Bayesian Source Discrimination in Radio Interferometry

Hague, P. R., Ye, H., Nikolic, B., Gull, S. F.
Astrophysics Group, Cavendish Laboratory, University of Cambridge

6 July 2018

ABSTRACT

Methods currently in use for locating and characterising sources in radio interferometry maps are designed for processing images, and require maps to be preprocessed in such a way as to resemble images. We demonstrate a Bayesian method that takes into account the visibility data despite working with more manageable data products. This method proves to be more capable of discriminating nearby sources than the commonly used SExtractor, and shows potential even in the more complicated analysis.

1 INTRODUCTION

Maps of surface brightness derived from interferometry differ from ordinary images more fundamentally than is immediately apparent, because such maps have been processed for the benefit of algorithms designed to deal with conventional images. The incompleteness of $uv$ coverage leads to more complex and difficult point spread functions (PSF) than are assumed by such algorithms. Any inference of sources from maps must take this issue into account.

In Hobson & McLachlan (2003) a Bayesian approach was shown to be valuable in detecting sources in a noisy image. The problem of extracting point sources from interferometric maps has been covered mathematically in Tan (1987)

The correct identification and characterisation of sources in images - and maps - is an important and continuing problem in astronomy. At present it is standard practice to apply tools designed for use on optical images (such as the SExtractor Bertin & Arnouts (1996)) to interferometric maps that have been through the CLEAN algorithm, (Hög bom 1974; Schwarz 1978) so as to resemble images.

We have developed instead a Markov Chain Monte Carlo (MCMC) process for the detection and characterisation of sources in maps. The method involves marginalisation over a parameter space defined by possible locations and fluxes of sources, in order to compute likelihoods for those sources. The resulting likelihoods can then be fed into a Bayesian statistical model. We demonstrate the effectiveness of this approach using simulated maps. We also introduce a public BASC1 python library that implements this method.

2 MARKOV CHAIN MONTE CARLO

MCMC is a method of finding the likelihood distribution of a parameter space via a random walk. The algorithm produces a chain of samples $X_{0..k}$ from the parameter space, such that

$$X_{n+1} = \begin{cases} X_n & \text{if } P(D|X') < P(D|X) \\
\text{and } \text{rand} > \frac{P(D|X')}{P(D|X)} & \\
X' & \text{otherwise} \end{cases}$$

where $D$ is the data being considered, $\text{rand}$ is a random number between 0 and 1, and $X'$ is a proposed new sample, which due to the Markov chain principle will be a random function of the previous sample. If new samples are selected from a symmetric function, so that the likelihood of selecting a sample $X_\text{k}$ given a current sample $X_\text{n}$ is the same as that of selecting $X_\text{n}$ given a current sample $X_\text{k}$, then the chain will always converge towards a state in which the density of samples in a given volume of parameter space is proportional to the likelihood. Typically, samples are selected from a Gaussian function centred on the current sample; this is the Metropolis-Hastings algorithm Metropolis et al. (1953); Hastings (1970).

The MCMC driver used here is BayeSys2, developed by Skilling (2004). This program differs from other MCMC algorithms through its use of atoms. Instead of each model being a set of parameters $\{x_i\}$, the model consists of an atom count $n$ and $n$ copies of the parameters $\{\{x_i\}_n\}$. (In this case, the parameters are coordinates of cells on the dirty maps, together with a flux $f$.) A conventional implementation of MCMC would be equivalent to using a single atom of this sort throughout the chain. The atom count is adjusted dynamically by proposing models with one additional atom or one fewer, at random intervals calculated such as to impose

1 https://github.com/petehague/BASC
2 http://www.inference.phy.cam.ac.uk/bayesys/

© 2018 The Authors
a specific prior on the atom count. In all cases in this paper, that prior is
\[ P(n) = (n!e)^{-1} \]  
(2)

Further details of the operation of BayeSys and the control of the number of atoms are set out in Skilling (2004).

