Chapter 12

Deconvolution in sythesis imaging–an introduction

Rajaram Nityananda

12.1 Preliminaries

These lectures describe the two main tools used for deconvolution in the context of radio aperture synthesis. The focus is on the basic issues, while other lectures at this school will deal with aspects closer to the actual practice of deconvolution. The practice is dominated by the descendants of a deceptively simple-looking , beautiful idea proposed by J. Högbom (A&A Suppl. 15 417 1974), which goes by the name of CLEAN. About the same time, another, rather different and perhaps less intuitive idea due to the physicist E.T. Jaynes was proposed by J.G. Ables (A&A Suppl 15 383 1974) for use in astronomy. This goes by the name of the Maximum Entropy Method, MEM for short. MEM took a long time to be accepted as a practical tool and even today is probably viewed as an exotic alternative to CLEAN. We will see, however, that there are situations in which it is likely to do better, and even be computationally faster. The goal of these lectures is to give enough background and motivation for new entrants to appreciate both CLEAN and MEM and go deeper into the literature.

12.2 The Deconvolution Problem

12.2.1 Interferometric Measurements

An array like the GMRT measures the visibility function $V(u,v)$ along baselines which move along tracks in the $u-v$ plane as the earth rotates, For simplicity, let us assume that these measurements have been transferred onto a discrete grid and baselines are measured in units of the wavelength. The sky brightness distribution $I(l,m)$ in the field of view is a function of l,m which are direction cosines of a unit vector to a point on the celestial sphere referred to the u and v axes. The basic relationship between the measured visibility function V and the sky brightness I is a Fourier transform.

$$V(u,v) = \int\int I(l,m)\exp(-2\pi i(lu+mv))\ dl\ dm.$$

This expression also justifies the term “spatial frequency” to describe the pair (u, v), since u and v play the same role as frequency plays in representing time varying signals.

Many things have been left out in this expression, such as the proper units, polarisation, the primary beam response of the individual antennas, the non-coplanarity of the baselines, the finite observing bandwidth, etc. But it is certainly necessary to understand this simplified situation first, and the details needed to achieve greater realism can be put in later.

Aperture synthesis, as originally conceived, involved filling in the $u - v$ plane without any gaps upto some maximum baseline b_{max} which would determine the angular resolution. Once one accepts this resolution limit, and writes down zeros for visibility values outside the measured circle, the Fourier transform can be inverted. One is in the happy situation of having as many equations as unknowns. A point source at the field centre.(which has constant visibility) would be reconstructed as the Fourier transform of a uniformly filled circular disk of diameter $2b_{max}$. This is the famous Airy pattern with its first zero at $1.22/(2b_{max})$. The baseline b is already measured in wavelengths, hence the missing λ in the numerator. But even in this ideal situation, there are some problems. Given an array element of diameter D (in wavelengths again!), the region of sky of interest could even be larger than a circle of angular diameter $2/D$. A Fourier component describing a fringe going through one cycle over this angle corresponds to a baseline of $D/2$. But measuring such a short baseline would put two dishes into collision, and even somewhat larger baselines than D run the risk of one dish shadowing the other. In addition, the really lowest Fourier component corresponds to $(u, v) = (0, 0)$, the total flux in the primary beam. This too is not usually measured in synthesis instruments Thus, there is an inevitable “short and zero spacings problem” even when the rest of the $u - v$ plane is well sampled.

12.2.2 Dirty Map and Dirty Beam

But the real situation is much worse. With the advent of the Very Large Array (VLA), the majestic filling in of the $u - v$ plane with samples spaced at $D/2$ went out of style. If one divides the field of view into pixels of size $1/(2b_{max})$, then the total number of such pixels (resolution elements) would be significantly larger than the number of baselines actually measured in most cases. This is clearly seen in plots of $u - v$ coverage which have conspicuous holes in them. The inverse Fourier transform of the measured visibility is now hardly the true map because of the missing data. But it still has a name - the “dirty map” I^D. We define a sampling cum weighting function $W(u, v)$ which is zero where there are no measurements and in the simplest case (called uniform weighting) is just unity wherever there are measurements. So we can get our limited visibility coverage by taking the true visibilities and multiplying by $W(u, v)$. This multiplication becomes a convolution in the sky domain. The “true” map with full visibility coverage is therefore convolved by the inverse Fourier transform of W which goes by the name of the “dirty beam” $B^D(l, m)$.

