
Total number “ground truth” sources in image	2067	3079
Number of sub-windows used	13	12
Mean (std dev) sources per window	34.7 (11.7)	37.2 (8.7)
Min, max number of sources per window	16, 53	20, 49

To obtain the set of windows used for evaluation, 50 windows were selected at random from the images (25 each). Windows were discarded from the evaluation set if they were outside the area used in construction of the ground truth catalogue. This yielded 25 images for evaluation of the DMR (13 from ELAIS and 12 from CDFS).

Note that for the construction of the ground truth catalogue, 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. This is a key point — some of the most scientifically important objects in astronomy are those that are dim, with intensities in the range of background noise [53]. Current source detection methods restrict search such that dim sources cannot be found. This is due to the nature of the pixel-intensity based thresholding methods that make up the vast majority of current search detection algorithms. If the search is not restricted in this way, the methods will identify far too many background regions as “sources”, and the output will be unusable. The DMR does not restrict search in this way, and so can be expected to be more effective at finding dim objects.

Simulated data were created by generating three “sources” and then overlaying the sources with Gaussian noise⁴ (zero-mean and, fixed-variance

²Australia Telescope Large Area Survey European Large Area ISO Survey S1

³Australia Telescope Large Area Survey Chandra Deep Field-South.

⁴This is typical of how simulated data were created for evaluation of Bayesian source detection methods, for examples, see: [9, 20, 24, 29, 63], each of which simulated sources

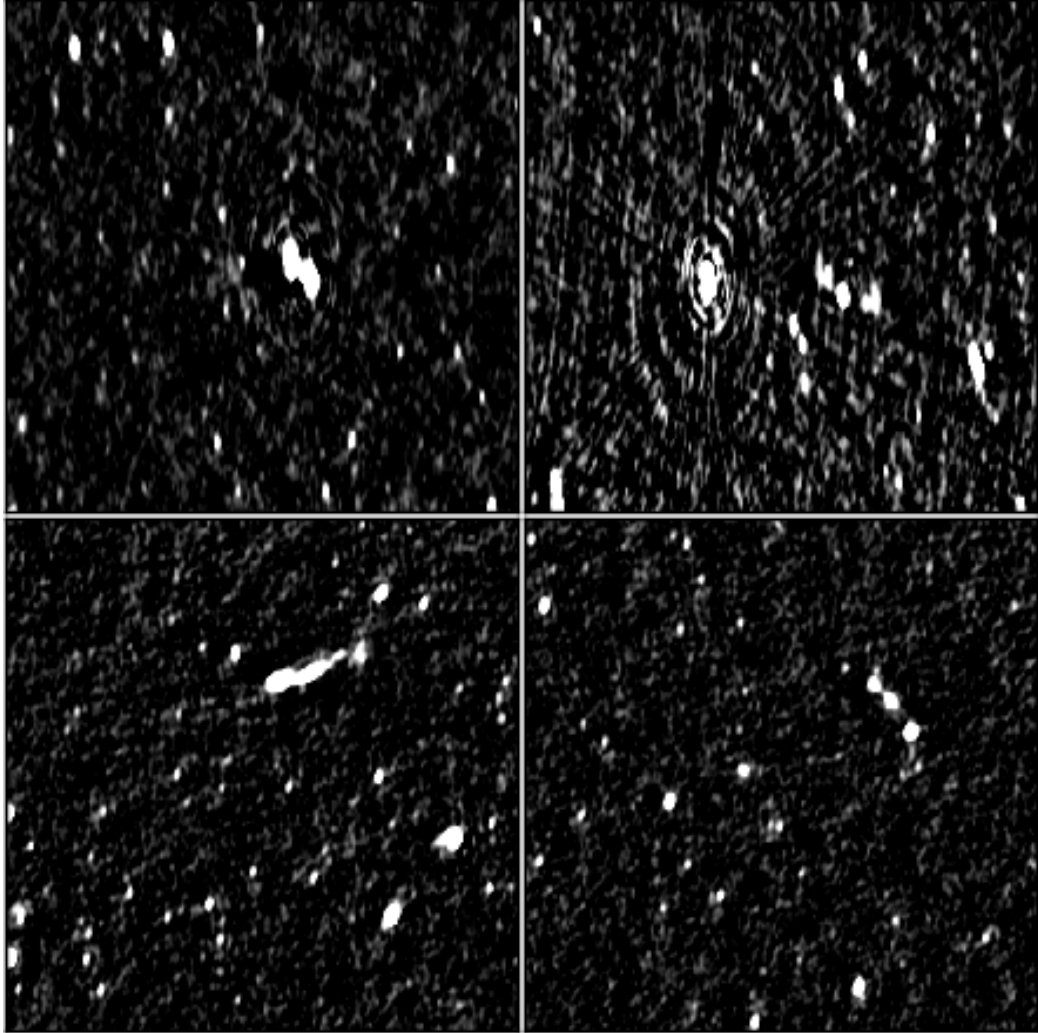


Figure 5.4: **Examples of astronomical data used for evaluation of the DMR.** The images shown are 500×500 windows from (top) CDFS and (bottom) ELAIS [54]. Note that all images are contrast adjusted to show sources.

with the same variance of background pixels for one source, variance that increases proportionally with the height of the source for another source, and such that faint sources have lower than average variance for the third source). The sources were generated using the Gaussian ellipse function (with center coordinates m_x and m_y , approximate half-widths in orthogonal direction σ_x and σ_y , and rotation parameter ϕ , and A controlling the height of the peak):

$$A \exp \left(- \left(a(x - m_x)^2 + 2b(x - m_x)(y - m_y) + c(y - m_y)^2 \right) \right) \quad (5.4)$$

with:

$$a = \frac{\cos^2(\phi)}{2\sigma_x^2} + \frac{\sin^2(\phi)}{2\sigma_y^2} \quad (5.5)$$

$$b = \frac{-\sin(2\phi)}{4\sigma_x^2} + \frac{\sin(2\phi)}{4\sigma_y^2} \quad (5.6)$$

$$c = \frac{\sin^2(\phi)}{2\sigma_x^2} + \frac{\cos^2(\phi)}{2\sigma_y^2} \quad (5.7)$$

Some examples of the simulated data are shown in Figure 5.5.

Performance of the DMR on simulated two dimensional data is given in Table 5.3, however these results should be read with caution: the sources in the simulated data were generated by a Gaussian ellipse function — the same function used to generate test regions in an image — and this may cause the success of the Dirichlet-multinomial ratio to be artificially inflated. It is worth noting, however, that authors of other Bayesian source detection methods have used a two dimensional Gaussian to both generate simulated data and as a model for sources in their detection algorithms; for examples, see: [20, 29, 63].

using the Gaussian formula and overlaid the sources with uncorrelated Gaussian noise.

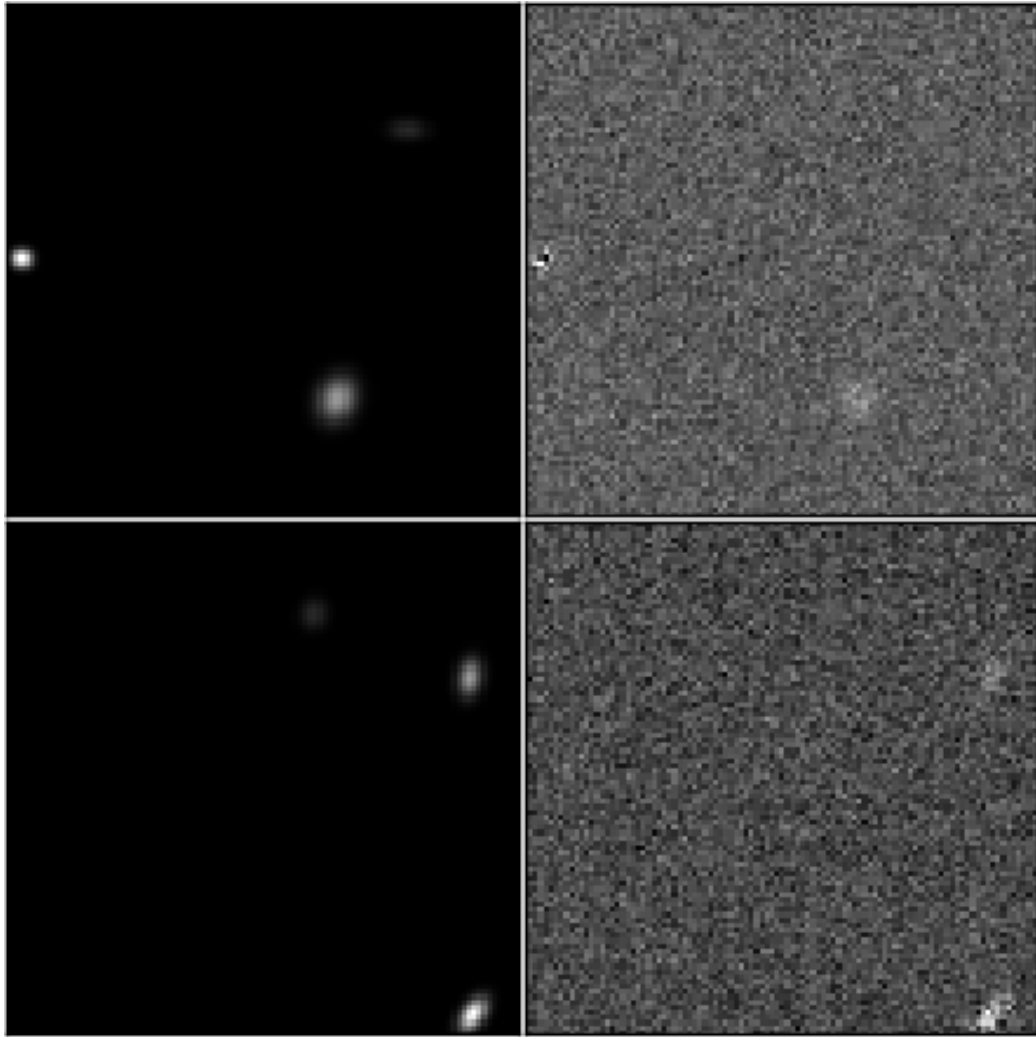


Figure 5.5: **Simulated astronomical images.** Two images with three simulated sources each are shown at left. The images overlaid with noise are shown at right.

A note comparing simulated data with real data for evaluation

Using simulated data for evaluation means that “ground truth” is known exactly: that is, the parameters of sources in the image are precisely known because they are generated; they do not need to be manually labelled by an astronomer (a time-consuming task), or found by an existing source detection package, or some combination of both methods. This allows precise evaluation of results. However, simulated data do not perfectly mimic the statistical properties of real (astronomical) data⁵.

On the other hand, labelled data may be incorrectly labelled. In particular, dim sources are often missed by existing source detection packages. If labelled data is missing such sources which are found by the source detection method being evaluated, the detection of these scientifically important sources will be designated as false detections.

In this thesis, early exploration work was done on both simulated and real data. Final evaluation of the DMR on one dimensional data was done using simulated data, and on two dimensional data was done on both simulated data and real data labelled by the BLOBCAT source detection package with manual postprocessing.

Test region

For one dimensional data, both “hard” (Equation 4.4) and “soft” (Equation 4.5) test regions were used in early exploration (see Figure 5.10). A soft test region was used for evaluation.

For two dimensional data, a Gaussian elliptical test region was used for both early exploration and evaluation; its weighting function $W_{x,y}^{(\theta)}$ is given in Equation 4.14.

⁵Simulated data that do emulate the properties of real data well were sought but not available in time for inclusion in this thesis.

Binning strategies

Equal occupancy and equal width bin border strategies with $K = 10$ (Chapter 2, Section 2.1) were used, and the Dirichlet-border “softening” of both types of bin borders (Chapter 2, Section 2.2) was also used with both strategies. Dirichlet bin border strategies used a symmetric α -vector with $\alpha_i = 10 \forall i \in K$, to make 50 sets of bins.

5.1.2 Derivation of α^S and α^B

The Dirichlet-multinomial ratio (Equation 4.27) compares the binned counts in a region with theoretical “background” and “source” distributions. More precisely, it compares the α -vectors of Dirichlet distributions from which multinomial distributions over bins for “background” and “source” distributions are drawn.

The α -vectors may be derived from a process such as latent Dirichlet allocation (Chapter 3), where the distribution over bins in “source” and “background” topics may be used for the respective α -vectors. This approach is explored⁶ in Chapter 6.

Alternatively, the data itself may be treated as pseudocounts, and be used directly as α -vectors.

As an example of using the data directly as α -vectors, a human expert (such as an astronomer) might manually select a source-free area of “typical background” within an astronomical image. Binning the pixels in this region into bins defined on the whole image would yield an α^B -vector that could be used in the DMR in Equation 4.27. Such manual selection is

⁶Note that topic distributions inferred by LDA are multinomial distributions, not Dirichlet distributions; however they may be used to approximate each other [32, 33]; this is discussed more in Chapter 6.

however costly and impractical, and may be biased by the idiosyncrasies of the human perceptual system.

Obviating the need for manual selection of a “typically background” region, the *whole image* may be used as the α^B -vector. That is, the whole image is binned according to a particular binning strategy, and the binned-counts are used as the α^B -vector. In terms of a generative model for background, distributions over bins for background regions are drawn from a Dirichlet with this α^B -vector. This is particularly appropriate for astronomical images in which background pixels far outnumber source pixels — in any particular background region, one would expect relatively few source pixels. This “whole image as background” approach is taken in this chapter.

The α^S -vector may similarly be derived from a process such as latent Dirichlet allocation or manually defined by an astronomer.

A source model with the symmetric α -vector with $\alpha_i = \alpha_j = 1 \ \forall i, j \in 1, \dots, K$ does not incorporate any knowledge about what a source distribution might be; all multinomial distributions over K bins are equally likely to be drawn from the Dirichlet distribution with this α -parameter [21]. In this chapter, α^S was set to this α -vector.

Given these parameterisations for α^B and α^S , background regions are expected to have a similar distribution over bins as the whole image does; while a source region may have any distribution over K bins. A high DMR value using Equation 4.27 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.