At the end of the process, atoms from each model are clustered so that they can be associated with sources, after which source properties are extracted from each cluster of atoms.

3 LIKELIHOOD CALCULATION

Following Chapter 1 of Tan (1987), we start with a set of \( N \) point sources having fluxes \( \{F_j\} \) at sky positions \( \{x_j\} \), giving a flux \( F(x) \) at position \( x \) as follows:
\[ F(x) = \sum_j F_j(x - x_j) \]  
(3)

By means of Fourier transform, this gives a visibility model at a particular visibility \( V \) of the form
\[ M_k = \sum_j F_j e^{2\pi i u_k x_j} \]  
(4)

where \( u_k \) is the wavevector of the visibility. For Gaussian errors on the visibilities, the likelihood of the observed visibilities given the model is
\[ P(\{V_k\}|\{M_k\}) = \prod_k \frac{1}{2\pi\sigma_k^2} e^{-|V_k - M_k|^2 / 2\sigma_k^2} \]  
(5)

The exponential term is equivalent to \( \exp(-\chi^2/2) \). This calculation for \( M \) visibilities would require \( M \times N \) calculations and may be prohibitively expensive. Alternatively, we can perform only \( N^2 \) calculations by using the equivalent expression for \( \chi^2 \),
\[ \chi^2 = \sum_k \frac{|V_k|^2}{\sigma_k^2} - \frac{1}{\sigma^2}(2D^T F - F^T BF) \]  
(6)

where \( D \) denotes the values of the dirty map at those points, and \( B \) is the corresponding value of the beam,
\[ B(x) = \Re \left[ \sum_k \frac{1}{\sigma_k^2} e^{-2\pi i u_k x} \right] \sigma^2 \]  
(7)

Since the MCMC chain is constructed entirely from the ratio of log likelihoods \( L_n - L_{n-1} \), the first term always cancels and does not need to be calculated in order to generate the chain. The calculation need be performed only at the spatial positions of the proposed points. To make use of this likelihood a thorough search of the parameter space is needed, which MCMC is capable of.

In BayeSys, parameters are always in the range \( (0, 1) \). To meet this condition the parameters \( x \) and \( y \) are simply given a flat prior and scaled to the size of the map. For the flux parameter, a prior is used such that
\[ F = A s \frac{f}{1 - f} \]  
(8)

where \( f \) is the flux parameter and \( F \) is the actual flux of the atom. Here, \( A \) is a scaling constant which places the centre of the parameter range \( (f = 0.5) \) at \( F = A \). The value of \( A \) can be varied to ensure the flux prior encloses the full range of fluxes; setting it equal to the dirty map noise usually gives good results, and this is used as the default.

BayeSys is allowed to burn in for as long as is necessary, after which an ensemble of 20 samples per step is generated for the same number of steps as it took to burn in. This typically leads to less than 200 models included in the final output, which is sufficiently accurate for this application – although a much larger number of models will have been explored internally by BayeSys without being included. The atoms are then clustered using the DBSCAN algorithm (see Appendix A), in order to identify them with sources.

4 TEST DATA

The two test sets consist of a set of 6 maps containing a single point increasing in brightness. The aim of this set is to study the effect of the signal-to-noise ratio on position accuracy. There are also 100 maps containing pairs of points in which one point is at the centre of the map and the other is offset by a random distance and angle. In some cases the offset is zero, giving a single point of double brightness. The set is repeated with the central point brightness increased by various factors, but with the distance and angle to the second point unchanged. The standard brightness is 10\( \sigma \), where the map noise \( \sigma = 5.6 \times 10^{-4} \text{Jy/beam} \).

