Bayesian inference for radio observations

Michelle Lochner,[1,2,3,4]★ Iniyan Natarajan,[5] Jonathan T. L. Zwart,[5,6] Oleg Smirnov,[7,8] Bruce A. Bassett,[1,2,3] Nadeem Oozeer[1,8,9] and Martin Kunz[1,10]

[1]*African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg 7945, South Africa*
[2]*Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch, Cape Town 7700, South Africa*
[3]*South African Astronomical Observatory, Observatory Road, Observatory, Cape Town 7935, South Africa*
[4]*Department of Physics and Astronomy, University College London, London WC1E 6BT, UK*
[5]*Astrophysics, Cosmology and Gravity Centre (ACGC), Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa*
[6]*Department of Physics and Astronomy, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa*
[7]*Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown 6140, South Africa*
[8]*SKA South Africa, 3rd Floor, The Park, Park Road, Pinelands 7405, South Africa*
[9]*Centre for Space Research, North-West University, Potchefstroom 2520, South Africa*
[10]*Département de Physique Théorique and Center for Astroparticle Physics, Université de Genève, Quai E. Ansermet 24, CH-1211 Genève 4, Switzerland*

ABSTRACT

New telescopes like the Square Kilometre Array (SKA) will push into a new sensitivity regime and expose systematics, such as direction-dependent effects, that could previously be ignored. Current methods for handling such systematics rely on alternating best estimates of instrumental calibration and models of the underlying sky, which can lead to inadequate uncertainty estimates and biased results because any correlations between parameters are ignored. These deconvolution algorithms produce a single image that is assumed to be a true representation of the sky, when in fact it is just one realization of an infinite ensemble of images compatible with the noise in the data. In contrast, here we report a Bayesian formalism that simultaneously infers both systematics and science. Our technique, Bayesian Inference for Radio Observations (BIRO), determines all parameters directly from the raw data, bypassing image-making entirely, by sampling from the joint posterior probability distribution. This enables it to derive both correlations and accurate uncertainties, making use of the flexible software MEQTREES to model the sky and telescope simultaneously. We demonstrate BIRO with two simulated sets of Westerbork Synthesis Radio Telescope data sets. In the first, we perform joint estimates of 103 scientific (flux densities of sources) and instrumental (pointing errors, beamwidth and noise) parameters. In the second example, we perform source separation with BIRO. Using the Bayesian evidence, we can accurately select between a single point source, two point sources and an extended Gaussian source, allowing for 'super-resolution' on scales much smaller than the synthesized beam.

Key words: methods: data analysis – methods: statistical – techniques: interferometric.

1 INTRODUCTION

The high sensitivity of the Square Kilometre Array (SKA) (up to 50 times more sensitive than current instruments; Carilli & Rawlings 2004) combined with a relatively cheap antenna design means a far more careful and detailed treatment of systematics will be required to fully exploit this telescope (Noordam 2000). The current approach to this calibration problem iteratively applies deconvolution methods such as CLEAN (Högbom 1974), alternating with sky and instrumental modelling to determine the best-fitting calibrated image (Pearson & Readhead 1984; Bhatnagar et al. 2008; Kazemi et al. 2011; Kazemi & Yatawatta 2013). This provides only a point estimate of the model parameters which will in general differ from the true parameters due to random noise (Enßlin et al. 2014).

A more rigorous approach is to infer the science and instrumental parameters simultaneously, deriving accurate uncertainties and correlations between them. Work in this direction includes improvements on the self-calibration algorithm (Pearson & Readhead 1984; Enßlin 2014; Enßlin et al. 2014; Dorn et al. 2015) and some extensions to the RESOLVE algorithm (Junklewitz et al. 2013; Junklewitz, Bell & Enßlin 2014). There has also been considerable effort in this direction in producing a maximum posterior image for the data and dealing with certain calibration parameters (Sutton & Wandelt

2006; Sutter et al. 2014). These works each solve specific aspects of the calibration and deconvolution problem, but so far do not explore the full posterior distribution, giving an inaccurate estimation of the uncertainties and correlations, and still rely on producing a single image (i.e. a point estimate).

We propose instead a new technique, called Bayesian Inference for Radio Observations (BIRO), which is able to: include any source of instrumental uncertainty, such as ionospheric effects, pointing errors and primary beam uncertainties, jointly determine the science and instrumental parameters and provide reliable estimates of the uncertainties and correlations on these parameters, in a holistic and mathematically rigorous manner.

A simultaneous analysis requires the full posterior probability distribution of the parameters, which can naturally be sampled in the Bayesian formalism by using (for example) MCMC (Metropolis et al. 1953; Hastings 1970) or nested sampling (Skilling 2004). Our new technique, BIRO, fits models including both instrumental and science parameters directly to the raw visibility data. We use the MEQTREES (Noordam & Smirnov 2010) software, which implements the Radio Interferometry Measurement Equation (RIME; Hamaker, Bregman & Sault 1996), for the modelling of the sky and instrumental effects. This technique thus obviates the need for intermediate imaging and map-making. The rigorous statistical use of all available information allows this technique to open new discovery windows, solving previously intractable problems, and is applicable to all interferometers and problems in radio interferometry.

This paper is arranged as follows: in Section 2, we provide an introduction to Bayesian statistics and illustrate the use of the RIME for modelling in the BIRO algorithm in Section 3. We then apply BIRO to two key simulated data sets to demonstrate its power: In Section 4, we jointly fit all scientific (source flux densities) and instrumental parameters (pointing errors, primary beam parameters and receiver noise) to a data set suffering from direction-dependent instrumental effects. In Section 5, we focus on the problem of reliably distinguishing between an extended source, point source and a pair of close point sources, for sources on sub-synthesized-beam scales. We conclude in Section 6.

2 BAYESIAN STATISTICS

The problem of obtaining the most information possible from an incomplete data set, such as obtained by an interferometer, is perfectly suited to the application of Bayesian statistics. These allow the fitting of arbitrarily complex models to data, providing reliable uncertainty estimates for the parameters. Bayes' theorem allows the use of a familiar quantity, the likelihood, to answer the question one is really interested in: What is the probability of a hypothesis, given the data in hand? This probability is known as the posterior and indicates by how much our degree of belief in the hypothesis has been updated by the new data. Simple application of Bayes' theorem also allows a robust and intuitive way to compare models, which we will require for the second example problem in this paper. What follows here is a brief overview of Bayesian theory, see Trotta (2008) for a more in-depth review.

From Bayes' theorem, the probability distribution, $\mathcal{P}(\Theta|D, H)$, of the values of parameters Θ, the quantity that is actually sought, given the data D that are in-hand and a model H (hypothesis plus any assumptions), is

$$\mathcal{P}(\Theta|D, H) = \frac{\mathcal{L}(D|\Theta, H)\,\Pi(\Theta|H)}{\mathcal{Z}(D|H)}. \tag{1}$$

This is known as the posterior probability distribution. The likelihood $\mathcal{L}(D|\Theta, H)$, which encodes any constraints imposed by observations, is the probability distribution of the data given parameter values and a model.

The prior $\Pi(\Theta|H)$ includes any prior knowledge of or prejudices about the parameter values. $\mathcal{Z}(D|H)$ is the integral of $\mathcal{L}(D|\Theta, H)\,\Pi(\Theta|H)$ over all Θ, not simply normalizing the posterior $\mathcal{P}(\Theta|D, H)$, but also allowing selection of different models by comparing their values quantitatively. This so-called evidence, $\mathcal{Z}(D|H)$, automatically includes an Occam's razor effect, penalizing models with a large number of parameters that are not preferred by the data. By computing the evidence for a range of models, we can select the best model by maximizing the evidence.

For this work, the likelihood function is

$$\mathcal{L}(D|\Theta, H) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\left(\sum_{i=1}^{N}(\mathbf{V}_i(\Theta) - \widetilde{\mathbf{V}}_i)^2\right)/2\sigma^2\right], \tag{2}$$

where $\mathbf{V}_i(\Theta)$ are the model visibilities produced by MEQTREES (see Section 3), with the parameters Θ as input, $\widetilde{\mathbf{V}}_i$ are the data visibilities, N is the number of data points. Here we assume the uncertainties on the visibilities are Gaussian and have the same value, σ, for all data points. The best-fitting model corresponds to maximum posterior.

The inferred posterior distributions are full probability distributions rather than a summary mean/median value and a (perhaps covariant) uncertainty, since this represents the total inference about the problem at hand. These distributions may be highly non-Gaussian, making such summary parameters inaccurate.

The application of Bayesian statistics allows one to marginalize out the effects of nuisance parameters, which are parameters such as the beam shape and pointing errors that are not of primary interest, but are unknown and can affect the estimates of the parameters of interest (i.e. science parameters) because of correlations and degeneracies. The marginalized posterior can be written as a function of the parameters of interest, Φ, the nuisance parameters, Ψ, and the data, D:

$$P(\Phi|D, H) = \int P(\Phi, \Psi|D, H)\mathrm{d}\Psi, \tag{3}$$

where the integral is performed over the parameter space of Ψ.