Because any unusual distribution conforms better to the source model than the background model using this parametrisation, a region with more dark pixels than is typical would get a higher DMR value than a region that conforms perfectly to the background model. Similarly, regions of lower than usual variance in an image would conform better to the “agnostic” source model than the whole-image-as-background model (for examples of simulated low variance sources, see Figures 5.7 and 5.10).

This broad definition of non-background like regions in an image allows a great deal of flexibility in the use of the DMR to find different types of features in astronomical images.

However in the context of identifying sources (which are generally brighter than background pixels), such dark or low variance regions are, although possibly worthy of attention, unlikely to be real sources. The α^S -vector could therefore be adapted to have high values in the region of bright pixels, for example $a_i > 1$ where $i > \epsilon$, and ϵ is some brightness threshold. Similarly, the α^S -vector could otherwise be set to identify features of an image with particular known characteristics. This approach was not taken in this chapter (however see Chapter 6 in which the α^S -vector is set according to the source model derived by LDA).

5.1.3 Dirichlet-multinomial ratio

A heuristic value for the prior for the ratio of source to background pixels in the image of 0.05/0.95 was used in place of the second term in Equation 4.27. This reflects the prior belief that radio astronomy images are dominated by background.

Exhaustive calculation of the DMR for one dimensional data

For illustrative purposes, the DMR and its gradient was exhaustively calculated over a range of values for parameters θ on a number of simulated one dimensional images.

Figures 5.6 to 5.8 show the DMR and its gradient over $m \in [0..500]$ and $\sigma \in [0..250]$. Each of the six plots has three subplots: 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); and a plot of the function-space, with parameter m on the x -axis and σ on the y -axis. Each point in the space is coloured along a continuum indicating whether background or source is favoured. Gradient lines (black) are superimposed.

Note that in each of the three figures, in general, 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.

Unsurprisingly, bright and large sources appear to be the easiest to find: the peak at $m = 38$ in Figure 5.6, at $m = 410$ in Figure 5.7, and at $m = 275$ in Figure 5.8 correspond with the darkest blue (most source-like) regions in the DMR space. Note however that the dim, low variance source at $m = 466$ in Figure 5.7 is also well found.

Background-compensation correction to the DMR

The parametrisation of α^B and α^S described in Section 5.1.2 biases the Dirichlet-multinomial ratio towards higher values for smaller regions.

This is due to the fact that the whole (binned) image is taken as pseudo-counts for the background model, and the source model is constructed to give all multinomial distributions over K bins equal probability. When

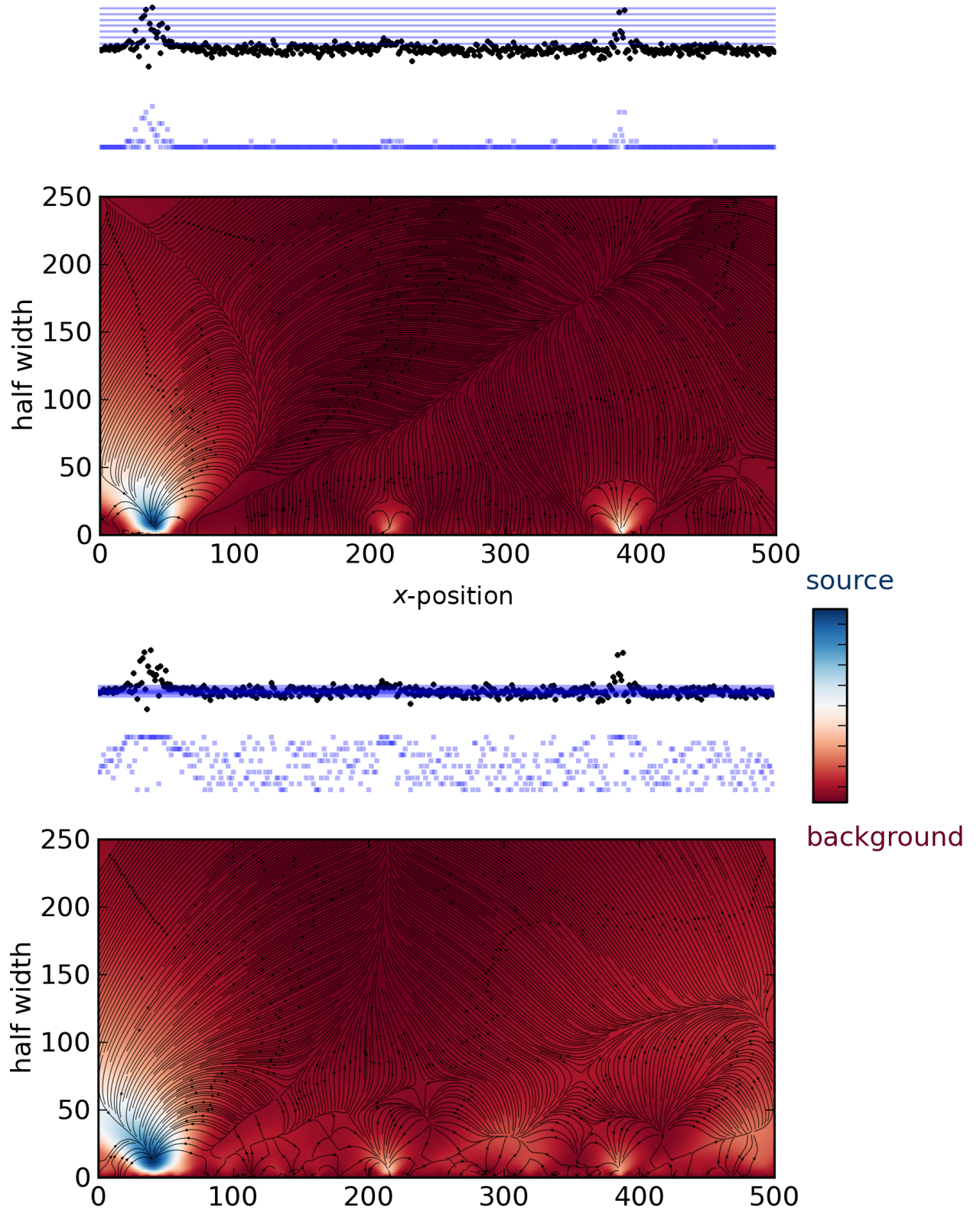


Figure 5.6: **Exhaustive calculation of the DMR and its gradient.** Three sources have parameter $m = (38, 219, 389)$. The top plot uses Dirichlet equal width bins; the bottom Dirichlet equal occupancy. Details are in text in Section 5.1.3.

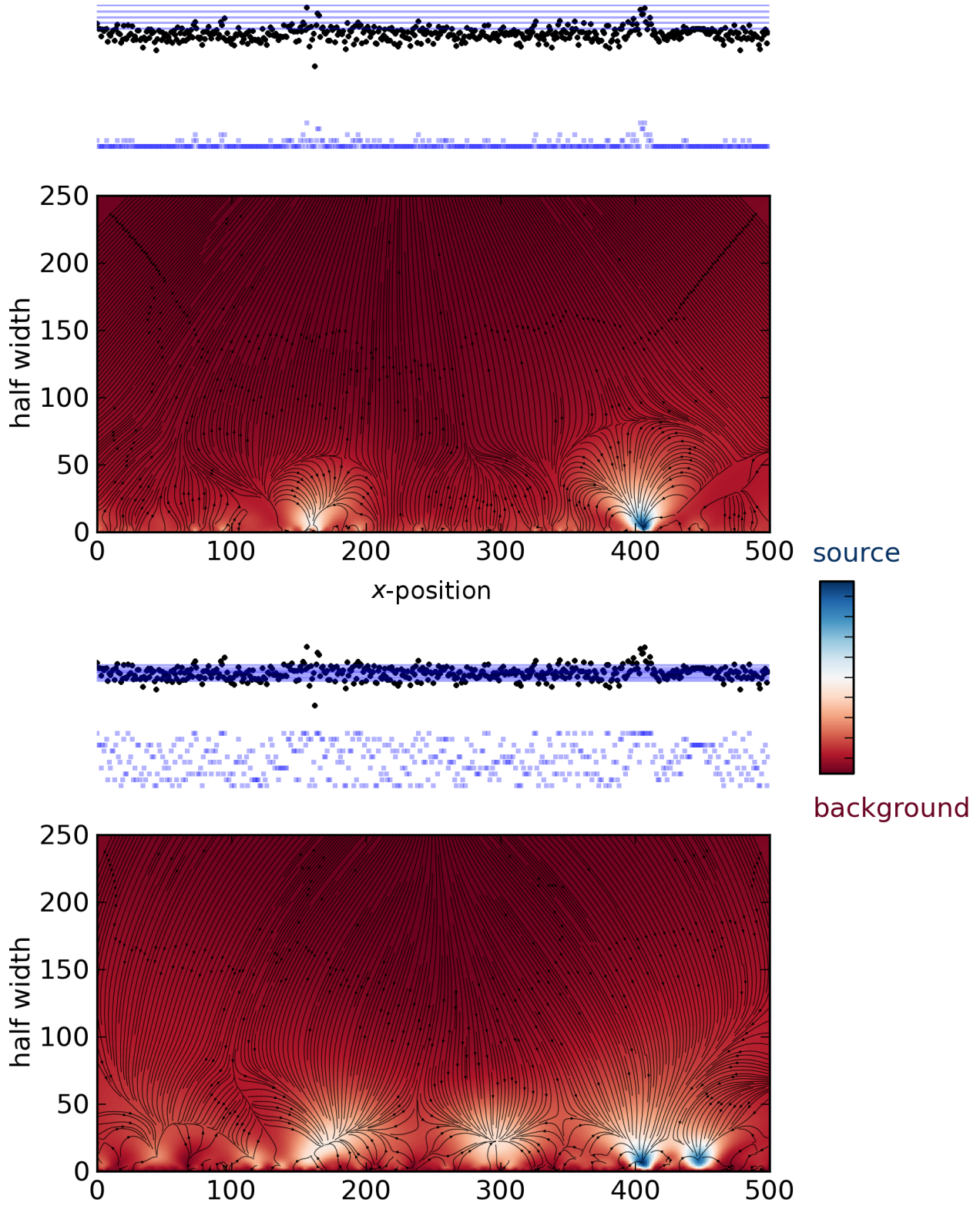


Figure 5.7: **Exhaustive calculation of the DMR and its gradient.** Three sources have parameter $m = (165, 410, 446)$. The top plot uses equal width bins, the bottom Dirichlet equal occupancy. Details are in text in Section 5.1.3. Note that the source at $m = 446$ has lower variance than is typical of the image.

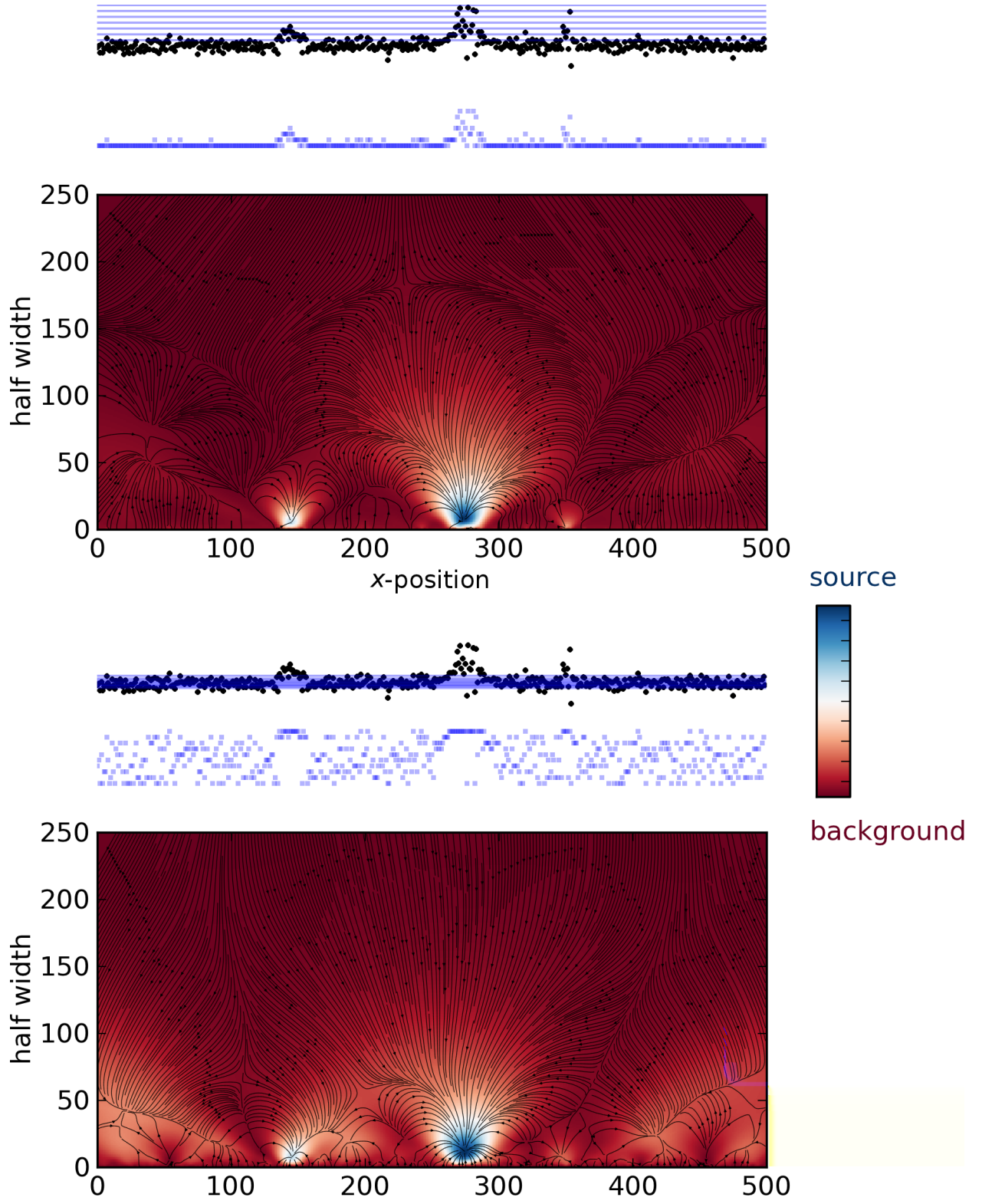


Figure 5.8: **Exhaustive calculation of the DMR and its gradient.** Three sources have parameter $m = (145, 275, 354)$. The top plot uses Dirichlet equal width bins, the bottom equal occupancy. Details are in text in Section 5.1.3.

the number of pixels N in a region \mathbf{n} gets small, the binned counts in that region are more likely to be dissimilar to the binned counts in the whole image by chance alone. In fact, where the number of pixels N in a region \mathbf{n} is smaller than the number of bins K , there is guaranteed to be at least one empty bin (a bin with a count of 0); in the extreme, a region of one pixel will have one bin occupied and $K - 1$ bins empty. Such small regions look dissimilar to the background model. As all multinomial distributions are equally favoured by the source model, these small regions are favoured by the source model. Therefore the DMR's value in such regions is artificially inflated.