Using the simalma task in CASA, we generate a set of maps that contain two point sources at varying distances of the order of the size of the central peak of the PSF, in order to test the ability of the method to discriminate between them. Maps are simulated by inputting a surface brightness map, which in this case is zero everywhere except for cells containing the point sources. This is then converted into an observation using a standard antenna configuration, which also yields both a dirty beam and a primary beam. We use ALMA antenna configuration 2.6 observing for 698\( \alpha \), based on an observation of NGC 1380 as detailed in Salak et al. (2017). This gives a map that is 20.48 arcseconds across, substantially smaller than the primary beam.

Figure 1 shows a sample of the main test set, zoomed in on the central region where the two points are located. In each case the points are separated by a uniformly random distance between 0.1 and 1 arcsecond. The angle of the line between the two points is distributed uniformly. The combination of a uniform distance and angle produces a centrally concentrated distribution of points; because the final result will be binned by radius, this distribution is desirable so that each bin has the same number of samples.

We have produced further sets using the same source positions but with the brightness of the source at the centre of the map increased by factors of 10, 20, 30 and 40. We have also used the same set of points to generate simulated
Figure 1. Cutouts of the central area of the 16 examples from the testing set (indices 68-83, top left to bottom right). Sample number 68 at top left was only detectable as two distinct points for the case in which one point is 40 times brighter than the other. Sample 70 (top row third from left) is a single point. The colour map is CubeHelix (Green 2011)

Figure 2. Plot of the separation of sources in the input file versus the separation inferred using each source finder. The dotted green line represents agreement with the input, solid red shows the results from BaSC, and dashed blue shows the results from SExtractor. Error bars are 90% confidence intervals. The vertical lines show the sizes of the minor and major axes of the CLEAN restoring beam.

VLA observations, so as to determine how a more complex PSF influences the performance of the two methods.

4.1 Bayesian resolving power

Since the maps of atom positions show structure on a substantially smaller scale than the restoring beam or the central peak of the dirty beam, it is necessary to investigate the limit of the resolution using this method.

We first consider the ability of the method to constrain the position of a single source. We begin with the expression for a dirty map $D(x)$ containing a single source of flux $F$ at the origin with visibilities $V_k$:

$$D(x) \approx FB(x) = F \sigma^2 \sum_k \frac{1}{\sigma_k^2} \cos(2\pi u_k \cdot x)$$

where $B(x)$ is the dirty beam. Truncation of the Taylor expansion of the expression for the beam (equation 7) at the origin gives a Gaussian approximation of the probability that the point is at position $x$,

$$P(x) \propto \exp \left[ -2\pi^2 F^2 \sum_{ij} \frac{(u_k)_i (u_k)_j}{\sigma_k^2} x_i x_j \right]$$

which implies a covariance of

$$< \delta x_i \delta x_j > = \frac{\sigma^2}{F^2} \Delta x^2_{ij}$$

where

$$(\Delta x^2)^{-1}_{ij} = -\frac{\partial^2 B(x)}{\partial y^2} |_{0} = 4\pi^2 \sigma^2 \sum_k \frac{(u_k)_i (u_k)_j}{\sigma_k^2}$$

This expression describes the characteristic width of the central part of the beam, for which a sufficient approximation is provided in terms of an ellipse $a, b, \theta$ by the CASA pipeline. The size of the resolving ellipse is therefore just $a/S, b/S$, where $S$ is the signal-to-noise ratio $F/\sigma$.

To test this result, we generated test maps containing a single point, with signal-to-noise ratios ranging from 10 to 160 in powers of 2, as shown in Table 4.1. The standard deviation in the derived position of the point along each axis was a good match to the theoretical prediction, and tended towards $1/\sqrt{12}$ of the cell width at the highest $F/\sigma$ i.e. the expected uncertainty within a single cell.

