MULTIGRID METHODS FOR TOMOGRAPHIC RECONSTRUCTION

T. J. Monks

1. INTRODUCTION

In many inversion problems, the random nature of the observed data has a very significant impact on the accuracy of the reconstruction. In these situations, reconstruction techniques that are based on the known statistical properties of the data, are particularly useful. In a report [1], we considered methods for the reconstruction of an object from its projection data, and showed that maximum likelihood methods are particularly successful when the data is very noisy. Although maximum likelihood methods were first applied to emission tomography [2], they have also found utility in other imaging modalities, including transmission [3], diffraction [4] and limited angle transmission tomography [5, 6]. The major problem with maximum likelihood methods for image reconstruction, is that the normal technique used, the expectation-maximisation (EM) algorithm, is excruciatingly slow to converge to the desired objective. In this paper, we demonstrate the use of multigrid methods to overcome this handicap.

In order to facilitate subsequent discussion, we shall give a brief review of the expectation-maximisation (EM) algorithm, as described in [1]. The EM algorithm is an algebraic reconstruction technique where we resolve the object to be reconstructed into a grid-like array of unknowns and set up algebraic equations for these unknowns in terms of the measured projection data. The reconstruction region is sub-divided into pixels numbered 1, ..., B. In pixel j, the parameter to be reconstructed is assumed to have a constant value λ_j. If the projections are measured at sites numbered 1, ..., T, then this linear system of equations can then be written as :

$$\mathbf{n} = \mathbf{A}\boldsymbol{\lambda} + \boldsymbol{\varepsilon} \tag{1.1}$$

where A is a sparse matrix connecting the object vector, $\boldsymbol{\lambda} \in \mathbb{R}^B$, to the measured data vector, $\mathbf{n} \in \mathbb{R}^T$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^T$, is the vector of errors due to inaccuracies in both the measurement and the discretization processes.

The reconstruction problem is to find a method of estimating the object vector, $\boldsymbol{\lambda}$, given only the measurements, **n**, and some estimate of the connection matrix, **A**. In transmission tomography, the matrix element A_{ij} linking pixel j to detector i is calculated as a function of the intersection area of the ray path through pixel j to the relevant detector. In emission tomography, A_{ij} is usually calculated as some approximation to the probability that detector i measures activity in pixel j. The algorithms to solve (1.1) can require a large amount of computation as a result of the characteristics of the matrix **A** :

- it is huge, possibly as large as $10^5 \times 10^5$ or more,
- it is reasonably sparse, perhaps only 10% of the elements are non-zero,
- the pattern or structure of the non-zero elements in the matrix is irregular and therefore difficult to exploit,
- the location and value of the non-zero elements may be efficiently generated either by rows or by columns, depending on the definition of A, but it is generally not possible to do so for both rows and columns.

The direct solution of sparse systems of this size is not only numerically unwise, it is also inefficient, as the sparse matrix is generally filled into a full one during the course of the method. Consequently, we usually resort to iterative methods to obtain the reconstructed image.

In situations where the object vector, $\boldsymbol{\lambda}$, is well modelled as a set of independent Poisson processes, the EM algorithm can be formulated as a multiplicative, iterative method that finds a solution to (1.1) that maximises the likelihood function :

$$\begin{aligned} l(\boldsymbol{\lambda}) &= \ln\left[\mathbf{P}(\mathbf{n} \mid \boldsymbol{\lambda})\right] \\ &= \sum_t \left[-\sum_{b'} A_{b',t}\lambda_{b'} \quad + n_t \ln\left[\sum_{b'} A_{b',t}\lambda_{b'}\right] - \ln\left[n_t\,!\right] \right] \end{aligned} \tag{1.2}$$

through the iterations :

$$\lambda_b^{(i+1)} \leftarrow \lambda_b^{(i)} \sum_{t=1}^{T} \frac{n_t A_{b,t}}{\sum\limits_{b'=1}^{B} \lambda_{b'}^{(i)} A_{b',t}} \tag{1.3}$$

The reconstructions obtained with the EM algorithm, have superior resolution, signal-to-noise and contrast compared to those obtained from the more conventional filtered-backprojection technique. Unfortunately, the computational requirements of the EM algorithm are more than an order of magnitude greater than backprojection methods and convergence to the asymptotic solution becomes increasingly slow as the iterations progress. A more complete description of the algorithm and its behaviour is given, for example, in [1, 7].

Large, sparse systems of equations, such as (1.1), are usually solved with iterative methods. In general, we find that the asymptotic approach to a solution becomes extremely slow, in fact, the convergence rate diverges exponentially as the solution is approached — this has been called *critical slowing down*. The inherent sluggishness of local iteration algorithms can be studied from a spatial frequency perspective. A Fourier analysis of the error function from one iteration to the next shows that while the components of the error decrease rapidly in some frequency bands, they decay slowly in other bands. The region of rapid decay depends on the specific relaxation method, but typically consists of the high frequency components which have wavelengths of the same order as the mesh spacing. On the other hand, the region of slow decay always includes the low to mid-frequency components, implying that errors about the global information persist through many iterations [8].

For the EM iteration of (1.3), the update term is given by :

$$\Delta\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i+1)} - \boldsymbol{\lambda}^{(i)} \tag{1.4}$$

We can examine the spectral content of this term by taking a 2D Fourier transform of $\Delta\boldsymbol{\lambda}$, then converting this 2D transform to a 1D function by summing it into bins with the same radial frequency. A plot of the resulting function at each iteration of a typical reconstruction problem is shown in figure 1.1. This figure shows that most of the energy in the update term of equation 1.4 is concentrated around two frequencies: the zero spatial frequency that represents overall brightness in the image, and the pixel spacing frequency that represents differences between a pixel and its closest neighbours. The information bearing parts of the image will be the shapes and blobs that span many pixels and will consequently have frequency components that lie between these two extremes. Unfortunately, the

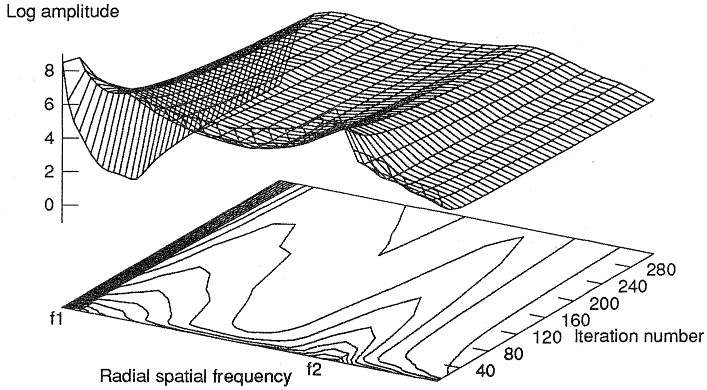


Figure 1.1 *The average spectral content of the EM update term (equation 1.4) shown per iteration. f1 is the zero spatial frequency, whilst f2 is the inter-pixel spacing frequency.*

figure shows that a fixed grid algorithm is inefficient at updating these components; many iterations are required to develop and alter the information-bearing regions in the reconstruction.

Since the step size of a single update must be limited to maintain stability at short distances, the speed of evolution of large scale features is greatly slowed. Common ($\mathbf{L}_2$ or $\mathbf{L}_1$) error norms decrease quickly during the first few iterations, so long as there are high-frequency components to be annihilated, but soon degenerate to a slow, asymptotic diminution when only the low-frequency components remain. This suggests that while the relaxations may be inefficient at completely cancelling the error function, they can be very efficient in smoothing it.

The multigrid method [9] has been very successful in overcoming the critical slowing problem. It is a paradigm of the general class of *divide-and-conquer* algorithms where the solution of a set of sub-problems is collectively faster than a whole problem tackled at once. The basic idea underlying multigrid approaches is to solve the problem on multiple grid scales, where processing on a given grid generally depends on coarser grid corrections and finer grid residual transfers. Hence, as the iterations progress, a solution evolves which is consistent over a large range of length scales. Empirical studies on model problems indicate that multigrid methods are *asymptotically optimal*, that is, the computational workload required to attain a fixed accuracy is proportional to the number of discrete unknowns (convergence in essentially $O(N)$ number of operations, where N is the number of unknowns). Single-grid methods on the other hand, show $O(N^3)$ convergence, so that the number of iterations needed for the system to relax to its steady state solution grows much faster than the number of mesh points. Hence in image processing applications, where N can be very large (order 10^4 to 10^6), multigrid methods offer potentially dramatic increases in efficiency over standard single-grid methods. We refer the reader to [9] for a description of the MV and FMV cycle algorithms that will be used in this paper.

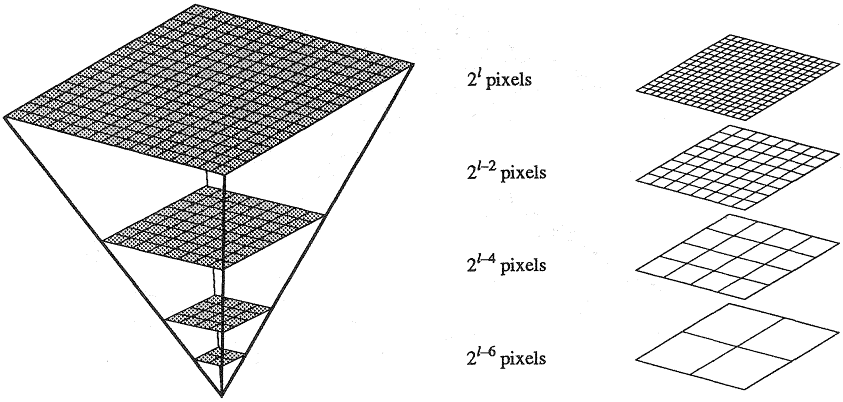


Figure 1.2 : *Two representations of a multi-resolution decomposition of an image. The left hand figure is a constant pixel size presentation showing the pyramidal neighbourhood relationships between pixels in adjacent layers. The right hand figure has a constant spatial dimension for all the layers and highlights the resolution changes. The transfer between layers is via the processes of restriction and prolongation, as discussed in the text.*

The use of multigrid schemes in image reconstruction should be particularly fruitful as there is ample psychological evidence that early processing within the human visual system takes place on multiple scales [10, and references therein]. Burt [11, 12] introduced the idea of a pyramid of related images (figure 1.2), which may be formed either by smoothing followed by restriction (known as a low-pass pyramid) or subtraction of the smoothed image from the original followed by restriction (known as a band-pass pyramid). This scheme can be generalised to encompass the idea of a pyramidal data structure which, at each descending level, provides successively condensed representations of the information in the image. This allows construction of sub-trees in the pyramid whose leaves are pixels (or local features) in the image, and whose roots represent global features of various types. Thus pyramids provide a possible means of bridging the gap between pixel-level and region-level image analysis since local computations on a coarse mesh can correspond to more global computations on a fine mesh. The decomposition of an image on multiple scales provides an intermediate representation between uni-grid spatial and Fourier domain descriptions. These multi-resolution schemas have been applied to image segmentation, region matching, feature and shape analysis, surface interpolation, optical flow, sub-band decomposition and compression.