This biasing of the DMR towards small regions can be seen in Figure 5.9.

To compensate for this biasing, a background correction can be added to the DMR:

$$\text{DMR}' = \text{DMR} - \text{DMR}^B \quad (5.8)$$

Where DMR^B is the Dirichlet-multinomial ratio on the vector \mathbf{n}^B , defined to be:

$$\mathbf{n}^B = \frac{\boldsymbol{\alpha}^B}{A^B} N \quad (5.9)$$

That is, the vector \mathbf{n}^B is constructed to have proportionally the same counts in bins as $\boldsymbol{\alpha}^B$, scaled by the size of the region. This therefore compensates for the effect of smaller windows having an artificially inflated DMR value by virtue of size alone.

The gradient of this corrected DMR is readily calculated:

$$\text{Gradient}' = \text{Gradient} - \text{Gradient}^B \quad (5.10)$$

as \mathbf{n}^B simply replaces \mathbf{n} in the DMR and gradient calculations, both of which are binned counts in a region.

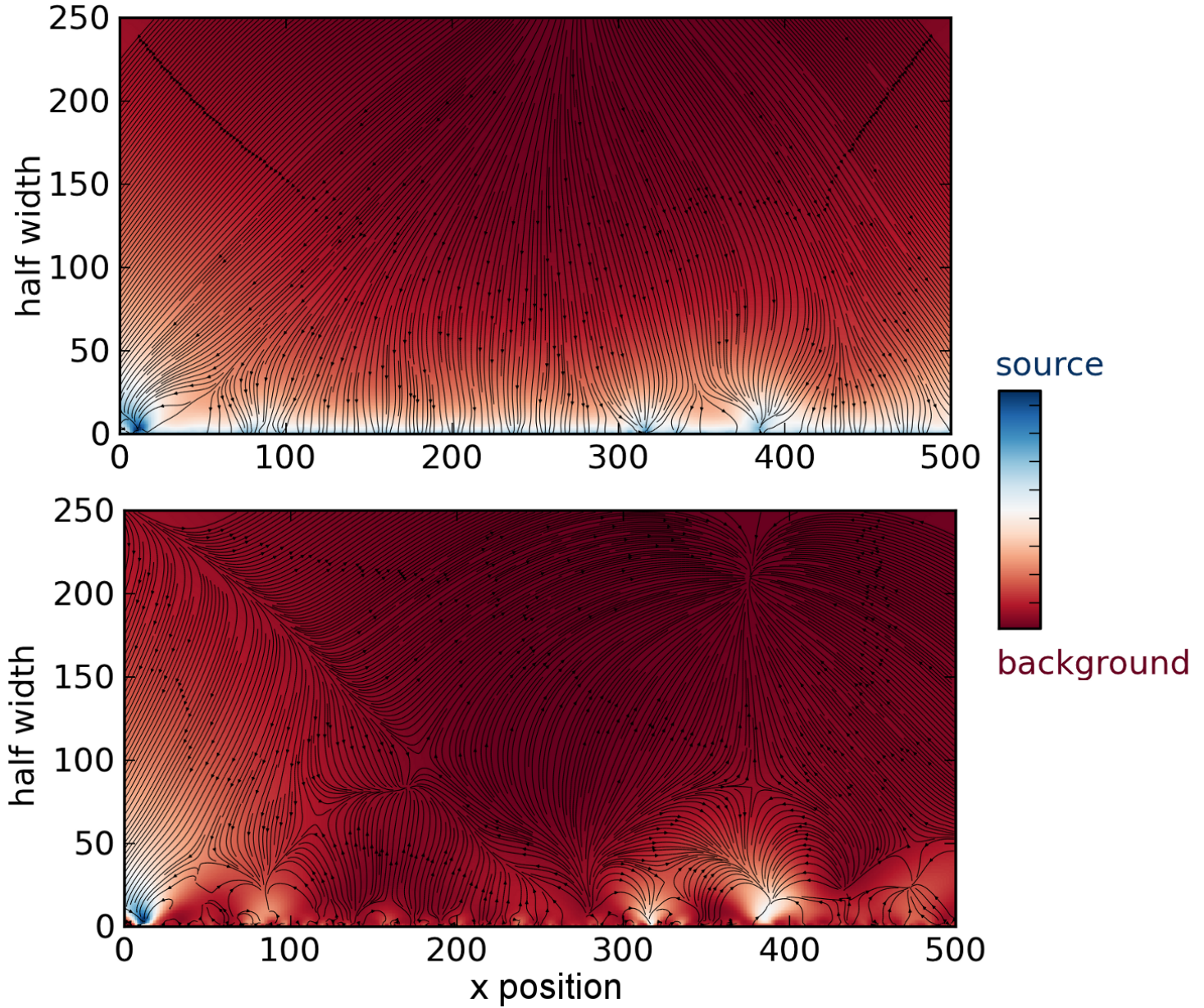


Figure 5.9: **Background compensation correction to the DMR.** Both plots show an exhaustive calculation of the DMR and its gradient over parameters $\theta = (m, \sigma)$ on simulated one dimensional data. The DMR is shown as a red to blue colour gradient, the gradient of the DMR as black lines. All parameters are the same in both plots. The upper plot shows the uncorrected DMR, while the lower shows the corrected function using Equation 5.8. Numerous spurious peaks in the DMR into which gradient lines converge can be seen in the plot of the uncorrected function as peaks at low y -values (that is, small test regions). In the corrected version these have disappeared.

This background compensation to the DMR was used for one dimensional data only; early exploration indicated that it was not needed for two dimensional data. It is possible that this is due to the fact that “small regions” in one dimensional space are more likely to have fewer pixels than such regions in two dimensional space. A thorough exploration is warranted but was not done for this thesis.

5.1.4 Gradient ascent

For one dimensional data, the gradient of the DMR was calculated using the Gaussian function parametrisation of the window function $W^{(\theta)}$ (Equation 4.5) used to define a test region. For the derivative with respect to m and σ for this parametrisation, see Equations 4.42 and 4.44.

For illustrative purposes, the gradient was exhaustively calculated over a range of values for parameters $\theta = (m, \sigma)$ on a number of images, real and generated; see Figures 5.6 to 5.8.

Similarly, for two dimensional data, the gradient of the DMR was calculated using the Gaussian function parametrisation of the window function $W^{(\theta)}$ (Equation 4.14) for a test region. The gradient with respect to θ for this parametrisation can be found in Section 4.3.1 of Chapter 4.

The `fmin_tnc` function from the Python package `scipy.optimize` [36] was used to implement gradient ascent. This function uses a Newton Conjugate-Gradient algorithm, and allows each parameter to have bounds: user-defined minimum and maximum values that a parameter can take [36, 50, 52]. For example, the parameter m_x would have its minimum value = 0 and its maximum value = X where there are X pixels in a one dimensional image.

In the case of one dimensional data, one “round” of source finding consisted of 50 trials of gradient ascent, each with random initial values for

parameters θ .

For two dimensional data 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 time). For both real and simulated two dimensional data, initial values for parameters m_x, m_y were set to the coordinates of the brightest pixel in the image⁷, with random initial values for σ_x, σ_y , and ϕ . Initialising search at the brightest point saved computational time and made use of the domain knowledge that sources are brighter than background.

The peak with the highest value in each round was recorded as a found source. That source was then removed from the data by replacing the pixels of the found source with NaNs. The pixels of the found source were defined by $m \pm 1.5\sigma$ for one dimensional data, and similarly, the pixels in the ellipse defined by the found source’s parameters $(m_x, m_y, 1.5\sigma_x, 1.5\sigma_y, \phi)$ for two dimensional data.

After each source removal, the 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 two dimensional real data, rounds of gradient ascent continued until the top-scoring source of a round had a value that indicated it was more likely to be background than source, or after 53 rounds (the maximum number of sources in any window, see Table 5.1), whichever came first.

⁷In the case where there was more than one pixel with the maximum intensity value in the image, one of these pixels was chosen at random to set the initial values for m_x, m_y .

5.1.5 Evaluation methods

Because the test regions are continuous in the image, for the purposes of evaluation the region of a found source is restricted to $m \pm 2\sigma$ for one dimensional data, and the ellipse defined by the parameters $\theta = (m_x, m_y, 2\sigma_x, 2\sigma_y, \phi)$ for two dimensional data. This is equal to 95 per cent of the area under the curve generated by the Gaussian function [72].

In the case of real data, “true sources” were those identified by the source-finding package BLOBCAT [26] with manual postprocessing by an astronomer. These sources have position parameters including the x, y coordinates of a source’s centre, and minimum and maximum pixel coordinates in both directions: $min_x, max_x, min_y, max_y$. These coordinates can be used to describe a “bounding box” around the source. Note that the sources in this ground truth catalogue are restricted to those at least 5σ above rms noise — because the DMR has no such restriction this means that any sources below 5σ above rms noise found by the DMR will be designated false positives when comparing them against the ground truth catalogue.

In the case of simulated one dimensional data, a “true source” was defined as $m \pm 2\sigma$ for Gaussian sources, and $[x_a, x_b]$ for skewed generalised Gaussian sources where x_a and x_b are the minimum and maximum pixel coordinates for the region where each pixel is $\geq 0.05 \times x_{max}$ (and x_{max} is the maximum intensity value of the curve).

Precision and recall were calculated in order to evaluate the performance of the Dirichlet-multinomial ratio. Precision is the proportion of true sources of all found sources (that is, all found sources that are really sources):

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad (5.11)$$

and recall is the proportion of found sources of all true sources in the im-

age (all true sources that are found):

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (5.12)$$

where **tp** = “true positive”, **fp** = “false positive”, and **fn** = “false negative”⁸ [56].

A true positive is defined where the location of at least 50 per cent of a found source’s pixels overlap with at least 50 per cent 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. In the case of simulated data, because there are three sources per image, and three rounds of gradient ascent producing three found sources per image, **fp** = **fn** and so precision = recall.

For simulated one dimensional and two dimensional data, precision and recall was additionally calculated over a range of proportions from $prop = 0.05$ to $prop = 0.95$ where a true positive is defined where at least the given proportion of a found source’s pixels overlap with at least that same proportion of a real source’s pixels, in order to compare the performance of different binning strategies at different thresholds. These results are shown in Figures 5.12 and 5.12.

5.2 Results and discussion

5.2.1 One dimensional data

Early exploration on one dimensional data suggested soft-bordered test regions improved the performance of the DMR as compared to hard-bordered

⁸Precision and recall are sometimes called “completeness” and “reliability” in the astronomical source detection literature [27].

Table 5.2: Performance of the DMR on simulated one dimensional data

Binning method	Precision	Recall
Equal width	0.84	0.84
Dirichlet equal width	0.77	0.77
Equal occupancy	0.68	0.68
Dirichlet equal occupancy	0.71	0.71

test regions. Figure 5.10 shows representative plots illustrating an improvement to the DMR’s performance on simulated one dimensional data using soft-bordered, as compared to hard-bordered test regions. For this reason, soft bordered test regions were used for the remainder of the analysis of the DMR.

Figure 5.11 shows an exhaustive calculation of the DMR for a one dimensional “slice” of an astronomical image, as well as the top scoring peaks of five rounds of gradient ascent, for each of the four different binning strategies used in this chapter.

Table 5.2 shows the performance of the DMR on 50 simulated one dimensional images (such as those shown in Figure 5.3), each of which contains three sources. Note that the rates for precision and recall are identical within each binning method, due to the fact that three rounds of gradient ascent were performed for each image. Therefore the number of false positives = the number of false negatives = (1 – the number of true positives). Precision and recall are moderate to high for each of the binning strategies. It is worth noting that the Dirichlet softening of bin borders improved results only for the equal occupancy strategy, and actually worsened performance in the case of equal width bin borders. An investigation into why this might be would be worth pursuing but was not done for this thesis.