4.2 Comparison

We use SExtractor as a comparison for our tests, with the detection threshold set to $3\sigma$ and the minimum deblending contrast parameter set to 0.001. This parameter indicates the ratio in the integrated flux between two proposed sources if the same island of detected pixels would be considered separate sources. This is the best performing set of parameters that we were able to find.

| $S/N$ ratio | $\sigma_x$ | $\sigma_y$ | Prediction |
|-------------|------------|------------|------------|
| 160         | 0.292      | 0.288      | 0.147      |
| 80          | 0.574      | 0.547      | 0.294      |
| 40          | 0.670      | 0.634      | 0.587      |
| 20          | 1.394      | 1.190      | 1.175      |
| 10          | 2.188      | 2.386      | 2.351      |

Table 1. Accuracy of positions in processed maps versus theoretical predictions.
Other source finding algorithms, such as \texttt{AEGEAN}, for which the performance depends on the image being cleaned (Hancock et al. 2012), can be expected to perform comparably in this specific test. Indeed, \texttt{AEGEAN} no longer detected the second point at the same distance as \texttt{SExtractor}.

5 RESULTS

The clustered output of the MCMC chain yields a distance between the two sources, which can be compared to the input parameter. Figure 2 shows the results, in comparison with the performance of \texttt{SExtractor}.

Out of 100 samples, \texttt{SExtractor} was able to discriminate two distinct points in 27 cases, with a minimum discrimination distance of 0.6 arcseconds, whereas \texttt{BaSC} was able to find two points in 51 cases, with a minimum discrimination distance of 0.3 arcseconds. The major and minor axes of the \texttt{CLEAN} restoring beam are respectively 0.63 and 0.59 arcseconds.

Figure 3 shows the effect of brightening one of the sources. The ability to discriminate increases with the brightness of the central source, because the constraint of that source becomes tight and largely independent of the dimmer source. The detection of the dimmer source can then be considered an independent problem, with an additional source of noise. As the signal to noise ratio of the brighter source increases, the removal of the main source becomes much more accurate, leaving the uncertainty in the flux as the only significant source of error. For a dim source close to a sufficiently bright source, the detection problem becomes a single point problem with a map noise equal to $2\sigma$. The dotted line in the Figure 3 shows the resolving power for the dim source if the noise is $2\sigma$. 

\footnote{https://github.com/PaulHancock/Aegean}
5.1 VLA simulations

Dirty maps produced from VLA observations are less clean than those produced from ALMA as a result of differences in their \( \text{uv} \) coverage, and we therefore expect source detection methods that are dependent on CLEAN to perform less well.

Figure 5 confirms that this is indeed the case; as well as BaSC being better able to differentiate the sources, there is a systematic underestimation of the distance between them using SExtractor. This is not a consequence of incorrectly fitting Gaussians to the clean map, but is because CLEAN itself has not represented the targets accurately. In Figure 4 we see that parts of the dirty PSF have been left in the map, creating a spurious ‘bridge’ of flux between the two points which causes SExtractor to fit the Gaussians with closer centres.

It is likely that a user of CLEAN could produce maps that represent the sources more accurately in interactive mode. In contrast, BaSC produced the result shown with no interactivity; the only human input was an initial estimate of the map noise, and basic parameters for the MCMC chain that remained unchanged in all of the present tests.

6 CONCLUSION

We have shown that BaSC is superior to SExtractor at the specific task of discriminating between nearby points in interferometric maps. This is hardly surprising, given that SExtractor is not designed for that task. The CLEAN process produces maps that are readily understandable to humans, but this is not a necessary nor even desirable property when the map is then to be processed by an algorithm not subject to human perceptual norms. Such a method produces a product designed to be interpreted visually only at the final step.

The theoretical resolving power for the Bayesian method has been shown to be realised for a single point by using BaSC. The ability to discriminate two points is improved when a bright source is close to a dim source. This is due to the better constraint and consequently better subtraction of the brighter source, and the ability to discriminate in this case has been confirmed by testing to be close to the theoretical limit.

