EVLA Memo – 102
Monte Carlo Methods for Bayesian Image Reconstruction and Analysis in Radio Astronomy

Urvashi R. V.[*] and T. J. Cornwell[†]

February, 2006

Abstract

Maximum a-posteriori methods used in image restoration usually result in a single most-probable image, with no additional statistical information. Further, complex formulations of the problem, resulting in poorly conditioned systems, can make direct solution and even regularization techniques infeasible. Markov Chain Monte Carlo (MCMC) techniques sample the posterior distribution to obtain statistical information about the reconstructed image and can potentially provide a solution to both these problems.

Sections 1–4 of this report describe the Bayesian interpretation of image reconstruction and discuss an application using the Maximum Entropy image prior. Section 5 discusses the application of MCMC techniques to obtain error estimates in the context of component-fitting of radio interferometric images. Such image analysis has traditionally been based on likelihood techniques applied to deconvolved images. This analysis usually ignores uncertainties arising from fitting components to extended emission as well as from the process of deconvolution itself. We present an approach in which a Bayesian image analysis is performed to fit elliptical gaussian components to sub-regions of the dirty image, taking full account of the point spread function. Our method samples the posterior distribution to estimate the relative probabilities and uncertainties associated with the number of components and their parameters. This information can augment the process of object detection and characterization.

[*] National Radio Astronomy Observatory/New Mexico Tech, Socorro,NM 87801, USA.
email:*rurvashi@nrao.edu*

[†] Australia Telescope National Facility, PO Box 76, Epping, NSW 2121,Australia.
email:*tim.cornwell@csiro.au*

Contents

1 Bayesian Inference

1.1 Bayes theorem

Bayes theorem can be used to formulate the problem of image reconstruction in terms of images modeled as random variates of a multi-dimensional conditional probability distribution. It shows how existing (a priori) information about an image can be modified by new information (in the form of observed data) to generate information about the actual image that was observed. The conditional probability $P(I^M|D, M)$ of an image I^M given the observed data D and a priori information M can be calculated from the following posterior distribution.

$$P(I^M|D,M) \;=\; \frac{P(D|I^M,M)P(I^M|M)P(M)}{\int_M P(D|M)P(M)} \;\propto\; P(D|I^M,M)P(I^M|M) \qquad (1)$$

where $I^M = \Sigma_i P_i$ denotes the model image as a collection of flux components.

$P(D|I^M, M)$ is the likelihood distribution, and gives a measure of the distance between an image I^M and the data D. The noise in the data measurements is taken to be Gaussian ($N(0, \sigma^2)$ and uncorrelated. The residual (noise) per data point i is given as $D_i - A(I^M)_i$ where I^M is an image, and A is a transform from the image to data space. The probability of obtaining this residual will be

$$P(D_i|I_i^M, M) = e^{-\frac{1}{2}\frac{(D_i - A(I^M)_i)^2}{\sigma_i^2}} \qquad (2)$$

Since each data point is independent, the joint probability $P(D|I^M, M)$ over all measured data points is given as follows.

$$P(D|I^M, M) = e^{-\frac{1}{2}\sum_i \frac{(D_i - A(I_i^M))^2}{\sigma^2}} \;\propto\; e^{-\frac{1}{2}\chi^2} \qquad (3)$$

For deconvolution in radio interferometry,

$$\chi^2 \;=\; \Sigma\left(V^{obs} - S.V^M\right)^2 \qquad (4)$$

where V^{obs} represents the observed visibilities, S is the visibility sampling function and $V^M \;=\; F.I^M$ represents the Fourier transform of the model image I^M to the visibility (data) space. Maximizing the likelihood function is equivalent to an unconstrained minimization of χ^2.

Equivalently, in the image domain,

$$\chi^2 \;=\; \Sigma\left(I^{D^2} - I^M\left[I^D + I^R\right]\right) \qquad (5)$$

where I^D is the dirty image and I^R is the residual image.

$\int_M P(D|M)P(M)$ is the marginal distribution of the data and serves as a normalizing constant for the posterior distribution. For a known model M, $P(M)$ is also a constant.

$P(I^M|M)$ is the prior distribution which provides a measure of how well the model image I^M conforms to known information about the component parameters. It serves to bias the posterior distribution towards a priori information when the data is inconclusive. A well known prior is the entropy prior, which is pixel-based and gives the probability of a model image I^M based on a measure of distance between I^M and an a-priori image I^P. Another kind of prior suitable for a parameterized flux component based image representation involves separate probability distributions for each type of parameter. For instance, position parameters of the flux components could have a uniform distribution within the region of interest, and scale parameters could have a non-uniform distribution reflecting the actual observed distribution of scale in a typical image.

The most probable image can be obtained by maximizing the posterior distribution function. This is equivalent to a constrained minimization of χ^2 in a search space whose dimensionality is given by the total number of component parameters being fitted.

1.2 Maximum Entropy Formulation

The classical Maximum Entropy deconvolution algorithm is based on the Bayesian interpretation of image restoration, where a priori information is used to constrain the reconstruction of an image from observed data. Maximum Entropy methods aim to obtain the most probable non-negative image consistent with the data, based on the number of ways in which such an image could have arisen (Narayan & Nityananda 1986; Cornwell & Evans 1985; Skilling & Bryan 1984).

$P(I^M|M)$ is the probability density function corresponding to a priori information. Given a total flux, the joint probability of all possible ways this flux could be distributed over the image in the form of 'flux quanta' leads to the following form of 'relative entropy'. Maximizing entropy corresponds to finding the image instance that could have arisen in the most number of ways, given the model bias.

$$H = -\sum_i I_i^M \; ln\left(\frac{I_i^M}{M_i}\right) + \sum_i (I_i^M - M_i) \tag{6}$$

The first term is derived from a combinatorial argument of the number of ways the total flux is distributed (as integral multiples of a flux quantum) across bins (pixels). Stirling's approximation applied to the natural logarithm of the probability of a particular configuration, results in the above form of entropy. The logarithm ensures that the reconstructed image I^M has the same sign as the model image M. Choosing M to be positive, therefore ensures positivity in the reconstruction. The second term is a total flux constraint.

The probability of a particular configuration (an image I^M) given the model M is therefore

$$P(I^M|M) = e^{\frac{H}{\alpha}} \tag{7}$$

where α is a scaling constant that represents the magnitude of a flux quantum.

Combining the model and data terms as in Eqn 1,

$$P(I^M|D,M) = e^{-\frac{1}{2}\chi^2}\ e^{\frac{H}{\alpha}} \tag{8}$$

In a numerical maximization, the data term is the goodness-of-fit criterion and the model based entropy and total flux terms are regularizers.

MEM (purely entropy based) is a zero-scale deconvolution algorithm where each pixel in the image is treated as an independent degree of freedom. Pixel based correlations enter the system only via the transfer function of the instrument (smoothing by the psf), in the computation of χ^2 for the data distribution. In the case of a zero scale point spread function, the joint N^2-dimensional probability density function over all pixels in the image is the product of N^2 one-dimensional pdfs corresponding to individual pixels. Figure 1 is a plot of the probability density functions corresponding to the model (entropy prior), the data (χ^2) and the posterior distribution, for a zero-dimensional image (a single pixel). The plots in Fig 1 and Fig 2 are for $M = 30, D = 70$ with $\sigma_{noise} = 2.0, 5.0, 12.0, 20.0$ and have been normalized to unit area.

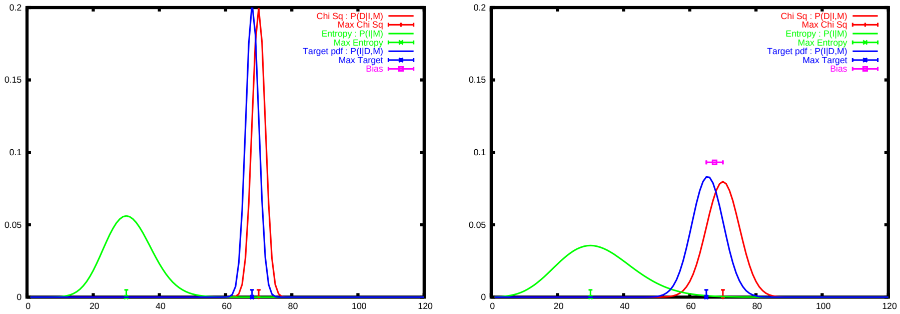

Figure 1: Prior and posterior density functions : $\sigma_{noise} = 2.0, 5.0$

$P(I^M|M) = e^{H/\alpha}$ is centred at the value of the model image M. $P(D|I,M) = e^{-\frac{1}{1}\chi^2}$ is a gaussian centred at D. The posterior distribution $P(I^M|D,M)$ is the product of these distributions and its mode will be biased according to the relative heights,widths and locations of the two contributing distributions.

These plots show that when the data is reliable (low noise), the posterior distribution closely follows the data. With higher noise the probability density functions are wider, and the bias towards the model is greater. This demonstrates the regularizing effect of the prior pdf, which biases the posterior pdf when the data is noisy and possibly

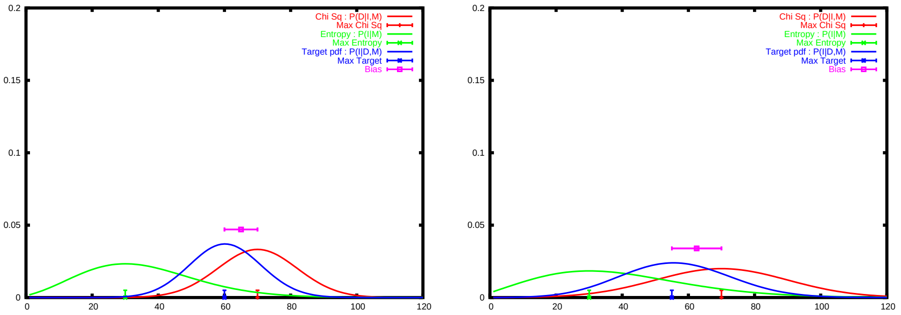

Figure 2: Prior and posterior density functions : $\sigma_{noise} = 12.0, 20.0$

unreliable. This allows the reconstructed image I to deviate from the model M only if evidence for this deviation is present in the actual measured data D. A flat default image (I^M) (corresponding to the most probable image with the entropy model) will therefore have the effect of smoothing high frequency ripples in the reconstruction I.

Fig 3 shows similar plots with the noise level fixed at $\sigma_{noise} = 20$, but with different magnitudes of D and M corresponding to changing signal to noise ratio.