Surprisingly, multigrid techniques have rarely been applied to image reconstruction problems. Terzopoulos [13, 14] has applied multigrid approaches to four problems in image reconstruction: surface fitting to sparse data, image reflectance from intensity, shape from shading and optical flow. Each of these problems can be formulated as a linear system, and the underlying procedure Terzopoulos used to solve each corresponds to the full multigrid V-cycle (FMV-cycle). Terzopoulos initially proposed an accommodative approach, where the method uses an internal check, based on the computation of the residual norm, to determine when to switch between grids. He found that for many problems, the accommodative algorithm behaves in a fairly fixed fashion, performing a similar number of iterations at each level before switching. It is then possible to abandon the computationally expensive calculation of the dynamic residual norm in favour of a pre-assigned, fixed-flow approach. For each problem, Terzopoulos obtained substantial improvements in efficiency over single-grid methods, and recommended that the multigrid approach be tried for all image reconstruction problems.

Herman *et al.* [15] applied a pseudo-multigrid approach to the problem of transmission tomography. The authors make the point that multigrid can change either one or both of the image discretization (the size of $\boldsymbol{\lambda}$, B) or data sampling (the size of $\mathbf{n}$, T). Full multigrid approaches require changes in both T and B, but the authors claim that "in image reconstruction the overhead associated with changing the picture digitization outweighs any potential savings, (consequently) we discuss in precise fashion only the special case in which data sampling is variable, but the picture digitization does not change". This is really a uni-grid approach using sub-sets of the measured data. The sub-sets were chosen as simple partitions of the measured data set. The conclusion they reached was that such a *partitioned uni-grid* implementation of an algebraic reconstruction technique (ART) was superior to standard ART in terms of both asymptotic error norm and efficiency, but was still less efficient than filtered backprojection.

Kaufman [16] applied a nested iteration to the EM algorithm reconstruction for emission tomography. In this method, an estimate on a coarse grid was obtained which was used as the starting solution on a fine grid after interpolation. No use was made of fine-to-coarse residual transfers. Kaufman obtained disappointing results using piecewise constant interpolation on a quad-tree pyramid. She found that interpolation artefacts persisted for many iterations on the finer grid. In hindsight, this form of interpolation introduces an entire range of new frequency components into the starting solution on each new grid that would be avoided with a smoother interpolant. Ranganath *et al.* [17] also used coarse-to-fine nested iteration on the EM algorithm and reported much better results than did Kaufman. Indeed, they achieved convergences an order of magnitude faster than a single-grid EM algorithm. The method is not described completely in the paper, however we know that a quad-tree pyramid was used. The interpolation scheme was not specified, other than stating that it satisfied a non-negativity constraint. Again, no use was made of fine-to-coarse residual transfers.

2. MULTIGRID IMPLEMENTATIONS OF THE EM ALGORITHM

In this chapter we shall give details of what we have found to be the most promising improvement to the EM algorithm, namely the use of multigrid methods. We have developed a natural extension of the work reported in the introduction [13-17], and tried both nested-iteration and a full multigrid method on the EM iterations. Three techniques will be described. The first two are nested-iteration algorithms where we perform EM iterations on a coarse grid until slow convergence appears, and then transfer this solution to a finer grid and repeat the process. The third method is a full multigrid approach.

2.1 Algorithm I — A nested-iteration implementation of the EM algorithm

An effective, and well established strategy [18] for finding the solution to a system such as $\mathbf{A}\mathbf{u} = \mathbf{f}$, is to first solve the problem on a coarse grid (this discrete space is denoted by Ω^h) to a desired level of accuracy and then form an initial guess on a finer grid by a process of prolongation (*i.e.* interpolation). This can be written as :

Step 1: Relax $\mathbf{A}^h \mathbf{u}^h = \mathbf{f}^h$ to obtain an estimate on $\Omega^h_,$
Step 2: Interpolate solution onto Ω^{h+1}: $\mathbf{u}^{h+1} \leftarrow \mathbf{I}_h^{h+1} \mathbf{u}^h$
Step 3: Relax $\mathbf{A}^{h+1} \mathbf{u}^{h+1} = \mathbf{f}^{h+1}$ to obtain an estimate on Ω^{h+1} (2.1)

Where a superscript h+1 denotes our finer grid variables and $\mathbf{I}_h^{h+1}$ is the prolongation operator between the two grids[1]. If our true solution is smooth this will be very effective because such a guess can eliminate many of the early relaxation sweeps generally required for naïve guesses. Thus the residual error usually starts off much smaller than it would with the naïve guess. This approach can be repeated to give a nested iteration of relaxations and prolongations. However, *asymptotic convergence rates are generally independent of the initial guess*, so the slow rates will quickly reappear on each level. Secondly, although it is suitable for generating a single accurate solution on the finest grid, it cannot generate solutions having the finest-level accuracy over the hierarchy of coarser grids. Finally, this method would be unsatisfactory if the true solution was highly oscillatory; smooth errors can be well approximated by coarse interpolants, but oscillatory errors cannot.

For our first algorithm, we employ a nested-iteration approach as described above, except that here we only sub-sample (restrict) the image data vector, $\boldsymbol{\lambda}$ and the connection matrix $\mathbf{A}$, we *do not* sub-sample the measured data vector, $\mathbf{n}$, on the coarser grids. This first algorithm can be written as:

Step 0: *Initialisation*:
$h = h_0$ (the coarsest grid)
${}^{h_0}\boldsymbol{\lambda} = \mathbf{c}$ (initial solution is constant valued)
Step 1: Perform v_1 relaxations of EM algorithm on grid Ω^h

$${}^h\lambda_b^{(i+1)} \longleftarrow {}^h\lambda_b^{(i)} \sum_{t=1}^{T} \frac{n_t \, {}^hA_{b,t}}{\sum\limits_{b'=1}^{B} {}^h\lambda_{b'}^{(i)} \, {}^hA_{b',t}}$$

Step 2: If $h =$ finest grid exit, otherwise interpolate solution onto Ω^{h+1}
${}^{h+1}\boldsymbol{\lambda} = \mathbf{I}_h^{h+1} \, {}^h\boldsymbol{\lambda},$
$h \leftarrow h+1,$
repeat step 1. (2.2)

Piecewise bilinear interpolation is used for prolongation of $\boldsymbol{\lambda}$ between grids, and injection (point sampling) is used for restriction of the connection matrix. Linear prolongation can be considered as a two stage process (although it may not be implemented as such). Firstly, we up-sample from Ω^h to Ω^{h+1}, assigning a zero value to pixels not in the set Ω^h. This operation mirrors the frequency spectrum of Ω^h onto the new high-frequency components introduced into Ω^{h+1}, a process sometimes called imaging. The second logical stage of prolongation is low-pass filtering, which attempts to reduce the

[1] *Most multigrid schemes employ meshes which the grid spacing on a coarse mesh is twice the spacing on the next finest mesh. This is nearly universal practice, since there seems to be no advantage in using grid spacings with ratios other than 2, which would incur the disadvantage of increased complexity in the prolongation and restriction operators.*

imaging effect. Similarly, linear restriction can be decomposed into low-pass filtering and down-sampling stages, where the filtering is now aimed at reducing the high-frequency components in Ω^h that would otherwise be aliased to low frequencies in Ω^{h-1}. A large filter support allows for good suppression of the aliasing and imaging effects. Smaller kernels however, preserve local image features and have a lower computational cost. We seek a filter kernel that is compact in both the spatial and the frequency domains. Wavelet methods are suggestive [10], the approximate Gaussian [11] and Garbor [19] functions are ideal. However, the interpolation of λ between grids must not introduce negative values, nor must the total sum over all of λ be altered (to conserve emission activity with the recorded projection counts). With these extra constraints, the tradeoff amongst the possible kernels is rather marginal, hence, we have chosen the simplest feasible schemes: bilinear interpolation and injection[2]. We must also consider the method used for extension at the periphery. We wish to extend the Ω^h image whilst avoiding artefacts in the high frequency components of Ω^{h+1} that would be introduced by artificial discontinuities. We have a choice of replication of the boundary pixels, a zero or constant valued extension, or a wrap-around from the opposite boundary. We have decided, for this problem, that the first extension has the minimal impact on the prolongation operation.

We initially implemented an accommodative algorithm with switching to a finer mesh based upon the rate of change in the likelihood function. The onset of critical slowing is signalled once this rate starts to diminish. However, we found that the calculation of the likelihood norm was computationally expensive, and the benefits of this adaptive switching were not significant. In the final version of the program, we therefore used a fixed switch method, with v_1 set between 5 and 20 iterations.

2.2 Algorithm II — Another nested-iteration implementation

This algorithm is almost the same as the first, except that we now sub-sample the measured data vector, $\mathbf{n}$, on the coarser grids as well as the image data vector, λ and the connection matrix $\mathbf{A}$. The new algorithm can be written as:

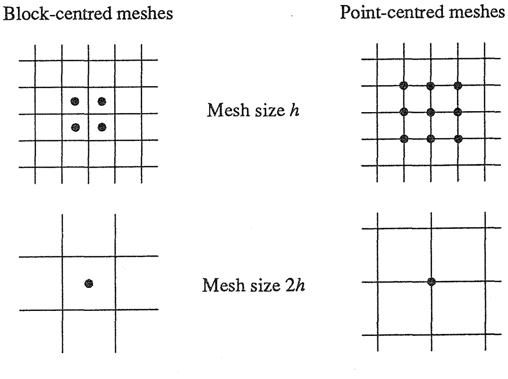


Figure F2.1 : *Grid systems for a discrete representation.*

[2] *There are two contrasting grid systems that we could use as a basis for a discrete representation. In a block-centred mesh, x(i,j) is identified with the average value of the underlying continuous function in block (i,j), whereas in a point-centred mesh it is considered to be the value at the intersection of the (i,j) grid lines.*

When we transform between different mesh sizes, we should employ filter kernels with even-sized supports for a block-centred grid and with odd-sized supports for a point-centred grid to prevent spatial shifts between the mesh and the underlying continuous function.

In this work, a block-centred representation is appropriate for the vectors $\mathbf{n}$ *and* λ*, consequently we must employ even-sized operators in the interpolation and restriction functions. A point-centred representation is appropriate for the connection matrix,* $\mathbf{A}$.

```
Step 0:  Initialisation:
             $h = h_0$        (the coarsest grid)
             ${}^{h_0}\boldsymbol{\lambda} = \mathbf{c}$       (initial solution is constant valued)
             Calculate the projection data sets, $\mathbf{N}^h$, for all grid sizes.
Step 1:  Perform $\nu_1$ relaxations of EM algorithm on grid $\Omega^h$
             ${}^{h}\lambda_b^{(i+1)} \longleftarrow {}^{h}\lambda_b^{(i)} \sum_{t \in \mathbf{N}^h} \frac{{}^{h}n_t \, {}^{h}A_{b,t}}{\sum_{b'=1}^{B} {}^{h}\lambda_{b'}^{(i)} \, {}^{h}A_{b',t}}$
Step 2:  If $h$ = finest grid exit, otherwise interpolate solution onto $\Omega^{h+1}$
             ${}^{h+1}\boldsymbol{\lambda} = \mathbf{I}_h^{h+1} \, {}^{h}\boldsymbol{\lambda}$,
             $h \leftarrow h+1$,
             repeat step 1.                                                    (2.3)
```

where $\mathbf{N}^h$ is the set of measured data values used on grid h.

Remembering that we have a fixed number of measured projection samples, we cannot restrict this set down through a large number of levels because we will end up with no data at all. Instead, we make the measured data vector, ${}^{h}\mathbf{n}$, the same size as the unknown image vector, ${}^{h}\boldsymbol{\lambda}$, on the coarsest grid. This should be beneficial because the equations on this grid will hopefully be consistent and give a unique solution. On successively finer grids we increase the projection data set size in line with the number of image pixels until we reach a stage where we run out of data and the equations become under-determined.