Precision and recall was also calculated over a range of proportions from

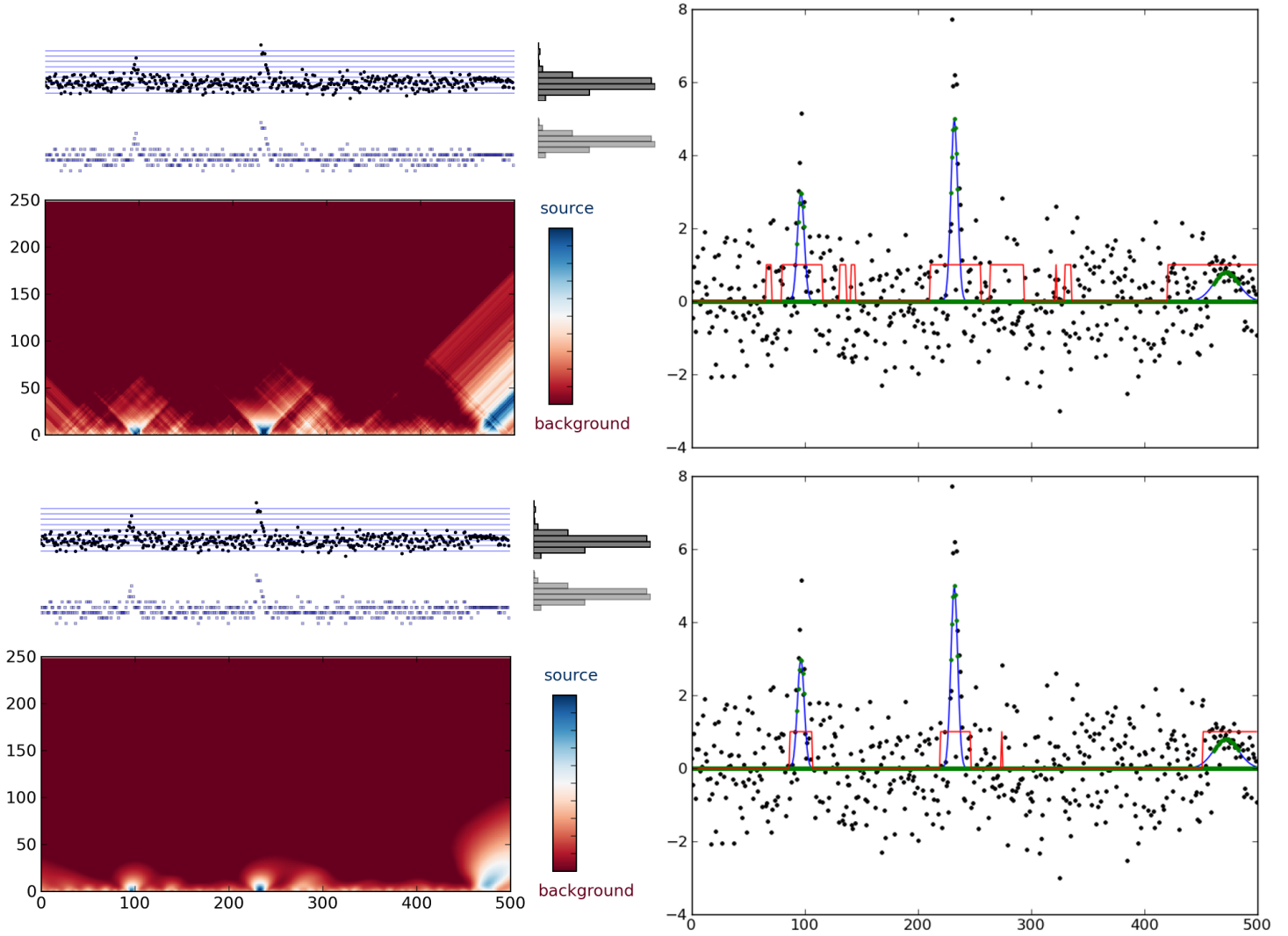


Figure 5.10: **Hard and soft borders.** Results of the DMR with hard-bordered (Equation 4.4; top) versus soft-bordered (Equation 4.5; bottom) test regions on one dimensional simulated data. The plots at left have three subplots; from top to bottom: the data (black dots) with bin borders superimposed (blue lines; with histogram binning at right), the binned data (blue dots), and a plot of the DMR over a range of values for $\theta = (m, \sigma)$. The plots at right show the data (black dots) with sources shown as blue curves (with $m \pm 1\sigma$ shown as green dots). The red lines are at 1 where the DMR's value > 0 , and 0 elsewhere. Note that number of false positives drastically reduces when using soft-bordered test regions. Note also the low variance source (far right of each plot), found by the DMR.

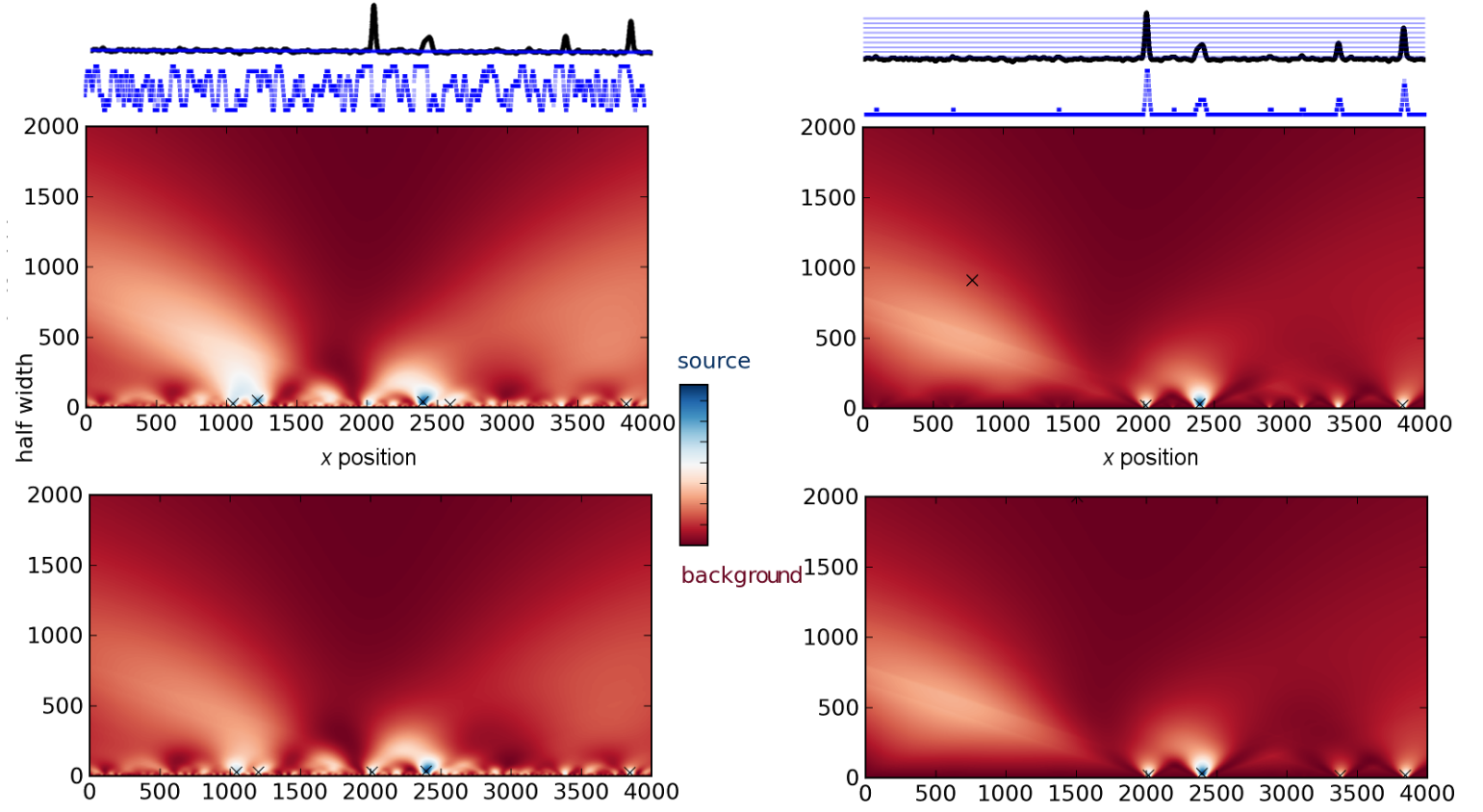


Figure 5.11: **The DMR on a one dimensional “slice” of an astronomical image** (ATLSB survey region A at 50” resolution [62, 68]; see Figure 5.1). Results using equal occupancy (top left), equal width (top right), Dirichlet equal occupancy (bottom left), and Dirichlet equal width (bottom right) histogram binning strategies. The main body of each plot shows an exhaustive calculation of the DMR over parameters $\theta = (m, \sigma)$ with $m \in [0..X]$ and $\sigma \in [0..\frac{X}{2}]$. The top scoring maxima from five rounds of gradient ascent with 50 trials each are shown as black **x** marks. In the top two plots, the top sub-plot shows the data in black with blue bin borders superimposed; the next sub-plot down is the binned data (blue dots). Note that for this example peaks in gradient best correspond to the actual peaks in data in the case of the Dirichlet equal width binning strategy.

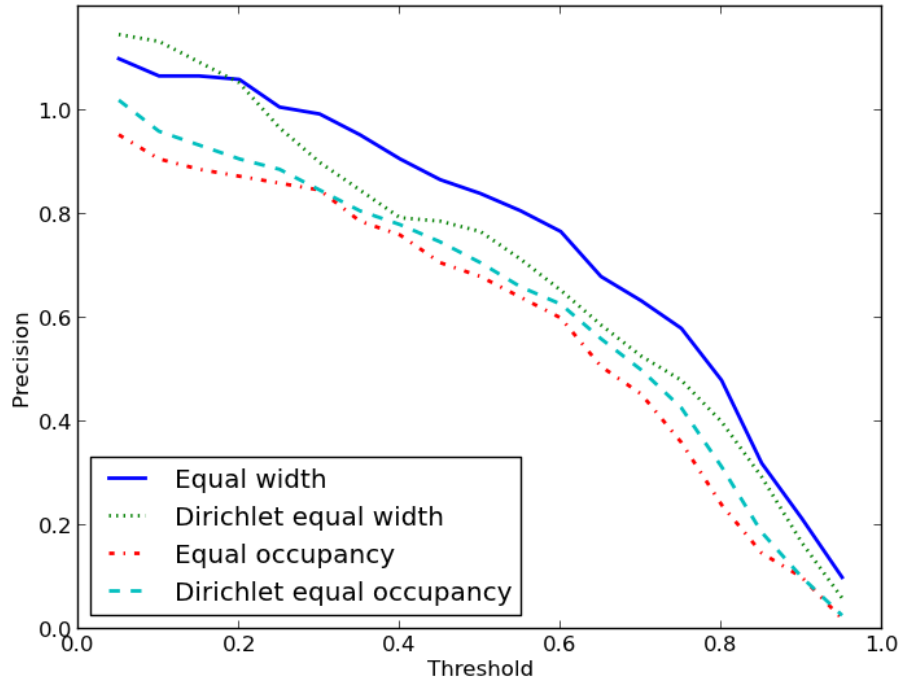


Figure 5.12: **The performance of the DMR at different thresholds** (one dimensional data, 50 simulated images, three sources each). Thresholds are defined as the proportion of a true source and an estimated source that must overlap to be defined as a true positive. For example at a threshold of 0.5, a true positive is defined where the location of at least 50 per cent of a found source's pixels overlap with at least 50 per cent of a real source's pixels. Note that the precision exceeds 1.0 for some binning strategies at thresholds ≤ 0.3 , as each estimated source may meet the definition of true positive for more than one true source at such a low threshold.

$prop = 0.05$ to $prop = 0.95$ (defining the proportion of pixels that must overlap between a true source and found source to be considered a true positive). These results are shown in Figure 5.12. The performance of each binning strategy drops off as the proportion increases. The relative performance of each of the four binning strategies remain relatively even across all thresholds, except that at very low thresholds the Dirichlet equal width binning strategy outperforms the equal width binning strategy.

5.2.2 Two dimensional data

Early exploration of the DMR on two dimensional data used simulated images (such as those shown in Figure 5.5).

Figure 5.13 illustrates an exhaustive calculation of the DMR on two dimensional 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.

Precision and recall were calculated for the 50 simulated images (with three sources per image), with results⁹ shown in Table 5.3. Values were high for three of the four binning strategies, but perplexingly low for the Dirichlet equal width binning strategy. Figure 5.14 shows the performance of the four binning strategies over a range of thresholds defining a true positive. This reveals that the poor performance of the Dirichlet equal width strategy held only at thresholds ≥ 0.4 ; under this threshold, the Dirichlet equal width strategy performed in the range of (and often better than) the other strategies.

A closer inspection of the results shows that this may be explained by the

⁹Note that as for the simulated one dimensional data, the rates for precision and recall are identical within each binning method, as three rounds of gradient ascent were performed for each image; therefore the number of false positives = the number of false negatives = $(1 - \text{the number of true positives})$.

Table 5.3: Performance of DMR on simulated two dimensional data

Binning method	Precision	Recall
Equal width	0.83	0.83
Dirichlet equal width	0.57	0.57
Equal occupancy	0.79	0.79
Dirichlet equal occupancy	0.84	0.84

Table 5.4: Performance of the DMR on real two dimensional data

Binning method	Precision	Recall
Equal width	0.51	0.74
Dirichlet equal width	0.59	0.75
Equal occupancy	0.30	0.30
Dirichlet equal occupancy	0.32	0.34

fact that Dirichlet equal width binning strategy resulted in found sources that are in most cases much smaller than the actual sources, but located in the correct position (see Figure 5.15) for examples. This may be due to the fact that equal width binning places the brightest pixels alone in bins, when these bright pixels are outliers and the vast majority of pixels lie in a small dim intensity range (as with radio astronomy images). The Dirichlet softening of equal width bin borders may shift bright pixels that are close to bin borders pixels down to dimmer bins. It is possible that the background compensation correction to the DMR, described in Section 5.1.3 may address this issue (although this correction was not applied to two dimensional data for this thesis).

The performance of the DMR on 25 windows from two large astronomical images [54] is shown in Table 5.4.

Results obtained using equal occupancy binning (with and without Dirichlet softening of bin borders) are very low, in comparison both to results obtained on simulated data, and equal width results on real data. This may

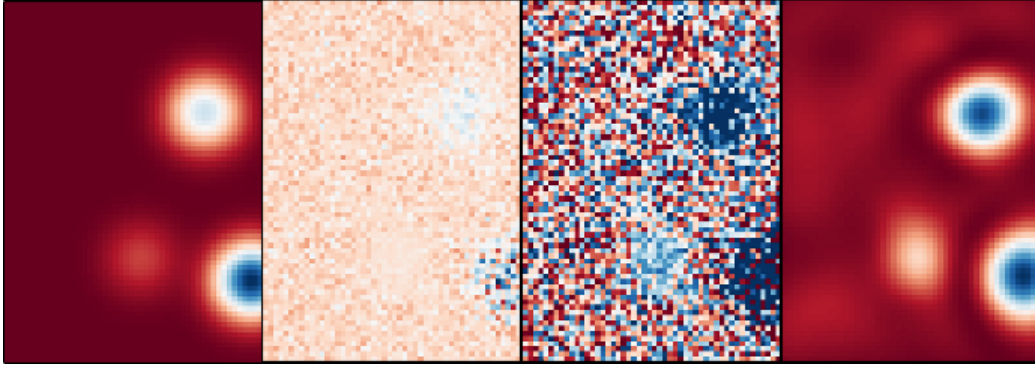


Figure 5.13: **An exhaustive calculation of the DMR** on two dimensional simulated data over x and y positions $\theta = (m_x, m_y)$ with parameters σ_x, σ_y and ϕ kept constant. From left to right: three simulated sources; the image overlaid with Gaussian noise; the noisy image binned with Dirichlet equal occupancy bins; the DMR over (m_x, m_y) calculated using these bins, shown using a colour gradient from red (background) to blue (source). Note that 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 DMR'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.