The posterior is, thanks to advances in modern computing, fairly easily determined using numerical techniques. In this paper, we use the Markov Chain Monte Carlo (MCMC) Metropolis–Hastings (Metropolis et al. 1953; Hastings 1970) algorithm for the joint scientific and instrumental parameter inference example. We chose MCMC due to its simplicity and the ease with which it handles large numbers of parameters (we have 103 parameters for the first example problem). For our second example, that of model selection related to source separation and extended structure, we require efficient calculation of the Bayesian evidence, something provided naturally by the nested sampling algorithm. We utilize the public code MULTINEST (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009) to determine both the parameters and the evidence for model comparison (Jeffreys 1998; Trotta 2008), but for a smaller set of parameters as nested sampling grows rapidly in complexity with increasing number of parameters.

3 USING THE RIME FOR MODELLING

Previous Bayesian visibility analyses (Lancaster et al. 2005; Feroz et al. 2009; Zwart et al. 2011; AMI Consortium 2012; Sutter, Wandelt & Malu 2012) focused on the sky model and were not generalized to include arbitrary instrumental effects (or were attempting to solve for a much more general sky model resulting in many more parameters, thus needing to fix instrumental parameters). The RIME (Hamaker et al. 1996; Smirnov 2011a,b) provides a powerful framework to easily describe exactly what happens to a signal as it travels from source to telescope, where it is converted into voltages. The RIME is a natural way to model the instrumental and scientific effects that we are inferring through our Bayesian technique. For example, the RIME for a single point source is given by

$$\mathbf{V}_{pq} = \mathbf{J}_p \mathbf{B} \mathbf{J}_q^H, \tag{4}$$

where $\mathbf{B}$ is the brightness matrix, which describes the sky flux distribution, $\mathbf{J}_p$ is the Jones matrix (Jones 1941) for antenna p, containing all instrumental and atmospheric effects that interfere with the signal, $\mathbf{J}_q$ is the Jones matrix for antenna q, H indicates the Hermitian of a matrix and $\mathbf{V}_{pq}$ are the visibilities, the outputs of the telescope correlator for baseline pq.

The effects that interfere with the signal on its route to the output of the telescope can each be described by a Jones matrix, with each effect adding a pair of Jones matrices in the 'onion' form of the RIME:

$$\mathbf{V}_{pq} = \mathbf{J}_{pn} \left(\ldots \left(\mathbf{J}_{p2} \left(\mathbf{J}_{p1} \mathbf{B} \mathbf{J}_{q1}^H \right) \mathbf{J}_{q2}^H \right) \ldots \right) \mathbf{J}_{qm}^H. \tag{5}$$

We can go a few steps further and consider the full-sky RIME by integrating over the direction cosines, l and m:

$$\mathbf{V}_{pq} = G_p \left(\int \int_{lm} \mathbf{E}_p \mathbf{K}_p \mathbf{B} \mathbf{K}_q^H \mathbf{E}_q^H \, \mathrm{d}l \mathrm{d}m \right) G_q^H. \tag{6}$$

Here, $\mathbf{K}_p$ and $\mathbf{K}_q$ are the Jones matrices describing the geometric delay between antennas p and q, G_p represents the *direction-independent* gains for antenna p, which we set to unity for all antennas, and $\mathbf{E}_p$ is the Jones matrix containing all the *direction-dependent* effects for antenna p. We focus in this paper on the more difficult to handle direction-dependent effects, but direction independent can also be handled with our technique. As with all other Jones matrices, $\mathbf{E}_p$ can be written as a product of Jones matrices, each describing a different effect. In Section 4, we consider both primary beam effects and pointing errors as examples of direction-dependent effects each with their own Jones matrix.

The RIME is implemented in the general, flexible software MEQTREES (Noordam & Smirnov 2010; Smirnov & de Bruyn 2011; Smirnov 2011c), which allows us to apply it to any sky model and for any telescope. MEQTREES has been useful for predicting the capabilities of future experiments and for understanding the intricacies of current telescopes. Here, we go a step further and use MEQTREES as the modelling step in our Bayesian analysis. In order to test BIRO and compare it with the standard deconvolution approach, we use data sets simulated with MEQTREES over which we have complete control and thus would know if we were correctly recovering the true input parameters.

MEQTREES takes from the user a sky model (such as the number and distribution of sources, their fluxes, shapes, etc.) as well as instrumental details (such as the telescope configuration, primary beam pattern, pointing errors, noise, atmospheric effects, ionospheric effects, etc.) and uses the measurement equation to produce realistic simulated visibilities that such a telescope would observe.
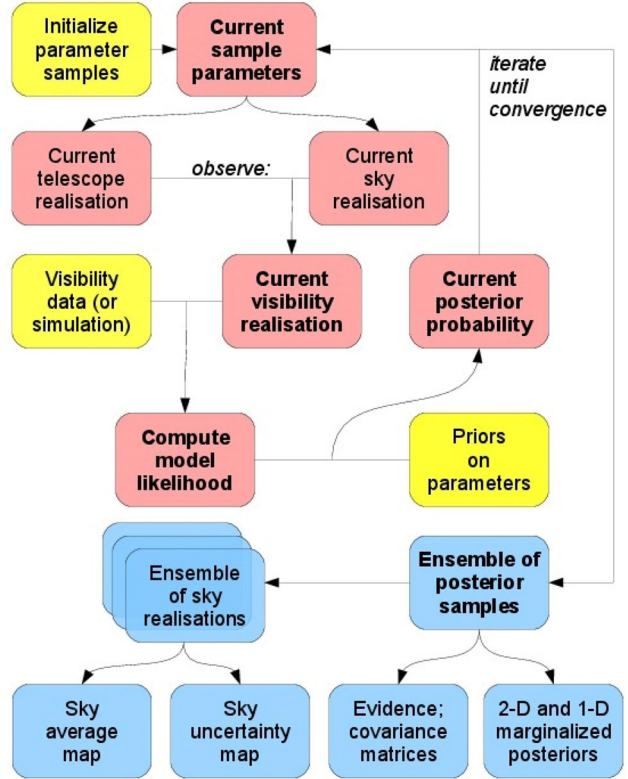

Figure 1. The BIRO algorithm. Fixed or initialized inputs are shown in yellow, while the sampling loop is represented by the pink boxes. Data products are in blue. The main iteration loop occurs within either the MCMC or MULTINEST algorithm (depending on the problem), where new parameters are used in each iteration to compute the likelihood. The initial parameters are drawn from the prior, which generally restricts the parameter ranges. In the final step, the ensemble of sky realisations can be generated with MEQTREES using the parameter samples in the posterior, if required.

In order to test the validity of our technique, we only work with simulations in this paper. We use MEQTREES to simulate the data and also to model the sky, to test if we recover the input parameters. MEQTREES can be used to model any telescope configuration and any sky and instrumental effects that can be described with the RIME. While we only concentrate on primary beam and pointing error effects in Section 4, in principle, a wide variety of source types and instrumental corruptions can be added in MEQTREES.

Fig. 1 shows a schematic overview of the BIRO approach. At each step in the chain of MCMC or MULTINEST, MEQTREES is called with new values for the parameters. MEQTREES then returns a visibility set that can be compared directly with the simulated data, to determine how well the parameters fit. This iterative process allows the determination of the full posterior for the parameters. We do not as yet have a public release of the BIRO code, but plan to in the future where we will integrate MontBlanc, a GPU implementation of the RIME (Perkins et al. 2015) with BIRO. MontBlanc is already publicly available meaning it can be combined with any sampler to allow the user to implement BIRO for themselves.

4 EXAMPLE 1: JOINT INFERENCE OF SCIENTIFIC AND INSTRUMENTAL PARAMETERS

In this example, we use BIRO to jointly estimate the scientific parameters and nuisance instrumental parameters. Below we describe

the model and simulated data set used, and details of the MCMC analysis, and show that the instrumental parameters studied are tightly correlated with the scientific parameters, a fact that cannot be ignored when determining these parameters.

4.1 Simulated data and parameters of the model

4.1.1 Telescope configuration

We use MEQTREES to simulate observations with the Westerbork Synthesis Radio Telescope (WSRT; Högbom & Brouw 1974), a 14-element east–west array with 25 m diameter dishes. All our WSRT simulations use an integration time of 30 s and a total observation time of 12 h at a frequency of 1.4 GHz. We use a narrow bandwidth of 125 kHz, a single channel (for simplicity) and include noise with a standard deviation of 0.1 Jy/visibility. At this frequency, WSRT has a field of view of 0°.5–0°.6 and a synthesized beamwidth of around 13 arcsec FWHM (full width at half-maximum).[1]

4.1.2 Scientific parameters

The simulated field consists of 17 unpolarized, point sources with known positions. The science goal was to determine the flux densities of these sources. We based the simulation on an existing field observed by WSRT, consisting of sources with a range of fluxes (from 0.03 to 3.13 Jy). This is a very simple sky model, consisting only of point sources, whereas in the second example of the paper, we address modelling of extended sources. We do not explore the possibility of extended sources of arbitrary shapes, as this is out of the scope of this paper, but this should be possible using shapelets, such as employed in the existing PYBDSM software.[2] The brightness matrix in equation (6) for an unpolarized point source is written as

$$\mathbf{B}^{\text{POINT}} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \tag{7}$$

where I is the intensity.