ACKNOWLEDGEMENTS

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653477

REFERENCES

Bertin E., Arnouts S., 1996, AAS, 117, 393
Campillo R. J. G. B., Moulavi D., Sander J., 2013, in Pei J., Tseng V. S., Cao L., Motoda H., Xu G., eds, Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp 160–172
Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD’96. AAAI Press, pp 226–231, http://dl.acm.org/citation.cfm?id=3001460.3001507
Green D. A., 2011, Bulletin of the Astronomical Society of India, 39, 269
Hancock P. J., Murphy T., Gaensler B. M., Hopkins A., Curran J. R., 2012, MNRAS, 422, 1812
Hastings W. K., 1970, Biometrika, pp 97–109
Hobson M. P., McLachlan C., 2003, MNRAS, 338, 765
Högboom J. A., 1974, A&AS, 15, 417
Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, Journal of Chemical Physics, 21, 1087
Sakai D., Tomiyasu Y., Nakai N., Kuno N., Miyamoto Y., Kaneko H., 2017, ApJ, 849, 90
Schubert E., Sander J., Ester M., Peter Kriegel H., Xu X., 2017, ACM Transactions on Database Systems, 42, 1
Schwarz U. J., 1978, A&A, 65, 345
Skilling J., 2004, BayeSys manual. http://www.inference.org.uk/bayesys/manual.tex
Tan S. M., 1987, PhD thesis, University of Cambridge, http://hooke.lib.cam.ac.uk/cgi-bin/bib_seek.cgi?cat=man&bib=14401

APPENDIX A: CLUSTERING

Because of the random nature of the MCMC algorithm, atoms in each model are not necessarily assigned to the same source in every instance. It is therefore necessary to adopt a clustering method in order to study individual sources on the map.

The clustering algorithm must be robust enough to disregard these outliers and provide the number of clusters correctly. In simulations the number of clusters is known, such as 2 in the example set out above. For real observations, this number would not be known in advance. We also wish to apply this method without large-scale human intervention. We must therefore choose a clustering algorithm that does not require the number of clusters to be known in advance.

The density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm meets these requirements. DBSCAN was proposed by Ester et al. (1996), and has been one of the most widely used clustering algorithm ever since. DBSCAN assigns atoms to a cluster or tags them as outliers, based upon the local density. Schubert et al. (2017) explain how DBSCAN can be adapted to process MCMC output:

(i) For each atom in the MCMC output, find the points which are within a radius \( \epsilon \) of it.
(ii) The data point is marked as a seed if the neighbour count is greater than \( \text{min}_{\text{samples}} \), the minimum grouping number.
(iii) For each seed, find all points connected to it by steps no larger than \( \epsilon \). Mark as outliers those unconnected to a seed in this way.
(iv) Each group of points containing one or more seeds is considered as a cluster, where MCMC considers there to be a peak in the likelihood distribution. The location of the source is calculated as the mean of the location of points within this cluster, while the outliers are removed in order to avoid distortion of the data.

Although DBSCAN does not require the number of clusters and does not constrain their shape, it does need the radius of the cluster searching area \( \epsilon \) as well as the mini-
Hague et al.

Figure 4. Example of a CLEANed (left) and a dirty (right) map from the VLA simulations.

Figure 5. Plot of the separation of points in the input file versus the separation inferred using each source finding method, in the case of the VLA simulations compared against the performance of SExtractor.

Figure A1. The x-axis represents the distances of all data points from their 4 nearest neighbours in ascending order, and the y-axis represents the distances. The sorted k-dist graph is used to provide an estimate of $\epsilon$ by classifying outliers and non-outliers according to a threshold, shown by the broken line. The distance value at the broken line is the recommended value of $\epsilon$.

The choice of $\epsilon$ should not be too large, or else two sources close to each other would be considered as one; whereas if $\epsilon$ is too small then outliers would be considered as sources, or even single sources would be divided into two or more.

The minimum grouping number $\text{min}_{\text{samples}}$ should not be too small, or else outliers would be considered as sources; whereas if the number is too large then two close-by sources would be considered as one, and in the worst case even the sources would be considered as outliers.