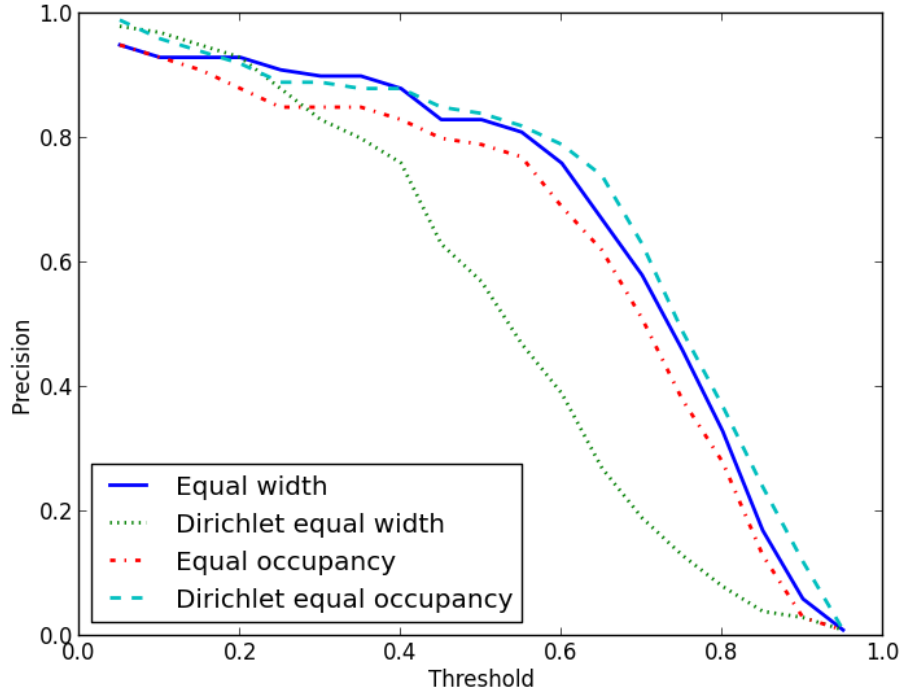


Figure 5.14: **The performance of the DMR at different thresholds** (two dimensional data, 50 simulated images, three sources each). Thresholds are defined as the proportion of a true source and an estimated source that must overlap to be defined as a true positive. For example at a threshold of 0.5, a true positive is defined where the location of at least 50 per cent of a found source's pixels overlap with at least 50 per cent of a real source's pixels.

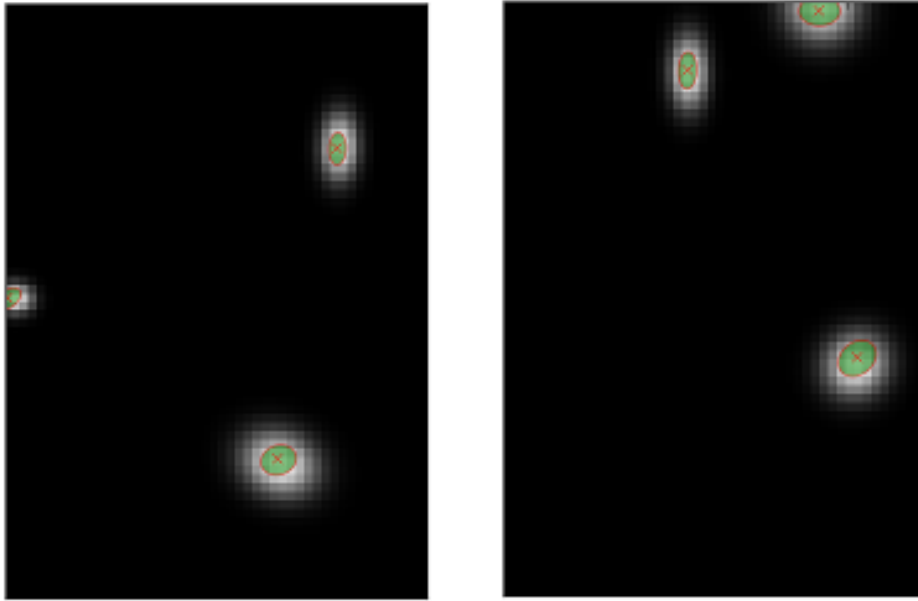


Figure 5.15: **Results of gradient ascent on the DMR** using Dirichlet equal width bins on simulated two dimensional images (shown without noise for clarity). The found sources are shown as green ellipses with a red cross at the central x, y coordinates of the found source. Note that the found sources are much smaller than the actual sources, which may account for the poor performance of this binning strategy at thresholds over 0.4, where thresholds are defined as the proportion of a true source and an estimated source that must overlap to be defined as a true positive.

possibly be attributed to the source finding process, in which each round of gradient ascent is initialised at a bright point in the data, and finishes with the removal of the found source and recalculation of the α^B -vector and bin borders. For radio astronomy data where the vast majority of pixels fall into a low intensity range with a small number of bright outliers, this process lends itself to equal width binning strategies. In such strategies, the brightest pixels will be binned alone; iterative removal of these pixels and recalculation of the bins will reveal the next brightest pixels, and so on. Figure 5.16 illustrates this point.

With regards to the equal width binning strategies, the precision and recall of the DMR on the real two dimensional data is lower than the results on the simulated one dimensional and two dimensional data. Though the rate of recall is good (0.74 and 0.75), precision, in particular, is very low (< 0.6) on the full set of sources found.

Low precision can in part be explained by the fact that the DMR found more sources than appear in the ground truth catalogue. While some of these additional found sources are spurious, some do 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 the 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 5.17.

The Dirichlet bin border softening improved results for both equal width and equal occupancy strategies on this data.

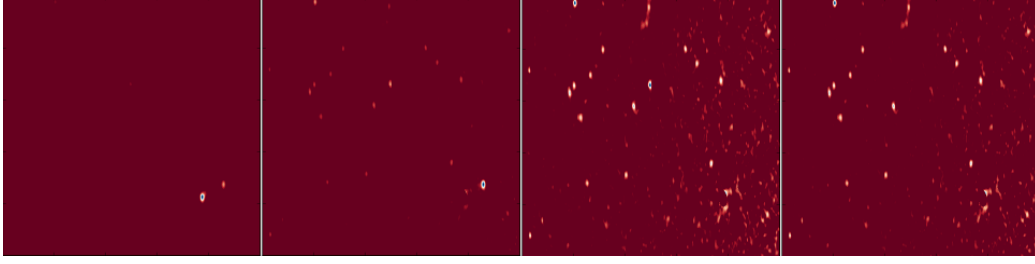


Figure 5.16: **Iterative source removal and rebinning.** After each round of gradient ascent, the found source is removed from the data and bin borders and the α^B -vector are recalculated on the remaining image. From left to right: the binned view of the initial image (a window from CDFS using Dirichlet equal width bins), the binned image after one, two and three rounds of gradient ascent. Notice that removing sources and rebinning the remaining data reveals sources that were previously hidden in background bins.

5.3 Conclusions and future work

5.3.1 Simulated data

Precision and recall rates for the DMR on both one and two dimensional simulated data were moderate to good, especially given the nature of the simulated data — care was taken to generate sources that differed from the DMR’s model for sources. In comparison, the simulated data used by other authors of Bayesian source detection methods generate sources with the same parameters as the model of source used in the detection methods. For example, many authors generate circularly symmetric Gaussian sources of a particular size, and use a circularly symmetric Gaussian of that size as a source model (see: [20, 29, 63]). Using the same parameters for simulated source and the method’s source model is likely to artificially inflate the success of the model, and is unjustified in the case of radio astronomy data, where sources can take a great variety of shapes

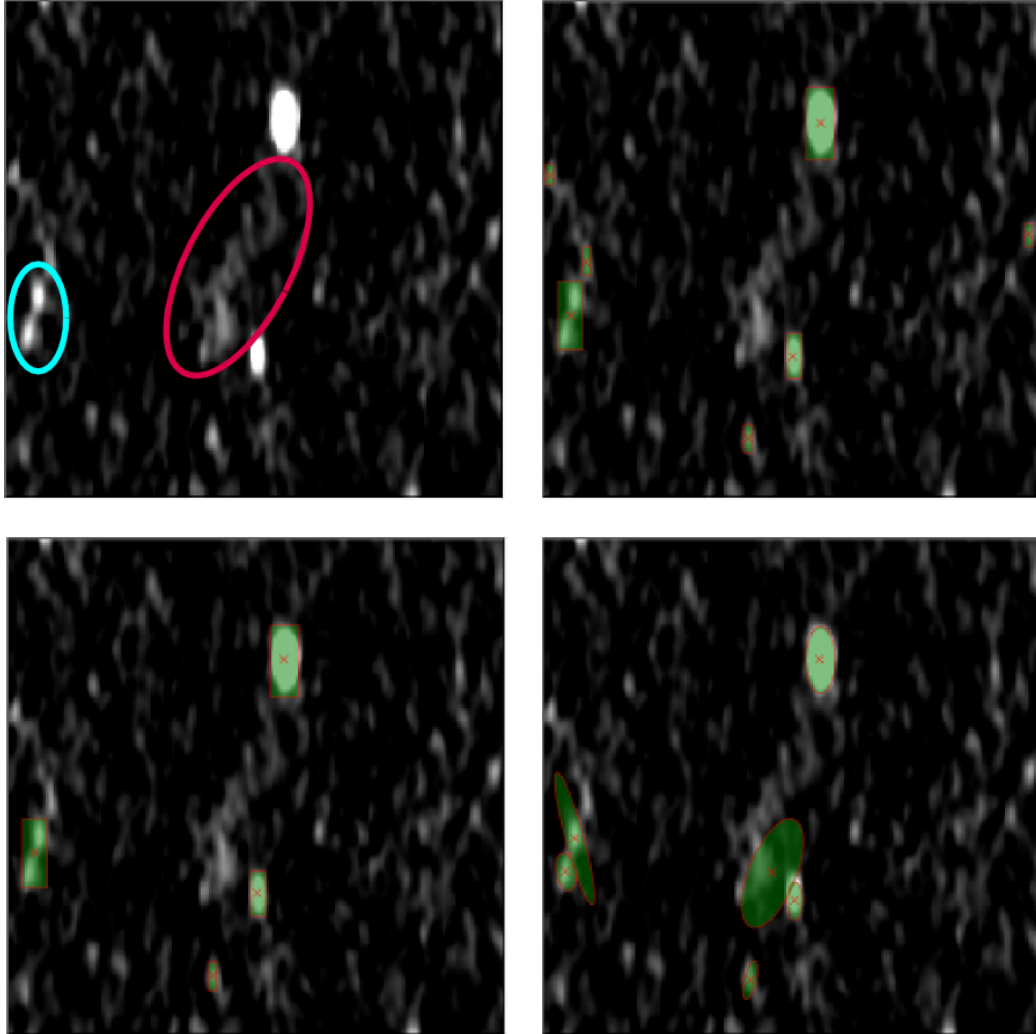


Figure 5.17: **Comparison with the ground truth catalogue.** A section of one of the CDFS test-windows is shown at top left. Top right shows the sources identified by BLOBCAT (raw, unprocessed output). Bottom left shows the sources in the ground-truth catalogue, and bottom right the sources identified by the DMR. Note that the DMR has identified all the sources in the ground truth catalogue, and in addition, has found part of a low brightness source (radio galaxy tail; pink ring). The DMR has also correctly identified two separate sources that the ground truth catalogue identified as one source (blue ring), though one of the DMR's width parameters for the upper source is incorrect.

and sizes.

A particular strength of the DMR is illustrated in Figures 5.7 and 5.10: the identification of low variance sources. The ability to set the source model to an α -vector that is completely “agnostic” creates a broad definition of non-background like regions in an image that allows a great deal of flexibility in the use of the DMR to find different types of features in astronomical images.

With regards to the background-compensation correction to the DMR, initial investigation indicates that this correction improves results for one dimensional but not two dimensional data. The reasons for this are not clear and further investigation is warranted.

5.3.2 Real data

The performance of the DMR using the Dirichlet equal width binning strategies on 25 windows from two large astronomical images yielded good recall (0.75), and moderate precision (0.59).

This moderate precision can in part be attributed to the fact that the ground truth catalogue was restricted to sources at least 5σ above rms noise. Though some of the DMR’s false positives were regions of noise, many were real sources below this threshold (Figure 5.17).

As noted earlier in this chapter, some of the most scientifically important objects in astronomy are dim, with intensities in the range of background noise [53]. The nature of pixel-intensity based thresholding algorithms such as BLOBCAT restricts their ability to find such dim sources without also finding such a large number of noise regions that the results are unusable.

The DMR does not restrict the objects found on the basis of some threshold above rms noise, and produces raw results containing most real ob-

jects, including dim sources not well found by other other algorithms, while returning only a moderate number of false positives in regions of noise.

Though further development of the DMR is needed, the DMR finds sources well and is in many cases better at finding scientifically important dim sources. It is approaching a usable algorithm that does not require the intensive manual parameter tuning of existing algorithms — the need for source finding software with a minimum of manual tuning is pressing, given the huge volumes of data to be produced by the next generation of radio astronomy telescopes [53].