Fig. 2 shows an image of the true input model without any instrumental effects, while Fig. 3 shows the dirty image of the sky.

4.1.3 Instrumental parameters

Beamwidth

Knowing the primary beam pattern is critical for any astronomical survey. Current practice is to determine the primary beam pattern using a technique such as holography (Scott & Ryle 1977), then fix a beam model, without propagating any uncertainty information into the estimates of the science parameters. Since the primary beam directly attenuates the flux distribution of the sky, even a small error in the beam model can lead to large biases. We thus include beam parameters in our analysis. WSRT commonly adopts a simple model for the primary beam, 1 namely: $\cos^3(c\nu\theta)$, where ν is the observing frequency (in GHz), θ is the distance from the pointing centre in degrees and c is the beam factor (in 1/GHz). The beam factor (or beamwidth) is known to vary slightly with frequency. As a proof of concept, we assume it is unknown, and include it as a
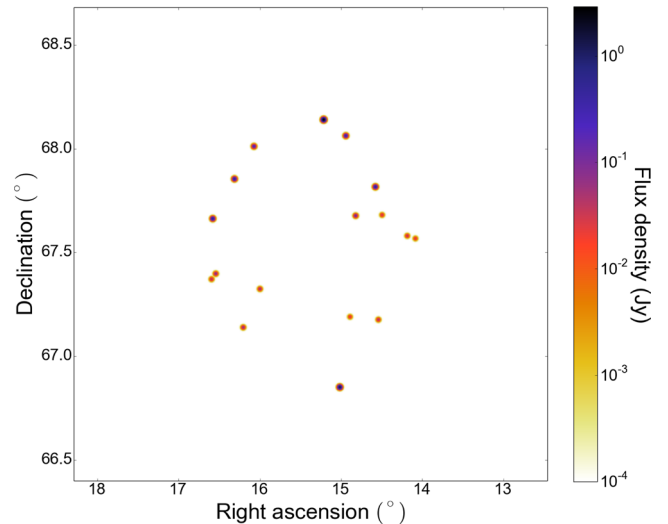
Figure 2. The simulated, noise-free sky model with 17 sources with flux densities varying between 0.03 and 3.13 Jy.
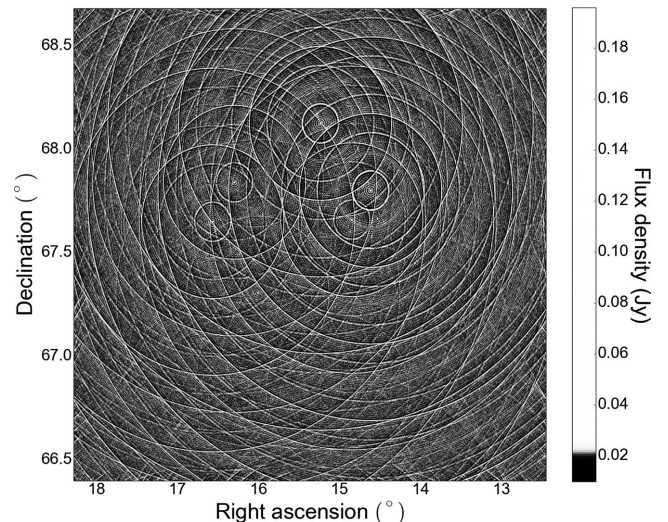

Figure 3. The dirty data set for the model of Fig. 2, as the telescope would see it (the colours are histogram-equalized to improve contrast). The image is produced directly from the visibilities and shows the typical ring structure around bright sources that is seen in interferometric data, due to the missing angular-scale information in the data set. The rms noise in flux density is about 0.28 mJy.

further instrumental parameter. One could provide a more complex model for the primary beam and easily fit those parameters with this technique as well, comparing the models with the Bayesian evidence. The model for the beam enters the RIME of equation (6) as a direction-dependent Jones matrix:

$$\mathbf{E}^{\text{BEAM}}(l, m) = \cos^3(c\nu\sqrt{l^2 + m^2})\,\mathbf{I}, \tag{8}$$

where $\mathbf{I}$ is the identity matrix.

Pointing errors

Pointing errors can substantially corrupt radio observations and are known to be a limiting factor in deep observations with WSRT (Smirnov & de Bruyn 2011) and other telescopes. The greatest effect is on sources on the flank of the primary beam, where the gradient of the beam pattern is steep, and a small pointing error

produces a larger error in apparent flux (compared to the centre of the beam). Since the errors can be different from antenna to antenna, this produces errors on the observed visibility amplitudes, which translates into artefacts in the image. Essentially, each source is 'defocused' in a complicated way. Thus, we can immediately suspect that there will be a correlation between the pointing errors and source flux densities. Two prior approaches to inferring pointing errors directly from the data have hinged on maximum-likelihood estimates. These are the pointing self-cal algorithm (Bhatnagar, Cornwell & Golap [2004]) and direct fitting with MEQTREES (Smirnov [2011c][3]). Neither approach estimates the correlation between pointing errors and source parameters, which the Bayesian approach naturally provides. We inject time-varying polynomial pointing errors for each of the 14 WSRT antennas. We use a second-order polynomial for each pointing error and fit for the coefficients. A polynomial pointing error in each orthogonal direction for each antenna results in a total of 84 pointing-error parameters. The pointing errors are written as a Jones matrix in equation (6):

$$\mathbf{E}_p^{\text{PE}}(l, m) = \mathbf{E}^{\text{BEAM}}(l + \delta l_p, m + \delta m_p),\tag{9}$$

where δl_p and δm_p are the pointing errors in the right ascension and declination direction, respectively, for antenna p. The pointing errors are taken to be time-varying polynomials, written as

$$\delta l_p = c_2 t^2 + c_1 t + c_0,\tag{10}$$

and similarly for δm_p, where t is time (rescaled over the observation) and c_k are the coefficients we determine with MCMC.

Noise

The noise on the visibilities is expected to be Gaussian, stationary and uncorrelated. Noise level can be estimated with some precision from the known system temperature, here however we show than it can also be inferred accurately directly from the data. We thus included one final parameter for the standard deviation of the noise on the visibilities.

4.1.4 Resulting measurement equation

The RIME for this example problem is thus

$$\mathbf{V}_{pq} = \sum_s \left(\mathbf{E}^{\text{BEAM}}(l_s + \delta l_p, m_s + \delta m_p) \mathbf{K}_p^{(s)} \mathbf{B}_s^{\text{POINT}} \mathbf{K}_q^{(s),H} (\mathbf{E}^{\text{BEAM}})^H \right.$$
$$\left. \times (l_s + \delta l_q, m_s + \delta m_q) \right),\tag{11}$$

where s runs from 1 to 17 over all the sources. This brings the total to 103 parameters: 17 scientific (the flux densities of the sources) and 86 instrumental (84 pointing error parameters, the beamwidth and the noise). The full model can be visualized in the Bayesian factor graph of Fig. [4] and a more detailed description of factor graphs is given in Appendix A.

4.2 Using MCMC for joint parameter inference

The initial step of our analysis was to choose an appropriate sky model in MEQTREES (specifying the brightness matrix in equation 6) and select the telescope configuration corresponding to the data set including all known sources of interference and instrumental errors (the Jones matrices in equation 6). We vary all the parameters within the model – the flux densities, pointing errors, beamwidth and
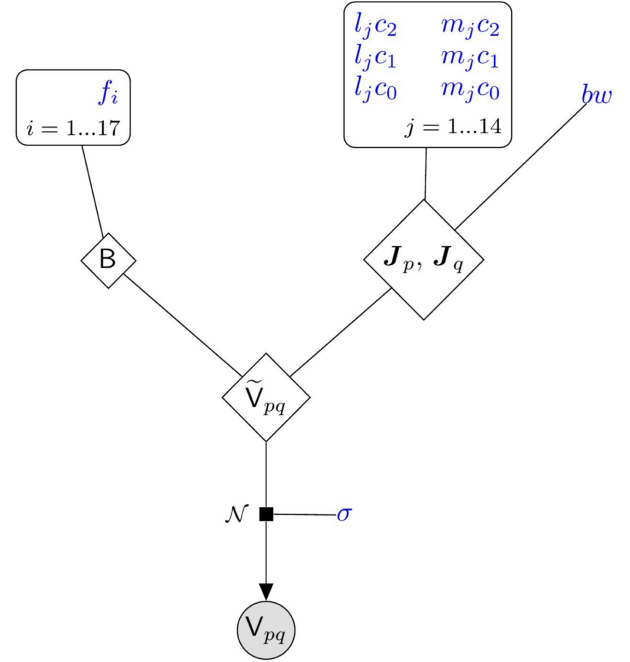
3 https://indico.skatelescope.org/getFile.py/access?contribId=20&sessionId =9&resId=0&materialId=0&confId=171

