SIPI Report #130

A Generalized EM Algorithm for 3-D Bayesian Reconstruction from Poisson Data using Gibbs Priors

Tom Hebert and Richard Leahy

Signal and Image Processing Institute
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089

This work was supported by the National Science Foundation
under Grant No. M1P-8708708

submitted to *IEEE Trans. Med. Imaging* - June 20, 1988

A Generalized EM Algorithm for 3-D Bayesian Reconstruction from Poisson Data using Gibbs Priors

Tom Hebert and Richard Leahy

Signal and Image Processing Institute
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089

This work was supported by the National Science Foundation
under Grant No. M1P-8708708

key words: *Bayesian reconstruction, emission tomography, EM algorithm, Markov random field*

Abstract

For independent Poisson observations having a complete/incomplete data representation, a generalized expectation-maximization (GEM) algorithm is developed for Bayesian reconstruction based upon locally correlated Markov random field priors in the form of Gibbs functions. For the M-step of the algorithm, a form of coordinate gradient ascent with an initial step-size resembling that of the EM likelihood algorithm is employed. Implementation closely follows that of the EM likelihood algorithm In addition, as the prior tends towards a uniform distribution, this algorithm reduces to the EM likelihood algorithm. Three different Gibbs function priors are examined. The generalized EM Bayesian approach is applied to estimating the 3-D image parameters in the Poisson model of single photon emission computer tomography (SPECT).

1. Introduction

The maximum likelihood (ML) approach to estimation is equivalent to a maximum a posteriori (MAP) approach in which the prior is assumed uniform over the feasible parameter space. While a ML approach to emission tomography has received considerable interest since the application of the expectation-maximization (EM) formulation by Shepp and Vardi [16], in practice two drawbacks are encountered: 1) due to the ill-conditioned nature of the reconstruction problem, reconstructions tend to take on an increasingly, non-smooth quality as the ML solution is approached; 2) convergence can be increasingly slow as the ML solution is approached. These drawbacks are often compensated for by initializing the algorithm with a smooth estimate and terminating the algorithm before convergence [7],[18]. Alternatively, a regularization approach to the problem has been taken by Snyder and Miller in [17] and Miller, Snyder, and Moore in [12]. A Bayesian approach based on a particular Gibbs prior, which we examine later, was taken by Geman and McClure. In [5] they impose the constraint that the unknown image pixels take values from a known interval and use gradient ascent to arrive at 2-D MAP estimates. In [1], Besag reported that his ICM algorithm had been applied in a preliminary way to gamma camera scans, but no details were provided. The ICM algorithm [1] is equivalent to coordinate descent [11] of the negative log of the posterior function. This approach updates a single pixel at a time by maximizing a univariate function which is conditioned on all data values to which that pixel contributes, all pixels which contribute

to those data values, and all neighbors of the pixel to be updated. In emission tomography, each source pixel contributes to many data values so that this algorithm can be more computationally demanding than the approach presented in this paper. More recently, several authors have investigated Gaussian and Poisson priors applying the EM approach to generate Bayesian reconstruction algorithms. Liang and Hart examined uncorrelated Gaussian and Poisson priors in [6] and in section 2 of [10] and correlated Gaussian priors in section 3 of [10]. The algorithms developed in those sections initially follow an EM derivation. However, in order to develop a closed form M-step, they replace a set of parameters in the prior distribution with a set of uncomputed image pixels values. Since this does not result in a true EM algorithm, as noted in [9], the question of whether the posterior distribution is increased at each step and the question of convergence for all possible data sets remain open. Their algorithm was shown to out-perform the EM likelihood algorithm in a 1-dimensional simulation in [10] and later in 2-D simulations in [6]. Levitan and Herman [9] derive a valid EM algorithm for an uncorrelated Gaussian prior and demonstrate a marked improvement over ML estimates in 2-D simulations. For the mean of the prior distribution Levitan and Herman used a smoothed filtered-backprojection reconstruction. Since it is generally accepted that the structure of images is one of non-stationary mean and local correlations [14], their prior satisfies the first of these two important image attributes. Markov random fields [14], described by Gibbs distributions, capture the property of local correlation and do not require specification of the mean.