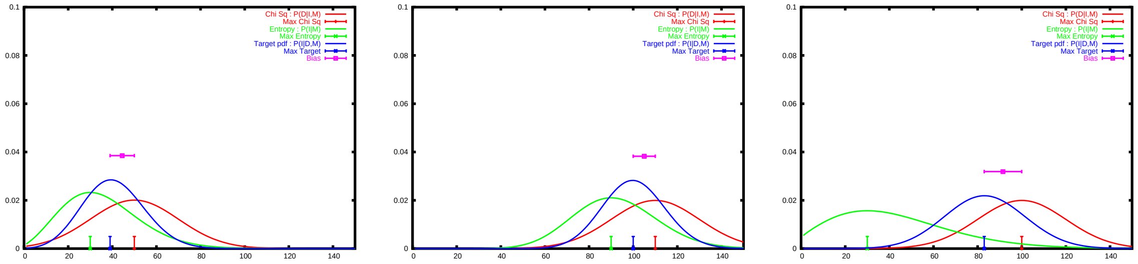

Figure 3: Prior and posterior density functions at different SNR : $\sigma_{noise} = 20.0$

The first two plots show that the position of the peaks of the pdfs change as a function of signal to noise ratio (based on absolute magnitude of D and M) but their widths and heights remain the same. This indicates that the error in the reconstruction (width of posterior distribution) depends on the absolute noise in the system and not on the signal to noise ratio when $D-M$ is fixed. The third plot shows that this width increases with increasing distance between the data D and the model M, with the noise at the same level of $\sigma_{noise} = 20$.

Observing these psfs with various combinations of σ_{noise} as well as varying relative strengths of D and M illustrate the shape of the posterior distribution in zero dimensions. Its behaviour in higher dimensions will be similar, but will also depend on the degree of correlation between pixels introduced by the point spread function.

1.2.1 Maximum Entropy Solution

Finding the most probable image, given the model and observed data, corresponds to maximizing the objective function J derived from Eqn 8.

$$J = \frac{H}{\alpha} - \frac{1}{2}\chi^2 \tag{9}$$

The solution image I that maximizes J is numerically evaluated via a quasi Newton Raphson algorithm.

Differentiating Eqn 9 with respect to I_i gives the gradient ∇J and Hessian $\nabla^2 J$ as follows

$$\nabla J = \frac{\delta J}{\delta I_i^M} = \frac{1}{\alpha} ln \left(\frac{I_i^M}{M_i} \right) - \sum_k {P^k}_i \left[(P \circ I^M)_i - D_i \right] \tag{10}$$

$$\nabla^2 J = \frac{\delta^2 J}{\delta I_i^M \delta I_j^M} = - \left[\frac{1}{\alpha I_i^M} \delta_{ij} + \sum_k {P^k}_i {P^k}_j \right] \tag{11}$$

where ${P^k}_i$ is the PSF centred at point k and evaluated at i (or vice-versa for a symmetric psf). The diagonal of the Hessian will therefore have values equal to the square of the area under the point spread function. Let q be an estimate of the area under the psf.

A diagonal approximation to the Hessian is used to calculate $(\nabla^2 J)^{-1}$. Maximizing J corresponds to minimizing $-J$ and the minimization step that results is given by

$$\Delta I_i^M = (-\nabla^2 J)^{-1} \times \nabla J \ = \ \frac{\alpha I_i^M}{1 + \alpha q^2 I_i^M} \times \nabla J \tag{12}$$

This diagonal approximation to the Hessian results in an inaccuracy that must be corrected. A linear interpolation is performed along the ∇J direction to calculate the length of a correcting step along ∇J. Iterations continue until chi-square converges. This adaptation of the quasi Newton Raphson non linear least squares technique is described in detail in Cornwell and Evans (1983).

2 Monte Carlo Methods

Direct maximization techniques gives a single solution , the image I^{MAP} that has the highest probability $P(I^M|D, M)$. This corresponds to the mode of the posterior distribution. It gives no information about any other statistics of the posterior distribution.

Analytical and numerical integration methods to calculate these statistics, are often infeasible, especially when the number of parameters being estimated (dimension) is high. Monte Carlo methods are therefore used to sample the posterior distributions and to compute posterior quantities of interest like posterior means,modes, standard deviations,etc. Markov Chain Monte Carlo methods are used to favourably constrain the sampling process.

A Markov Chain is a stochastic process in which the i^{th} state depends only on the $(i-1)^{th}$ state. A Markov Chain is stationary if the transition probabilities between two states stays constant in time. The matrix of these transition probabilities is called the Markov Matrix, and will be symmetric if the probability of moving either way between two states is the same. Such a transition matrix can be proved to have a unique limiting distribution.

Given a target distribution, there are several algorithms that can be used to construct Markov Chains with limiting distributions equal to the target distribution. This means that moving from state to state in this chain creates an ensemble that converges to the target distribution. One such algorithm is the Metropolis Hastings Sampler, described below[1]. Another frequently used algorithm is the Gibbs sampler (Skilling 1998) (not discussed here).

2.1 Metropolis Hastings Sampling Algorithm

This algorithm uses a trial distribution to generate steps. The form of this trial distribution must be chosen such that obtaining random samples from it is straightforward. An immediate choice for a trial distribution for each parameter would be a gaussian distribution with standard deviation proportional to the uncertainty in the estimation of that parameter.

1. Choose an initial position in the multi-dimensional space of images as the current image I^c.

2. Using samples from a symmetric trial distribution, generate a trial image I^t, as a step from I^c.

3. In order that this step be part of a Markov chain, the following procedure is used to either accept or reject this trial image I^t.

[1] See http://public.lanl.gov/kmh/talks/maxent00b.pdf

4. The *acceptance probability* is calculated as follows

$$a(I^c, I^t) = min \left[1, \frac{P(I^t|D)/P(I^c|I^t)}{P(I^c|D)/P(I^t|I^c)} \right] \tag{13}$$

   For symmetric of the Markov matrix, this reduces to

$$a(I^c, I^t) = min \left[1, \frac{P(I^t|D)}{P(I^c|D)} \right] \tag{14}$$

   A random number p is then chosen from $U[0,1]$ and if $p < a$, the trial solution is accepted, otherwise it is rejected.

5. If the *trial* is accepted, this becomes the new current image I^c and the loop repeats from (2).

After a large number of such steps, the ensemble of images can be used to evaluate moments of the posterior distribution.

Intuitively, in zero dimensions (1-D pdf), this algorithm corresponds to starting at a certain point I^c, centering the trial distribution on this current point, taking a random sample from this trial distribution and using it to calculate a trial position I^t. If I^t is in a more probable region of the target distribution ($P(I^t|D) > P(I^c|D)$), it is accepted, and made the current position I^c. If I^t is in a less probable region, that trial is accepted with a probability equal to the ratio of $\frac{P(I^t|D)}{P(I^c|D)}$.

A zero-dimensional simulation of the Metropolis Hasting's algorithm gave the sample distribution a histogram of which is shown in Fig 4. 8000/10000 trial samples were accepted.

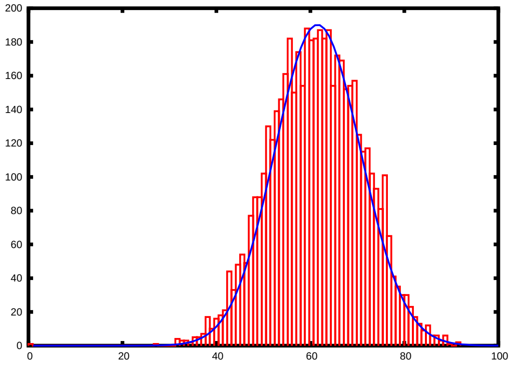

Figure 4: Histogram of the ensemble of samples obtained for a single pixel using the Metropolis-Hastings algorithm

2.1.1 Convergence

When a Markov Chain has generated a sequence of images, the ensemble of which satisfies the target distribution, it is said to have converged. The rate of convergence is a measure of the frequency of obtaining statistically independent samples from this sequence.

1. One diagnostic of the optimality of this sampling is the autocorrelation function (ACF) of the sequence of samples generated. A measure of efficiency computed from the ACF as $\eta = (1 + \textit{area under the } ACF)^{-1}$ can indicate the rate of convergence of the sampling. For example, if the efficiency is 1%, 10000 samples would be the equivalent of 100 statistically independent samples.(Hanson, MCMC Tutorial)

   For a fixed number of parameters to be estimated, the sampling efficiency is a function of the widths of the trial distribution. The following are plots of the trials I^t as a function of time. Fig 5 is optimal, and uses an appropriate trial distribution width. Fig 6 corresponds to a trial distribution that is too narrow. It restricts the samples from moving far enough to randomly sample the target distribution and induces correlations in successive samples. It will therefore take many more samples to generate a statistically independent sample. In this case, all trial samples may get accepted and this may be deceptive. Fig 7 shows the trials when the width of the trial distribution was too large. The steps were too large most of the time and hence rejected. Successive rejections make parts of this trace flat.