Figure 4. Bayesian factor graph (see Appendix A) of the model for the first simulated data set. All parameters we estimate with MCMC are the constants, without any circles around them, coloured blue. The $\mathbf{V}_{pq}$ are the observed visibilities, which are drawn from a normal distribution of mean $\widetilde{\mathbf{V}}_{pq}$ (the unobserved, true visibilities) and standard deviation σ, which is one of the parameters we estimate with MCMC. These 'true' visibilities are governed by the RIME, which is here simplified graphically to two components, the brightness matrix, $\mathbf{B}$, and the Jones' matrices of the antennas, $\mathbf{J}_p$, $\mathbf{J}_q$. The flux densities of the 17 sources are represented by f_i, which form components of $\mathbf{B}$. The coefficients of the polynomial time-varying pointing errors, $l_j c_k$ and $m_j c_k$ (where j represents the antenna number and k is the number of polynomial coefficient) enter the Jones matrices, along with the beam width, bw.

noise – using MCMC. Fig. [1] illustrates how the sampling algorithm repeatedly calls MEQTREES with new parameter values and evaluates the likelihood. MCMC uses the likelihood (equation 2) to determine the best-fitting parameter values and to explore the surrounding parameter space, thus determining the uncertainties and correlations for all parameters.

4.3 Technical details and priors

Due to the large volume of the parameter space, we use a standard, gradient-based optimization algorithm to get close to the best-fitting parameter values and provide a good starting point for the MCMC. We run several chains in parallel, each of around 500 000 steps, repeatedly computing and diagonalizing the covariance matrix to improve convergence, and we test convergence using the Gelman–Rubin statistic (Gelman & Rubin [1992]). The estimated parameters and their uncertainties are determined by finding the mean and standard deviation (using percentiles) from the marginalized one-dimensional posterior for each parameter. For this particular setup, MEQTREES takes about 0.4 s for one likelihood calculation, parallelized using four cores of 2.2 GHz each. As 10 chains were run, 40 cores in total were used resulting in approximately 55 CPU hours for convergence per data set.

We apply a uniform prior to the pointing error parameters, restricting them to the broad range of ± 200 arcsec. We also restrict

the beamwidth to be positive, and vary the noise on the visibilities in logarithmic space (with an infinitely broad prior in log-space). We do not restrict the ranges of the flux densities.

4.4 Comparison with CLEAN plus Source Extraction

To compare our technique with the standard approach, we apply CLEAN followed by a Source Extraction algorithm to determine the flux densities of the sources (we call this combination CLEAN+SE), without any instrumental calibration. We do not use any calibration algorithms such as self-cal, because it would have no benefit: our data set only has direction-dependent instrumental effects, whereas self-cal can only correct for direction-independent effects. Current approaches to direction-dependent calibration are of no help here because:

(i) Direction-dependent solutions (such as peeling, or differential gains) can in principle solve for the variable gains induced by pointing error, given a prior source model. However, this destroys information on the source, since deviations between the true sky and the prior model are completely absorbed by such gain solutions.

(ii) Pointing self-cal should in principle improve the CLEAN maps and thus produce better source model estimates. However, implementations of this remain unavailable to the public.

(iii) MEQTREES should in principle be able to do a maximum-likelihood solution for the source parameters and pointing errors simultaneously. However, only solutions for the latter has been demonstrated to work in practice and as we have argued, a maximum-likelihood solution produces a point estimate for the parameters which may be biased due to correlations.

Instead, we apply a naïve CLEAN algorithm, followed by Source Extraction, to compare with BIRO as a worst case scenario in the case of time-varying pointing errors. Note that we do provide prior information on the positions of the sources to CLEAN, in the form of CLEAN boxes.

We use the CLEAN implementation (specifically the Cotton–Schwab algorithm) in the software package CASA[4] to image the simulated data sets. The images were made with robust weighting with a robustness parameter of −1.0. We did 1000 iterations of CLEAN with a loop gain of 0.1. Interactive cleaning was performed on the visibility data twice, once with masks defined around known source positions and then with masks defined around only those sources that were found during the cleaning procedure. The source extraction was performed interactively using PYBDSM to ensure that the artefacts were not wrongly identified as sources.

4.5 Results

To illustrate fitting a model to the raw data, we plot a subset of the visibilities in Fig. 5 with the best-fitting visibilities as obtained by BIRO. Fig. 6 (with numerical details in Table 1) shows the comparison between the flux densities obtained by CLEAN+SE and those by BIRO. The flux densities of CLEAN+SE are on average biased due to undealt-with correlations with the pointing errors and underestimated uncertainties. Additionally, because of the time-varying pointing errors corrupting the data, CLEAN+SE only manages to find 5 of the 17 sources. With polynomial pointing errors included in the simulations, bright artefacts dominated the final image resulting in the weaker sources being swamped. In contrast, because these
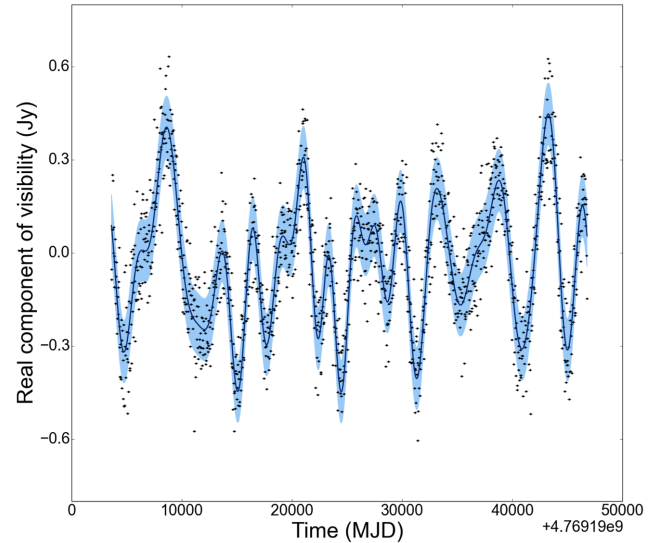
---

[4] Common Astronomy Software Applications, http://casa.nrao.edu/.

Figure 5. Example of fitting a model to the raw data. Plotted are the real component of the visibilities for a single baseline (between antenna 0 and 1) and for the single channel of the data, in black. The best-fitting model line is overplotted in dark blue, with a band of uncertainty of 0.1 Jy (the original noise added to the simulation) in pale blue.
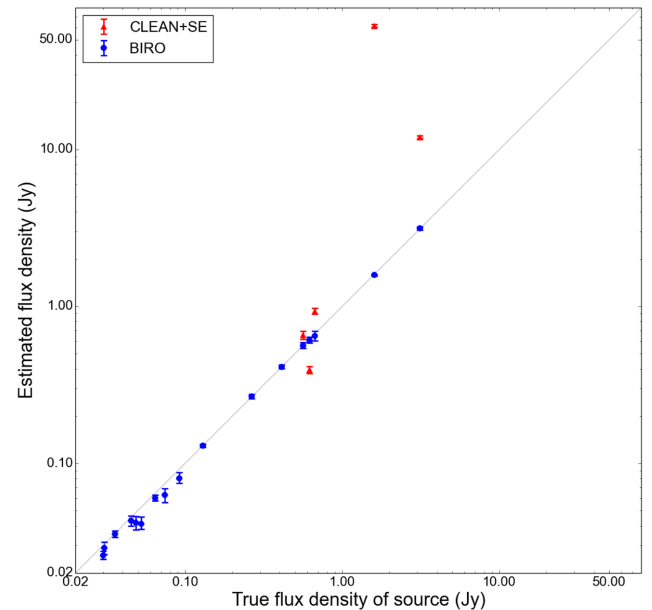

Figure 6. Estimated versus true flux densities of the sources with error bars as estimated by BIRO (blue circles) and by a CLEAN+SE algorithm (red triangles). Note that CLEAN+SE only detects 5 out of 17 sources. The BIRO error bars are the standard deviation of marginalized one-dimensional posterior for each flux parameter. While the BIRO results are unbiased, CLEAN+SE has two problems: it underestimates the error bars and yields biased estimates of the flux densities of up to 44σ. The reader is reminded that this data set contains no direction-independent effects that may normally cause biases in a CLEAN analysis; these biases are instead due entirely to the complexities in the data set introduced by the time-varying pointing errors.

Table 1. Comparison between the CLEAN+SE results (shortened to C+SE) and the BIRO results for the flux densities (in Jy) of the sources in the data set. The bias in terms of number of standard deviations away from the true flux density is given in brackets. For the five sources CLEAN+SE found, the error on the position was less than 10^{-4} deg.