We wish to smooth some of the noise that is strongly present in the measured data set, so some form of low-pass filtering is appropriate. Because we consider $\mathbf{n}$ to be a second order tensor, indexed by projection angle and detector position, we use full weighted averaging with a 2×2 kernel. The filtering necessary to construct the sets of projection data is performed once during the initialisation and saved to file for recall as necessary.

2.3 Algorithm III — A full multigrid implementation of the EM algorithm

The third algorithm we have implemented is the full multigrid V-cycle, which with reference to [9], and for the solution to the system $\mathbf{Au} = \mathbf{f}$, can be defined as $\mathbf{u}^h \leftarrow \mathbf{FMV}^h(\mathbf{u}^h, \mathbf{f}^h)$:

```
Step 0:  Intialisation :
            Find an approximate solution for $\mathbf{u}^{h_0}$,
            Set $\mathbf{u}^h = \mathbf{0}$ for $h > h_0$
Step 1:  Repeat until sufficient accuracy is attained:
         do
         {
            $\mathbf{u}^{h+1} = \mathbf{u}^{h+1} + \mathbf{I}_h^{h+1} \mathbf{u}^h$,
            Perform $\nu_0$ cycles: $\mathbf{u}^{h+1} \leftarrow \mathbf{MV}^{h+1}(\mathbf{u}^{h+1}, \mathbf{f}^{h+1})$,
            $h \leftarrow h+1$
         }
         until $(h = \text{finest grid})$                                      (2.4)
```

where the V cycle is defined recursively as follows $\mathbf{u}^h \leftarrow \mathbf{MV}^h(\mathbf{u}^h, \mathbf{f}^h)$:

```
Step 1:  Relax $\nu_1$ times on $\mathbf{A}^h \mathbf{u}^h = \mathbf{f}^h$ to obtain an estimate on $\Omega^h$
Step 2:  If $h > h_0$ (the coarsest grid)
         {
             $\mathbf{f}^{h-1} = \mathbf{I}_h^{h-1}\left(\mathbf{f}^h - \mathbf{A}^h \mathbf{u}^h\right)$,        (transfer residual to coarser grid)
             $\mathbf{u}^{h-1} = \mathbf{0}$,
             $\mathbf{u}^{h-1} \leftarrow \mathbf{MV}^{h-1}\left(\mathbf{u}^{h-1}, \mathbf{f}^{h-1}\right)$,   (find solution on coarser grid by recursion)
             $\mathbf{u}^h \leftarrow \mathbf{u}^h + \mathbf{I}_{h-1}^h \mathbf{u}^{h-1}$               (correct approximation on $\Omega^h$ )
         }
Step 3:  Perform $\nu_2$ relaxation sweeps on $\mathbf{A}^h \mathbf{u}^h = \mathbf{f}^h$                  (2.5)
```

where $\mathbf{I}_h^{h-1}$ denotes the process of restriction (or decimation as it is known in a signal processing context) in which residuals on a fine grid are used to correct solutions on a coarser grid.

An immediate problem that we face in implementing the FMV algorithm is that the EM relaxation given in (1.3) is formulated for the maximum likelihood objective for a set of Poisson-distributed variables. For any particular pixel, we can write the prior distribution for the value of that pixel, λ, as :

$$f(\lambda) = \frac{\exp(-\Lambda)\,\Lambda^{\lambda}}{\lambda!},$$
$$\Lambda = \mathrm{E}(\lambda) \tag{2.6}$$

Then if we consider the case where the set of equations (1.1) are consistent, the maximum likelihood estimate of λ is just its expected value, Λ, and the prior distribution of the error $e = \Lambda - \lambda$ is [3]:

$$f(e) = \frac{\exp(-\Lambda)\,\Lambda^{(\Lambda-\lambda)}}{\Gamma(\Lambda - \lambda + 1)} \tag{2.7}$$

and is shown in figure 2.1. It is clear that the error does not have a prior Poisson distribution, so that

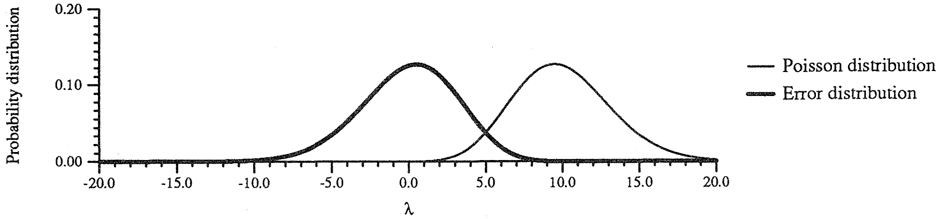


Figure 2.1 : *The distributions* $f(\lambda)$ and $f(e)$ *for* $\Lambda = 10$.

[3] *This is derived as follows: The cumulative distribution of the error is given by :*

$$\begin{aligned} \mathrm{F}(e \le e_1) &= \Pr(e \le e_1) \\ &= \Pr(\Lambda - \lambda \le e_1) \\ &= \Pr(\lambda \le \Lambda - e_1) \\ &= \int_{\Lambda - e_1}^{\infty} \frac{\exp(-\Lambda)\,\Lambda^x}{\Gamma(x+1)}\,dx \end{aligned} \tag{F3.1}$$

So the error distribution is :

$$f(e_1) = \frac{d\,\mathrm{F}(e \le e_1)}{de_1} \tag{F3.2}$$

(2.7) follows using Leibnitz's rule for differentiation of integrals.

the EM relaxation of (1.3) is invalid for the set of residual equations. Indeed it is not obvious that likelihood function of the error will be a concave function. Consequently, the Kuhn-Tucker conditions that were used to derive (1.3), may be neither the sufficient nor the necessary conditions for a maximum.

Clearly, we cannot use the EM iterations in the form of (1.3) for relaxation of the residual equation. We have been unable to derive a closed form EM-type algorithm for the distribution in (2.7) as it stands. It may be possible to do so if the error distribution is approximated by a Gaussian, however, a problem remains with the unknown value for Λ. This could be overcome if we admit the possibility of using $\lambda \approx \Lambda$, but we find this unpalatable, given the large errors that occur during reconstruction. As we cannot derive a closed form expression of the EM algorithm, we have been forced to drop the maximum likelihood objective for the residual equations and look for an alternative method.

We have found that simple iterative techniques, such as Gauss-Seidel or successive over-relaxation, which attempt to solve either the residual equations or their associated normal form, have erratic and very slow convergence for this problem. The primary reason for this behaviour is that the residual equations are even more inconsistent than the direct equations. The secondary reason is that matrix **A** is not diagonally-dominant. Hence, we feel that such a solution of the residual equations is not feasible and that instead, we must opt for some form of minimum-norm solution.

A workable alternative we have used is the regularized least-squares objective :

$$L(e) = \|r - Ae\|_2 + \gamma^2 \|e\|_2 \qquad (2.8)$$

where the parameter γ can be altered to reflect the degree of smoothing we wish to impose on the solution. We know of several methods that can be used to solve this objective: QR factorisation, truncation of the singular value decomposition (SVD) and variants on the conjugate gradient method.

The direct methods based on QR, or the more computationally expensive, SVD factorizations are popular for small dense, or large sparse structured matrices, because we can write the solution to the least squares problem directly in terms of the factored matrices. If **A** has a regular structure, the columns of **A** can be pre-ordered to also make **R** sparse. For unstructured sparse matrices, the dense rows of **A** will cause **R** to fill-in, and the storage requirements of this direct method will blow out. The same problem affects SVD methods; again it is not possible to preserve the same sparsity pattern for the decomposed arrays as for the original matrix. Also, if **A** is not of full rank, numerical errors will usually cause the SVD algorithm to break down, if the matrix is not diagonally dominant.

When **A** is Hermitian positive definite, the conjugate gradient method is perhaps the most effective iterative solver [20]. The popularity of this method is due in part to its optimality; at each step the **A**-norm of the error is minimised over some sub-space. By minimising in other than the **A**-norm, different conjugate gradient methods result, some of which are applicable to non-positive definite matrices. A review of these is given in [21]. These methods are characterised by their need for only a few vectors of working storage, and their theoretical convergence in at most N iterations (where **A** is square $N \times N$). The methods work well when **A** is well conditioned and has many nearly equal singular values.

This least squares objective in (2.8) can be solved using an analytically equivalent method to the original conjugate gradient method, but with more favourable numerical properties. This so-called *LSQR* algorithm has been published by Paige and Saunders [22, 23]. It is based upon a Lanczos process applied to a symmetric matrix derived from A[4]. The method is iterative, storage-efficient and only requires the user to supply kernels for multiplication by A and A^T.

The use of this unbalanced scheme, with one relaxation method applied to the direct equation, $\mathbf{A\lambda} = \mathbf{n}$, and a different method applied to the residual equation, $\mathbf{Ae} = \mathbf{r}$, does have some side effects:

- Firstly, the convergence rates and smoothing properties of the two relaxation processes will be different, making analysis of the performance of the algorithm difficult;
- Secondly, the rate of convergence and the final reconstructed image will be strongly influenced by the regularization parameter, γ. As $\gamma \to 0$, the data component of the regularized objective dominates and the solution vector, **e**, found on application of *LSQR* will be highly oscillatory. Application of such an solution for **e** in the residual correction will degrade, rather than enhance, the current direct solution $\boldsymbol{\lambda}$. As γ is increased, the solution from *LSQR* is smoothed yielding a less deleterious effect on $\boldsymbol{\lambda}$, however convergence is slower;
- Finally, each iteration of the EM algorithm automatically conserves the image intensity :

  $$\sum_{b=1}^{B} \lambda_b^{(i+1)} = \sum_{b=1}^{B} \lambda_b^{(i)} = \sum_{t=1}^{T} n_t \tag{2.9}$$

  We would like the solution of the residual equation to satisfy :

  $$\sum_{b=1}^{B} e_b^{(i)} = \sum_{t=1}^{T} r_t^{(i)} = 0 \tag{2.10}$$

  So that the residual update to $\boldsymbol{\lambda}$ will not violate (2.9). However, this is not the case with *LSQR*. With additive algorithms such as the FMV cycle, we are faced with the unpleasant choice between enforcing the non-negativity constraint (which imposes a considerable computational burden on the relaxation algorithm, see for instance Bertero and Dovì [24]) or ignoring it, which leads to problematic negative intensities that are difficult to remove from the reconstructed image without leaving artefacts.

To conclude this discussion, we shall state the third algorithm as :

[4] *We can write equation (2.8) equivalently as :*

$$\min \left\| \begin{bmatrix} \mathrm{A} \\ \gamma \mathrm{I} \end{bmatrix} \mathrm{e} - \begin{bmatrix} \mathrm{r} \\ 0 \end{bmatrix} \right\|_2 \tag{F4.1}$$

The solution to equation (F4.1) satisfies the symmetric, but indefinite system :

$$\begin{bmatrix} \mathrm{I} & \mathrm{A} \\ \mathrm{A}^{\mathrm{T}} & -\gamma^2 \mathrm{I} \end{bmatrix} \begin{bmatrix} \rho \\ \mathrm{e} \end{bmatrix} = \begin{bmatrix} \mathrm{r} \\ 0 \end{bmatrix} \tag{F4.2}$$

where ρ *is the residual vector* $\rho = \mathrm{r} - \mathrm{Ae}$. *If we write (F4.2) as :*

$$\mathbf{Bx} = \mathbf{b} \tag{F4.4}$$

then given the symmetric matrix **B** *and starting vector* **b** *the Lanczos process is a method of reducing* **B** *to tridiagonal form.*

```
Step 0:  Initialisation:
             $h = h_0$        (the coarsest grid)
             ${}^{h_0}\boldsymbol{\lambda} = \mathbf{c}$     (initial solution is constant valued)
             Calculate the projection data sets, $\mathbf{N}^h$, for all grid sizes.
Step 1:  do
         {
             ${}^{h+1}\boldsymbol{\lambda} = {}^{h+1}\boldsymbol{\lambda} + \mathbf{I}_h^{h+1}\,{}^{h}\boldsymbol{\lambda}$,
             Perform $\nu_0$ cycles: ${}^{h+1}\boldsymbol{\lambda} \leftarrow \mathbf{MV}^{h+1}\left({}^{h+1}\boldsymbol{\lambda},\ {}^{h+1}\mathbf{n}\right)$,
             $h \leftarrow h+1$
         }
         until $(h = \text{finest grid})$
```

(2.11)

where the *MV*-cycle shown in (2.5) is modified to perform the appropriate relaxation process depending on the current variables.