A significant drawback of the implementation of the DMR in this thesis however, is that the gradient ascent process was too slow to be practical (or even possible) on full-sized images. Future work to improve this is needed. One aspect that could be improved to speed up this process is the use of a test region with finite support. The use of a Gaussian distribution for test regions mean that regions extend potentially across the whole image, slowing down the gradient ascent process. This was also an issue for at least one other Bayesian source detection algorithm [20] for which investigation into alternative distributions to the Gaussian is ongoing.

An issue with the process within which the DMR is used in this chapter is that the iterative removal of found sources sometimes leaves artefacts around the removed source. This can cause multiple “found sources” where in fact only one source lies (see Figure 5.18 for an example). Addressing this issue may reduce the number of false positives caused by multiple found sources on one true source.

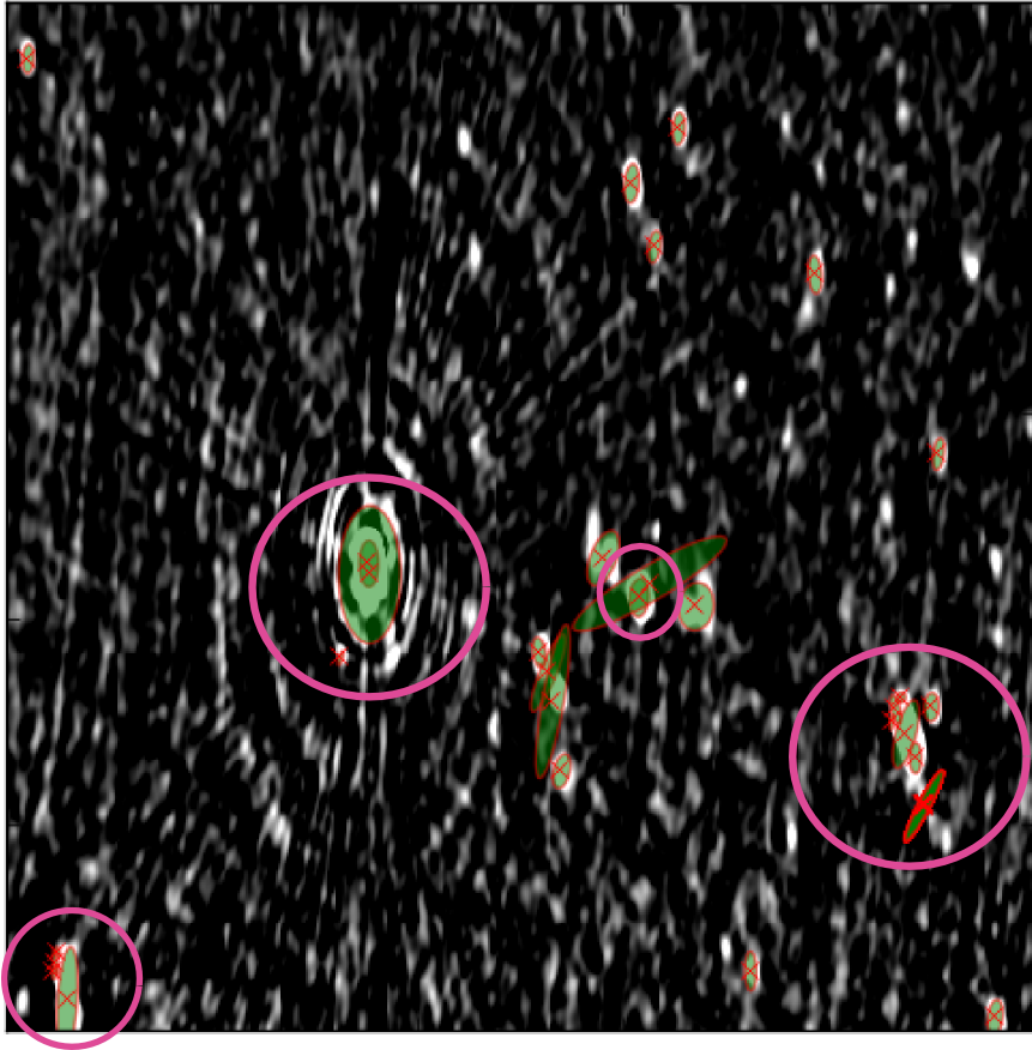


Figure 5.18: **False positives due to multiple peaks on one source.** Iterative removal of sources can leave artefacts around the removed source, which can cause the DMR to find more “sources” than there actually are in an image. The pink circles indicate regions where this has occurred.

5.3.3 Binning strategies

The results of the DMR in this chapter are inconclusive with regards to which binning strategy is best, and whether or not the novel method for incorporating uncertainty about bin border locations — Dirichlet softening — improves the performance of the equal width and equal occupancy binning strategies.

While the width-based strategies were superior for simulated one dimensional and real two dimensional data (improving results more than two-fold for the latter), the occupancy-based strategies yielded better results in the case of simulated two dimensional data. A fuller comparison of binning strategies is warranted.

In the case of simulated data, the Dirichlet softening of bin borders improved performance for equal occupancy binning but deteriorated performance for equal width binning strategies (however, note the impact of changing the threshold at which true positives are defined; Figure 5.14). For real data, Dirichlet softening improved both equal width and equal occupancy strategies.

The Dirichlet softening of bin borders in this chapter used a symmetric α -vector with $\alpha_i = 10$. The Dirichlet with this α -vector produces more uniform distributions over K bins than one with values < 10 , for example. Lower α values would introduce more variation into the location of bin borders. The effects of different α -vectors in softening bin borders is worth investigating further.

5.3.4 Potential extensions to the DMR

The second term in the DMR (Equation 4.27) weights the likelihood of a region being source or background by the relative proportion of each in the image. This term was set heuristically in this chapter, to reflect the

prior belief that astronomical images are dominated by background pixels. However, the DMR could be extended to incorporate a Dirichlet hyperparameter over the source and background mixing proportions in an image or regions within that image. The mixing proportions could be modelled as a multinomial distribution, similar to the mixing proportions over topics in LDA (Chapter 3).

In this chapter, the α^S -vector of the DMR was set to the symmetric α -vector with $\alpha_i = 1$, in order to give equal likelihood to all multinomial distributions over K for source regions (so that source regions were defined as “non-background”). It would be useful to explore ways of setting the α^S -vector such that particular distributions are favoured over others. Chapter 6 explores one such setting of the α^S -vector, by using the source and topic distributions output by LDA as α -vectors in the DMR.

Chapter 4 included a discussion of how the DMR could be extended to higher dimensional data. In practical terms, however, this would not be possible until the speed of the gradient ascent is addressed.

Chapter 6

The Dirichlet-multinomial ratio using LDA output

This chapter presents an exploration into combining latent Dirichlet allocation (Chapter 3) with the Dirichlet-multinomial ratio (Chapters 4 and 5) on astronomical images discretised by histogram binning (Chapter 2).

LDA and the DMR were combined by using the unnormalised source and background distributions¹ derived by LDA as the α^S and α^B parameters in the DMR. These unnormalised topic distributions are the per-bin counts of pixels for each topic, as estimated by Gibbs sampling.

Because the Dirichlet distribution is the conjugate prior to the multinomial distribution², the unnormalised topics can be viewed as pseudocounts in K bins, which can be treated as an α -vector from a Dirichlet distribution³ [40, 51, 67].

¹Where source and background distributions are multinomial distributions over bins.

²That is, if a data point x comes from a multinomial distribution, and the prior for its parameter $p(\theta)$ comes from a Dirichlet distribution, then the posterior distribution of its parameter given the data point $p(\theta|x)$ also comes from a Dirichlet distribution.

³This is how the topic distributions are treated in the Gibbs sampling process in LDA.

6.1 Methods

Test data comprised five windows of 500×500 pixels from the two large astronomical images [54], chosen randomly from the set of 25 used for the evaluation of the DMR in Chapter 5. Sources were labelled by the source-finding package BLOBCAT [26] with manual postprocessing by an astronomer. The windows are shown in Figure 6.1.

The images were discretised by the equal width histogram binning method described in Chapter 2.

Latent Dirichlet allocation, as described in Chapter 3, was run to extract multinomial distributions over bins for source and background.

These distributions over bins were then used as α^S and α^B parameters in the DMR, and the overall topic proportions derived by LDA were used in place of the second term in Equation 4.27.

In contrast to the symmetric α^S -vector with $\alpha_i = 1 \ \forall i$ used in the DMR in Chapter 5, the α^S -vector derived by LDA is not symmetric, but skewed towards brighter pixel intensities relative to the α^B -vector (while still having counts in the darker intensities). In the DMR in Chapter 5, all multinomial distributions over K bins are equally likely to be drawn from the Dirichlet distribution for source, while the α^S -vector derived by LDA is more likely to produce distributions with higher likelihood for brighter pixels.

Similarly, the α^B -vector derived by LDA is skewed towards darker pixels, even more so than using the whole binned image as the α^B -vector, which includes source pixels as well as background pixels.

Gradient ascent, as described in Chapter 5, was performed to find peaks in the DMR.

Sources found at the end of each round of gradient ascent were removed from the data, but in contrast to the gradient ascent procedure for DMR in

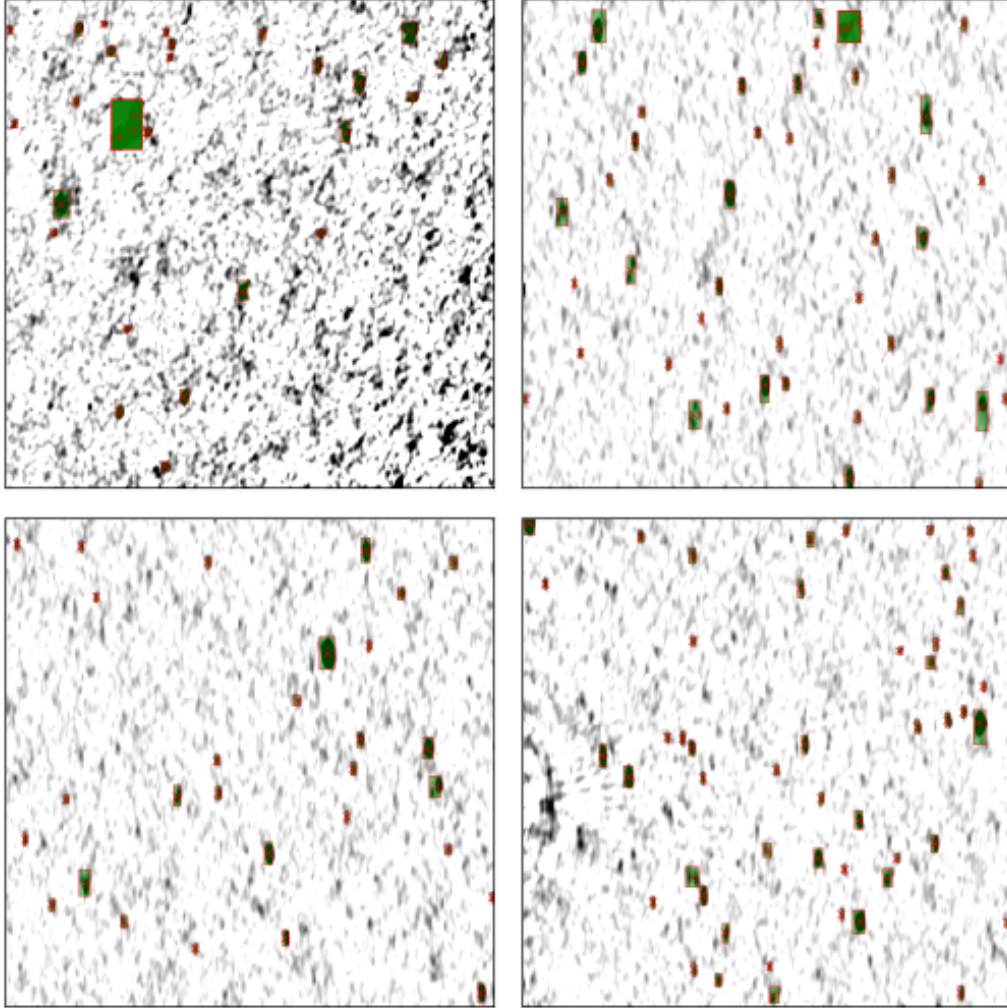


Figure 6.1: **Astronomical images used for evaluation.** The central coordinates and a bounding box around each ground-truth source shown in green. The images are 500×500 windows from ELAIS (top left: ELAIS W1) and CDFS (top right: CDFS W1, bottom left: CDFS W2, bottom right: CDFS W3) [54]. Images are contrast adjusted with colour inverted to show sources.

Chapter 5 the α^B -vector and bin borders were not recalculated. In Chapter 5 the entire image was used as a proxy for background. In this context, removing source pixels from the α^B parameter should improve the background model. However if the background distribution derived from LDA is regarded as a good model, the α^B parameter should not need to be recalculated.

The performance of the combined method was compared to the performance of the DMR using equal width bin borders⁴ using the procedure for real two dimensional data described in Chapter 5.

Precision and recall were calculated as in Chapter 5, with the same definitions for true positives, false positives and false negatives.

6.2 Results

Some preliminary results of this technique are presented in Table 6.1. In general, precision and recall are higher than the results for the DMR alone (that is, the DMR without LDA topic distributions as α -vectors), as well as higher than the results in Chapter 5. Given the small sample size of five images, this can only be taken as a preliminary indication of success of the combined method. More investigation and comparison between the two methods, and of the DMR with different α^S and α^B -vectors, is needed.

In particular, precision for the DMR with LDA output is much higher than the precision for the DMR alone. Note however that the ground truth catalogue included only sources at least 5σ above rms noise. The combined method in this chapter therefore found *no* sources below this threshold

⁴The equal width binning strategy is the strategy used for LDA in this chapter. Using the same binning strategy to compare the results for the DMR using LDA output with the DMR as described in Chapter 5 allows a fair comparison of the two methods.