Name	RA (°)	Dec (°)	True flux density (Jy)	C+SE flux density (Jy)	BIRO flux density (Jy)
F1	15.216	68.139	3.1285	$11.9723 \pm 0.2240\,(39.5\sigma)$	$3.1410^{+0.0828}_{-0.0710}\,(0.2\sigma)$
F2	15.017	66.849	1.6030	$61.4220 \pm 1.3557\,(44.1\sigma)$	$1.5834^{+0.0185}_{-0.0185}\,(1.1\sigma)$
F3	16.583	67.662	0.6719	$0.9345 \pm 0.0415\,(6.3\sigma)$	$0.6476^{+0.0478}_{-0.0410}\,(0.5\sigma)$
F4	14.578	67.815	0.6170	$0.3942 \pm 0.0191\,(11.7\sigma)$	$0.6114^{+0.0249}_{-0.0249}\,(0.2\sigma)$
F5	16.315	67.853	0.5648	$0.6579 \pm 0.0372\,(2.5\sigma)$	$0.5635^{+0.0272}_{-0.0227}\,(0.0\sigma)$
F6	14.943	68.061	0.4115	*Not found*	$0.4127^{+0.0135}_{-0.0112}\,(0.1\sigma)$
F7	16.075	68.010	0.2640	*Not found*	$0.2671^{+0.0093}_{-0.0093}\,(0.3\sigma)$
F8	14.822	67.676	0.1293	*Not found*	$0.1295^{+0.0031}_{-0.0031}\,(0.1\sigma)$
F9	16.546	67.396	0.0919	*Not found*	$0.0805^{+0.0071}_{-0.0059}\,(1.6\sigma)$
F10	16.207	67.138	0.0742	*Not found*	$0.0629^{+0.0064}_{-0.0064}\,(1.8\sigma)$
F11	16.001	67.323	0.0645	*Not found*	$0.0599^{+0.0029}_{-0.0023}\,(1.6\sigma)$
F12	14.541	67.175	0.0524	*Not found*	$0.0413^{+0.0041}_{-0.0031}\,(2.7\sigma)$
F13	16.598	67.370	0.0487	*Not found*	$0.0419^{+0.0040}_{-0.0040}\,(1.7\sigma)$
F14	14.186	67.579	0.0451	*Not found*	$0.0430^{+0.0031}_{-0.0031}\,(0.7\sigma)$
F15	14.496	67.680	0.0357	*Not found*	$0.0354^{+0.0017}_{-0.0015}\,(0.2\sigma)$
F16	14.085	67.566	0.0306	*Not found*	$0.0289^{+0.0026}_{-0.0026}\,(0.6\sigma)$
F17	14.891	67.188	0.0301	*Not found*	$0.0259^{+0.0017}_{-0.0013}\,(2.5\sigma)$

correlations are taken into account, the Bayesian approach is able to recover the true flux densities for all sources and to determine error bars that include the effects of all nuisance parameters. Without the instrumental errors, BIRO achieves similar flux estimates to CLEAN+SE.

Fig. 7 shows a subset of the covariance matrix between parameters and Fig. 8 shows an example 1σ and 2σ contour plot between pairs of parameters. The key result of Fig. 7 is that it highlights the significant and complex correlations between the pointing errors and flux densities, i.e. the instrumental and science parameters, which therefore need to be estimated jointly allowing for the correlations.

The (anti)correlations between pointing errors and flux densities are easy to understand qualitatively. Consider a source on the flank of the main lobe of the primary beam, e.g. on the half-power point. If a given antenna mispoints *towards* the source, the source will be subject to a higher primary beam gain, in other words, it will be perceived as brighter by all baselines involving that antenna. Mispointing *away* from the source has the opposite effect. The nature of the correlation will also strongly depend on the position of the source with respect to the pointing centre. For example, a source near the centre of the main lobe (i.e. on a 'flat' part of the primary beam pattern) will correlate very weakly with pointing error, while a source on the inner flank of the first side-lobe will correlate with mispointing *away* rather than *towards*. Since different baselines contribute to different Fourier mode measurements, pointing error will also have a complicated interaction with perceived source structure. Similar arguments apply to beamwidth.

Deriving the exact quantitative nature of this correlation analytically is highly impractical, which is why a technique like BIRO proves so powerful. This covariance matrix could be used to assist in calibration, study calibration parameters or as input to future MCMC analyses on similar data sets.

5 EXAMPLE 2: MODEL COMPARISON

In this example problem, we show that BIRO is able, using model selection (Jeffreys 1998; Trotta 2008), to choose the correct model in each of three different cases, distinguishing between an extended source, an unresolved point source and two close (sub-synthesized-beam) sources. The sources recovered are all smaller than the synthesized beam. This is known as super-resolution and has recently been shown to be possible with compressive sensing (Wiaux et al. 2009; Li, Cornwell & de Hoog 2011; Carrillo, McEwen & Wiaux 2012, 2014; Honma et al. 2014, and to some extent Martí-Vidal, Pérez-Torres & Lobanov 2012). Here we use the Bayesian evidence to determine the correct model of these sub-synthesized-beam sources, with statistical significance. Although in this example problem we exclude instrumental effects, they can, in general, be included as in example 1.

5.1 Simulated data sets and models

The data sets for this example use the same frequency, bandwidth, integration time and noise characteristics as the data set simulated in Section 4. We simulate three data sets with three different sky models with all the sources away from the phase centre: a point source, a sub-synthesized-beam extended source modelled as a Gaussian and two point sources separated by the distance the size of that Gaussian. No instrumental effects were included in the model-selection simulations and the beamwidth and noise were assumed to be known. Fig. 10 shows the input model for all three cases in the left-hand column.

The point sources are parametrized by the Stokes *I* flux density and the position as the distance from the phase centre, along two mutually perpendicular axes, *l* and *m*. The extended Gaussian source has three more parameters in the form of the projections of the major
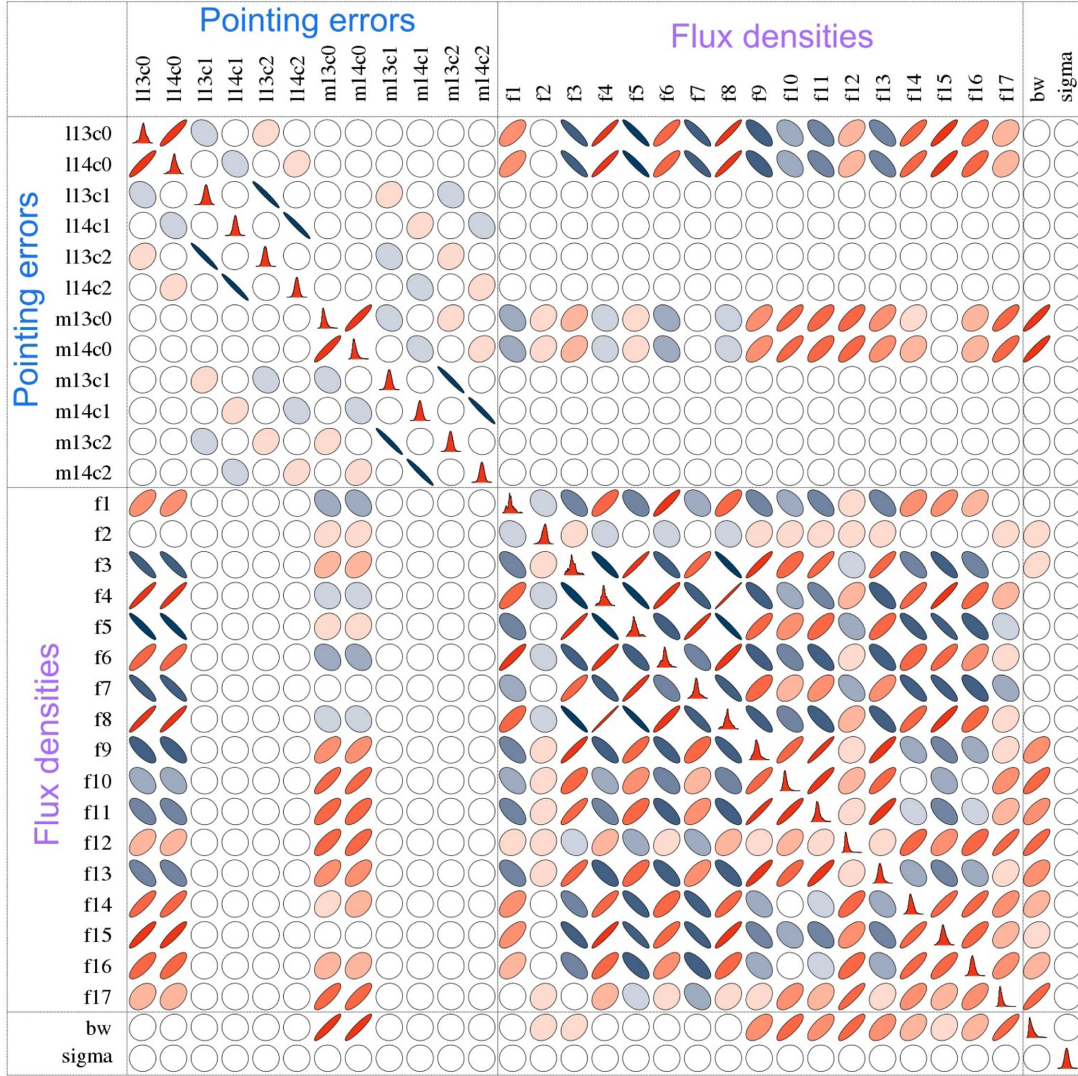
Figure 7. Covariance matrix between a subset of parameters illustrating the strong correlations between the science and instrumental parameters that must be accounted for to achieve unbiased results. The parameters are listed on each axis with the correlations between them represented by a coloured ellipse, either positive (red ellipse angled to right) or negative (blue ellipse angled to left). The leading diagonal shows the one-dimensional marginalized posterior for each parameter. For the pointing errors, l_jc_k refers to the kth coefficient of the polynomial time-varying pointing error in the right ascension direction for the jth antenna and m_jc_k is the same for the declination direction. The flux densities of the 17 sources are given by f_i, ordered from brightest to faintest, and bw and sigma represent the beamwidth and noise on the visibilities, respectively.