We chose to do nothing about the negative intensities that *LSQR* will introduce. Instead we renormalised $\boldsymbol{\lambda}$ to have the appropriate sum (as given by 2.9) every time an interpolation or restriction was performed. This leads to negative intensities in the final reconstruction, to be imaged by applying a linear warp to force the range of all pixels to lie in [0,255]. Artefacts from this can be seen in the reconstruction, to be shown later.

3. RESULTS AND DISCUSSION

3.1 The model problem

We propose to test the reconstruction algorithms using simulated data. This has the advantages that we have exact knowledge of the source function to compare to the reconstructions, and can make the problem as complicated as the simulation allows. The choice of a model for $\boldsymbol{\lambda}$ is important. We use a function which is relatively simple to calculate, yet represents the type of object we are trying to image. It is important to avoid using a phantom that is too smooth, or too well-matched to the algorithm, which would make the reconstruction perform too well, thereby biasing our results. We use a phantom that can be represented by a set of ellipses of arbitrary size, orientation and emitter density. In [25] we describe details of the full Monte Carlo simulation of the photon transport equation. Here it is suffice to say that we track a large number of trajectories accounting for the physics of emission, scattering, attenuation and detection to arrive at a representative set of projections.

In the model problem for this paper, the object under test is represented as a body with a low level of emitter activity, that encloses number of higher activity sources. The low level sources mimic background radiation and leakage, whilst the internal sources are the main targets that we wish to image. The model has been designed to test the resolution of the reconstruction and the relative accuracy of the activity of different sources, whilst highlighting any interference effects that may occur between sources, finding any problems with attenuation compensation, and indicating if "ringing disturbances" near the boundary are going to pose a problem.

The projections are gathered at 64 rotational positions enscribing 360° around the object with 64 detectors per position, using the forward modelling program, *python*, described in [25]. Figure 3.1 shows the geometry of the parallel hole collimator, together with one of the projections at a particular projection angle and a Cartesian reconstruction grid of size 32 × 32 superimposed upon the model phantom. The projections have a relatively high level of Poisson noise, 0.5%, that corresponds to a total observed count of 40×10^3 photons, a value that is at the low end of the range of actual measurements in commercial systems.

The spatial distribution of emission sites generated by *python* in the forward simulation is shown in figure 3.3*a* in the form of a 2D histogram. In figure 3.3*b* these primary sites are shown together with the secondary emission sites where scattering takes place. These secondary sites are concentrated around the rim of the phantom because the probability that scattering has occurred increases exponentially with the distance travelled by the photon packet through the scattering media. Figures 3.1 and 3.3*b* provide dramatic evidence of the twin problems that reconstruction algorithms must overcome in emission tomography: the high level of noise that is present in the measured projections, and the indirect paths taken by a large fraction of the photons between their source and their sink.

3.2 Calculation of the transfer matrix

Until now we have not given any details of the calculation of the matrix A. This is described in more detail in [1], and we shall only give a summary here. We define $A_{b,t}$ as the probability that a photon emitted from box *b* is detected by detector *t*. This will depend on a number of factors. Some of these factors we can model, including the spatial correspondence between the detector and the pixel, the detector geometry and the scattering and absorption of photons. Other factors we can safely ignore by assuming an arbitrary scale for our reconstruction, such as the activity and rate of decay of the radio-isotope and the exposure time. A few remaining factors, such as the mobility and rate of elimination of the radio-isotope from the body, are imponderable and it is difficult to see how they could ever be modelled for an inversion. The terms we incorporate into the calculation are dependent on the geometry of the experiment and the mechanics of photon transport. We can incorporate all these items if we calculate the matrix element as follows.

$$A_{b,t} = k \cdot \frac{\mathrm{S}(\Omega_{b,t})}{2\pi} \cdot \exp(-\hat{\mu}\, d_{b,t}) \tag{3.1}$$

where k is the normalising constant to make the cumulative probability of detection over all the detectors unity,

$d_{b,t}$ is the distance from the centre of pixel *b* to the periphery of the object, along the ray path to the centre of detector *t*, and,

$\Omega_{b,t}$ is the angle subtended by detector *t* from the centre of the pixel *b*.

The two free terms in this equation are, $\hat{\mu}$, the uniform linear attenuation and the function $\mathrm{S}(\Omega_{b,t})$ that represents the calculation of acceptance angles using a detector with an aperture that has been artificially widened by a scalar multiplier A_m. These two terms are used to approximate the acceptance of scattered photons that originate outside the horizon cone of the detector. We could ignore this process, but in doing so we exaggerate the ill-posedness of the problem. To perform a full-blown scatter inversion is impractical, nevertheless, we know from simulations of the forward problem that the majority of scattered photons detected undergo only a single scattering event. To find appropriate values for the free parameters (A_m, $\hat{\mu}$) we fit the geometric model for $A_{b,t}$ given by equation (3.1) to the point spread function generated by the forward modelling program. The parameters we chose are those that minimise the RMS error between these two calculated functions at a number of discrete *(b,t)* coordinates.

It is of interest to look at the sparsity structure of the A matrix as this has an impact on the efficiency of our matrix calculations. Until now we have not mentioned the ordering of elements in A. Actually, A is best thought of as a fourth order tensor, with two indices for the x and y coordinates in the reconstruction grid (the b index in $A_{b,t}$) and two more for the detector position and rotation angle (the t index in $A_{b,t}$). It is not obvious how to map the four indices down to two. If we use a straightforward lexicographic mapping :

$$\begin{aligned}
&b = N_x y + x,\\
&t = N_d \theta + d\\
&x, y = \text{coordinates of pixel } b,\\
&N_x = \text{X dimension of reconstruction grid},\\
&\theta, d = \text{coordinates of detector } t,\\
&N_d = \text{number of detectors per projection angle.}
\end{aligned} \qquad (3.2)$$

Then positions and pixels that are physically close are widely separated in the A matrix. This shows up as a complicated structure when we plot the non-zero elements for a 64 × 64 × 64 × 64 matrix, as shown in figure 3.2, which would be difficult to exploit in a structured sparse algorithm. We could look at more appropriate mappings than the lexicographic one of (3.2) to give a pattern for the non-zero matrix elements that conforms to one of the standard sparse models. However, it would appear difficult to find a mapping of sufficient generality for all geometries and grids we wish to employ. An alternative to this is to exploit the block structure that is evident in figure 3.2, that shows that A can be partitioned into 64 × 64 sub-matrices each having a tight banded structure with varying orientation. A block partitioned matrix calculation would then be more efficient than a straightforward full matrix approach, particularly in the Lanczos decomposition employed in *LSQR*.

3.3 Error metrics

We are interested in several aspects of quality of the reconstruction. Obviously the resolution and signal to noise ratio, but also the more subjective aspects are relevant: these include the coherence of the reconstruction, the uniformity of the background and the main emission zones, and the intensity of (any) false sources. So, how can we judge the quality of a reconstruction or compare two reconstructions derived from two different algorithms? Since we generate the projection data via simulation, we know the exact form of the original phantom. A perfect reconstruction of the phantom is shown in figure 3.3*c,d*. Thus it seems natural to take some metric on the space of images and to say that the best algorithm is the one giving the reconstruction closest to the original according to our chosen metric. The most commonly quoted error metrics are the $\mathbf{L}_2$ norm (otherwise known as the normalised root mean square distance) and the $\mathbf{L}_1$ norm (known as the normalised mean absolute error measure). We define these norms as follows :

$$\mathbf{L}_1 \text{ norm} \equiv \frac{1}{N} \frac{\sum_{j=1}^{N} \left|\tilde{\lambda}_j - \lambda_j\right|}{\sum_{j=1}^{N} \left|\tilde{\lambda}_j\right|} \qquad \mathbf{L}_2 \text{ norm} \equiv \frac{1}{N} \sqrt{\frac{\sum_{j=1}^{N} \left(\tilde{\lambda}_j - \lambda_j\right)^2}{\sum_{j=1}^{N} \left(\tilde{\lambda}_j - \bar{\tilde{\lambda}}\right)^2}} \tag{3.3}$$

where $\tilde{\lambda}_j$ is a pixel in the perfect reconstruction, λ_j is the corresponding pixel in the actual reconstruction, $\bar{\tilde{\lambda}}$ is the average pixel value for the perfect reconstruction and N is the number of pixels. Normalisation by the number of pixels is required to compare images on different grids. We calculate the perfect reconstruction, $\tilde{\boldsymbol{\lambda}}$, on a 512 × 512 mesh and derive the coarser grid $\tilde{\boldsymbol{\lambda}}$ by restricting this with a 2 × 2 full averaging operator.

Although computationally simple, these $\mathbf{L}_2$ and $\mathbf{L}_1$ norms do not always convey the correct measure of quality in the reconstruction as would be judged by a human observer because they both implicitly assume that all pixels are equally important irrespective of their information content. In a study comparing the performance of these two measures against the rankings of a trained observer, Yeung and Herman [26] found that the $\mathbf{L}_1$ relative error norm correlated reasonably well with the observer ratings, but that the $\mathbf{L}_2$ distance norm was less well correlated. This was attributed to the stronger effect that outliers have on the $\mathbf{L}_2$ norm, that would be ignored by the observer.

It is well known that a trained observer will deliberately ignore extraneous artefacts to concentrate on the information-bearing regions of an image. It is very difficult to model this type of cognitive response. In place of a better error metric that incorporates this functionality, we must use some subjective judgement. Consequently, for this work we shall give three assessments: the mechanistic, though possibly misleading, $\mathbf{L}_1$ and $\mathbf{L}_2$ norms, alongside my subjective opinion.

3.4 Results from the model problem

There are some practical limitations on the range of grid sizes we use in these multigrid reconstructions. The smallest grid should not be too small to annihilate all global features from the image, which sets a lower limit of about 8 × 8. We would like to use a very fine mesh on the maximum grid size to obtain high resolution. However, the finer we make the mesh, the fewer the number of emissions per pixel and consequently the noisier the reconstruction. Vardi *et al.* [7] conjecture that there is a maximal grid resolution for the maximum likelihood objective. This is achieved with the largest B that gives a unique solution, whilst increasing B beyond this limit will not yield any additional information, indeed the reconstruction will become noisier as the problem becomes more ill-conditioned. However, we cannot determine the limit analytically, although we do know that it is bounded by T. Nonetheless, we can refine the grid size until no further improvement is shown because we are employing a multigrid strategy. In any case, we set a practical limit of 512 × 512 on the maximum mesh in order to have reasonable computation times on the Silicon Graphics 4D/240GTX, 64Mbyte machine.