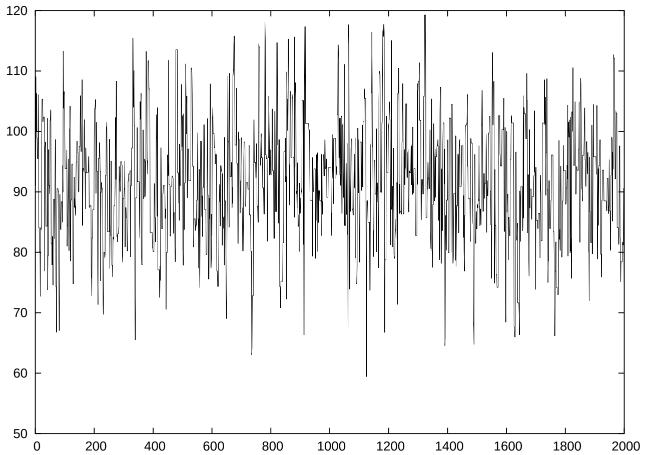

Figure 5: MCMC sequence of trials with optimal trial distribution width

2. One choice for the widths of the trial distributions per parameter (pixel) would be to estimate it by analysing the one-dimensional pdfs obtained for a single pixel. This however may not be accurate when the parameters are correlated. A more practical choice for the widths of the trial distribution for each parameter

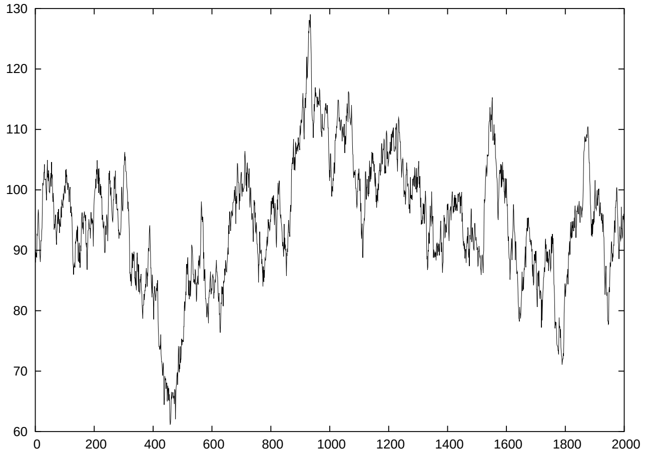

Figure 6: MCMC sequence of trials with trial distribution width reduced by a factor of 10

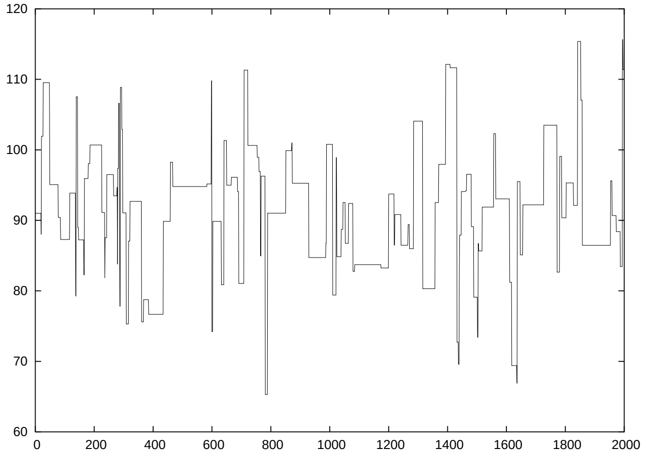

Figure 7: MCMC sequence of trials with trial distribution width increased by a factor of 10

is an estimate of the level of uncertainty expected for that parameter obtained from the covariance matrix of the objective function J calculated with respect to the mode (I^{MAP}) of the posterior distribution. Trial steps can be generated by multiplying a noise vector (gaussian random normal N(0,1)) by the Cholesky decomposition of the covariance matrix. This results in a noise vector with the same variance and correlation structure as in I^{MAP}, and this can be used as a trial step.

3. Another scheme to control the trial steps is to restrict the length of the step to be unity. This corresponds to normalizing a calculated step by the total number of parameters. This decreases the step length along individual directions, and thus attempts to guard against stepping too far in any one dimension.

4. The efficiency drops reciprocally with the number of dimensions (parameters). In high dimensions ($N > 100$) the theoretical optimal efficiency with fixed univariate gaussian trial distribution widths is very low ($\eta \propto 0.3/N$) (Hanson,MCMC tutorial) . An acceleration technique that alters these widths during the sampling process can increase this efficiency.

   The ratio of the number of samples accepted to the total number of trial samples, is defined as the acceptance ratio. This can be used as a metric to control the widths of the trial distribution. Small steps result in an increased rate of acceptance and large steps tend to be rejected more often. Therefore, increasing the width, when the acceptance ratio gets too high and decreasing it when the acceptance ratio is too low can keep the acceptance ratio approximately constant. It has been empirically found that for $N \leq few\ hundred$, optimal convergence can be achieved by controlling the acceptance ratio to be approximately 0.23.

5. If the initial image is far from the mode of the distribution, a large number of initial MCMC steps will be required to move the current image into a region around the peak of the distribution. Once it has reached this region of higher probability, the sampling becomes more random, and the rate of convergence increases. Discarding the first few 40% of the samples is called $burn - in$ and is advisable, when the initial image is not guaranteed to be favourably located with respect to the probability distribution. In addition, to reduce the level of correlation between successive MCMC samples, a *thinning* procedure of retaining only every k samples (where k is some measure of correlation length) is used.

3 Monte Carlo techniques with the Maximum Entropy Formulation

The approach discussed here uses a Markov Chain Monte Carlo algorithm to sample the maximum entropy posterior distribution and obtain estimates of its moments. The Metropolis-Hastings sampling algorithm is used in which a Markov Chain is followed to generate a sequence of images, the ensemble of which satisfies the posterior distribution. Various statistical quantities are then estimated from this ensemble of images.

Currently this algorithm has been implemented for deconvolution purely in the image domain (psf corresponding to a filled aperture). This can be extended to practical deconvolution for radio interferometry, as well as to more complex formulations that could include uv sampling and calibration as separate probability density functions. This problem can then become infeasible to solve via standard maximization methods and will require MAP analysis on the combined probability density functions.

We have implemented two related Monte Carlo techniques to generate image instances.

1. Sampling in the data space corresponds to obtaining statistics about how the peak of the posterior distribution varies with different realizations of the data D. This was done with multiple measurements of the data, which differ from each other only by noise.

2. Sampling in the image space corresponds to sampling the posterior distribution given by Eqn 8. We do this via a Markov Chain Monte Carlo simulation - a random walk around in the target distribution which 'correctly' samples it.

3.1 Data Realizations

3.1.1 Algorithm

Different data realizations are obtained by adding different realizations of random noise to the true image convolved with the psf.

MEM solutions are then found for each of the realizations and their resulting solutions (modes of the posterior distributions) are averaged to calculate $< I^{MEM} >$.

3.1.2 Simulation Results and Interpretation

Simulations were performed on 128x128 images where the data was created by adding gaussian noise at various levels to the trueimage convolved with the PSF(gaussian of half width = 5 pixels).

The simulations were run with 1000 realizations (samples) and with noise levels $\sigma_{noise} = 30, 10, 1, 0.1$. The images obtained are listed in Appendix A.

The mean image $< I^{MEM} >$ is smoother than any single MEM solution showing that the mean of this distribution is a good estimator of the true solution. This is most apparent with higher noise levels, and just demonstrates the effect of obtaining a higher signal to noise ratio by averaging over multiple measurements.

There is no systematic bias between the mean image and any single instance. Differences between instances are correlated across scales which depend on the noise in the system. At high noise levels, the bias towards smoothness (and hence apparent correlation between pixels in the reconstructed image) increases. As a function of decreasing noise levels in the data, the individual MEM solutions get closer to the true data, the bias between the mean image and any single MEM image is correlated across smaller scales. The variance image becomes flatter (possibly less correlated with the actual noise in the system and more due to the dependence of the width of the posterior density function on the distance of the data from the flat model). The effect of ringing around the reconstruction of a point source on a bright background is also emphasized.

3.2 Image Realizations - Markov Chain Monte Carlo

The MEM solution I^{MEM} is the peak (mode) of the posterior distribution, and is used as the initial (current) position for a random walk in the multidimensional parameter space. Trial solutions are obtained from the current solution using a step calculated via a trial distribution. An check is done to either accept or reject this step. Every accepted step results in a Monte Carlo sample.

3.2.1 Implementation of the Metropolis-Hastings sampler with MEM

In our Markov Chain simulation, the initial image is I^{MEM} (the mode of the posterior distribution). At every stage, $P(I^t|I^c) = P(I^c|I^t)$ and this results in a symmetric transition matrix.

In this implementation, the trial distribution is a multi-dimensional gaussian distribution, with the standard deviation for each degree of freedom, calculated from the inverse of the Hessian. This was initially done by evaluating the complete Hessian from the known I^{MEM} solution, inverting it to form the covariance matrix, computing its Cholesky decomposition, and filtering an $N(0,1)$ noise vector through it. For large images, the Hessian calculation and inversion became impractical, and approximations to the inverse Hessian diagonal were made. A diagonal approximation to the Hessian diagonal was directly inverted, and then scaled by an approximation to the combined effect of the diagonal and the first off diagonal of the Cholesky decomposition. This scaling resulted in a noise vector of absolute amplitude corresponding to the result from the complete Cholesky decomposition, but had no correlation structure in it.

Tests on small images (16x16) showed that this approximation produced results comparable to those produced by the complete Hessian calculation, its inversion, and the Cholesky decomposition. The total length of the steps was normalized by dividing it by the number of pixels in the image.

Burn-in of 40% was used and the widths of the trial distributions were collectively scaled to keep the acceptance ratio between 0.15 and 0.25.

3.2.2 Simulation Results and Interpretation

Data was simulated by smoothing a 256x256 M31 image with a gaussian PSF and noise was added at a peak signal to noise ratio of 50.

The true image and the data are shown in Fig 8

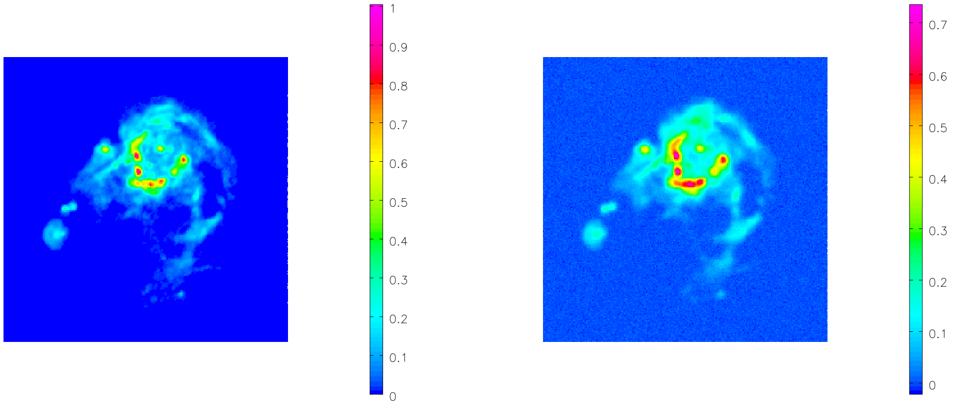

Figure 8: The true M31 image and the data obtained by convolving the true image with a 3 pixel wide point spread function and $\sigma = 0.005$ noise added.

The MEM solution obtained was Fig 9

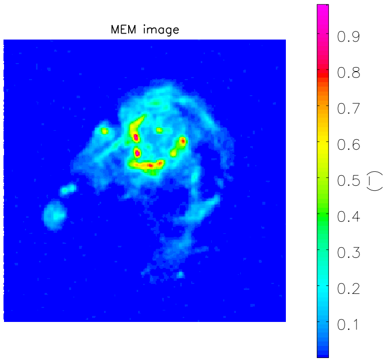

Figure 9: MEM solution : Mode of the posterior distribution

This algorithm was run on this data, using the diagonal approximation to the Hessian and approximate scaling, for 200000 realizations.