axis on the l and m axes and the ratio of the minor to major axis, defined as

$$l_\perp = e_{\text{maj}} \sin(\alpha) \qquad (12)$$

$$m_\perp = e_{\text{maj}} \cos(\alpha) \qquad (13)$$

$$r = e_{\text{min}}/e_{\text{maj}}, \qquad (14)$$

where e_{maj} and e_{min} are the major and minor axes of the Gaussian source and α is the position angle (the angle of rotation of the extended source). See Fig. 9 for a visual description. The brightness matrix of equation (6) for an extended Gaussian is simply the product of a Gaussian and the brightness matrix for a point source. The RIME is simple in this example, since there are no instrumental effects apart from the usual phase shift between antennas:

$$\mathbf{V}_{pq} = \sum_s \left(\int \int_{lm} \mathbf{K}_p^{(s)} f(l, m) \mathbf{B}_s^{\text{POINT}} \mathbf{K}_q^{(s),H} \mathrm{d}l \, \mathrm{d}m \right), \qquad (15)$$

where $f(l, m)$ is a Gaussian in l and m for the extended source case and f is a delta function for the one- and two-source models. Also in the one- and two-source models, l and m reduce to single points l_s and m_s, as in equation (11).

5.2 Using MULTINEST for model selection

We use MULTINEST for calculating the Bayesian evidence (see Section 2) and MEQTREES for predicting the model visibilities from the sampled source parameters from which the likelihood is computed iteratively. The likelihood is computed according to equation (2). The posterior probability distributions are obtained as a by-product along with the uncertainties in the best-fitting parameter values and the Bayesian evidence.

For the single-point-source model, we vary three parameters: the flux density and relative source position, l and m. We similarly vary the flux densities and positions of the two sources in the two-source
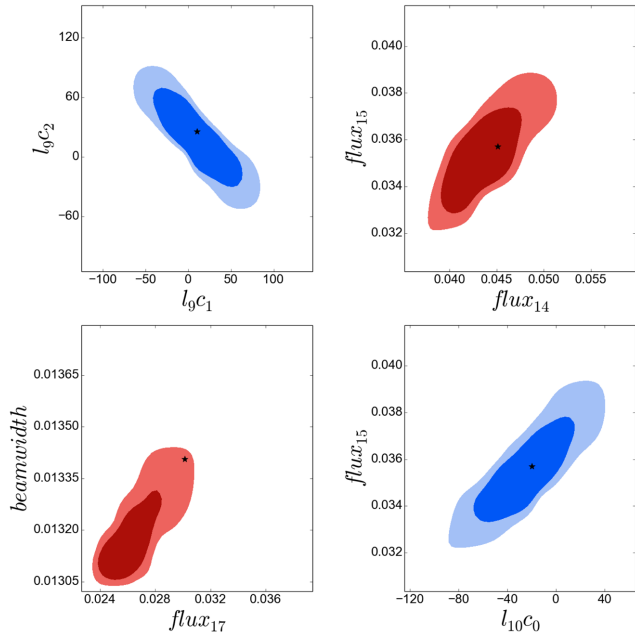
Figure 8. Credible interval contour plots between a subset of parameters. The 1σ and 2σ probability densities are shown in dark and light colours, respectively. The true (input) parameters are marked with a black star. The pairs of parameters are as follows. Upper left: the two highest order coefficients of the pointing error in the right ascension direction for antenna 9. Upper right: flux densities of two of the sources. Lower left: the flux density of the 17th source versus the beamwidth. Lower right: the constant term from the polynomial pointing error in the right ascension direction for antenna 10 versus the flux density of the 15th source.
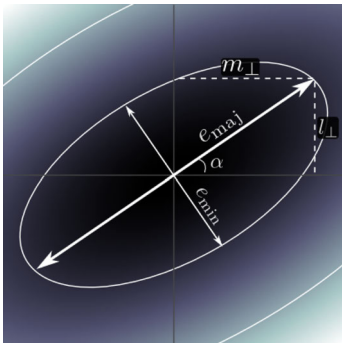

Figure 9. The parametrization of a Gaussian extended source in MEQTREES. Here, e_{maj} and e_{min} are the major and minor axes of the Gaussian and α is the position angle. MEQTREES uses $l_\perp$, $m_\perp$ and $r = e_{min}/e_{maj}$ in its parametrization of a Gaussian.

model. The Gaussian extended source model has six parameters: the flux density, position coordinates and the shape parameters ($l_\perp$, $m_\perp$, r).

We generate a unique, simulated data set for each of the three cases and then fit each of the three models to them, to see if the correct model is selected in each case. MULTINEST fits for the parameters, their uncertainties and correlations (just as MCMC does in example 1), but also returns the evidence, $\mathcal{Z}(\boldsymbol{D}|H)$ (the probability of the data, given the hypothesis). By taking the ratio of evidences, one can determine whether one model is favoured over another, and by how much. The Jeffrey scale (Jeffreys 1998; Trotta 2008) provides an intuitive way of deciding whether the evidence is strong

Table 2. Relative evidences for each model in each simulated data set. A is the two-source model, B is the extended source model and C is the one-source model. The pieces of evidence are relative to the model used to generate the data set (so, for example, for the two-point-source data set, the evidence for each model is compared to the two-point-source model). The maximum error in log-evidence is 1.5. High odds indicate that the input model is favoured (as it is in all three cases), showing that nested sampling selects the correct model at high significance (at an SNR of 1000).

Simulation	Hypothesis		
	A	B	C
A	1 : 1	10^{593} : 1	10^{7200} : 1
B	10^{993} : 1	1 : 1	10^{5079} : 1
C	62 : 1	857 : 1	1 : 1

enough to select a model, based on odds derived directly from the evidence.

5.3 Technical details and priors

We use uniform priors for all the source parameters. The flux density is restricted to the range 0 to 2 Jy. The position parameters are allowed to be both positive and negative in the range -25 to 25 arcsec since the position is measured relative to the phase centre. For the shape parameters of the extended source, ($l_\perp$ and $m_\perp$), we allow the prior ranges to be big enough to encompass the point spread function of the interferometer and no more, since we are dealing with sub-synthesized-beam sources. This translates to a range of 0–20 arcsec for $l_\perp$ and -20 to 20 arcsec for $m_\perp$. Finally, we restrict the minor-to-major axis ratio (r) to be positive, but less than unity to be physically meaningful. We found that using 1000 live points achieved good results from MULTINEST.

5.4 Results

The relative logarithmic evidences are computed for each model giving the relative confidence with which one model is preferred over another (see Table 2). We find that the correct hypothesis is selected in all cases, at odds of 10^{593}:1, 10^{993}:1 and 62:1, for the two-point-source, extended-source and single-point-source models, respectively. Using model selection, BIRO is able to select the correct model in all three cases (the model with the highest evidence), showing it can perform source separation even on sub-synthesized-beam scales.

We computed a 'best-fitting' image by running MEQTREES with the maximum posterior model and parameters in each of the three cases, to compare with the CLEANed image (see Fig. 10). We use the same CLEAN parameters as in Section 4.4. The CLEANed images (at least in this case, without an enforced smaller beam size) are unable to reach the sub-synthesized-beam scales achievable by BIRO.

In Fig. 11, we determine the point at which model selection fails to distinguish an extended source from a point source for different source sizes and signal-to-noise ratios (SNRs). Any evidence lower than 'strong' is not usually considered high enough to say either way which model is correct. Perhaps obviously, at high SNR extremely small sources can be detected (around 1.0 arcsec) and sources become more difficult to distinguish as the SNR is reduced.