The algorithms described have been coded in the programs *em_nest*, *em_nest2*, and *em_mg*. Results for the model problem are given in figures 3.4 - 3.7. The first three figures show the reconstruction attained on all the grids (8 × 8 through to 512 × 512), whilst the fourth figure shows the final reconstruction on a 512 × 512 mesh for the three multigrid algorithms as well as the uni-grid EM algorithm. Error results are shown in table 3.1. The starting solution we used in each instance is a disk of uniform activity.

Parameter	uni-grid EM	nested-EM (Algorithm 1)	nested-EM (Algorithm 2)	multigrid-EM (Algorithm 3)
L_1 error norm	1.83×10^{-6}	2.08×10^{-6}	2.01×10^{-6}	2.33×10^{-6}
L_2 error norm	2.85×10^{-6}	2.89×10^{-6}	2.89×10^{-6}	2.92×10^{-6}
Algorithm parameters	50 iterations	$v_1 = 10$	$v_1 = 10$	$v_0 = 1$, $v_1 = v_2 = 5$, $\gamma = 1.0$
Computation time	32.6 hours	4.6 hours	6.4 hours	17.8 hours

Table 3.1 *Error norms for the EM reconstructions of the model problem*

Figure 3.8 shows the progression of both error norms at each instance that λ was updated by the algorithms. Also shown is the calculation time for a conventional deconvolved, filtered-backprojection reconstruction, as was discussed in [1]. Curiously, this figure also shows that the L_1 norm has several local minima, although we would judge the reconstruction at these points to be inferior to the final image. Clearly, the error norms are not very good indicators of the quality of the reconstructions. We would be very wary of any conclusions based solely on such metrics.

3.5 Discussion

We consider the reconstructions from Algorithm I to be superior on the very coarse grids having better resolution and contrast to those achieved using Algorithms II and III. This would indicate that it is better to find a maximum likelihood solution to our set of strongly over-determined equations (on a 8 × 8 grid we have 64 unknowns and 4096 data points) than it is to find the unique solution to a heavily restricted, smoothed and consistent set of equations. It is of interest to note that the use of blurring kernels for the EM algorithm has been advocated for the benefits of increased convergence rate and fewer artefacts in the reconstructions [27, 28]. There is however, a compromise between the degree of smoothing and the resolution that can be obtained in the final reconstruction. The filtering we have used for restricting the projection data is perhaps excessive, and an alternative with a narrower effective support may improve the reconstructions obtained with Algorithm II on the coarse grids[5].

[5] *The effective width of the filtering kernel in restricting the* 64 × 64 *projection set down to an* 8 × 8 *set, using nested* 2 × 2 *averaging is* 8 × 8.

Algorithms I and II give very similar reconstructions on the finer meshes. Indeed, the two results are virtually indistinguishable in a profile through the centre region of the reconstructions on the 512 × 512 grid, as shown in figure 3.9. On the other hand, the result for Algorithm III appears to be considerably different. The line profile reveals the cause; the linear warp that was applied to the reconstruction to image the negative intensities has shifted those pixels with values close to zero to more positive values. This increase in background intensity and concomitant reduction in contrast is responsible for much of the visible differences between the three reconstructions.

We could devise alternative imaging schemes for Algorithm III to reduce the impact of the negative intensities, however, our main concerns are the influence of the ridge parameter, γ, and the sluggish performance of the method. If γ is decreased to 0.1, the reconstructions are swamped by high frequency oscillations, whilst if γ is increased to 10.0, the residual correction has an unobservable effect on the reconstruction (while taking an even longer time to compute). The cause of this is the very slow convergence of *LSQR* in finding a regularized solution for the residual equation, which dominates the computation time compared with the EM relaxations. In a *pers. comm.* with one of the authors (Michael Saunders, mike@sol-michael.stanford.edu), he says:

> "...(*LSQR*) may prove to be slow unless your problem is well-conditioned or has clustered singular values. You might be able to speed things up if there is obvious structure that you can use to cook up a preconditioner, *e.g.* if A has a significant block-diagonal part".

Unfortunately, I have not been able to deduce an appropriate preconditioner from the rather randomly structured and singular form of A. The singular value spectrum of A on several small grids is shown in figure 3.10. These were calculated using a dense matrix routine[6] up to a maximum size of 1024 × 1024 that is set by the memory limitations. The indications from this figure are that the singular values are not tightly clustered, though the clustering and the range both appear to be increasing with the matrix size.

All the reconstructions show edge artefacts that surround discontinuities in the image. This can be explained with reference to the singular value spectrum of A. We have hypothesized that as the size of A increases, the range of the singular values will increase, and a significant number will be either zero, or below the "noise floor" caused by finite machine precision. Frequency components in the image beyond this point in the spectrum can *never* be recovered, regardless of the algorithm used. The truncation of the frequency spectrum will lead to a Gibbs-like phenomena around edge discontinuities. We also have a higher density of scattering sites around the periphery (as evidenced by the scatter histogram in figure 3.3*b*) and the effect of these will be magnified by ill-conditioning of A. We can alleviate the problem by windowing the projections prior to reconstruction, but a superior way is to reduce the condition number of the matrix.

[6] *KDSVDC from the* CLASSPACK *library. This is equivalent to DSVDC from* LINPACK. CLASSPACK *is a copyrighted product of Kuck & Associates.*

4. CONCLUSIONS

We have shown that nested iteration and multigrid approaches are much more computationally efficient than a uni-grid approach to the EM algorithm. For the particular problem discussed here, Algorithm I appears superior, being both simple to implement and giving the greatest speed-up. Algorithm II confers no real advantage over Algorithm I since the computational savings gained by restricting the data set has minimal consequence on the overall timings that are dominated by the fine grid computation rates. For problems where the conditioning and noise are not as severe as this model case, Algorithm III could be beneficial.

We conclude from these experiments that the system is too poorly conditioned for the full multigrid approach of Algorithm III to be truly effective. There are several approaches that can be taken to improve the stability of the inversion. We could re-formulate the geometry to make A better behaved, for instance by using a circularly symmetric grid, although such symmetry is not present in the original geometry. Smoothing the projection data could be improved. Finally, and perhaps most importantly, the *ad hoc* use of regularized least squares to solve to residual equations should be replaced with a maximum likelihood technique. The difficulties associated with this have been highlighted previously.

Figures 3.4 - 3.7 highlight the dangers of attempting to reconstruct super-resolution images. The best reconstruction for practical purposes is probably that on the 64 × 64 or the 128 × 128 grid. When we force the reconstruction on meshes that are inconsistent with the sampling and noise in the original data, then the inevitable consequences are spurious artefacts in the image. An advantage of all the multigrid algorithms considered here over a uni-grid method is that the optimum grid level is quickly evident.

One avenue for further work with considerable promise is the use of locally-adaptive mesh refinement. It is clear from figure 3.4, that the important features are present on the coarsest meshes, yet significant effort is wasted because we use global refinement. Thus the same computational effort is expended on uniform areas as on those regions with a higher level of local activity. We are currently investigating methods of incorporating into Algorithm I an adaptive mesh refinement based on an image activity measure.

5. ACKNOWLEDGEMENTS

The work described in this paper was supported and funded in part through National Research Fellowship (reference 87/1114) awarded to the author by the Australian Department of Employment, Education and Training.

I would like to thank my colleagues in the BHP Research Laboratories for useful discussions whilst I was working on these problems, and in particular, Angela Bowles for careful reading of the various drafts of this paper.

6. REFERENCES

[1] Monks, T. J. (1991) "Solving the Inverse Problem for Emission Tomography", *Report # BHPRML/KHMJM8/91/003*, BHP Research - Melbourne Laboratories.

[2] Shepp, L. A. & Vardi, Y. (1982) "Maximum Likelihood Reconstruction for Emission Tomography", *IEEE Trans.* **MI-1**, 2, pp. 113-122.

[3] Lange, K., Bahn, M. & Little, R. (1987) "A Theoretical Study of Some Maximum Likelihood Algorithms for Emission and Transmission Tomography", *IEEE Trans.* **MI-6**, 2, pp. 106-114.

[4] Devaney, A. J. & Tsihrintzis, G. A. (1991) "Maximum Likelihood Estimation of Object Location in Diffraction Tomography", *IEEE Trans.* **ASSP-39**, 3, pp. 672-682.

[5] Medoff, B. P., *et al.* (1983) "Iterative Convolution Backprojection Algorithms for Image Reconstruction from Limited Data", *J. Opt. Soc. Amer.* **73**, 11, pp. 1493-1500.

[6] Ollinger, J. M. (1990) "Iterative Reconstruction-Reprojection and the Expectation-Maximization Algorithm", *IEEE Trans.* **MI-9**, 1, pp. 94-98.

[7] Vardi, Y., Shepp, L. A. & Kaufman, L. (1985) "A Statistical Model for Postitron Emission Tomography (with comments)", *J. Amer. Statist. Soc.* **80**, 389, pp. 8-37.

[8] Kuo, C. & Levy, B. C. (1990) "Discretization and Solution of Elliptic PDEs — A Digital Signal Processing Approach", *Proc. IEEE* **78**, 12, pp. 1808-1842.

[9] Briggs, W. L. (1987) **A Multigrid Tutorial** SIAM, Philadelphia.

[10] Mallat, S. G. (1989) "Multifrequency Channel Decompositions of Images and Wavelet Models", *IEEE Trans.* **ASSP-37**, 12, pp. 2091-2110.

[11] Burt, P. J. (1981) "Fast Filter Transforms for Image Processing", *Comp. Graphics Image Proc.* **16**, pp. 20-51.

[12] — (1984) "The Pyramid as a Structure for Efficient Computation." pp. 6-35 in **Multiresolution Image Processing and Analysis** Ed. Rosenfeld, A. Springer-Verlag, Berlin. *This is an updated version of a paper that was presented at the Workshop on Multiresolution Image Processing and Analysis, Leesberg, VA. July 19-21 (1982).*

[13] Terzopoulos, D. (1984) "Multilevel Reconstruction of Visual Surfaces: Variational Principles and Finite-Element Representations." pp. 237-310 in **Multiresolution Image Processing and Analysis** Ed. Rosenfeld, A. Springer-Verlag, Berlin. *This is an updated version of a paper that was presented at the Workshop on Multiresolution Image Processing and Analysis, Leesberg, VA. July 19-21 (1982).*

[14] — (1986) "Image Analysis using Multigrid Relaxation Methods", *IEEE Trans.* **PAMI-8**, 2, pp. 129-139.

[15] Herman, G. T., *et al.* (1984) "Multilevel Image Reconstruction." pp. pp.121-135 in **Multiresolution Image Processing and Analysis** Ed. Rosenfeld, A. Springer-Verlag, Berlin. *This is an updated version of a paper that was presented at the Workshop on Multiresolution Image Processing and Analysis, Leesberg, VA. July 19-21 (1982).*

[16] Kaufman, L. (1987) "Implementing and Accelerating the EM algorithm for Positron Emission Tomography", *IEEE Trans.* **MI-6**, 1, pp. 37-51.

[17] Ranganath, M. V., Dhawan, A. P. & Mullani, N. (1988) "A Multigrid Expectation Maximisation Reconstruction Algorithm for Positron Emission Tomography", *IEEE Trans.* **MI-7**, 4, pp. 273-278.

[18] Southwell, R. V. (1946) **Relaxation Methods in Theoretical Physics** Oxford University Press, Oxford.