1. To test whether the Metropolis Hastings sampling algorithm did indeed produce an ensemble of images satisfying the correct target distribution for this problem, a MCMC sequence was initialized with a flat image. This ensured that the sequence would begin far from the area of interest, and would test the process of convergence to a more probable region. On small images this was immediately apparant, and discarding the first 20% of the samples resulted in a suitable ensemble of images. For the M31 image, the first 40% of the samples were discarded and the mean of the resulting ensemble was computed to form the image in Fig 10 (left)

   This shows that the sequence had not yet converged but that is would eventually converge to the correct distribution.

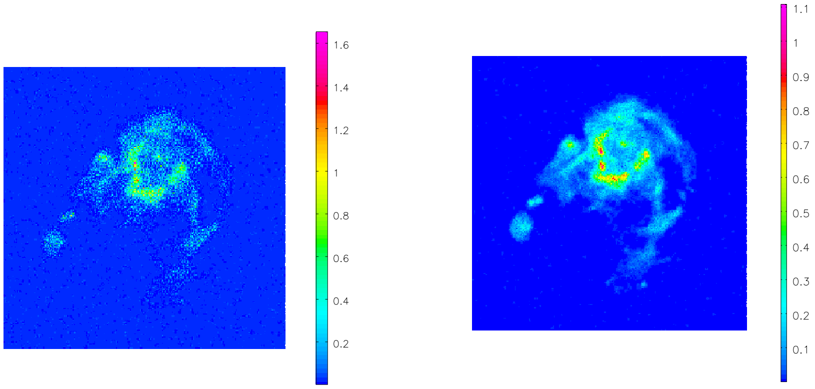

Figure 10: Mean image obtained after 100000 MCMC iterations , (left) starting from a flat image and (right) starting from the MEM image

   The default image used here was a flat image. A MCMC sequence generated using a smoothed version of the true image as the default image, gave much quicker convergence.

   Having seen that this algorithm did produce the correct ensemble at convergence, an MCMC sequence was generated using the MEM image as the initial image. This ensured that the sequence began in a region of high probability, and should converge quickly. The following statistics were obtained from the resulting ensemble of images.

2. The Mean image obtained from the ensemble generated with the MEM solution as the initial image is shown in Fig 10(right)

3. The difference between the Mean and the Mode(MEM) : Fig 11(left)
   This image showed no systematic bias between the mean and the mode. This is plausible, since the one-dimensional pdfs show that the posterior distribution deviates very slightly from a gaussian.

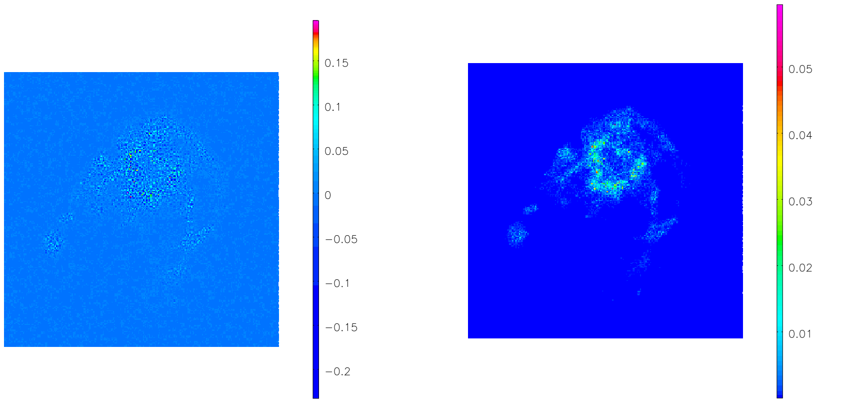

Figure 11: (left) Difference image : Mode - Mean , (right) Variance image obtained from the ensemble

4. The Variance image : Fig 11 (right)
   This has values of the order of the actual noise, in regions of high flux, and lower values elsewhere. This demonstrates a feature of the MEM where the intrinsic noise is retained in regions of high flux, while it is suppressed in other areas.

5. The Skewness image : Fig 12(left)
   Lack of any prominent structure in this image shows that the probability distributions are not far from being gaussian. (This structure may also be due to incomplete convergence)

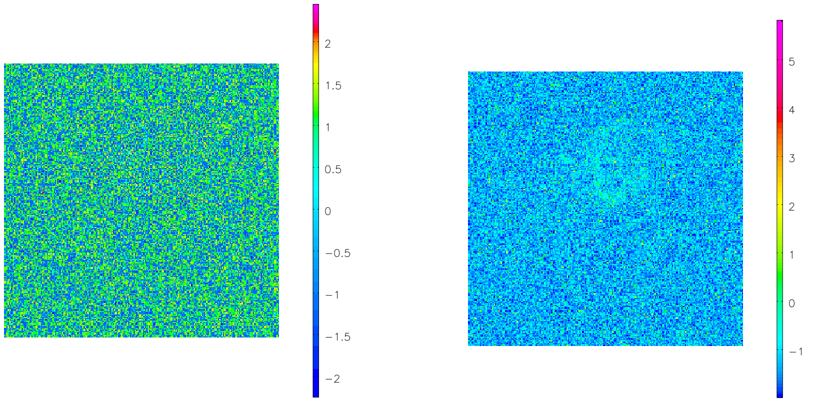

Figure 12: (left) Skewness Image and (right) Kurtosis image obtained from the ensemble

6. The Kurtosis image : Fig 12(right)
   This image with higher values in regions of high flux shows that these high flux regions have higher signal to noise ratios and are probably better constrained than pixels with lower flux. Also since these high flux pixels are further away from from the flat default image, their corresponding pdfs are likely to be more asymmetric than those of low flux pixels.

These results show that the MCMC method when applied to image reconstruction in this form, gives realistic statistics consistent with the parameters in the simulation. In more complex formulations where a direct solution to obtain the mode of the distribution is not feasible, this method can be used to generate a suitable ensemble of images. This MCMC sampling is however extremely inefficient and the above simulation ran for several hours before convergence.

Some areas of further investigation can now be identified as follows. Points 1,2 and 3 are addressed in section 4.

1. Combinations of Priors
   Treating image pixels as individual parameters results in too many dimensions for MCMC algorithms to be efficient. The dimensionality of the posterior distribution can be reduced by decomposing an image into components. A combination of a flat component, point source component, and a scale sensitive component, will regularize the reconstruction to give emphasis to features at these different scales, and can also reduce the total number of parameters to be estimated.

2. Efficient MCMC sampling
   This has to be studied for higher dimensions, and hybrid algorithms should be applied to accelerate convergence. Also suitable approximations to the Hessian (band diagonal) and specialized matrix inversion algorithms must be used to obtain better MCMC trial steps.

3. Extend to Interferometric Imaging
   This algorithm must be implemented for the case of practical interferometric imaging, where the PSF is derived from incomplete sampling in the visibility domain, with noise added in the visibility domain.

4. UV Sampling, Calibration and Imaging as PDFs
   Explore the feasibility of formulating the entire observation process of UV Sampling, Calibration and Imaging as probability distributions, and doing a combined maximum a-posteriori analysis on it. Direct maximization techniques may not be feasible for a problem of this complexity, and Monte Carlo algorithms may be only practical way to generate a solution.

4 Bayesian formulations with different priors

This is a short study of the effect of different image models (M) on the process of image restoration via the Bayesian formulation. The three different models examined are those of entropy, positivity and emptiness.

4.1 Entropy, Positivity, Emptiness

1. $Entropy = -I^M ln(I^M/M)$
   This is one form of entropy, derived from the combinatorial formulation of the total number of ways flux can be distributed over pixel bins to form an image. Given a positive default image, it enforces positivity and is a distance measure between and image I^M and the default image M.

2. $Positivity = ln(I^M)$
   This prior enforces the basic physical constraint of positivity on an image I^M.

3. $Emptiness = -ln[cosh((I^M - M)/\sigma)]$
   This form of prior is derived from the assumption of a mostly empty sky. An image can deviate from near zero, only if there is strong evidence present in the data.

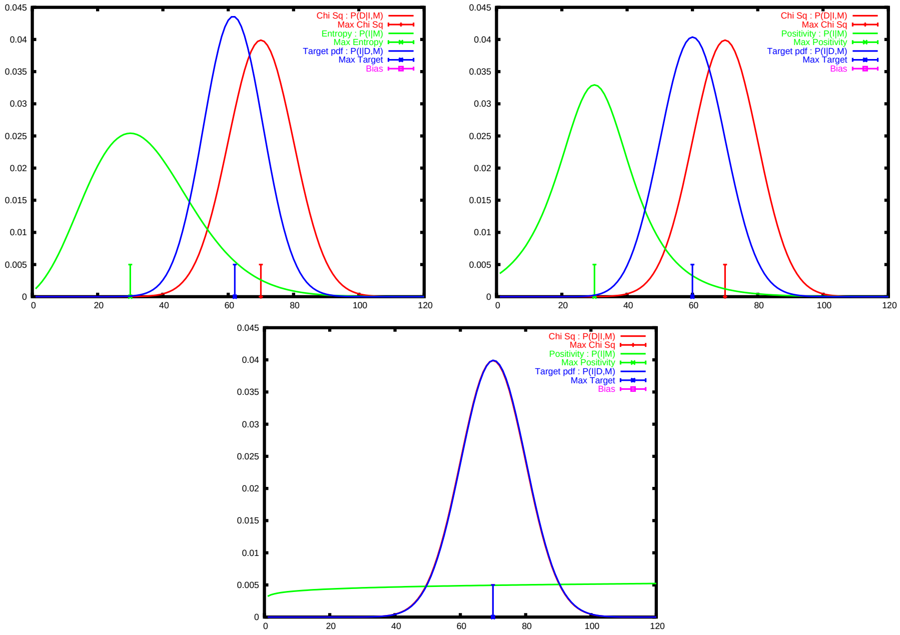

Figure 13: Prior and posterior density functions for the entropy, emptiness and positivity prior models

The probability density functions corresponding to these three priors for a single pixel (1-D pdfs) are shown in Fig 13. These PDFs correspond to $M = 30, D = 70, \sigma_{noise} = 10$ and have been normalized to unit area.