For independent Poisson data with a complete/incomplete data representation, we develop a generalized EM algorithm for Bayesian reconstruction based upon locally correlated Markov random field priors in the form of Gibbs functions. For the M-step of the algorithm, a form of coordinate gradient ascent with an initial step-size resembling the EM likelihood algorithm is employed. Implementation of this algorithm closely follows that of the EM likelihood algorithm. In addition, as the prior tends towards a uniform distribution, this algorithm reduces to the EM likelihood algorithm. The reconstructed image pixels are not constrained to any interval as in [5] and no approximations as in [6],[10] are used. Three different Gibbs function priors are examined. We apply this generalized EM Bayesian algorithm to estimate the 3-D image parameters in the Poisson model of single photon emission data.

A Note on Notation: In the following sections we represent vectors with emboldened lowercase characters and matrices with emboldened uppercase characters. Subscripts are

used to indicate a particular element of the vector or matrix. Vector superscripts are used to indicate the particular iteration at which the vector has been computed.

2. The Generalized EM Approach to Bayesian Reconstruction

The probability distribution of the image vector λ conditioned on the data vector $\mathbf{y}$ is formulated using Bayes rule.

$$f(\lambda \,|\mathbf{y}) = \frac{f(\mathbf{y} \,|\lambda)\, f(\lambda)}{f(\mathbf{y})}$$

The ML approach is to treat $f(\mathbf{y})$ as a constant for a particular data vector $\mathbf{y}$, to treat the prior distribution $f(\lambda)$ as uniform over the acceptable parameter range, and to define the likelihood function as any monotonic function of $f(\mathbf{y} \,|\lambda)$, e.g. $\log f(\mathbf{y} \,|\lambda)$. For Bayesian estimation, again $f(\mathbf{y})$ is a constant for a given data vector, but some a priori information is incorporated by specification of the prior $f(\lambda)$. Bayesian reconstruction thus requires solution of the problem

$$\max_{\lambda} \; B(\lambda \,|\mathbf{y}) = \log f(\mathbf{y} \,|\lambda) + \log f(\lambda) \tag{1}$$

Bayesian estimation offers a potential for reconstruction improvement when prior distributions, more meaningful than the uniform distribution, can be defined. In image reconstruction, priors are sought which lend increased probability to realizations which feature segmented, slowly changing regions and decreased probability to highly erratic images. Such image models should incorporate local interactions while allowing abrupt changes across edges or region boundaries. Gibbs functions, which will be used later in this paper, have been demonstrated to be useful in this application [1],[4].

The EM algorithm as presented by Dempster, Laird, and Rubin [3] is a general approach to iterative optimization of likelihood or Bayesian functions when the data can be formulated in a complete/incomplete framework. A complete/incomplete data formulation is applicable when data is missing or when the problem has a more natural formulation in terms of a set of unobserved data. At each iteration, the EM approach requires two steps: an expectation step (E-step) followed by a maximization step (M-step). Often, these two steps can be combined into one.

Let $\mathbf{x}$ be the vector of complete but unobserved data and $\mathbf{y}$ the vector of incomplete, but observed, data. In order to apply the EM approach, the relationship between the complete data and the incomplete data must be a many-to-one mapping. Each realization of $\mathbf{x}$ must correspond to only one possible realization of $\mathbf{y}$, while, many different

realizations of $\mathbf{x}$ may correspond to the same $\mathbf{y}$. That is, given a realization $\hat{\mathbf{x}}$, only one particular realization $\hat{\mathbf{y}}$ has non-zero probability of having occurred. Given a realization $\hat{\mathbf{y}}$, there is a feasible set $\{\mathbf{x}\}_{\hat{\mathbf{y}}}$ with non-zero probability of having occurred.