$$I^D(l, m) = \int \int I(l', m') B^D(l - l', m - m') \; dl' \; dm'$$

where

$$B^D(l, m) \propto \sum W(u, v) \exp(+2\pi i(lu + mv)).$$

For a patchy $u - v$ coverage, which is typical of many synthesis observations, B^D has strong sidelobes and other undesirable features. This makes the dirty map difficult to interpret. What one sees in one pixel has contributions from the sky brightness in neighbouring and even not so neighbouring pixels. For the case of $W = 1$ within a disk of radius b_{max} we get an Airy pattern as mentioned earlier. This is not such a dirty beam after all, and could be cleaned up further by making the weighting non-uniform, i.e. tapering the function W down to zero near the edge $|(u,v)| = b_{max}$. For example, if this weighting is approximated by a Gaussian, then the sky gets convolved by its transform, another Gaussian. This dirty map is now related to the true one in a reasonable way. But, as Ables remarked, should one go to enormous expense to build and measure the longest baseline and then multiply it by zero?

12.2.3 The Need for Deconvolution

Clearly, there has to be a better way than just reweighting the data to make the dirty beam look better, (and fatter, incidentally, since one is suppressing high spatial frequencies), But this better way has to play the dangerous game of interpolating (for short spacings and for gaps in the $u - v$ plane) and extrapolating (for values beyond the largest baseline) the visibility function which was actually measured. The standard terminology is that the imaging problem is “underdetermined” or “ill-posed” or “ill-conditioned”. It has fewer equations than unknowns. However respectable we try to make it sound by this terminology, we are no better than someone solving $x + y = 1$ for both x and y!. Clearly, some additional criterion which selects one (or a few) solutions out of the infinite number possible has to be used. The standard terminology for this criterion is “a priori information”. The term “a priori” was used by the philosopher Kant to describe things in the mind that did not seem to need sensory input, and is hence particularly appropriate here.

One general statement can be made. If one finds more than one solution to a given deconvolution problem fitting a given data set, then subtracting any two solutions should give a function whose visibility has to vanish everywhere on the data set. Such a brightness distribution, which contains only unmeasured spatial frequencies, is appropriately called an “invisible distribution”. Our extra- /inter- polation problem consists in finding the right invisible distribution to add to the visible one!

One constraint often mentioned is the positivity of the brightness of each pixel. To see how powerful this can be, take a sky with just one point source at the field centre. The total flux and two visibilities on baselines $(D/2, 0), (0, D/2)$ suffice to pin down the map completely. The only possible value for all the remaining visibilities is equal to these numbers, which are themselves equal. One cannot add any invisible distribution to this because it is bound to go negative somewhere in the vast empty spaces around our source. But this is an extreme case. The power of positivity diminishes as the field gets filled with emisssion.

Another interesting case is when the emission is known to be confined to a window in the map plane. Define a function $w(l, m) = 1$ inside the window and zero outside. Let $\tilde{w}(u, v)$ be its Fourier transform. Multiplying the map by w makes no difference. In Fourier space, this condition is quite non-trivial, viz $V(u, v) = V(u, v) * \tilde{w}(u, v)$. Notice how the convolution on the right transfers information from measured to unmeasured parts of the $u - v$ plane, and couples them.

12.3 CLEAN

12.3.1 The Högbom Algorithm

Consider a sky containing only isolated point sources. In the dirty map, each appears as a copy of the dirty beam, centred on the source position and scaled by its strength. However, the maxima in the map do not strictly correspond to the source positions, because each maximum is corrupted by the sidelobes of the others, which could shift it and alter its strength. The least corrupted, and most corrupting, source is the strongest. Why not take the largest local maximum of the dirty map as a good indicator of its location and strength? And why not subtract a dirty beam of the appropriate strength to remove to a great extent the bad effects of this strongest source on the others? The new maximum after the subtraction now has a similar role. At every stage, one writes down the co-ordinates and strengths of the point sources one is postulating to explain the dirty map. If all goes well, then at some stage nothing (or rather just the inevitable instrumental noise) would be left behind. We would have a collection of point sources, the so called CLEAN components, which when convolved with the dirty beam give the dirty map.

One could exhibit this collection of point sources as the solution to the deconvolution problem, but this would be arrogant, since one has only finite resolution. As a final gesture of modesty, one replaces each point source by (say) a gaussian, a so called “CLEAN” beam, and asserts that the sky brightness, convolved with this beam, has been found.

This strategy, which seems so reasonable today, was a real breakthrough in 1974 when proposed by J. Högbom. Suddenly, one did not have to live with sidelobes caused by incomplete $u - v$ coverage. In fact, the planning for new telescopes like the VLA must have taken this into account- one was no longer afraid of holes.