The emptiness prior density function has larger tails than that of the entropy prior. Both these priors bias the posterior distribution, but in different ways. The entropy prior allows the posterior distribution to deviate from a default flat image only if there is evidence for it present in the data, and this ensures smoothness in the reconstruction. Large scale structures are hence better reconstructed using this prior. The default image for the emptiness prior is the zero image, and any deviation from this is possible only with strong evidence from the data. This would lead to a sharper image and small scale structure would be better reconstructed. This emptiness prior is considered as the closest Bayesian representation of the CLEAN algorithm.

The positivity prior has no biasing effect on the posterior density function when the data is well above zero. If the data or the default image has values near zero, and the noise in the system allows these values to become negative, this prior will restrict the posterior density function to be completely positive.

4.2 Maximum A-Posteriori solutions

The effect of these three priors on a reconstructed image was analysed by deconvolving a 2D gaussian image, smoothed by a gaussian psf with noise added in the image domain. Cuts through the 2D reconstructions using these three priors, at different noise levels gave the following results. The posterior density functions are shown in Figs 14 and the 1-D cuts through the 2D reconstructed images are shown in Fig 15

1. At a low noise level, all three posterior distributions are almost the same (Fig 14(top left), and so are the reconstructions.(Fig 15(top left))

2. At a higher noise level, the difference between the posterior distributions is apparant(Fig 14(top right). In the reconstructions, the entropy and positivity prior both resulted in smooth reconstructions whose peaks did not reach that of the original true image. The emptiness prior however resulted in a much sharper reconstruction, demonstrating that it deviated from zero only in the presence of strong evidence in the data. The fact that this reconstruction did retrieve the correct amplitude at the peak, shows that this prior can result in better reconstruction of small scale features in the image.(Fig 15(top right))

3. At a much higher noise level, there is again little difference between the three posterior distributions, and all of them are equally wide(Fig 14(bottom). As is expected, none of the three solutions are good reconstructions. (Fig 15(bottom))

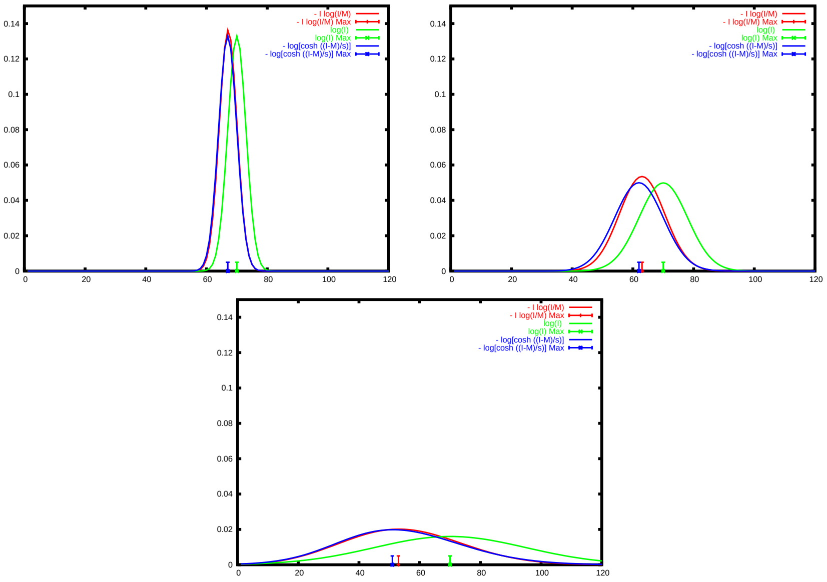

Figure 14: Posterior density functions for the three priors at a (top left)low, (top right)medium and (bottom)high noise level.

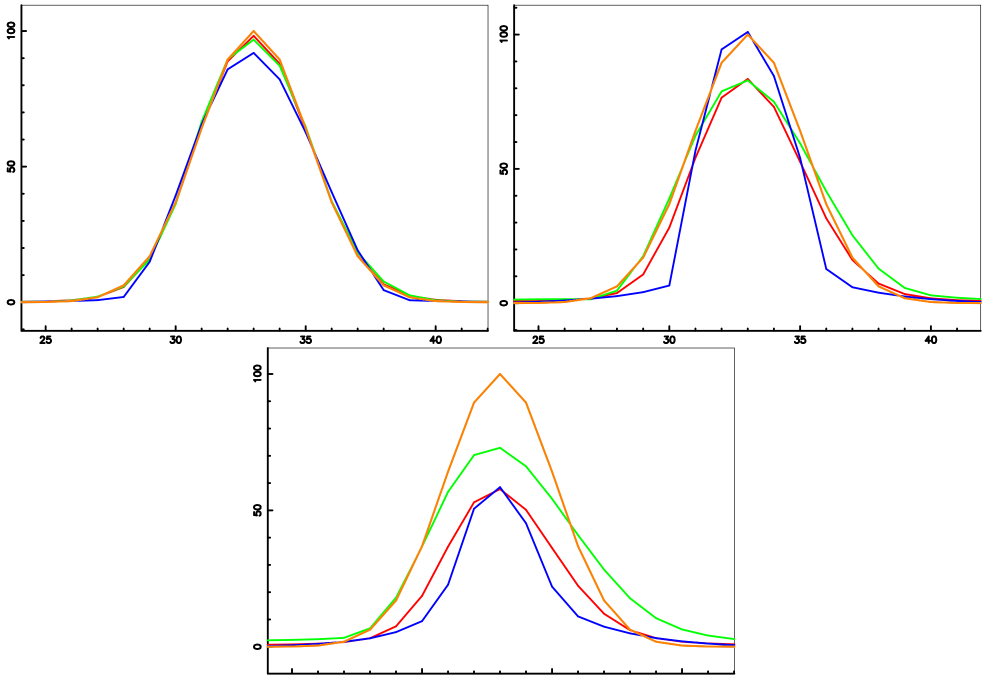

Figure 15: One dimensional cut through the 2D reconstruction using the three priors (Entropy:red, Emptiness:blue, Positivity:green, True image:orange), at a low (top left) medium (top right) and high (bottom) noise level.

5 Bayesian analysis using parameterized flux components

Image analysis in radio interferometry often involves component fitting applied to previously deconvolved images. Combinations of gaussian, disc or point source flux component models are used to approximate the emission patterns seen in a deconvolved image. There are two sources of uncertainty associated with this approach which traditional methods generally ignore.

1. Incomplete sampling of the spatial coherence function leads to an invisible distribution which makes deconvolution a non-linear inverse problem with no exact solution. Algorithms that use different approximations can converge to slightly different results[2]. Moreover, the only estimate of error due to deconvolution is the residual image which holds no information about the possible existence of multiple solutions with the same goodness-of-fit.

2. There is an uncertainty in the process of estimating the parameters of overlapping flux components in extended emission. The objective function involved in fitting a linear combination of elliptical gaussians is often multi-modal and any likelihood technique (unconstrained and constrained minimization) is subject to this uncertainty.

A Monte Carlo approach to this form of image analysis has the potential of providing information about relative probabilities and uncertainties associated with the number of flux components and their parameters by sampling the posterior distribution associated with the Bayesian formulation of deconvolution.

Due to the uncertainty arising from the process of deconvolution, one would like to estimate the properties of components representing the entire emission present in the raw dirty image. In practice however, a pixel based Monte Carlo method is not feasible given the large dimensionality of the parameter space. A scale sensitive representation of an image as a collection of gaussian, disk or point source flux components $I^M = \Sigma_i P_i$ can control this problem. An additional advantage of thus incorporating the flux component model into the image representation is that the deconvolution and the component fitting will then be treated together and the resulting uncertainty estimates will reflect the entire process applied to the raw dirty image.

We present the MC-FIT algorithm, in which an image is represented as a collection of elliptical gaussian flux components, and Monte Carlo sampling is performed on the number of flux components and their parameters. A-priori information is provided in the form of probability distribution functions for each type of parameter and the χ^2 goodness of fit criterion is used as the likelihood function. We used BayeSys (Skilling 2004) as the Monte Carlo sampling engine. Tests on simulated synthesis data as well as real data are presented.

[2] CLEAN,MEM,MS-CLEAN,ASP-CLEAN are each suited to different emission patterns.

5.1 Monte Carlo Sampling of $P(I^M|D, M)$: BayeSys

Various techniques exist for efficiently sampling a probability distribution in a multi-dimensional parameter space. Markov Chain Monte Carlo techniques generate samples by following a Markov Chain through the parameter space based on a transition matrix whose limiting distribution is the target posterior distribution.

BayeSys (Skilling 2004) is an application that samples the posterior distribution associated with an object represented as a collection of atoms, an atomic prior, and a suitable goodness of fit criterion. BayeSys implements an MCMC algorithm along with selective annealing and has the following features.

1. A BayeSys object is a mixture model comprised of a collection of flux components called Atoms, each represented by a set of parameters. Prior information can be supplied in the form of distribution functions per parameter.

2. Multiple parallel sample streams, called ensembles are allowed to communicate so that they can catalyse each other's progress and guard against any sequence being trapped in a local maxima.

3. There are several sampling engines used to create, destroy and move the atoms around efficiently by varying their parameters. Generality of the sampling engines is achieved by restricting the parameters to lie within a U[0,1] hypercube. These samples are then transformed according to their individual prior probability distributions to obtain sets of parameters that can be used along with the goodness of fit criterion to evaluate the posterior distribution.

4. Annealing is achieved by initializing the sequence to sample the prior distribution and gradually increasing the influence of the likelihood function. Annealing is usually stopped (and the sequence is said to have converged) when the likelihood and prior are equally weighted and the posterior distribution is being sampled. Further annealing will result in only the likelihood distribution being sampled.

5.2 MC-FIT Algorithm

We present the MC-FIT algorithm, which combines the two stages of component based image analysis (deconvolution and component fitting) and uses the Bayesian formulation of deconvolution to obtain estimates (with associated uncertainties) for parameters of extended flux components present in the dirty image. The main features of this algorithm are as follows.

1. An image is represented as a sum of elliptical gaussian flux components, each described by six parameters (two for position, two for scale, one for amplitude and one for position angle). The practical advantage of such a representation is three-fold. It allows for a scale sensitive representation of an image, is well

suited to component fitting based image analysis, and conforms to the object representation framework used by BayeSys.

2. The MC-FIT algorithm uses component based priors. The distribution functions for the position parameters are chosen to be uniform between the bounds of the image (or region of interest). The distribution functions for the scale parameters are chosen to be non-uniform and of the form Ase^{-Bs} where s represents the scale. This approximates the distribution of scales observed in a typical image, where the majority of features correspond to small scales and a comparatively smaller number of large scale features exist. The distribution for the amplitude is chosen to be uniform within a range, and the position angle (defined here as the angle between the horizontal axis and the first specified axis of the ellipse) is allowed to vary uniformly between 0 and 90 degrees.