It then follows, that due to this many $\mathbf{x}$ to one $\mathbf{y}$ mapping

$$f(\mathbf{x}\,|\mathbf{y}\lambda) = \frac{f(\mathbf{xy}\,|\lambda)}{f(\mathbf{y}\,|\lambda)} = \frac{f(\mathbf{x}\,|\lambda)\, I_{\mathbf{y}}(\mathbf{x})}{f(\mathbf{y}\,|\lambda)} \tag{2}$$

where $I_{\mathbf{y}}(\mathbf{x})$ is the indicator function which is equal to 1 if $\mathbf{x}$ results in $\mathbf{y}$ and equal to 0 otherwise. In addition, for any λ^k

$$E_{\mathbf{x}}\{\log f(\mathbf{y}\,|\lambda)\,\big|\mathbf{y}\lambda^k\} = \int_{\{\mathbf{x}\}_{\mathbf{y}}} \log f(\mathbf{y}\,|\lambda) f(\mathbf{x}\,|\mathbf{y}\lambda^k) d\mathbf{x} = \log f(\mathbf{y}\,|\lambda) \tag{3}$$

Combining (2) and (3) gives

$$\log f(\mathbf{y}\,|\lambda) = E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\lambda)\,\big|\mathbf{y}\lambda^k\} - E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\mathbf{y}\lambda)\,\big|\mathbf{y}\lambda^k\} \tag{4}$$

a function of λ and $\mathbf{y}$. It follows that with a complete/incomplete data formulation, $E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\lambda)\,\big|\mathbf{y}\lambda^k\}$ and $E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\mathbf{y}\lambda)\,\big|\mathbf{y}\lambda^k\}$ are each functions of λ^k but their difference is not. Substitution of (4) into (1) results in an expression for $B(\lambda\,|\mathbf{y})$ given a data vector $\mathbf{y}$

$$B(\lambda\,|\mathbf{y}) = Q(\lambda\,|\lambda^k) - E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\mathbf{y}\lambda)\,\big|\mathbf{y}\lambda^k\} \tag{5}$$

$$\text{where}\quad Q(\lambda\,|\lambda^k) = E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\lambda)\,\big|\mathbf{y}\lambda^k\} + \log f\,(\lambda)$$

To clarify how the EM approach works, we first note that from Jensen's inequality [15] it holds for any $\lambda^{k+1} \neq \lambda^k$

$$E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\mathbf{y}\lambda^{k+1})\,|\mathbf{y}\lambda^k\} \leq E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\mathbf{y}\lambda^k)\,|\mathbf{y}\lambda^k\}$$

with equality if and only if $\log f(\mathbf{x}\,|\mathbf{y}\lambda^{k+1}) = \log f(\mathbf{x}\,|\mathbf{y}\lambda^k)$ almost everywhere [3],[19]. It follows that a sufficient condition for $B(\lambda^{k+1}\,|\mathbf{y}) > B(\lambda^k\,|\mathbf{y})$ is $Q(\lambda^{k+1}\,|\lambda^k) > Q(\lambda^k\,|\lambda^k)$ since the second term on the right-hand side of (5) is guaranteed, from Jensen's inequality, not to decrease.

Beginning with some initial estimate $\lambda^0 > 0$ the EM Bayesian algorithm thus consists of the two steps:

The E–step: form $E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\lambda)\,|\mathbf{y}\lambda^k\}$ (6)

The M–step: solve $\max_{\lambda} Q(\lambda\,|\lambda^k) = E_{\mathbf{x}}\{\log f(\mathbf{x}\,|\lambda)\,|\mathbf{y}\lambda^k\} + \log f(\lambda)$ (7)

If the M-step is carried out to a global maximum of the E-step, the approach is termed an EM algorithm. If the M-step is only carried out to ensure $Q(\lambda^{k+1}|\lambda^k) \geq Q(\lambda^k|\lambda^k)$ the approach is termed a generalized EM (GEM) algorithm.