[19] Watson, A. B. (1987) "The Cortex Transform: Rapid Computation of Simulated Neural Images", *Computer Vision, Graphics, and Image Processing* **39**, pp. 311-327.

[20] Hestenes, M. R. & Stiefel, E. (1952) "Methods of Conjugate Gradients for Solving Linear Systems", *J. Res. Nat. Bur. Standards* **49**, pp. 409-435.

[21] Ashby, S. F., Manteuffel, T. A. & Saylor, P. E. (1990) "A Taxonomy for Conjugate Gradient Methods", *SIAM J. Numer. Anal.* **27**, 6, pp. 1542-1568.

[22] Paige, C. C. & Saunders, M. A. (1982) "LSQR: An Algorithm for Sparse Linear Equations and Sparse Least-Squares", *ACM Trans. Math. Software* **8**, 1, pp. 43-71.

[23] — (1982) "ALGORITHM 583 LSQR: Sparse Linear Equations and Least Squares Problems", *ACM Trans. Math. Software* **8**, 2, pp. 195-209.

[24] Bertero, M. & Dovì, V. (1981) "Regularized and Positive-constrained Inverse Methods in the Problem of Object Restoration", *Optica Acta* **28**, 12, pp. 1635-1649.

[25] Monks, T. J. (1991) "Simulation of the Forward Problem for Emission Tomography", *Report # BHPRML/KHMJM8/91/002*, BHP Research - Melbourne Laboratories.

[26] Yeung, K. T. D. & Herman, G. T. (1989) "Objective Measures to Evaluate the Performance of Reconstruction Algorithms" presented during the SPIE conference on Medical Imaging, 31 Jan. - 3 Feb. 1989, held at Newport Beach, CA. Published in *SPIE Proceedings Series: Medical Imaging III: Image Processing 1092,* , pp. 326-335. ed. Schneider, R. H., *et al.* SPIE Press, Bellingham, Washington.

[27] Snyder, D. L., *et al.* (1987) "Noise and Edge Artefacts in Maximum Likelihood Reconstructions for Emission Tomography", *IEEE Trans.* **MI-6**, 3, pp. 228-238.

[28] Silverman, B. W., *et al.* (1990) "A Smoothed EM Approach to a Class of Problems in Image Analysis and Integral Equations", *J. Roy. Statist. Soc.* **B**, *preprint.*

Image Processing & Data Analysis Group,
BHP Research - Melbourne Laboratories,
245 Wellington Road,
Mulgrave, Victoria, 3170, Australia
(tim@resmel.bhp.com.au)

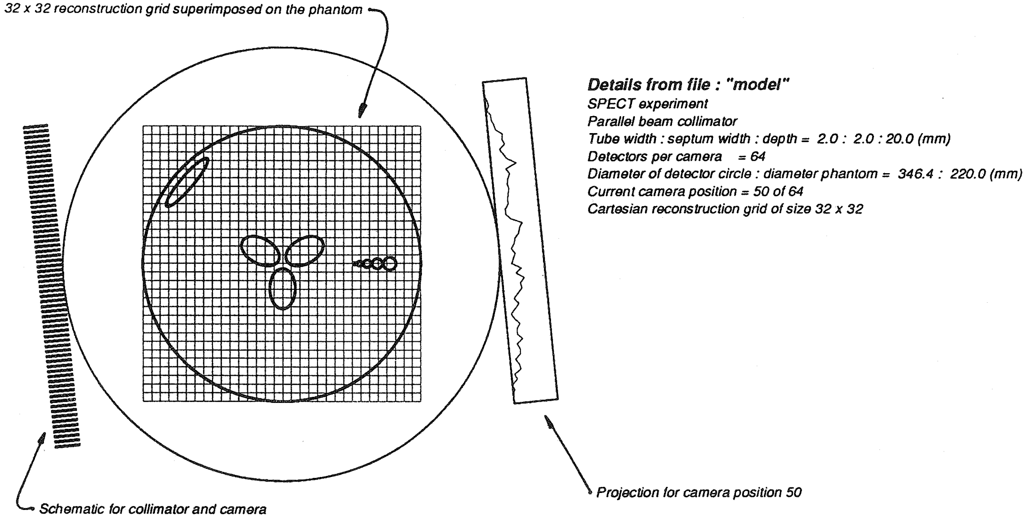

32 x 32 reconstruction grid superimposed on the phantom
Details from file : "model"
SPECT experiment
Parallel beam collimator
Tube width : septum width : depth = 2.0 : 2.0 : 20.0 (mm)
Detectors per camera = 64
Diameter of detector circle : diameter phantom = 346.4 : 220.0 (mm)
Current camera position = 50 of 64
Cartesian reconstruction grid of size 32 x 32
Projection for camera position 50
Schematic for collimator and camera

Figure 3.1 : *Geometry for the model problem, showing collimator,* 32 × 32 *reconstruction grid and a projection.*

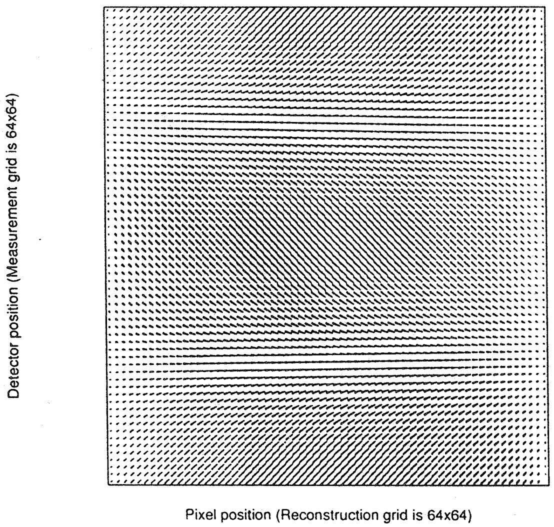


Figure 3.2 : *Sparsity structure (non-zero elements only) for a 4096 × 4096 A matrix using lexicographic ordering.*

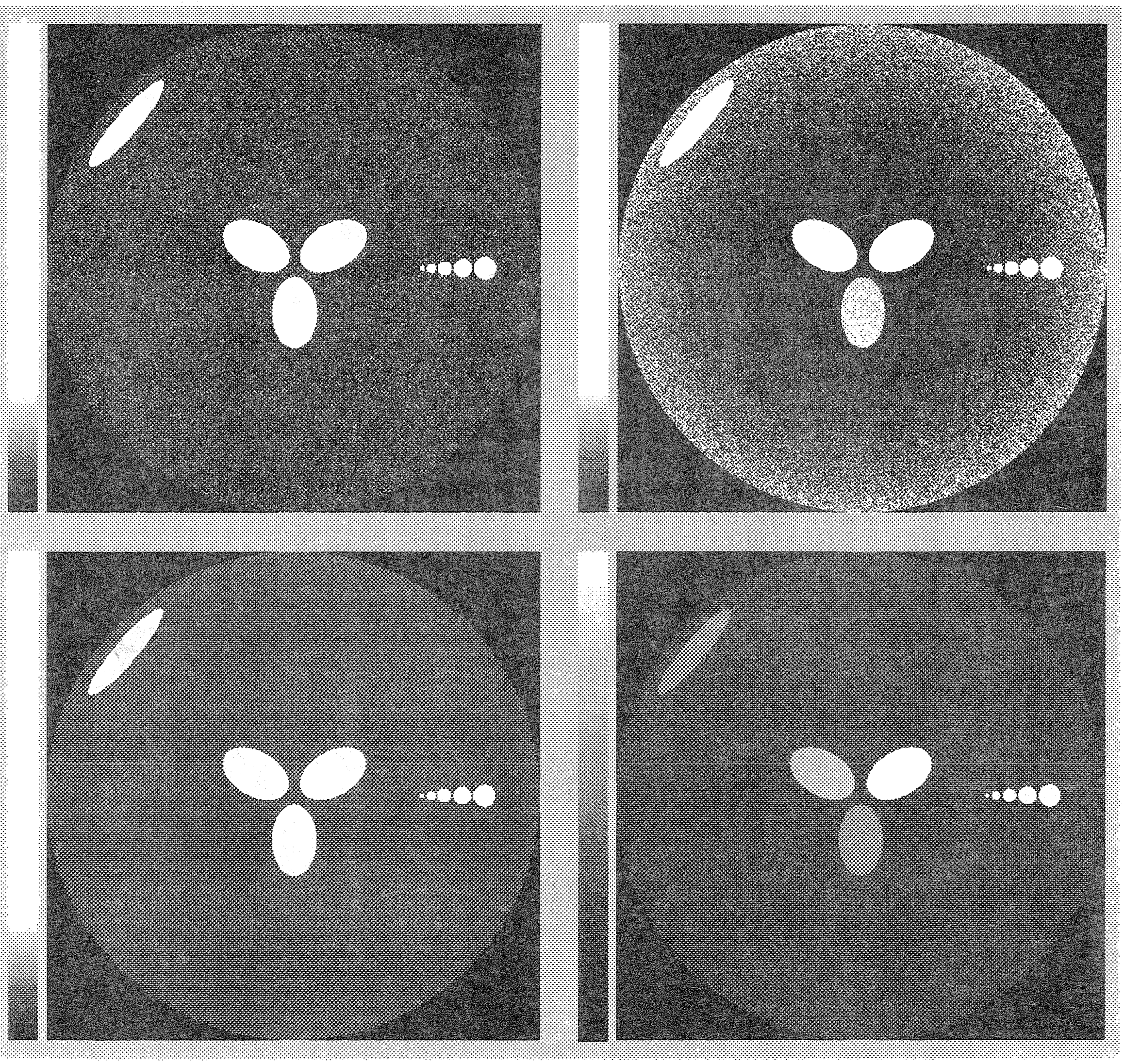


Figure 3.3 *a* : *Grey level histogram of the emission sites generated during simulation of the forward problem.*

b : *Grey level histogram of the emission sites and the secondary scattering sites. Note the concentration of scattering sites towards the periphery of the object.*

c : *Perfect reconstruction of the model problem - same colour map as figures a & b.*

d : *Perfect reconstruction of the model problem - stretched colour map, with background highlighted*