12.3.2 The Behaviour of CLEAN

With hindsight, one can say that the initial successes were also due to the simplicity of the sources mapped. It is now clear that one should not be applying this method to an extended source which covered several times the resolution limit (the width of the central peak of the dirty beam). Such a source could have a broad, gentle maximum in the dirty map, and subtracting a narrow dirty beam at this point would generate images of the sidelobes with the opposite sign. This would generate new maxima where new CLEAN components would be placed by the algorithm, and things could go unstable. One precaution which certainly helps is the “gain factor” (actually a loss factor since it is less than one). After finding a maximum, one does not subtract the full value but a fraction g typically 0.2 or less. In simple cases, this would just make the algorithm slower but not change the solution. But this step actually helps when sources are more complex. One is being conservative in not fully believing the sources found initially. This gives the algorithm a chance to change its mind and look for sources elsewhere. If this sounds like a description of animal behaviour, the impression being conveyed is correct. Our understanding of CLEAN is largely a series of empirical observations and thumb rules, with common sense rationalisations after the fact, but no real mathematical theory. One exception is the work of Schwarz (A&A 65 345 1978) which interpreted each CLEAN subtraction as a least squares fit of the current dirty map to a single point source. This is interesting but not enough. CLEAN carries out this subtraction sequentially, and that too with a gain factor. In principle, each value of the gain factor could lead to a different solution, i.e a different collection of CLEAN components, in the realistic case when the number of $u - v$ points is less than the number of resolution elements in the map. So what are we to make of the practical successes of CLEAN? Simply that in those cases, the patch of the sky being imaged had a large enough empty portion that the real number of CLEAN components neeeded was smaller than the number of data points available in the $u-v$ plane. Under such conditions, one could believe that the solution is unique. Current implementations of CLEAN allow the user to define “windows” in the map so that one does not look for CLEAN components outside them. But when a large portion of the field of view has some nonzero brightness, there are indeed problems with CLEAN. The maps show spurious stripes whose separation is related to unmeasured spatial frequencies (that's how one deduces they are spurious). One should think of this as a wrong choice of invisible distribution which CLEAN has made. Various modifications of CLEAN have been devised to cope with this, but the fairest conclusion is that the algorithm was never meant for extended structure. Given that it began with isolated point sources it has done remarkably well in other circumstances.

12.3.3 Beyond CLEAN

Apart from the difficulties with extended sources, CLEAN as described above is an inherently slow procedure. If N is the number of pixels, subtracting a single source needs of the order of N operations. This seems a waste when this subtraction is a provisional, intermediate step anyway! B.G. Clark had the insight of devising a faster version, which operates with a truncated dirty beam, but only on those maxima in the map strong enough that the far, weak sidelobes make little difference. Once these sources have been identified by this rough CLEAN (called a "minor cycle"), they are subtracted together from the full map using an fast fourier transform (FFT) for the convolution, which takes only $N \log N$ operations. This is called the "major cycle". The new residual map now has a new definition of "strong" and the minor cycle is repeated.

A more daring variant, due to Steer, Dewdney, and Ito, (hence SDI CLEAN) carries out the minor cycle by simply identifying high enough maxima, without even using CLEAN, which is kept for the major cycle. Other efforts to cope with extended sources go under the name of "multiresolution CLEAN". One could start with the inner part of the $u-v$ plane and do a CLEAN with the appropriate, broader dirty beam. The large scale structure thus subtracted will hopefully now not spoil the next stage of CLEAN at a higher resolution, i.e using more of the $u - v$ plane.

12.4 Maximum Entropy

12.4.1 Bayesian Statistical Inference

This method, or class of methods, is easy to describe in the framework of an approach to statistical inference (i.e all of experimental science?) which is more than two hundred years old, dating from 1763! Bayes Theorem about conditional probabilities states that

$$P(A|B)P(B) = P(B|A)P(A) = P(A, B).$$

As a theorem, it is an easy consequence of the definitions of joint probabilities (denoted by $P(A, B)$), conditional probabilities (denoted by $P(A|B)$) and marginal or unconditional probabilities (denoted by $P(A)$). In words, one could say that the fraction of trials A *and* B *both* happen ($P(A, B)$) is the product of (i) the fraction of trials in which A happens ($P(A)$) irrespective of B, and (ii) the further fraction of A-occurences which are also B-occurences ($P(B|A)$). The other form for $P(A|B)$ follows by interchanging the roles of A and B.