Video 1 in the online-only content shows visually how MULTINEST converges to the correct model, exploring the posterior as it goes,
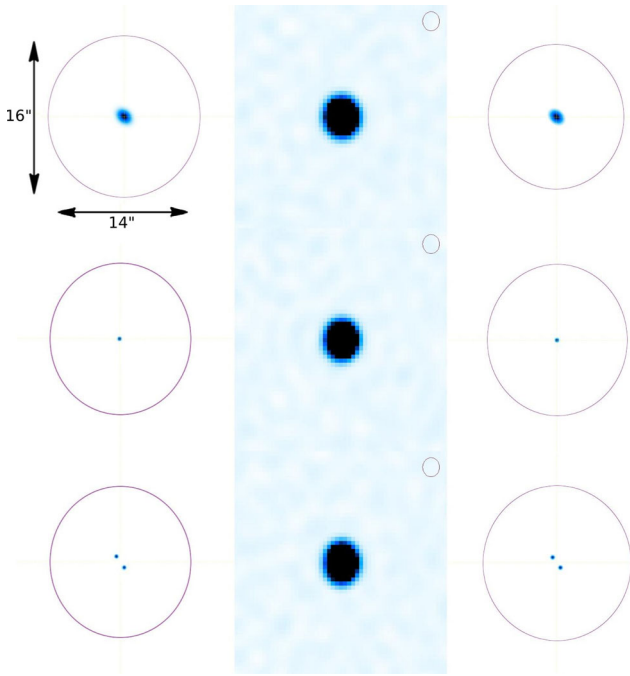
for the extended source model. Each frame is an image generated using the parameters from every 40th step of the chain.

6 DISCUSSION AND CONCLUSIONS

We have introduced the technique BIRO, a Bayesian approach to the deconvolution problem of radio interferometry. Instead of making an image and then performing source extraction, BIRO uses MCMC or nested sampling to fit models directly to the visibility data and obtain the posterior for the parameters of interest, as well as nuisance parameters.

In the first example problem, we focused on the relationship between scientific and instrumental parameters. It was found that all parameter estimates from BIRO were consistent within their error bars with the true values. As well as determining the uncertainties of the parameters, BIRO also returns the covariance matrix between them, as a by-product of the full posterior. Our work shows that these correlations are complicated and non-negligible. BIRO effortlessly incorporates the effects of the correlations in the estimates of the marginalized uncertainties on the individual parameters, as well as providing a way to study these correlations in the form of the covariance matrix. We compared our results to a standard CLEAN algorithm, without calibration (since our simulated data contains *only* direction-dependent effects and publicly available calibration algorithms only deal with direction-independent effects). Because of the time-varying pointing errors we introduce to the data set, CLEAN is only able to find 5 out of the 17 sources and returns biased flux densities for them, while BIRO returns unbiased flux densities for all sources. BIRO is also able to correctly determine the coefficients of the time-varying pointing errors, the primary beamwidth and the noise on the visibilities.

In the second example problem, we addressed the issue of how to determine the best sky model for the data. We worked with three models: a single point source, a Gaussian extended source and two point sources. We simulated data for each of the three models and then, for each data set, ran MULTINEST to fit each of the three
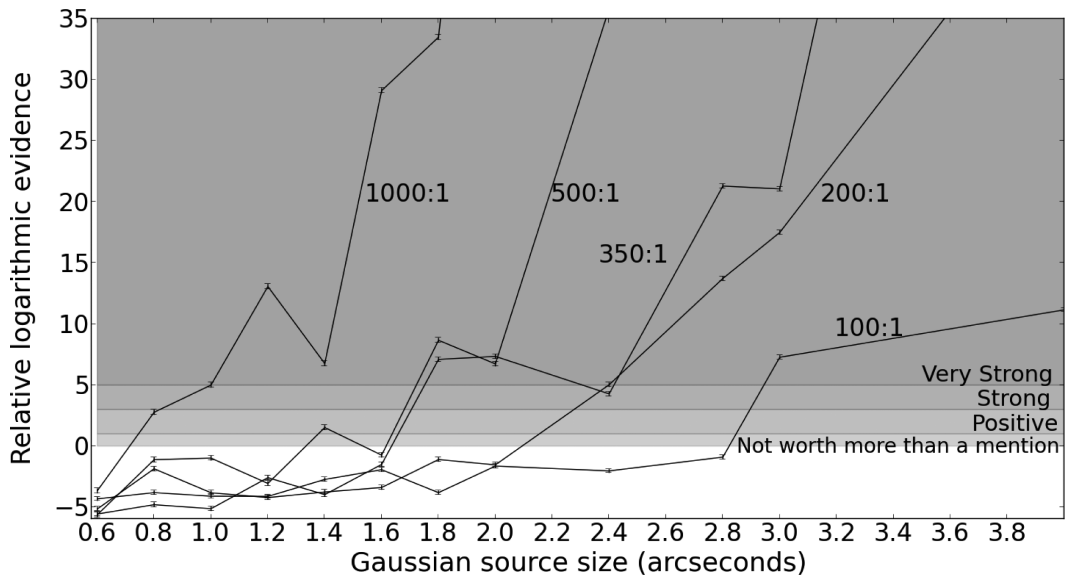

Figure 10. Left-hand column: the true sky for the extended Gaussian, single-point-source and two-point-source models (from top to bottom). Middle column: the CLEANed image for the three models. Right-hand column: the maximum posterior BIRO image for the three models. The purple contour in each image indicates the size of the synthesized beam, as returned by CLEAN (note that the sources are all much smaller than the synthesized beam). BIRO recovers the correct input model each time while CLEAN is unable to distinguish between the models at the same SNR (in this case the SNR was 1000).

Figure 11. Relative natural log-evidence (i.e. the natural logarithm of the ratio of the Bayesian evidence for the true model to that of a single point source) as a function of Gaussian source size, for the extended source input model, showing the evidence-crossover points for different source sizes and SNRs (peak flux to background noise). The horizontal axis gives the size of the circular Gaussian source in the input model (the reader is reminded that the FWHM of the synthesized beam is around 13 arcsec). The vertical axis gives the odds in favour of the Gaussian source model when model comparison is performed for the Gaussian model against a point source model. The more positive the relative log-evidence is, the more strongly is the Gaussian model favoured. Each curve on the graph is for a different noise level with the approximate (map) SNRs shown in the legend.

models and determine the Bayesian evidence. The evidence then determines the selection of the correct model. All of the sources detected were several times smaller than the synthesized beam, hence we successfully achieved super-resolution as well as source separation.

This paper constitutes a proof of concept but more work is required before the technique can be easily applied to interferometric data set.

(i) First, while using a WSRT simulation has relevance to the SKA due to the similar instrumental setup, the SKA will have many more antennas (of the order of a thousand) which will of course result in many more instrumental parameters (and indirectly more science parameters as the source count increases with sensitivity). Fortunately, while the number of instrumental parameters scales as the number of antennas, N, the number of data points scales as the number of baselines, i.e. $\mathcal{O}(N^2)$, meaning it is plausible that one could simultaneously determine the sky and instrumental parameters for large N. While the precedent for sampling an extremely large parameter space exists (Jasche & Wandelt 2012), new and sophisticated sampling techniques (Duane et al. 1987; Neal 2012; Goodman & Weare 2010; Foreman-Mackey et al. 2013) (which are also easily parallelized) will be required to improve convergence in the thousand-parameter regime, especially as the non-linear nature of the modelling makes sampling inefficient (as addressed in Jasche & Wandelt 2012).

(ii) Secondly, the Bayesian approach is far more computationally intensive than standard deconvolution, taking hours (55 CPU hours in the case of example 1) to converge to the correct posterior distribution. The complexity of the likelihood computation scales as the number of antennas squared (i.e. the number of baselines), making an SKA-like computation difficult with the current setup. However, the RIME is intrinsically highly parallelizable allowing an efficient implementation of MEQTREES on GPUs. Preliminary work on a GPU implementation indicates a speed-up of the likelihood computation of about 250 times (Perkins et al. 2015). This means this technique can be applied to data from existing telescopes such as ALMA (Hills & Beasley 2008) and LOFAR (van Haarlem et al. 2013), using current computer clusters.

(iii) Thirdly, we need to address the problem of not knowing the sky model beforehand, which is a common difficulty when dealing with calibration but is particularly important here, as a Bayesian analysis relies on a good model. There are a number of ways to tackle this issue which we hope to address in future publications. A simple, but computationally intensive, solution would be to run several different models (with increasing numbers of sources) and select between them using the Bayesian evidence. Another possible approach is to use a deconvolution algorithm, like CLEAN or RESOLVE, to get an initial set of sources and then iterate between deconvolution and the best fit of BIRO to get a subsequently better model. A more rigorous solution would be to use an algorithm such as birth–death (Stephens 2000) or reversible jump (Green 1995) MCMC, which is able to determine both the number of parameters required and the posterior for them simultaneously. A further possibility is to combine the more general approach proposed in Sutter et al. (2014) and Junklewitz et al. (2013), which divides up the field into many 'pixels' that are then allowed to vary, with the calibration capabilities of BIRO to produce estimates of the sky model. This is even more computationally challenging however, but would provide a more general and robust solution.