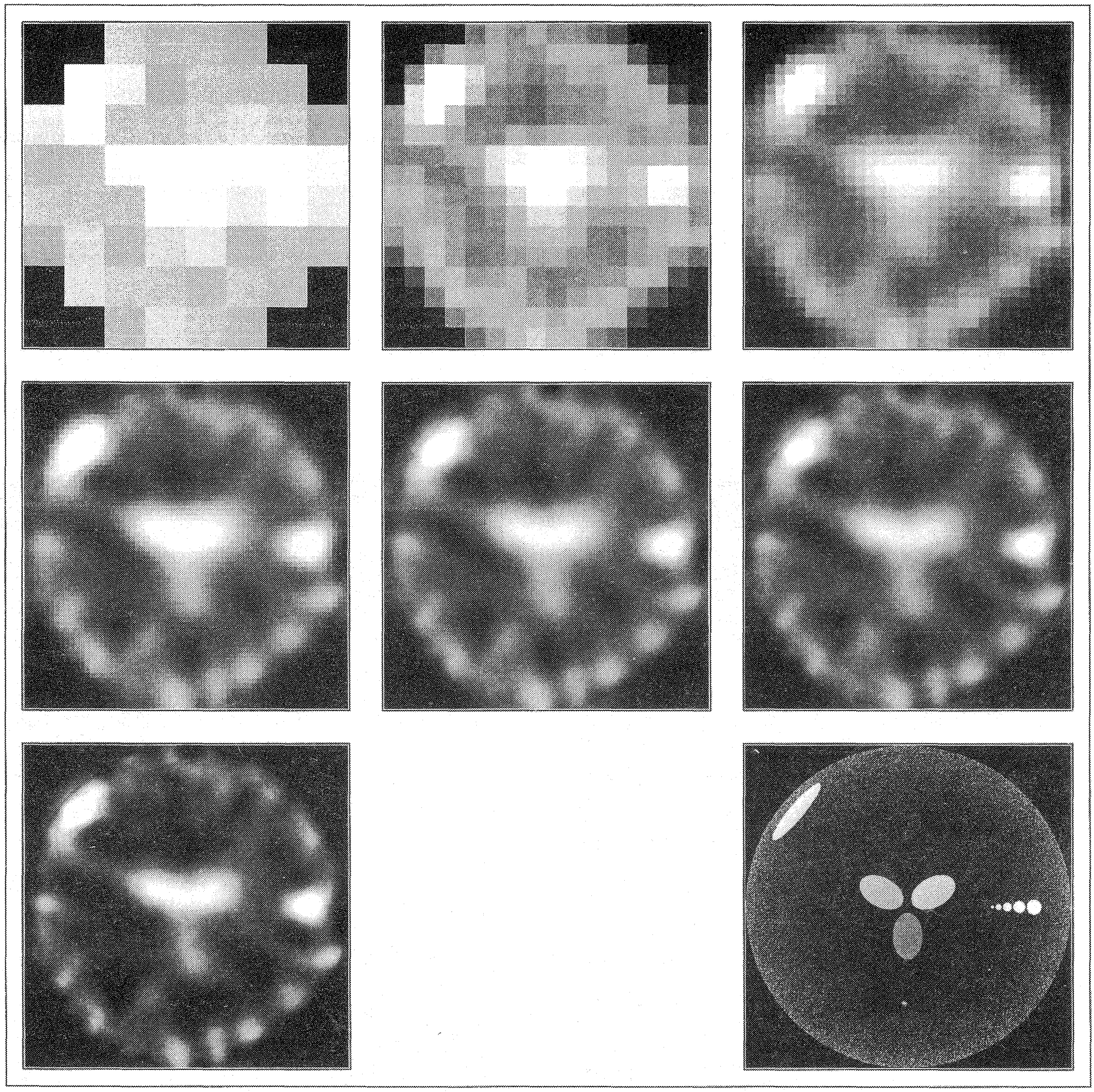


Figure 3.4 *Reconstruction of the model problem using Algorithm I and showing results on grids* 8 × 8 *through to* 512 × 512. *The grey level histogram of the emission scattering sites is also shown.*

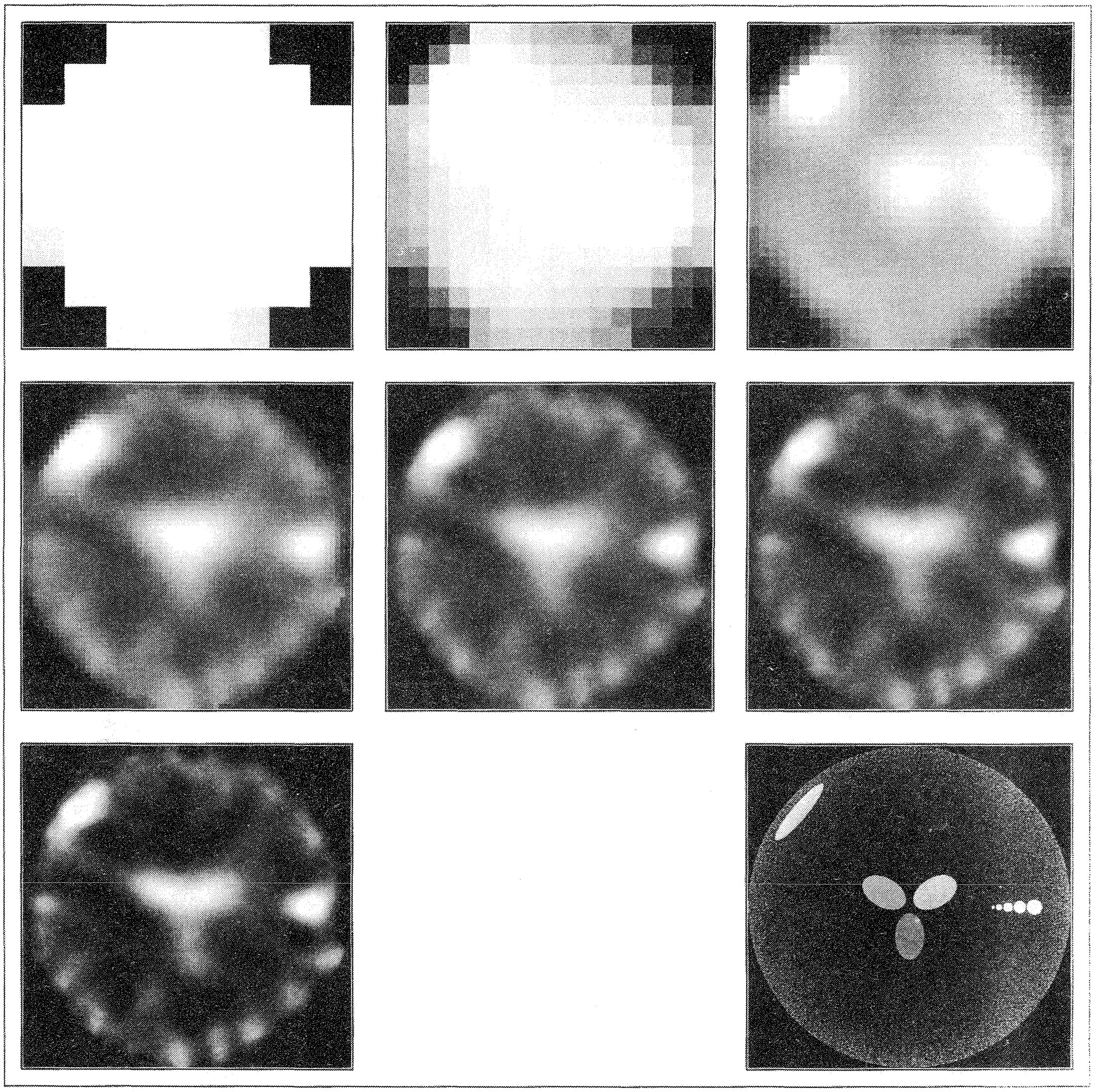


Figure 3.5 *Reconstruction of the model problem using Algorithm II and showing results on grids* 8 × 8 *through to* 512 × 512. *The grey level histogram of the emission scattering sites is also shown.*

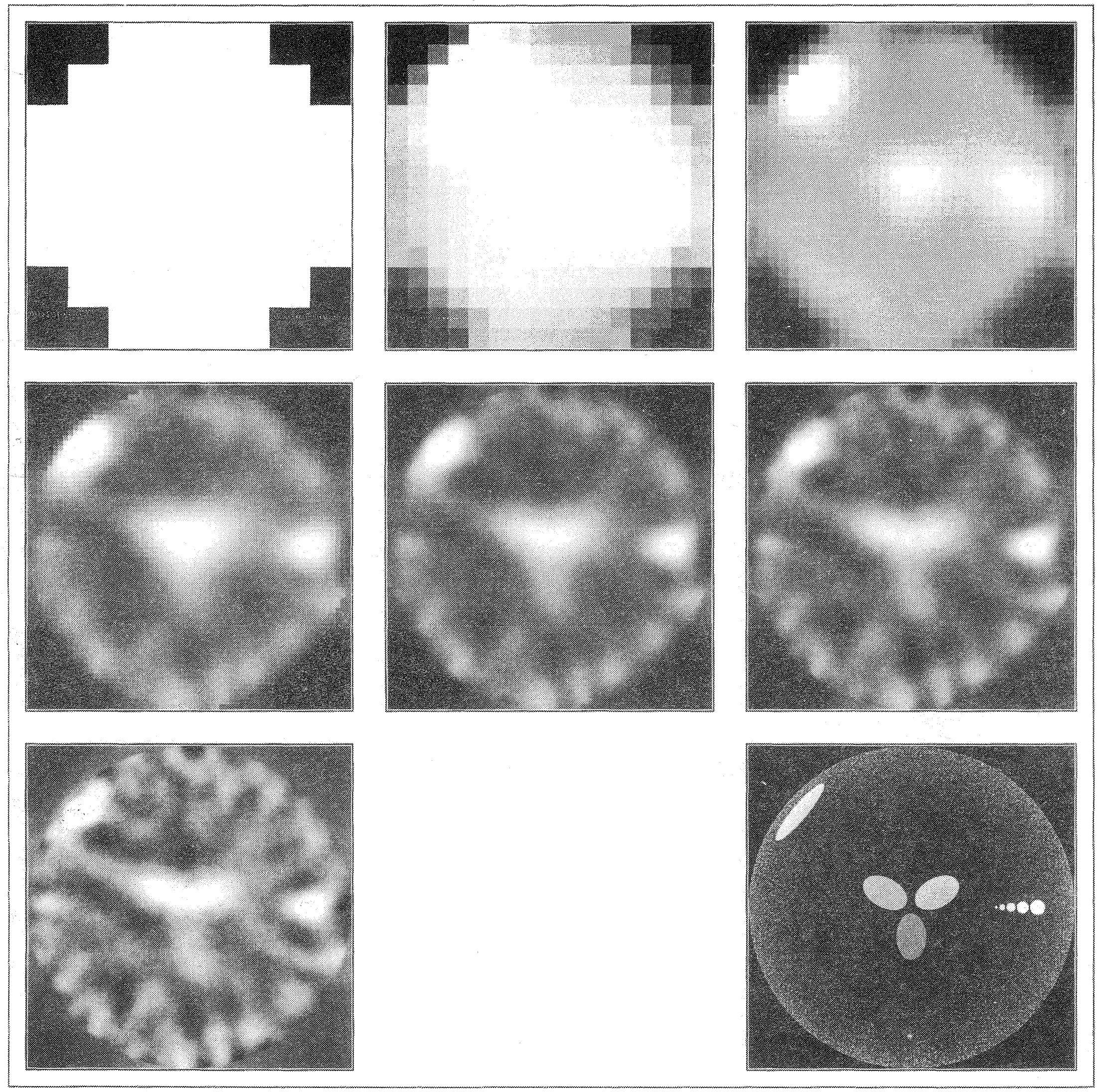


Figure 3.6 *Reconstruction of the model problem using Algorithm III and showing results on grids* 8 × 8 *through to* 512 × 512. *The grey level histogram of the emission scattering sites is also shown.*