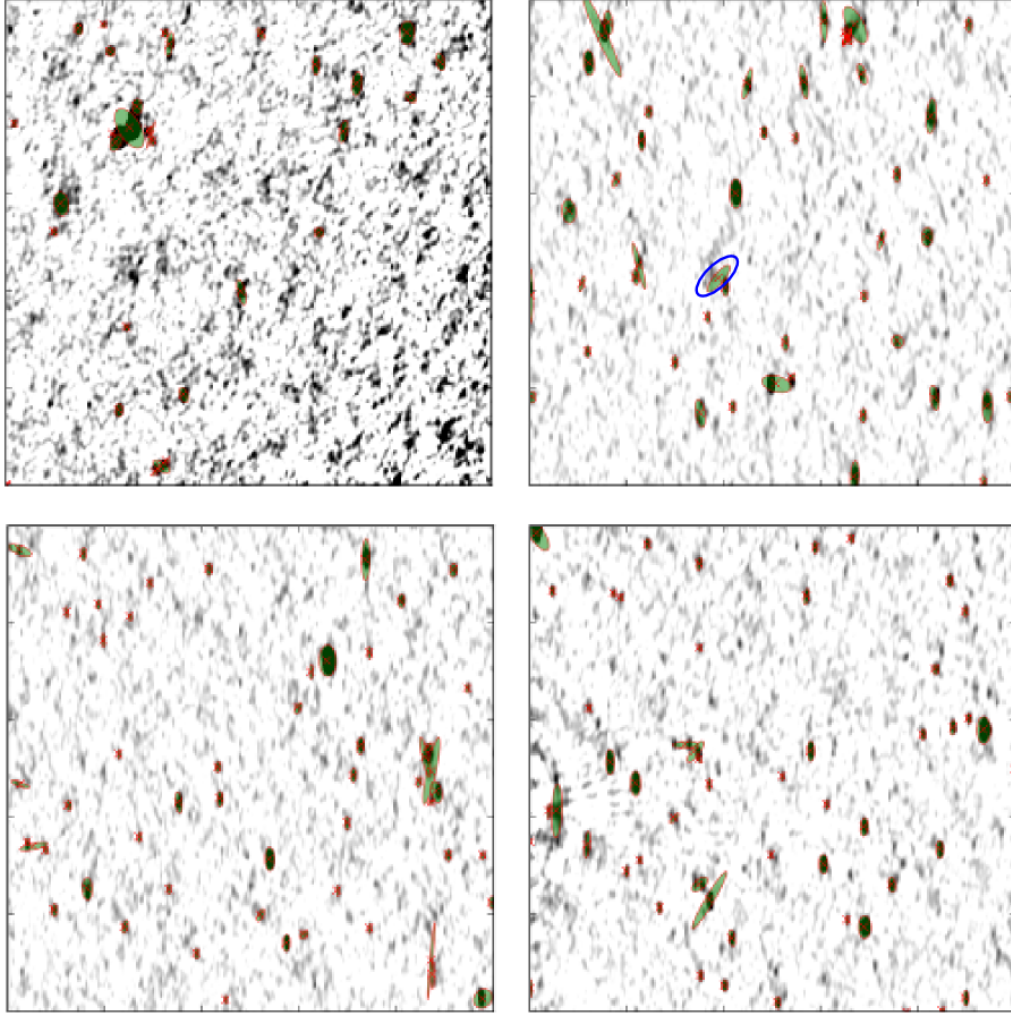


Figure 6.2: **Found sources.** Sources found by the DMR with α parameters set using source and background topics derived by LDA. The sources in the ground-truth catalogue in each image are shown in Figure 6.1. An example of a low surface brightness source found by the DMR but not in the ground truth catalogue is circled in blue.

Table 6.1: Performance of DMR using LDA output

Subwindow	DMR alone		using LDA output	
	Precision	Recall	Precision	Recall
ELAIS W1	0.71	0.59	0.87	0.90
CDFS W1	0.75	0.91	0.77	0.93
CDFS W2	0.51	0.93	0.55	1.0
CDFS W3	0.66	0.71	0.70	0.76
CDFS W4	0.29	0.86	1.0	0.71
Total (all windows)	0.58	0.85	0.72	0.87

in window CDFS W4, which yielded precision of 1.0 on the ground truth catalogue.

For the other images, however, in which precision is lower, there are examples of sources identified that are missing from the ground truth catalogue (though in some cases noise was identified as sources). Figure 6.2 shows a low surface brightness radio galaxy tail, with intensity less than 3σ above rms noise, identified by the methods in this chapter but missing from the ground truth catalogue.

6.3 Discussion and future work

The preliminary results in Table 6.1 show good results for recall and precision, much higher than DMR alone.

Like the DMR in Chapter 5, while some areas of noise were falsely identified as sources, many so called “false positives” — identified by the combined method in this chapter but not in the ground truth catalogue — are actual sources that are missing from the catalogue.

The ground truth catalogue contains only sources of intensity at least 5σ

above rms noise. The nature of pixel-intensity based thresholding algorithms such as BLOBCAT restricts the ability of the algorithms to find dim sources without also finding such a large number of noise regions that the results are unusable.

However, some of the most scientifically important objects in astronomy are dim, with intensities in the range of background noise [53]. Figure 6.2 shows an example of a low surface brightness object identified by the methods in this chapter but missing from the ground truth catalogue. This is an object that would likely to also be missed by LDA alone — the implementation of LDA in Chapter 3 had high rates of precision and recall for sources with high intensity pixels, but, as concluded in that chapter, was unlikely to identify low surface brightness sources.

In contrast to the low surface brightness sources identified, the combined method in this chapter found *no* sources below the 5σ threshold in window CDFS W4. This suggests that the values with which the α -vectors are set could be possibly manipulated to set thresholds in the source finding task.

Given the encouraging results of this preliminary investigation into the combination of LDA and the DMR, more experiments would be worthwhile pursuing.

Chapter 7

Conclusions

7.1 Summary

This thesis explored Bayesian methods for source detection that use Dirichlet or multinomial models for pixel intensity distributions in discretised radio astronomy images.

- A novel image discretisation method that incorporates uncertainty about how the image should be discretised was developed.
- Latent Dirichlet allocation — a method originally developed for inferring latent topics in document collections — was used to estimate source and background distributions in radio astronomy images.
- A new Dirichlet-multinomial ratio, indicating how well a region conforms to a well-specified model of background versus a loosely-specified model of foreground, was derived.
- Latent Dirichlet allocation and the Dirichlet-multinomial ratio were combined for source detection in astronomical images.

The methods developed in this thesis perform source detection well in comparison to two widely-used source detection packages, and importantly, find scientifically important dim sources not well found by other algorithms, while returning only a moderate number of false positives in regions of noise.

7.2 Future work

This thesis included initial explorations of a number of techniques. The results of these early experiments reveal some of the shortcomings of the techniques and suggest a number of avenues for future work. These are summarised below.

7.2.1 Image discretisation

In this thesis, only simple width and frequency based histogram binning methods (with and without the Dirichlet softening of bin borders) were considered. These approaches may be problematic when the underlying distribution of pixel intensities is not uniform [42], as is the case with radio astronomy images (Figure 2.1). These strategies can consign pixels that would otherwise be discriminatory into histogram bins such that their informativity is destroyed. For example, low brightness source pixels may become lost among background pixels.

Bright outliers can also skew the binned representation of the data¹.

¹In the case of equal width bins with iterative source removal and rebinning, as in the DMR, this may actually be an advantage. Isolating the brightest pixels highlights them, and removing the sources that contain these pixels allow the next brightest pixels to be revealed and so on until (ideally) all sources have been removed and only background pixels remain. This may explain the success of the equal width binning schemes in comparison to the equal occupancy schemes on astronomical images in Chapter 5.

It would be worth exploring other, potentially more principled, approaches to image discretisation. A binning scheme that preserves the underlying distribution of the data would be worth investigating. Such discretisation methods could be evaluated on the basis of whether using them improved source detection results in techniques such as LDA and the DMR.

With regards to the Dirichlet bin-border softening developed in this thesis, more exploration of the α -vector used to create the borders should be undertaken. The symmetric α -vector with $\alpha_i = 10$, which produces fairly uniform distributions ([21]; Figure 2.2) was used in all cases of Dirichlet softening in this thesis. Symmetric vectors with different values could be explored, as could asymmetric vectors with values that cause some bin borders to be more variable than others.

7.2.2 Latent Dirichlet allocation

Though there were some good early results of LDA, with LDA performing similarly to *Duchamp* in detecting sources for which the total flux (intensity) is less than 1.63 mJy, the implementation of LDA in this thesis is insufficient to find the scientifically important sources with intensities close to background. There are a number of avenues for exploration that could potentially address this issue.

In many cases, final background and source topics derived by LDA were simple thresholds in bins, with the background distribution over bins placing almost all weight on bins from 0 to some bin i , and the source distribution placing almost all weight from bin i to bin K . In this way, the implementation of LDA in this thesis is no better than pixel-intensity based thresholding algorithms that restrict search to those sources that contain pixels above a particular threshold.

This thresholding behaviour may be due to the binning scheme used (equal

width binning), and so an investigation into a better binning scheme may help. The implementation of LDA in this thesis does not allow for the use of Dirichlet bin borders, as the Gibbs sampling algorithm on the conditional distribution of topic allocations does not work with partial counts. The use of Dirichlet bin borders in LDA should be investigated.

Thresholding may also be due to the Dirichlet priors over topic distributions, which have α values < 1 . Such values produce sparse distributions ([21]; Figure 2.2) over bins; using an α -vector that encourages more even distributions over bins may help address the thresholding observed.

Pixels are treated as fundamental units (words) in the implementation of LDA in this thesis. As the resolution element of a radio telescope is generally several pixels, individual pixels may be better thought of as having sub-word size (perhaps “syllables”). The implementation of LDA may be adjusted accordingly.

The use of the final topic distributions derived by LDA — segmentation by assigning each pixel a hard topic label and source detection by flood-filling on the segmented image — is crude. A more nuanced approach would eliminate this hard assignment and take a more probabilistic approach to region labelling. For example, given the multinomial models for background and source, gradient ascent could be performed to find regions that have high likelihood under a particular model.

7.2.3 The Dirichlet-multinomial ratio

Again, an investigation into the optimal binning scheme is needed, as no clear conclusions can be drawn on this point given the results of the DMR in this thesis.

The α -vectors in the DMR could similarly be investigated. In this thesis, the α^S -vector was set to the symmetric α -vector with $\alpha_i = 1$, in order to

give equal likelihood to all multinomial distributions over K for source regions (defining source regions as any “non-background” region). It would be useful to explore ways of setting the α^S -vector such that particular distributions are favoured over others.

With regards to the background compensation correction to the DMR, initial investigation indicated that this correction improves results for one dimensional but not two dimensional data. The reasons for this are not clear and further investigation is warranted.

Procedural issues

The gradient ascent process used in this thesis was too slow to be practical (or even possible) on full-sized large images. If the DMR is to be usable on the large amounts of data produced by next generation telescopes, speed must be addressed. The use of a Gaussian distribution for test regions meant that regions extend potentially across the whole image, slowing down gradient ascent. One aspect that could be improved to speed up the gradient ascent process therefore is the use of a test region with finite support.

The iterative removal of sources found by the DMR sometimes left artefacts around the removed source. This can cause multiple “found sources” where in fact only one source lies (Figure 5.18). Addressing this issue may reduce the number of false positives the DMR yields.

Extensions

With regards to future extensions, the second term in the DMR (Equation 4.27) weights the likelihood of a region being source or background by the relative proportion of each in the image. This term is set heuristically in

this thesis, to reflect the prior belief that astronomical images are dominated by background pixels. However, a Dirichlet hyperparameter could be placed over the source and background mixing proportions in an image, and inferred, rather than set heuristically.

7.2.4 The Dirichlet-multinomial ratio using LDA output

Early experiments into the combination of LDA with the DMR suggest better performance than either algorithm alone. More evaluation to confirm this result would be worthwhile.

On the other hand, like LDA, the combination of LDA with the DMR seems to miss a number of low intensity sources. As with LDA, an investigation into an optimal binning scheme may address this.

Though sources were iteratively removed with this version of the DMR, there was no recalculation of the α -vectors or of the bin borders. Doing so may improve performance on dim sources. However unlike the DMR implementation in Chapter 5 where the initial α^B -vector and bin borders were calculated using the whole image as a proxy for background, there is no readily apparent way of performing this iterative recalculation. One possible approach would be to simply reduce the counts in the α^S and α^B -vector bins corresponding to the bins of the pixels of the removed source. Further exploration into rebinning schemes would be worthwhile.

7.2.5 From exploration to usable software

The need for new tools for source detection in radio astronomy is pressing: the sheer volume of data to be produced by next generation radio telescopes makes automated methods for the detection of astronomical objects essential. As discussed in the opening of this thesis, current meth-

ods require time-intensive parameter tuning and do not find all sources, particularly scientifically interesting low surface brightness objects.

A number of steps would be required to take any of the methods discussed in this thesis from exploration stage to software for public use. More testing and validation of the algorithms would be required on a greater set of test data.

Both Gibbs sampling (in latent Dirichlet allocation) and the gradient ascent process (DMR) were much too slow to be practically useful. Significant speed gains would need to be made before building software using these algorithms.

Once these issues are addressed, a software package using any combination of the algorithms described in this thesis could start to be developed. Design decisions would need to be made, including what parameters would be tunable by the end-user and how they may be tuned (e.g. in predefined steps or absolutely). A user-interface (whether command-line, graphical, browser-based, etc) would need to be designed and built. Any software that is developed would need to be thoroughly tested before being released publicly.

There are many existing algorithms that are sufficient to find point sources, which are abundant in radio astronomy images [27]. With this in mind, it may be prudent to make, for example, the DMR part of a pipeline in which one of the existing faster and simpler algorithms is first used to identify and remove all (or most) point sources in an image, with the DMR then used to find spatially extended and faint objects.