The theorem acquires its application to statistical inference when we think of A as a hypothesis which is being tested by measuring some data B. In real life, with noisy and incomplete data, we never have the luxury of measuring A directly, but only something depending on it in a nonunique fashion. If we understand this dependence, i.e understand our experiment, we know $P(B|A)$. If only, (and this is a big IF!), someone gave us $P(A)$, then we would be able to compute the dependence of $P(A|B)$ on A from Bayes theorem.

$$P(A|B) = P(B|A)P(A)/P(B).$$

Going from $P(B|A)$ to $P(A|B)$ may not seem to be a big step for a man, but it is a giant step for mankind. It now tells us the probability of different hypotheses A being true based on the given data B. Remember, this is the real world. More than one hypothesis is consistent with a given set of data, so the best we can do is narrow down the possibilities. (If "hypothesis" seems too abstract, think of it as a set of numbers which occur as parameters in a given model of the real world)

12.4.2 MEM Images

Descending now from the sublime to aperture synthesis, think of A as the true map and B as the dirty map, or equivalently its Fourier transform, the set of measured visiblilities. We usually want a single map, not a probability distribution of A. So we need the further step of maximising $P(A|B)$ with respect to A. All this is possible if $P(A)$ is available for a given true map $I(l, m)$. One choice, advocated by Gull and Daniell in 1978, was to take

$$\log P(\{I(l, m)\}) \propto -\int\int I(l, m) \ln I(l, m)\ dl\ dm.$$

The curly brackets around I on the left side are meant to remind us that the entropy is a single number computed from the entire information about the brightness, i.e the whole set of pixel values. Physicists will note that this expression seems inspired by Boltzmann's formula for entropy in statistical mechanics, and communication engineers will see the influence of Shannon's concept of information. It was E.T. Jaynes writing in the Physical Review of 1957 who saw a vision of a unified scheme into which physics, communication theory, and statistical inference would fall (with the last being the most fundamental!). In any case, the term "entropy" for the logarithm of the prior distribution of pixel values has stuck. One can see that if the only data given was the total flux, then the entropy as defined above is a maximum when the flux is distributed uniformly over the pixels. This is for the same reason that the Boltzmann entropy is maximised when a gas fills a container uniformly. This is the basis for the oft-heard remark that MEM produces the flattest or most featureless map consistent with the data - a statement we will see requires some qualification. But if one does not want this feature, a modified entropy function which is the integral over the map of $-I \ln(I/I^d)$ is defined. $I^d(l, m)$ is called a "default image". One can now check that if only total flux is given the entropy is a maximum for $I \propto I^d$.

The selection of a prior is, in my view, the weakest part of Bayesian inference, so we will sidestep the debate on the correct choice. Rather, let us view the situation as an opportunity, a license to explore the consequences of different priors on the "true" maps which emerge. This is easily done by simulation – take a plausible map, Fourier transform, sample with a function W so that some information is now missing, and use your favourite prior and maximise "entropy" to get a candidate for the true map. It is this kind of study which was responsible for the great initial interest in MEM. Briefly, what MEM seemed to do in simple cases was to eliminate the sidelobes and even resolve pairs of peaks which overlapped in the true map, i.e it was sometimes "better" than the original! This last feature is called superresolution, and we will not discuss this in the same spirit of modesty that prompted us to use a CLEAN beam. Unlike CLEAN, MEM did not seem to have a serious problem with extended structure, unless it had a sharp edge (like the image of a planet). In this last case, it was found that MEM actually enhanced the ripples near the edge which were sitting at high brightness levels; though it controlled the ripples which were close to zero intensity. This is perhaps not surprising if one looks at the graph of the function $= I \ln I$. There is much more to be gained by removing ripples near $I = 0$ than at higher values of I, since the derivative of the function is higher near $I = 0$.

Fortunately, these empirical studies of the MEM can be backed up by an analytical/graphical argument due to Ramesh Narayan, which is outlined below. The full consequences of this viewpoint were developed in a review article (Annual review of Astronomy and Astrophysics 24 127 1986), so they will not be elaborated here, but the basic reasoning is simple and short enough. Take the expression for the entropy, and differentiate it with respect to the free parameters at our disposal, namely the *un*measured visibilities, and set to zero for maximisation. The derivative of the entropy taken with respect to a visibility $V(u', v')$ is denoted by $M(u', v')$. The understanding is that u', v' have *not* been measured. The condition for a maximum is