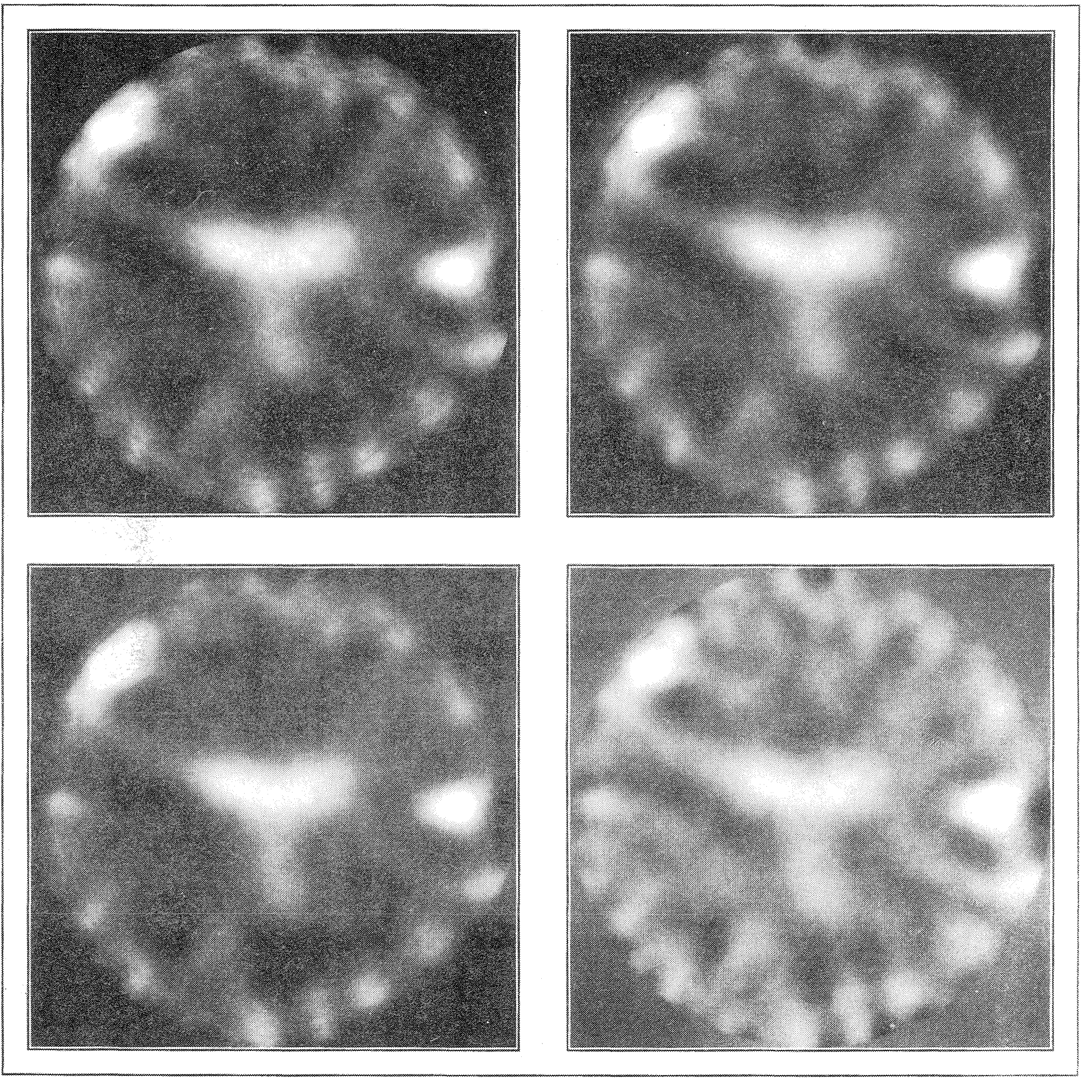
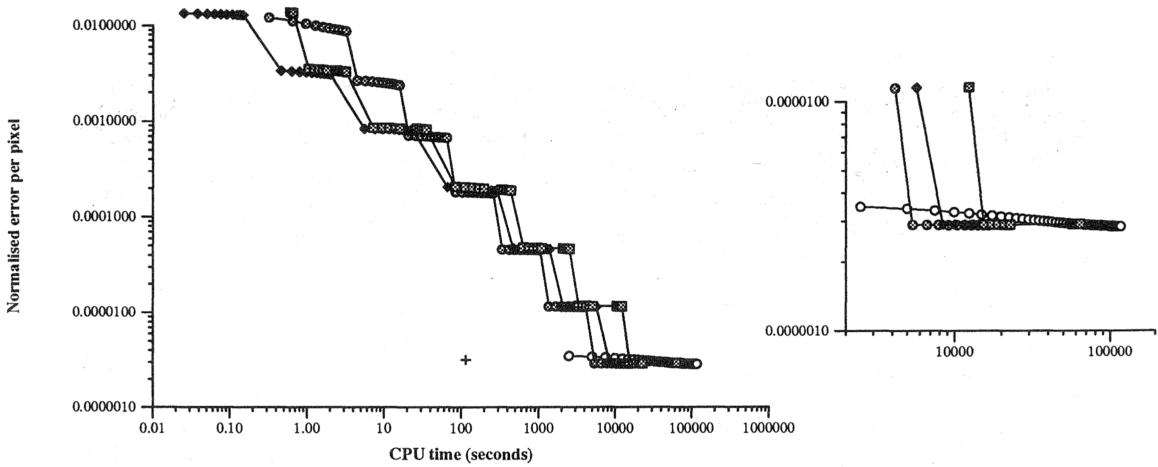
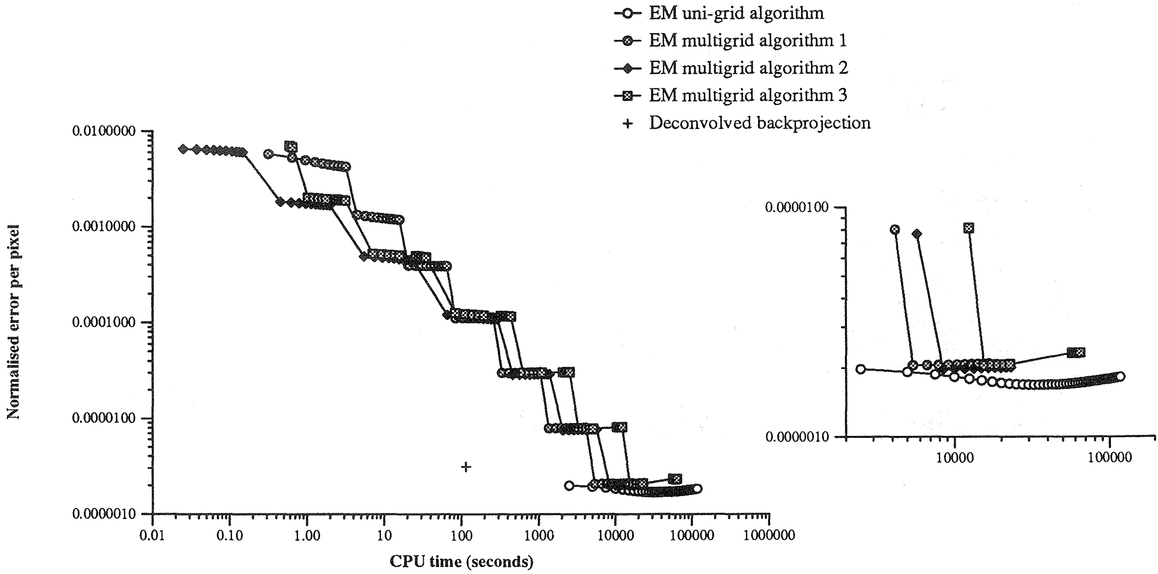


Figure 3.7 *Reconstructions on a 512 × 512 grid for the uni-grid EM algorithm (top-left), nested iteration EM, Algorithm I (top-right), Algorithm II (bottom-left) and multigrid EM (bottom-right).*



L2 error metric for various algorithms



L1 error metric for various algorithms

Figure 3.8 *Time series of the L1 and L2 error norms for all the algorithms. Small graphs to the right of the main plots show the final iterations on a magnified scale.*

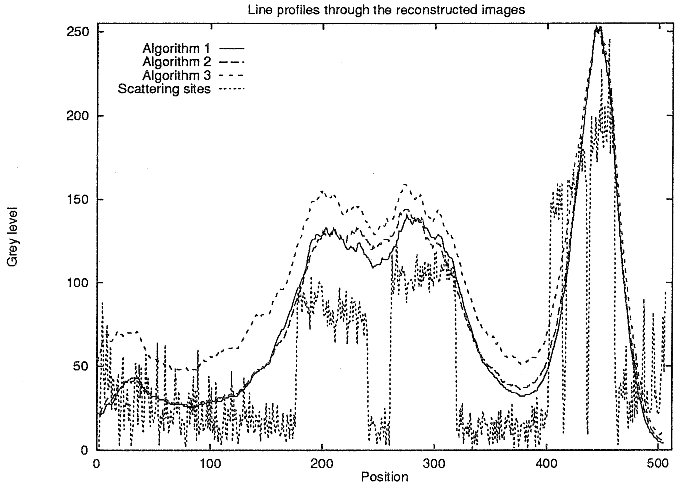


Figure 3.9 *Profile (approximately through the horizontal centre line) for the reconstructions in fig. 3.7 (algorithms I, II & III) as well as the scatter histogram.*

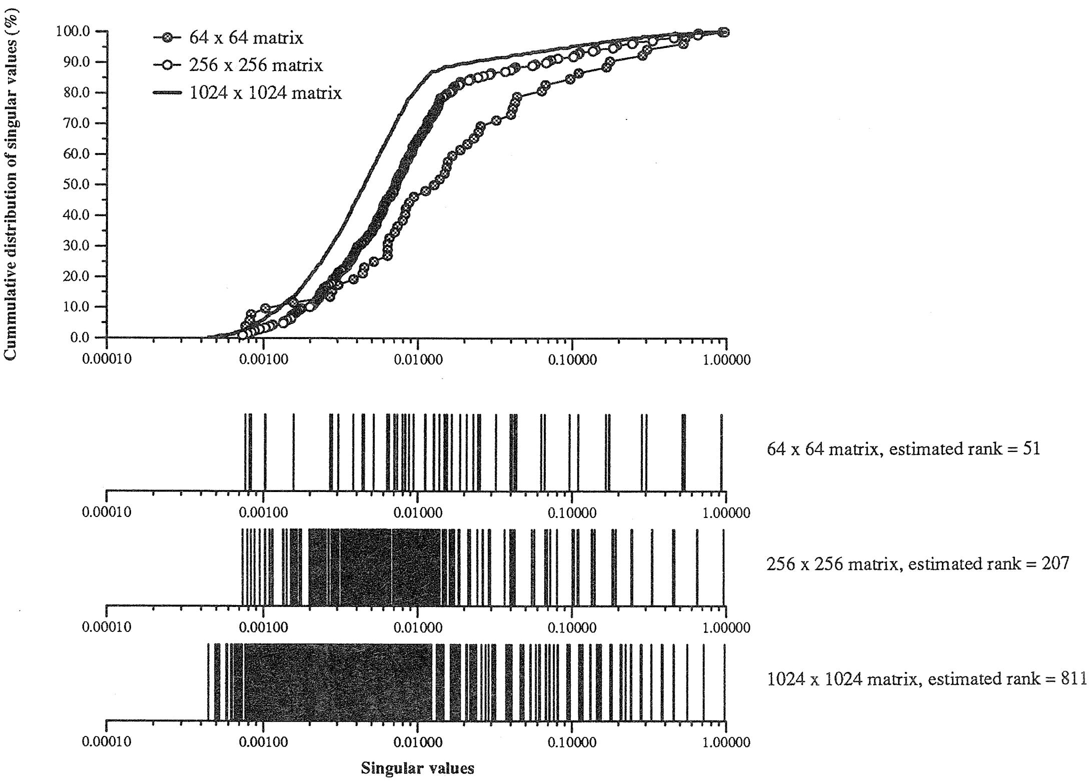


Figure 3.10 *Singular value spectrum and distribution for several sizes of the* A *matrices.*