BIRO is not only useful for dealing with systematics, which will become more important as telescopes become more sensitive, but it is also a powerful technique for lending statistical strength to topical scientific questions. Potential applications include structures of black hole systems, jet emission in active galaxies, time variability of objects and radio weak lensing. BIRO allows a holistic way to include instrumental effects while at the same time returning the science we are interested in. By leveraging the power of Bayesian statistics, BIRO uses all information available to get the most out of interferometric data sets.

Video 1: Online only (also available at https://vimeo.com/117391380). Images generated from the MULTINEST chain for the extended Gaussian source data set and model. At every 40th step in the chain, that step's parameters were used to generate an image of the field. The parameters are at first quite variable but soon converge to the correct shape, position and flux density for the source. The sample probability, which is the normalized posterior for that point, improves as the chain converges to the correct parameter values.

AUTHOR CONTRIBUTIONS

ML performed the research for the joint estimation example and wrote the majority of the paper. IN performed the research for the source separation problem and wrote the corresponding sections of the paper and did some CLEAN analysis and wrote the corresponding section. JZ assisted with the source separation research and wrote part of the paper. OS provided the simulated data sets and led the technical aspects of radio observation modelling. BB developed the original idea. NO performed some CLEAN analysis. MK contributed to the conceptual development of ideas and provided support with the statistical methods. All authors commented on the research and edited the paper.

REFERENCES

AMI Consortium, 2012, MNRAS, 423, 1463

Bhatnagar S., Cornwell T. J., Golap K., 2004, Tech. Rep., EVLA Memo 84: Solving for the Antenna Based Pointing Errors. NRAO, Green Bank, WV

Bhatnagar S., Cornwell T. J., Golap K., Uson J. M., 2008, A&A, 487, 419

Carilli C. L., Rawlings S., 2004, New Astron. Rev., 48, 979

Carrillo R. E., McEwen J. D., Wiaux Y., 2012, MNRAS, 426, 1223

Carrillo R. E., McEwen J. D., Wiaux Y., 2014, MNRAS, 439, 3591

Dietz L., 2010, Tech. Rep., Max Planck Institute for Informatics

Dorn S., Enßlin T. A., Greiner M., Selig M., Böhm V., 2015, Phys. Rev. E, 91, 013311

Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Phys. Lett. B, 195, 216

Enßlin T., 2014, in Niven R. K., Brewer B., Paull D., Shafi K., Stokes B., eds, AIP Conf. Proc. Vol. 1636, Proc. 33rd Int. Workshop on Bayesian

Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 49

Enßlin T. A., Junklewitz H., Winderling L., Greiner M., Selig M., 2014, Phys. Rev. E, 90, 043301

Feroz F., Hobson M. P., 2008, MNRAS, 384, 449

Feroz F., Hobson M. P., Zwart J. T. L., Saunders R. D. E., Grainge K. J. B., 2009, MNRAS, 398, 2049

Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Gelman A., Rubin D., 1992, Stat. Sci., 7, 457

Goodman J., Weare J., 2010, Commun. Appl. Math. Comput. Sci., 5, 65

Green P. J., 1995, Biometrika, 82, 711

Hamaker J. P., Bregman J. D., Sault R. J., 1996, A&AS, 117, 137

Hastings W. K., 1970, Biometrika, 57, 97

Hills R. E., Beasley A. J., 2008, in Stepp L. M., Gilmozzi R., eds, Proc. SPIE Conf. Ser. Vol. 7012, Ground-based and Airborne Telescopes II. SPIE, Bellingham, p. 70120N

Högbom J. A., 1974, A&AS, 15, 417

Högbom J. A., Brouw W. N., 1974, A&A, 33, 289

Honma M., Akiyama K., Uemura M., Ikeda S., 2014, PASJ, 66, 95

Jasche J., Wandelt B. D., 2012, MNRAS, 425, 1042

Jeffreys H., 1998, The Theory of Probability. Oxford Univ. Press, Oxford

Jones R. C., 1941, J. Opt. Soc. America, 31, 488

Junklewitz H., Bell M. R., Selig M., Enßlin T. A., 2013, preprint (arXiv:1311.5282)

Junklewitz H., Bell M. A., Enßlin T., 2014, preprint (arXiv:1401.4711)

Kazemi S., Yatawatta S., 2013, MNRAS, 435, 597

Kazemi S., Yatawatta S., Zaroubi S., Lampropoulos P., de Bruyn A. G., Koopmans L. V. E., Noordam J., 2011, MNRAS, 414, 1656

Lancaster K. et al., 2005, MNRAS, 359, 16

Li F., Cornwell T. J., de Hoog F., 2011, A&A, 528, A31

Martí-Vidal I., Pérez-Torres M. A., Lobanov A. P., 2012, A&A, 541, A135

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, J. Chem. Phys., 21, 1087

Neal R. M., 2012, in Brooks S., Gelman A., Jones G., Meng X., eds, Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, Boca Raton, FL

Noordam J. E., 2000, in Smolders A. B., Haarlem M. P., eds, Perspectives on Radio Astronomy: Technologies for Large Antenna Arrays. Conf. Proc., ASTRON, p. 307

Noordam J. E., Smirnov O. M., 2010, A&A, 524, A61

Pearson T. J., Readhead A. C. S., 1984, ARA&A, 22, 97

Perkins S., Marais P., Zwart J., Natarajan I., Smirnov O., 2015, preprint (arXiv:e-prints)

Scott P., Ryle M., 1977, MNRAS, 178, 539

Skilling J., 2004, in in Fischer R., Preuss R., von Toussaint U., eds, AIP Conf. Proc. Vol. 735, 24th Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 395

Smirnov O. M., 2011a, A&A, 527, A106

Smirnov O. M., 2011b, A&A, 527, A107

Smirnov O. M., 2011c, A&A, 527, A108

Smirnov O. M., de Bruyn A. G., 2011, preprint (arXiv:1110.2916)

Stephens M., 2000, Ann. Stat., 28, 40

Sutter P. M., Wandelt B. D., Malu S. S., 2012, ApJS, 202, 9

Sutter P. M. et al., 2014, MNRAS, 438, 768

Sutton E. C., Wandelt B. D., 2006, ApJS, 162, 401

Trotta R., 2008, Contemp. Phys., 49, 71

van Haarlem M. P. et al., 2013, A&A, 556, A2

Wiaux Y., Jacques L., Puy G., Scaife A. M. M., Vandergheynst P., 2009, MNRAS, 395, 1733

Zwart J. T. L. et al., 2011, MNRAS, 418, 2754

APPENDIX A: BAYESIAN FACTOR GRAPHS

Here we introduce Bayesian factor graphs, useful tools for visualizing Bayesian models, which we use to describe the model of

Table A1. Factor graph node types (adapted from Dietz 2010). The concept of a plate is worth an extra mention. Frequently in models variables are repeated, such as the 17 flux densities or 14 sets of pointing errors in our model in example 1. A plate in a factor graph allows one to easily show these variables are repeated, but each can have a unique value. So in the case of the source flux densities, m would range from 1 to 17, the value of $\mathcal{M}$.

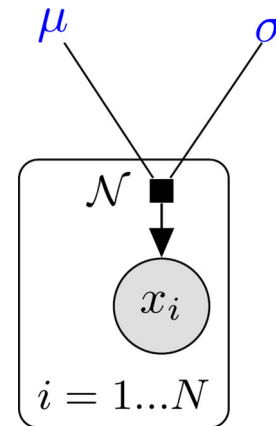| Type	Description	Output
Latent variable	An unobserved, undetermined variable	x
Observed variable	The data	y
Deterministic	A variable completely determined by other variables	z
Constant	A 'true' variable. We usually want to determine these	a
Factor	A distribution to be drawn from	$\mathcal{N}$ ■
Plate	Repeats variables (all the sources in the sky model for example)	x_m $m \in \mathcal{M}$

Figure A1. A simple example factor graph. In this model, the data are represented by a vector x_i, which we suspect is normally distributed. This is modelled by a normal distribution (represented by the factor labelled $\mathcal{N}$) which is governed by the parameters μ and σ. These constants would be the parameters we would want to estimate with an MCMC or MULTINEST analysis.

Section 4. We make use of the directed factor graph notation, developed in Dietz (2010), to visualize how the parameters in our models depend on one another. Table A1 defines the graphical primitives of a factor graph. Fig. A1 demonstrates the use of the factor graph notation in a simple example.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

video1.mp4 (http://mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stv679/-/DC1)

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.