3. Sampling on the number of components is achieved by defining a distribution for the number of components (uniform or Poisson) within a specified range.

4. Image based priors based on entropy, emptiness (and positivity constraints), can be applied by including them during the evaluation of the posterior distribution function.

5. Component fitting can be restricted to certain regions in the image. Prior to the fitting of emission inside a given region, the emission outside the region is removed from influence by masking out and subtracting from the observed data, visibilities corresponding to parts of a previously deconvolved image. This information is then incorporated into the prior distributions of the position parameters by restricting them to follow the mask.

6. The computation of χ^2 in the evaluation of the likelihood distribution is computed using Equations 4 and 5 for synthesis data involving a point spread function. For fitting components directly to images obtained without the use of a point spread function χ^2 is computed as $\chi^2 = \Sigma {I^R}^2$

7. BayeSys provides the user with control points for the actual sampling algorithm used. Choices about the sampling engine to be used, the number of parallel ensembles to run, and the annealing speed, allow the user to tune the sampler to obtain faster convergence and better sampling of the posterior distribution.

Following are the steps by which MCMC samples are generated and collected.

1. Define suitable priors for each of the parameters.

2. Generate a random sample set of parameters from the U[0,1] hypercube, using one of several sampling engines, and based on the current state of the system.

3. Transform the U[0,1] numbers to physical parameters using transfer functions for each of the parameter prior distributions.

4. Compute a model image I^M.
5. Compute the likelihood and an image based prior term (if present), to obtain the corresponding value of the posterior distribution.
6. Update BayeSys annealing parameters based on the result of Step 5.
7. Repeat steps 2 through 6 until annealing is complete.
8. Fix the annealing parameter and collect a large number of samples to form the desired ensemble reflecting the posterior distribution.
9. Analyse this ensemble to obtain (most probable) values for fitted parameters and their associated uncertainties. For a more accurate best-fit, the most probable parameter set can be used to initialize a regular likelihood maximization scheme (MEM + ASP).

5.3 Tests on Simulated Synthesis Data

The MC-FIT algorithm was applied to a simulated synthesis data set corresponding to a VLA C array observation of a source composed of four overlapping elliptical gaussian features.

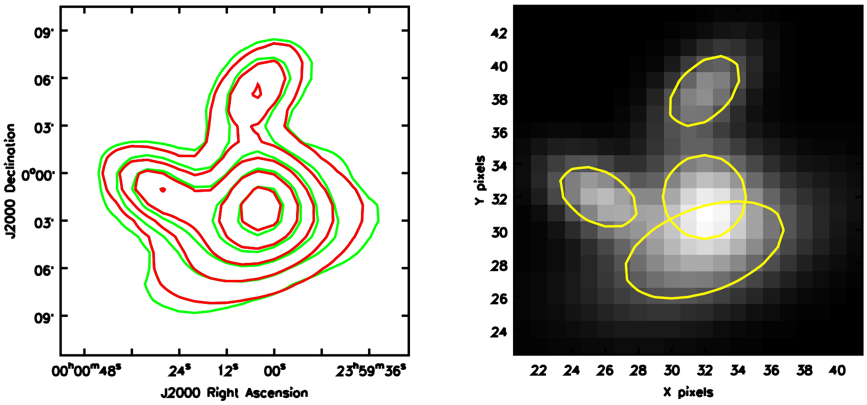

Figure 16: Tests on simulated synthesis data : [Left] original image (red), dirty image (green) with signal to noise ratio = 100. [Right] Most probable model image with 3σ contours for the corresponding set of gaussian components.

Component	x0	y0	Amplitude	σ_x	σ_y	Position Angle
1	5.0	4.5	80	1.0	0.5	20
2	4.0	5.0	100	0.3	0.5	60
3	5.0	6.0	80	0.5	0.3	45
4	5.0	5.0	100	0.5	0.5	45

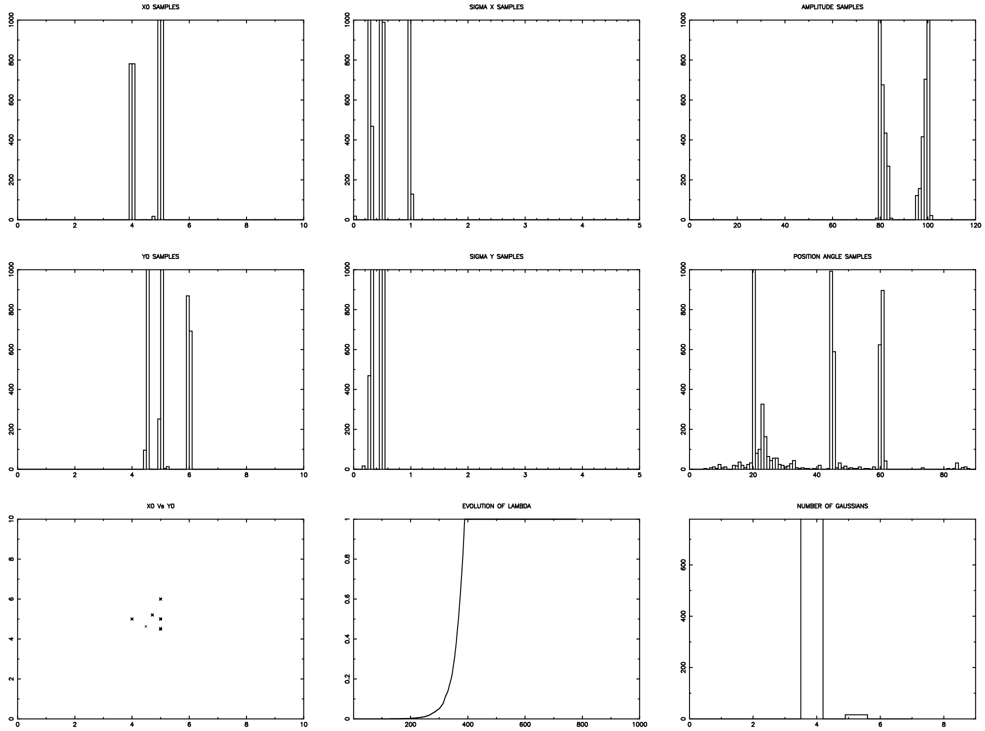

Figure 17: [Rows 1,2] Histograms of the obtained samples for each of the six parameters. [Row 3] (left) Scatter plot of the positions of the centres of the gaussian components, (middle) evolution of the annealing parameter between 0(sampling the prior) and 1(sampling the posterior), as a function of sample iteration number, and (right) histogram of the number of components over all samples.

The true parameters of the elliptical gaussian components used to compute the sample data are given in Table 5.3. Inspection of Figure 17 shows that all the histograms have peaks at the correct locations.

The value of normalized χ^2 computed using the most probable image obtained via the MC-FIT algorithm, was compared to the values obtained using images from the CLEAN and MEM algorithms, and with images formed from gaussian components fitted to the deconvolved images.

		Normalized χ^2
1	Gaussian fits to CLEAN restored image	2.023
2	Gaussian fits to MEM restored image	1.996
3	MC-FIT mode image	1.122
4	CLEAN model image	1.113
5	MEM model image	1.114

Table 5.3 lists values of normalized χ^2 for various model images. Rows 1 and 2 were obtained via the traditional method of fitting gaussian components to deconvolved,

restored images. Row 3 represents the corresponding value for the image formed from the most probable set of component parameters produced by MC-FIT. Rows 4 and 5 were computed from the CLEAN and MEM unrestored model images and represent the value of χ^2 obtained as a result of the respective minimization algorithms. This data shows that if an image can be decomposed into elliptical gaussian components, the MC-FIT algorithm is capable of producing a component list that represents the image almost as well as the minimum χ^2 CLEAN and MEM model images.

The effect of the noise level and different priors on the shape of the posterior distribution can be seen using histogram displays similar to Figure 17. Increased noise levels result in a slower convergence via annealing and wider distributions. The use of different priors affect the rate of convergence. In the absence of any significant flux components, the parameter histograms reflect the shape of the prior distributions.

5.4 Tests on Real Data

5.4.1 MC-FIT with G192.16-3.84 non-synthesis data

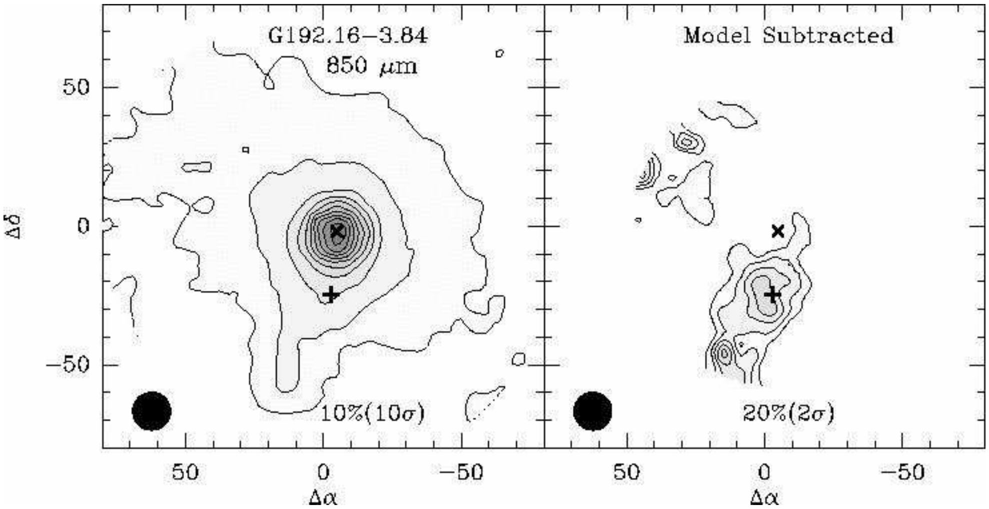

Figure 18: G192.16-3.84 - dust continuum map. (Image credits:Ref (Shepherd 2004))

A test was performed on a dust continuum map of the Early B protostar G192.16-3.84 shown in Figure 18. Evidence from NH_3 data shows the existence of a low-mass protostellar core southwest of the main protostar. The signal to noise ratios of the main peak and the peak of the off-centre core, in the dust continuum map are 7 and 1.4 respectively, making it difficult for an accurate detection. The authors had to model the main core and subtract it out to reveal with more clarity the position of the off-centre core. To obtain error estimates on the location and shape of the off-centre core, we ran the MC-FIT algorithm on the central 64x64 pixels of this image (Figure 19), and obtained the histograms shown in Figure 20.