Bibliography

- [1] BANFIELD, J. K., FRANZEN, T. M. O., HOPKINS, A., NORRIS, R. P., SEYMOUR, N., CHOW, K. E., HALES, C., HUYNH, M. T., LENC, E., MAO, M., AND MIDDLEBERG, E. The Australia Telescope Large Area Survey I: 1.4 GHz source catalogue and counts, in prep.
- [2] BERTIN, E., AND ARNOUTS, S. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series* 117, 2 (1996), 393–404.
- [3] BLEI, D. Probabilistic topic models. *Communications of the ACM* 55, 4 (2012), 77–84.
- [4] BLEI, D., AND LAFFERTY, J. Topic models. In *Text mining: Classification, Clustering, and Applications*, A. Srivastava and M. Sahami, Eds., vol. 10. Chapman & Hall/CRC, 2009, pp. 71–94.
- [5] BLEI, D., AND MCAULIFFE, J. Supervised topic models. *Neural Information Processing Systems* 21 (2007).
- [6] BLEI, D., NG, A., AND JORDAN, M. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3 (2003), 993–1022.
- [7] BREWER, B. J., FOREMAN-MACKEY, D., AND HOGG, D. W. Probabilistic catalogs for crowded stellar fields. *The Astronomical Journal* 146, 1 (2013), 7.

- [8] CAO, L., AND FEI-FEI, L. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (2007), IEEE, pp. 1–8.
- [9] CARVALHO, P., ROCHA, G., AND HOBSON, M. A fast Bayesian approach to discrete object detection in astronomical data sets—Powellsnakes I. *Monthly Notices of the Royal Astronomical Society* 393, 3 (2009), 681–702.
- [10] CHMIELEWSKI, M. R., AND GRZYMALA-BUSSE, J. W. Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning* 15, 4 (1996), 319–331.
- [11] CLARKE, E. J., AND BARTON, B. A. Entropy and MDL discretization of continuous variables for Bayesian belief networks. *International Journal of Intelligent Systems* 15, 1 (2000), 61–92.
- [12] DIACONIS, P., AND YLVISAKER, D. Conjugate priors for exponential families. *The Annals of Statistics* 7, 2 (1979), 269–281.
- [13] DICKEY, J. M. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association* 78, 383 (1983), 628–637.
- [14] DOUGHERTY, J., KOHAVI, R., AND SAHAMI, M. Supervised and unsupervised discretization of continuous features. In *Proceedings of the Nineteenth International Conference on Machine Learning* (1995), Morgan Kaufmann Publishers, Inc., pp. 194–202.
- [15] DOYLE, G., AND ELKAN, C. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), ACM, pp. 281–288.

- [16] ELKAN, C. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 289–296.
- [17] FEAIN, I., CORNWELL, T., EKERS, R., CALABRETTA, M., NORRIS, R., JOHNSTON-HOLLITT, M., OTT, J., LINDLEY, E., GAENSLER, B., MURPHY, T., ET AL. The radio continuum structure of Centaurus A at 1.4 GHz. *The Astrophysical Journal* 740, 1 (2011), 17.
- [18] FEI-FEI, L., AND PERONA, P. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), IEEE, pp. 524–531.
- [19] FERGUS, R., FEI-FEI, L., PERONA, P., AND ZISSERMAN, A. Learning object categories from Google’s image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (2005), IEEE, pp. 1816–1823.
- [20] FERROZ, F., AND HOBSON, M. Multimodal nested sampling: an efficient and robust alternative to Markov chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society* 384, 2 (2008), 449–463.
- [21] FRIGYIK, B. A., KAPILA, A., AND GUPTA, M. R. Introduction to the Dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washington, UWEETR-2010-0006* (2010).
- [22] GEMAN, S., AND GEMAN, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6 (1984), 721–741.
- [23] GONZALEZ, R., AND WOODS, R. *Digital Image Processing*, second ed. Prentice Hall, New Jersey, 2002.

- [24] GUGLIELMETTI, F., FISCHER, R., AND DOSE, V. Background–source separation in astronomical images with Bayesian probability theory–I. the method. *Monthly Notices of the Royal Astronomical Society* 396, 1 (2009), 165–190.
- [25] GULL, S., AND DANIELL, G. Image reconstruction from incomplete and noisy data. *Nature* 272 (1978), 686–690.
- [26] HALES, C. A., MURPHY, T., CURRAN, J. R., MIDDELBERG, E., GAENSLER, B. M., AND NORRIS, R. P. BLOBCAT: software to catalogue flood-filled blobs in radio images of total intensity and linear polarization. *Monthly Notices of the Royal Astronomical Society* 425, 2 (2012), 979–996.
- [27] HANCOCK, P. J., MURPHY, T., GAENSLER, B. M., HOPKINS, A., AND CURRAN, J. R. Compact continuum source finding for next generation radio surveys. *Monthly Notices of the Royal Astronomical Society* 422, 2 (2012), 1812–1824.
- [28] HOADLEY, B. The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association* 64, 325 (1969), 216–229.
- [29] HOBSON, M., AND MCLACHLAN, C. A Bayesian approach to discrete object detection in astronomical data sets. *Monthly Notices of the Royal Astronomical Society* 338, 3 (2003), 765–784.
- [30] HOLLITT, C., AND JOHNSTON-HOLLITT, M. Feature detection in radio astronomy using the circle Hough transform. *Publications of the Astronomical Society of Australia* 29, 3 (2012), 309–317.
- [31] ISHIBUCHI, H., AND YAMAMOTO, T. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. *Fuzzy Sets and Systems* 141, 1 (2004), 59–88.

- [32] JOHNSON, N. L. An approximation to the multinomial distribution: some properties and applications. *Biometrika* 47, 1/2 (1960), 93–102.
- [33] JOHNSON, N. L., KOTZ, S., AND BALAKRISHNAN, N. *Discrete multivariate distributions*, vol. 165. Wiley New York, 1997.
- [34] JOHNSTON-HOLLITT, M. *Detection of magnetic fields and diffuse radio emission in Abell 3667 and other rich southern clusters of galaxies*. PhD thesis, University of Adelaide, 2003.
- [35] JOHNSTON-HOLLITT, M., SATO, M., GILL, J., FLEENOR, M., AND BRICK, A.-M. Radio observations of the Horologium-Reticulum supercluster-I. A3158: excess star-forming galaxies in a merging cluster? *Monthly Notices of the Royal Astronomical Society* 390, 1 (2008), 289–303.
- [36] JONES, E., OLIPHANT, T., PETERSON, P., ET AL. SciPy: Open source scientific tools for Python. <http://www.scipy.org/>, 2001–.
- [37] KARIM, M. M. Zebra running Ngorongoro. http://en.wikipedia.org/wiki/File:Zebra_running_Ngorongoro.jpg.
- [38] KASS, R. E., AND RAFTERY, A. E. Bayes factors. *Journal of the American Statistical Association* 90, 430 (1995), 773–795.
- [39] KOTSIANTIS, S., AND KANELLOPOULOS, D. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering* 32, 1 (2006), 47–58.
- [40] KOTZ, S., BALAKRISHNAN, N., AND JOHNSON, N. L. *Continuous multivariate distributions, models and applications*, vol. 334. Wiley-Interscience, 2004.
- [41] KUETTEL, D., BREITENSTEIN, M. D., VAN GOOL, L., AND FERRARI, V. What’s going on? Discovering spatio-temporal dependencies in

- dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 1951–1958.
- [42] LIU, H., HUSSAIN, F., TAN, C. L., AND DASH, M. Discretization: An enabling technique. *Data Mining and Knowledge Discovery* 6, 4 (2002), 393–423.
- [43] LÓPEZ, L. M., RUIZ, I. F., BUENO, R. M., AND RUIZ, F. T. Dynamic discretization of continuous values from time series. In *Machine Learning: ECML 2000*. Springer, 2000, pp. 280–291.
- [44] MADSEN, R. E., KAUCHAK, D., AND ELKAN, C. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning* (2005), ACM, pp. 545–552.
- [45] MARTIN, D., FOWLKES, C., TAL, D., AND MALIK, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (2001), IEEE, pp. 416–423.
- [46] MASIAS, M., FREIXENET, J., LLADÓ, X., AND PERACLAULA, M. A review of source detection approaches in astronomical images. *Monthly Notices of the Royal Astronomical Society* (2012).
- [47] MILLER, N. A., BONZINI, M., FOMALONT, E. B., KELLERMANN, K. I., MAINIERI, V., PADOVANI, P., ROSATI, P., TOZZI, P., AND VAT-TAKUNNEL, S. The Very Large Array 1.4 ghz survey of the Extended Chandra Deep Field South: Second data release. *The Astrophysical Journal Supplement Series* 205, 2 (2013), 13.
- [48] MOSIMANN, J. E. On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika* 49, 1/2 (1962), 65–82.

- [49] NADARAJAH, S. A generalized normal distribution. *Journal of Applied Statistics* 32, 7 (2005), 685–694.
- [50] NASH, S. G. Newton-type minimization via the Lanczos method. *SIAM Journal on Numerical Analysis* 21, 4 (1984), 770–788.
- [51] NG, K. W., TIAN, G.-L., AND TANG, M.-L. *Dirichlet and Related Distributions: Theory, Methods and Applications*, vol. 888. Wiley, 2011.
- [52] NOCEDAL, J., AND WRIGHT, S. J. *Numerical optimization*. Springer Science+ Business Media, 2006.
- [53] NORRIS, R., HOPKINS, A., AFONSO, J., BROWN, S., CONDON, J., ET AL. EMU: Evolutionary map of the universe. *Publications of the Astronomical Society of Australia* 28, 3 (2011), 215–248.
- [54] NORRIS, R. P., AFONSO, J., AND APPLETON, P. N. Deep ATLAS radio observations of the Chandra Deep Field-South/Spitzer Wide-area Infrared Extragalactic Field. *The Astronomical Journal* 132, 6 (2006), 2409–2423.
- [55] NORRIS, R. P., AFONSO, J., BACON, D., BECK, R., BELL, M., ET AL. Radio continuum surveys with Square Kilometre Array pathfinders. *Publications of the Astronomical Society of Australia* 30 (2013), 20.
- [56] OLSON, D., AND DELEN, D. *Advanced data mining techniques*. Springer Verlag, 2008.
- [57] PERACAULA, M., LLADÓ, X., FREIXENET, J., OLIVER, A., TORRENT, A., PAREDES, J. M., AND MARTI, J. Segmentation and detection of extended structures in low frequency astronomical surveys using hybrid wavelet decomposition. In *Astronomical Data Analysis Software and Systems XX* (2011), vol. 442, p. 151.
- [58] PRATLEY, L., JOHNSTON-HOLLITT, M., DEGHAN, S., AND SUN, M. Using head-tail galaxies to constrain the intracluster magnetic field:

- an in-depth study of PKS J0334-3900. *Monthly Notices of the Royal Astronomical Society* 432, 1 (2013), 243–257.
- [59] QUELHAS, P., MONAY, F., ODOBEZ, J., GATICA-PEREZ, D., TUYTELAARS, T., AND VAN GOOL, L. Modeling scenes with local descriptors and latent aspects. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (2005), IEEE, pp. 883–890.
- [60] RENCHER, A. *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.
- [61] RUSSELL, B., FREEMAN, W., EFROS, A., SIVIC, J., AND ZISSERMAN, A. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), IEEE, pp. 1605–1614.
- [62] SARIPALLI, L., SUBRAHMANYAN, R., THORAT, K., ETERS, R., HUNSTEAD, R., JOHNSTON, H., AND SADLER, E. ATLAS extended source sample: The evolution in radio source morphology with flux density. *The Astrophysical Journal Supplement Series* 199, 27 (2012).
- [63] SAVAGE, R. S., AND OLIVER, S. Bayesian methods of astronomical source extraction. *The Astrophysical Journal* 661, 2 (2007), 1339.
- [64] SIVIC, J., RUSSELL, B., EFROS, A., ZISSERMAN, A., AND FREEMAN, W. Discovering object categories in image collections. Tech. Rep. AIM-2005-005, Massachusetts Institute of Technology, 2005.
- [65] SIVIC, J., RUSSELL, B., EFROS, A., ZISSERMAN, A., AND FREEMAN, W. Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (2005), IEEE, pp. 370–377.
- [66] SKILLING, J. Nested sampling. In *AIP Conference Proceedings* (2004), vol. 735, p. 395.

- [67] STEYVERS, M., AND GRIFFITHS, T. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*, T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, Eds. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA, 2007, pp. 424–440.
- [68] SUBRAHMANYAN, R., EKERS, R., SARIPALLI, L., AND SADLER, E. ATLAS: the Australia telescope low-brightness survey. *Monthly Notices of the Royal Astronomical Society* 402, 4 (2010), 2792–2806.
- [69] SUDDERTH, E., TORRALBA, A., FREEMAN, W., AND WILLSKY, A. Learning hierarchical models of scenes, objects, and parts. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (2005), IEEE, pp. 1331–1338.
- [70] TORRENT, A., PERACLAULA, M., LLADÓ, X., FREIXENET, J., SÁNCHEZ-SUTIL, J., MARTI, J., AND PAREDES, J. Detecting faint compact sources using local features and a boosting approach. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (2010), IEEE, pp. 4613–4616.
- [71] WANG, X., AND GRIMSON, E. Spatial latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20. MIT Press, Cambridge, MA, 2008, pp. 1577–1584.
- [72] WASSERMAN, L. *All of statistics: a concise course in statistical inference*. Springer Verlag, 2004.
- [73] WELLS, D., GREISEN, E., AND HARTEN, R. FITS — a flexible image transport system. *Astronomy and Astrophysics Supplement Series* 44 (1981), 363.
- [74] WHITING, M. Duchamp: a 3D source finder for spectral-line data. *Monthly Notices of the Royal Astronomical Society* 421, 4 (2012), 3242–3256.

- [75] YANG, Y., AND WEBB, G. I. Non-disjoint discretization for naive-Bayes classifiers. In *Proceedings of the Nineteenth International Conference on Machine Learning* (2002), pp. 666–673.
- [76] YANG, Y., WEBB, G. I., AND WU, X. Discretization methods. In *Data Mining and Knowledge Discovery Handbook*. Springer, 2010, pp. 101–116.
- [77] ZANARDO, G., STAVELEY-SMITH, L., NG, C.-Y., GAENSLER, B., POTTER, T., MANCHESTER, R., AND TZIOUMIS, A. High-resolution radio observations of the remnant of SN 1987A at high frequencies. *The Astrophysical Journal* 767, 2 (2013), 98.