$$M(u', v') = \int\int(-1 - \ln(I(l, m)) \exp(+2\pi i(lu' + mv')\ dl\ dm = 0.$$

This can be interpreted as follows. The *logarithm* of the brightness is like a dirty map, i.e it has no power at unmeasured baselines, and hence has sidelobes etc. But the brightness I itself is the exponential of this "band limited function" (i.e one with limited spatial frequency content). Note first of all that the positivity constraint is nicely implemented– exponentials are positive. Since the exponential varies rather slowly at small values of I, the ripples in the "baseline" region between the peaks are suppressed. Conversely, the peaks are sharpened by the steep rise of the exponential function at larger values of I. One could even take the extreme point of view that the MEM stands unmasked as a model fitting procedure with sufficient flexibility to handle the cases usually encountered. Högbom and Subrahmanya independently emphasised very early that the entropy is just a penalty function which encourages desirable behaviour and punishes bad features in the map (IAU Colloq. 49, 1978). Subrahmanya's early work on the deconvolution of lunar occultation records at Ooty (TIFR thesis, 1977) was indeed based on such penalties.

More properties of the MEM solution are given in the references cited earlier. But one can immediately see that taking the exponential of a function with only a limited range of spatial frequencies (those present in the dirty beam) is going to generate all spatial frequencies, i.e., one is extrapolating and interpolating in the $u - v$ plane. It is also clear that the fitting is a nonlinear operation because of the exponential. Adding two data sets and obtaining the MEM solution will not give the same answer as finding the MEM solution for each separately and adding later! A little thought shows that this is equally true of CLEAN.

If one has a default image I^d in the definition of the entropy function, then the same algebra shows that I/I^d is the exponential of a band-limited function. This could be desirable. For example, while imaging a planet, if the sharp edge is put into I^d, then the MEM does not have to do so much work in generating new spatial frequencies in the ratio I/I^d. The spirit is similar to using a window to help CLEAN find sources in the right place.

12.4.3 Noise and Residuals

The discussion so far has made no reference to noise in the interferometric measurements. But this can readily be accomodated in the Bayesian framework. One now treats the measurements not as constraints but as having a Gaussian distribution around the "true" value which the real sky would Fourier transform to. Thus the first factor $P(B|A)$ on the right hand side of Bayes theorem would now read

$$P(B|A) = \prod \exp(-(\int\int I(l, m)exp(-2\pi i(lu + mv))\ dl\ dm - V_m(u, v)|^2/2\sigma_{u,v}^2.$$

The product is over measured values of u, v. A nice feature of the gaussian distribution is that when we take its logarithm, we get the sum of the squares of the residuals between the model predictions (the integral above) and the measurements $V_m(u, v)$ – also known as "chi-squared" or χ^2. The logarithm of the prior is of course the entropy factor. So, in practice, we end up maximising a linear combination of the entropy and χ^2, the latter with a negative coefficient. This is exactly what one would have done, using the method of Lagrange multipliers, if we were maximising entropy subject to the constraint that the residuals should have the right size, predicted by our knowledge of the noise.

All is not well with this recipe for handling the noise. The discrepancy between the measured data and the model predictions can be thought of as a residual vector in a multidimensional data space. We have forced the length to be right, but what about the direction? True residuals should be random, i.e the residual vector should be uniformly distributed on the sphere of constant χ^2. But since we are maximising entropy on this sphere, there will be a bias towards that direction which points along the gradient of the entropy function. This shows in the maps as a systematic deviation tending to lower the peaks and raise the "baseline" i.e the parts of the image near zero I. To lowest order, this can be rectified by adding back the residual vector found by the algorithm. This does not take care of the invisible distribution which the MEM has produced from the residuals, but is the best we can do. Even in the practice of CLEAN, residuals are added back for similar reasons.

The term "bias" is used by statisticians to describe the following phenomenon. We estimate some quantity, and even after taking a large number of trials its average is not the noise-free value. The noise has got "rectified" by the non-linear algorithm and shows itself as a systematic error. There are suggestions for controlling this bias by imposing the right distribution and spatial correlations of residuals. These are likely to be algorithmically complex but deserve exploration. They could still leave one with some subtle bias since one cannot really solve for noise. But to a follower of Bayes, bias is not necesarily a bad thing. What is a prior but an expression of prejudice? Perhaps the only way to avoid bias is to stop with publishing a list of the measured visibility values with their errors. Perhaps the only truly open mind is an empty mind!

12.5 Further Reading

1. R.A. Perley, F.R. Schwab, & A.H. Bridle, eds., 'Synthesis Imaging in Radio Astronomy', ASP Conf. Series, vol. 6.

2. Thompson, R.A., Moran, J.M. & Swenson, G.W. Jr., 'Interferometry & Synthesis in Radio Astronomy', Wiley Interscience, 1986.

3. Steer, D.G., Dewdney, P.E. & Ito, M.R., "Enhancements to the deconvolution algorithm 'CLEAN"', 1984,A&A,137,159.