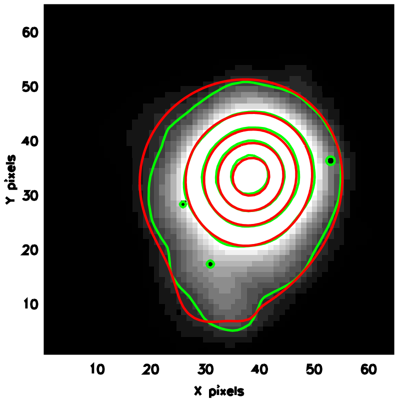

Figure 19: Central 64x64 pixels of the dust-continuum map

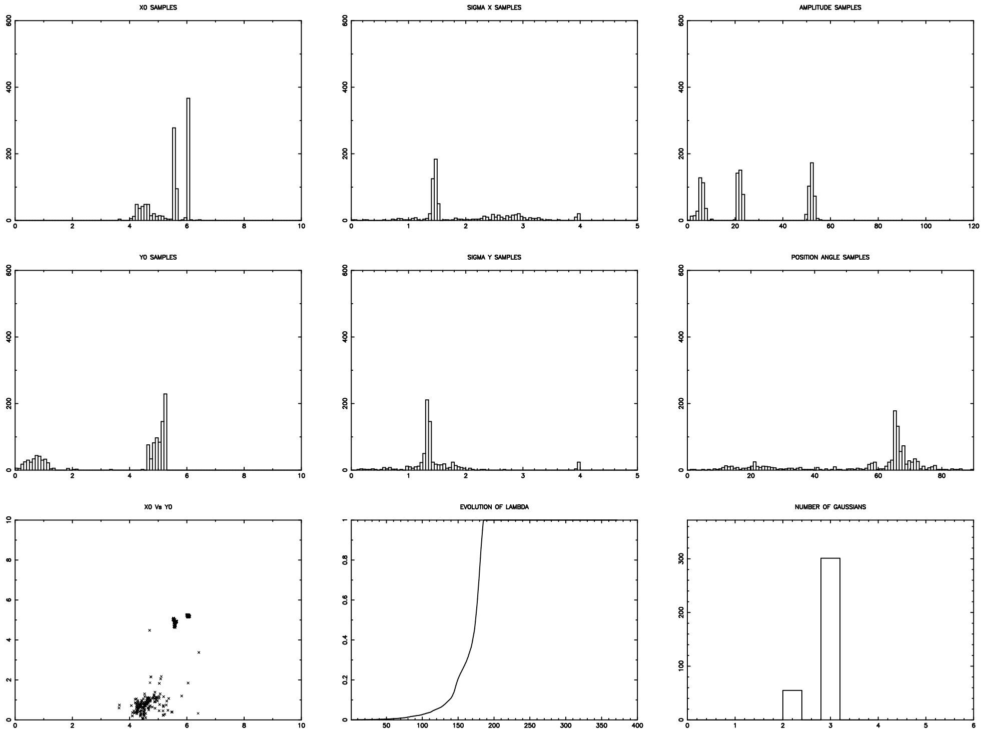

Figure 20: Sample histograms - G192.16-3.84 dust continuum detection of a low-mass protostellar core Southwest of the main core. A normalized χ^2 of 1.77 was obtained. The histograms for the position parameters show relative probabilities and uncertainties for the two peaks. The central core is slightly elongated and has been modeled by two gaussian components whose total amplitude matches the value at the peak in Figure 19. The scale histograms suggest that the off-centre core is relatively compact.

The histograms and plots in Figure 20 confirm the existence of a peak at the location of the low-mass protostellar core. A comparison of the heights and widths of the histogram peaks for the position parameters shows respectively the relative uncertainty in the existence of the peaks and a measure of the uncertainty in the actual position of the peaks. An analysis of the scale parameter histograms shows that the off-centre core is compact. The amplitude histogram peaks at values that correspond to the correct peak amplitudes in the image (Figure 19).

5.4.2 MC-FIT with 3C273 synthesis data

A second test was performed on 3C273 synthesis data. Here χ^2 was computed using Equation 5. Figure 21 shows the dirty image, and the unrestored model image corresponding to the parameter set whose probability was the maximum of all the samples. The histograms in Figure 22 show that the algorithm has found the central core and the extended radio lobe. Again, relative heights and widths of the parameter histograms give estimates of the uncertainty in the position, amplitude and shape of the gaussian components. The scale histograms shows the existence of a fainter elongated component. The amplitude histogram peaks at 30Jy which is the known peak flux for 3C273. The algorithm was allowed to create samples of upto five components, and it found that two components were sufficient to accurately describe the image.

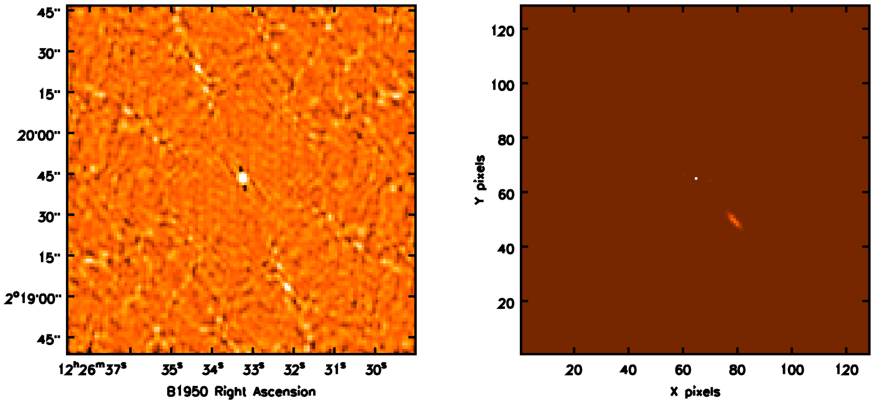

Figure 21: 3C273 : (left) dirty image, (right) MC-FIT mode image

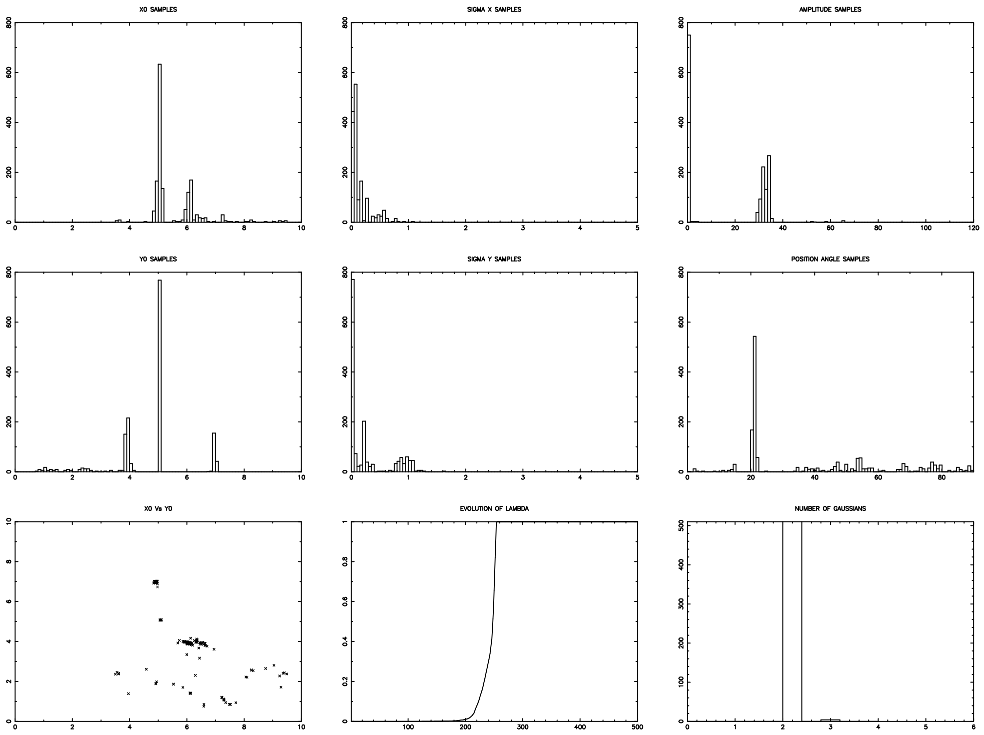

Figure 22: Sample histograms – 3C273. See Fig 17 for panel details and text for discussion.

5.5 Discussion

A Bayesian approach to component fitting in the analysis of images formed from synthesis and non synthesis data, can thus be used to obtain values of flux component parameters along with the associated uncertainty estimates. The running time of this algorithm is a directly proportional to the number of components being fitted ($O(N)$), the number of ensembles ($O(N)$), and the image size ($O(N^2 log_2 N)$ for an $N \times N$ image). The number of iterations required to reach convergence depends on the type of prior information provided to the algorithm, the signal to noise ratio of the data, the area under the clean beam, and the complexity of the brightness distribution.

Further work would involve the implementation of the MC-FIT algorithm as a usable tool in AIPS++, parallelizing the implementation, and providing additional ways of analysing and interpreting the parameter ensembles.

Acknowledgements

This work was supported by the NRAO Graduate Summer Student Research Assistantships over the summers of 2003 and 2004. We thank the NRAO for their support.