Since $\epsilon$ and $\text{min}_{\text{samples}}$ work as a pair, in this application of MCMC results the value of $\epsilon$ should not be less than 0.5, or half of the pixel size. This is to prevent atoms in the same pixel being separated into more than one cluster. Also, $\text{min}_{\text{samples}}$ is not less than 10 so that outliers are not considered as clusters. In using the program there will be a recommended value of $\epsilon$ for each dataset with $\text{min}_{\text{samples}} = 10$. The recommendation was made based on the “sorted k-dist graph” (Ester et al. 1996); see Figure A1. This figure plots the distances from $K$-nearest neighbours for all data points. In this case we set $K = 4$; the x-axis represents the distances of all data points from their 4 – nearest neighbours in as-
Figure A2. BaSC output of simulated two-source data; one source is 10 times brighter than the other. The black dots on the map are outliers, and the blue and orange dots are two distinct clusters with stars as their cluster centres. The two sources have a separation of 30 cells.

Figure A3. The black dots on the map are outliers, while the blue dots represent the only cluster with a star as its cluster centre. Upon increasing $\epsilon$ to 30, the two clusters detected in Figure A2 are considered as one.

A more recent clustering algorithm, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), has been developed by Campello et al. (2013) based on DBSCAN, and does not need the user to provide $\epsilon$. HDBSCAN is another density-based clustering algorithm, but unlike DBSCAN it can find clusters with varying densities. Only two parameters are required by the hdbscan library \(^4\). The first is $min_{\text{cluster}}$, which is identical to $min_{\text{samples}}$ in DBSCAN, the minimum grouping number. The second is $min_{\text{samples}}$, which controls the sensitivity of HDBSCAN for picking up less dense clusters; for larger $min_{\text{samples}}$, the algorithm is more likely to ignore a number of outliers which cluster together; for smaller $min_{\text{samples}}$ those outliers may be regarded as clusters. Consequently, HDBSCAN can be used in this application; it is recommended that $min_{\text{cluster}} = 10$ at first, in order to pick up as many clusters as possible, then set $min_{\text{samples}} = 30$ so that lower density agglomerations, which are more likely to be outliers, are not taken as clusters.

When dealing with a large number of maps without human supervision, we recommend DBSCAN with the recommended $\epsilon$, or HDBSCAN with $min_{\text{cluster}} = 10$ and $min_{\text{samples}} = 30$. After the clustering process, and if concerns become obvious over certain maps, parameters can be changed manually by the user so as to improve the clustering process in specific applications.

Figure A2 shows an example of implementing DBSCAN. The black dots on the map are outliers, and the blue and orange dots are two distinct clusters with amber stars and red stars as their respective cluster centres. The two clusters were readily identified by DBSCAN, with their centres at (512.50, 512.50) and (535.50, 488.50). DBSCAN has an outlier detection feature, but asks for the radius of cluster sizes as well as the minimum grouping number. The recommended radius parameter is 10, and the minimum grouping number was chosen as 10. There are 43 outliers, marked as black dots. If instead $\epsilon = 20$ and $min_{\text{samples}} = 10$, the two clusters are considered to be centred at the single location (519.50, 505.50). This is because the radius $\epsilon$ is too large, so that DBSCAN considers the two cluster centres separated by 33.24 ($< 40$) units as a single cluster; see Figure A3.

Figure A4 shows an example of implementing HDBSCAN. The black dots on the map are outliers, and the blue and amber dots are two distinct clusters with stars as their respective cluster centres. The two clusters were readily found by HDBSCAN, with their centers at (512.50, 512.50) and (535.50, 488.50). The two parameters chosen are $min_{\text{cluster}} = 10$ and $min_{\text{samples}} = 30$.
512.50) and (535.50,488.50). The two parameters chosen are \( \text{min}_{\text{cluster}} = 10 \) and \( \text{min}_{\text{samples}} = 30 \).

This paper has been typeset from a \TeX/\LaTeX\ file prepared by the author.