References

Cornwell, T. J. & Evans, K. F. 1985, A&A, 143, 77

Narayan, R. & Nityananda, R. 1986, ARA&A, 24, 127

Shepherd, D. S. e. a. 2004, ApJ, 614, 211

Skilling, J. 1998, Journal of Microscopy, 190, 28

Skilling, J. 2004, "Bayesys and Massinf", Tech. rep., Maximum Entropy Data Consultants Ltd

Skilling, J. & Bryan, R. K. 1984, MNRAS, 211, 111

A Data Realizations

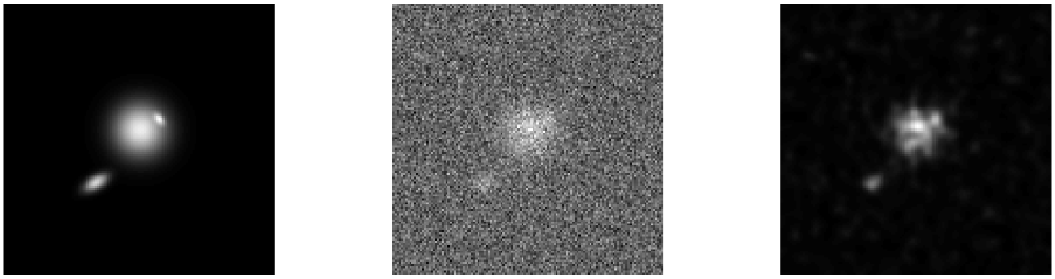

Figure 23: True image(left), data(middle) and MEM image(right) with $\sigma_{noise} = 30.0$

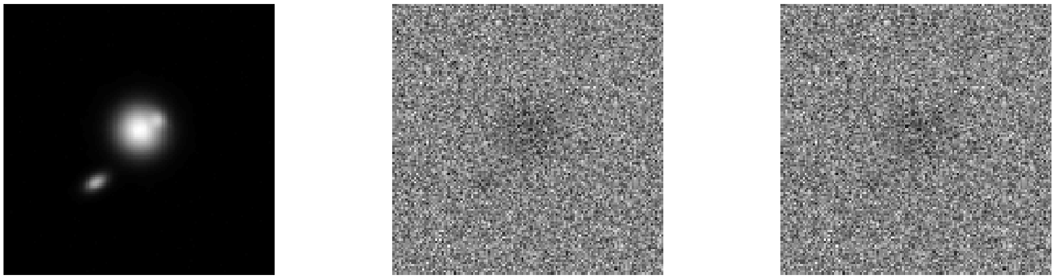

Figure 24: Mean image(left), MEM residual image(middle) and Mean residual image(right)

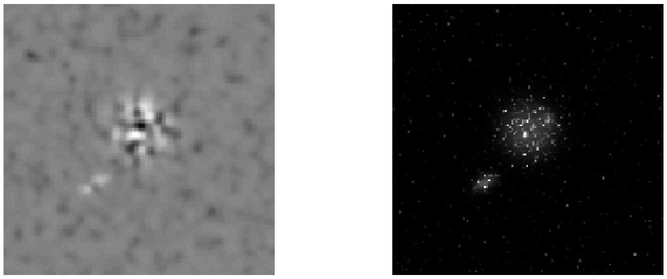

Figure 25: Difference image (MEM - Mean) and Variance image

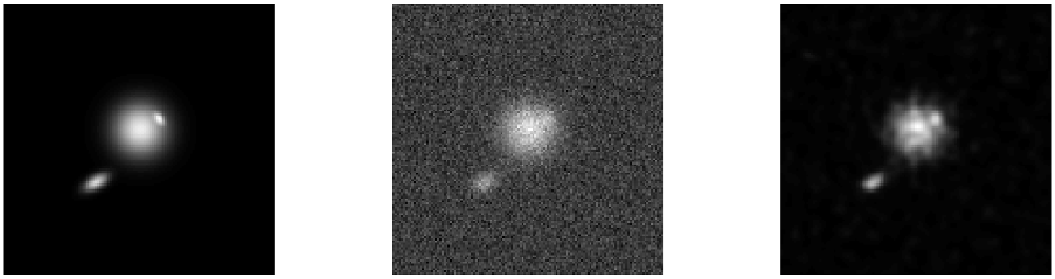

Figure 26: True image(left), data(middle) and MEM image(right) with $\sigma_{noise} = 10.0$

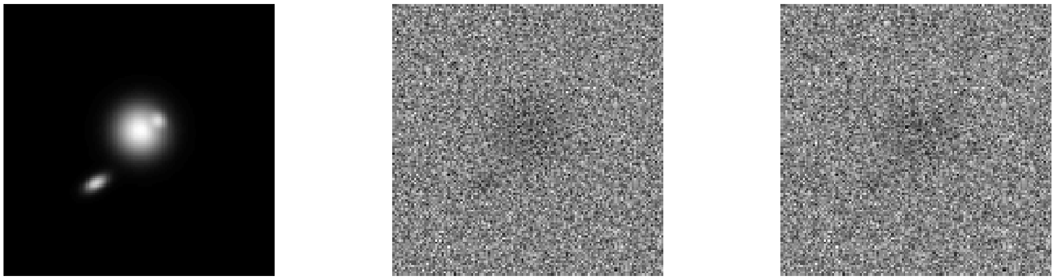

Figure 27: Mean image(left), MEM residual image(middle) and Mean residual image(right)

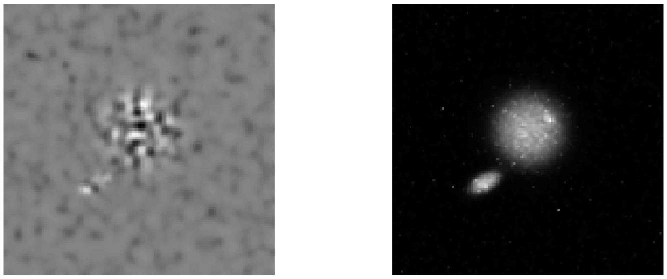

Figure 28: Difference image (MEM - Mean) and Variance image

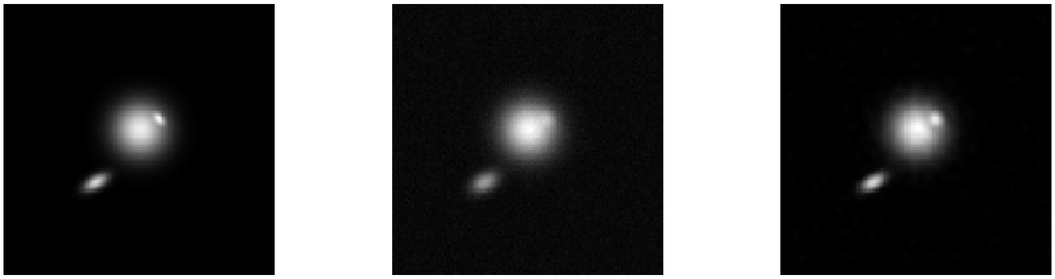

Figure 29: True image(left), data(middle) and MEM image(right) with $\sigma_{noise} = 1.0$

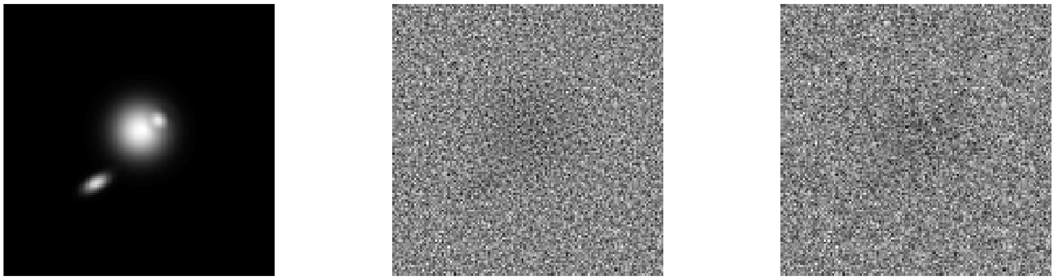

Figure 30: Mean image(left), MEM residual image(middle) and Mean residual image(right)

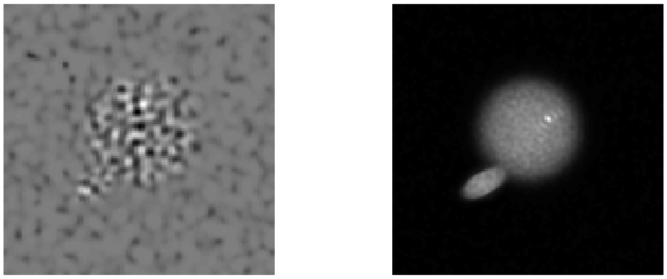

Figure 31: Difference image (MEM - Mean) and Variance image

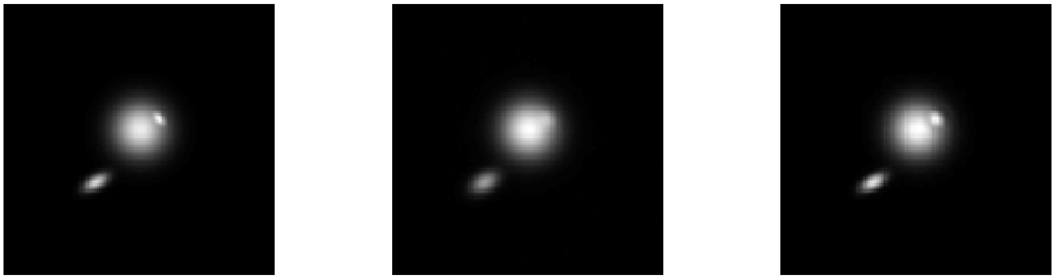

Figure 32: True image(left), data(middle) and MEM image(right) with $\sigma_{noise} = 0.1$

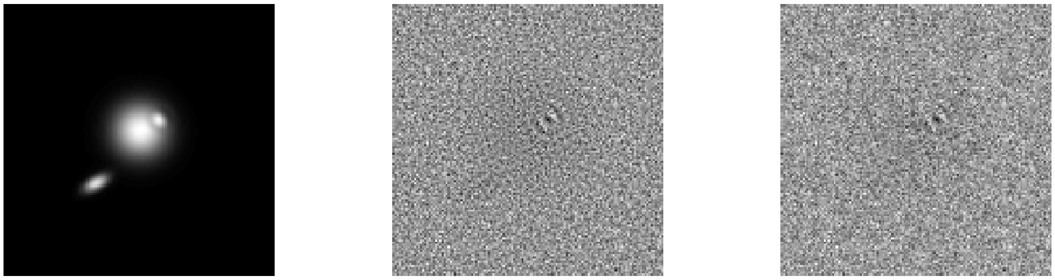

Figure 33: Mean image(left), MEM residual image(middle) and Mean residual image(right)

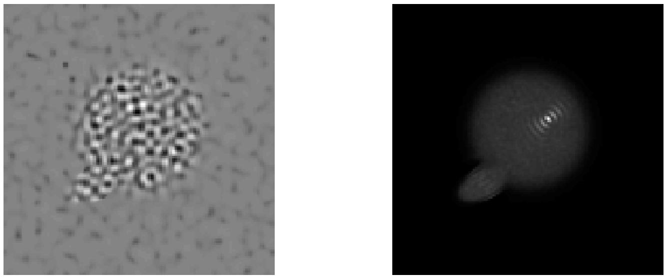

Figure 34: Difference image (MEM - Mean) and Variance image