We note here that a maximum of the posterior distribution (4) is not obtained by a single M-step since the M-step only involves maximization with respect to a portion of the posterior distribution with the guarantee that the remaining portion is increased but not maximized. This is also the key to why the EM/GEM approach is not guaranteed to achieve a global maximum even if the M-step involves a global maximization [19]. As shown above and in [3], an EM/GEM approach ensures an increase in $B(\lambda|\mathbf{y})$ so that, for $B(\lambda|\mathbf{y})$ bounded from above, convergence to some B^* is assured. Continuity of $Q(\lambda|\lambda^k)$ with respect to both λ and λ^k is sufficient to ensure that all limit points of the sequence $\{\lambda^k\}$ are stationary points of $B(\lambda|\mathbf{y})$ [19]. In general, if $B(\lambda|\mathbf{y})$ is not unimodal and the set of stationary points contains points which are not local maxima, the EM/GEM approach at best only assures convergence of the sequence $\{\lambda^k\}$ to a stationary value. As Wu states [19], this should not be suprising since in such a case no general optimization algorithms are guaranteed to converge to local maxima. Outside of a single exception, no added convergence result is obtained by an EM versus a GEM approach. This exception, which is only of theoretical interest, is provided by Wu [19]. It occurs when it can be demonstrated that any stationary point which is not a local maximum is additionally not a global maximum of the E-step. Since the M-step of an EM algorithm requires global maximization of the results from the E-step, the EM M-step would not arrive at a stationary point which was not a local maxima under the above condition. The difficulty in verifying this condition, were it in fact to hold true for a given problem, makes this condition mainly of theoretical interest. From a practical viewpoint, the question of whether to carry out a global maximization versus an increase within the M-step is solely one concerned with increasing the per-iteration convergence speed at an increased per-iteration computational cost. For a complete treatment of the EM/GEM approach and its convergence properties see [3] and [19].

Optimization of likelihoods or Bayesian functions with independent priors may result in closed forms for the M-step [9]. If the complete data $\mathbf{x}$ are independent, the complete data are a linear function of the incomplete data, and the image pixels λ are treated as independent, the M-step only requires optimization of a set of univariate functions. However, it is generally accepted that the structure of images of any content is one of local correlation [14]. It would therefore seem more desireable to examine the use

of locally correlated priors. When used with an EM approach, the use of correlated priors prohibit the existance of closed form solutions for an EM M-step. Performing each EM M-step thus requires an iterative optimization of an N-dimensional function, N being the dimension of λ. What results is an iterative optimization algorithm within each iteration of an iterative optimization algorithm. In this case, iterative maximization of the posterior distribution by a method such as ICM [1] without using an EM formulation would seem more sensible than an EM approach. The GEM algorithm presented in this paper offers an attractive alternative to both of these approaches.

It should be noted that the use of a non-uniform prior distribution can induce local minima as well as local maxima in both the posterior distribution and in the results from an E-step of an EM formulation. Therefore, setting the gradient of the E-step equal to zero and solving the resulting set of equations does not ensure a valid M-step and may result in jumps which drastically decrease the posterior distribution. Alternatively, using an approximation to achieve a closed-form M-step opens the question of convergence and the possibility at any stage of decreasing the posterior distribution. We prefer a generalized EM approach since it guarantees a monotonic increase of the posterior function and has proven convergence properties. Under this approach, each M-step may consist of 1 or more iterations of an algorithm to increase $Q(\lambda \,|\lambda^k)$ without the requirement of maximizing it. We would expect the per-iteration speed of convergence to be slower for a GEM versus an EM approach. However, where a closed form M-step does not exist, the GEM approach may result in a greater increase in the posterior distribution for a given amount of computation. This may often be the case since it is generally true that the first iteration of an iterative optimization algorithm produces the largest improvement. In this work, we perform Bayesian reconstruction using Markov random field priors in the form of Gibbs functions.

3. Neighborhood Systems and Gibbs Function Priors

A discrete Markov random field (MRF) defined on a lattice is a collection of random variables, corresponding to the sites of the lattice, for which the probability of a given site value conditioned on the values of all other sites in the lattice is equal to the probability of the site value conditioned on the values at a small subset of the lattice sites. This subset of the lattice sites is called the neighborhood of the given site. Let the set of indices of sites in the neighborhood of pixel j be denoted N_j. Then

$$P(\lambda_j \,|\lambda_i : i \neq j) = P(\lambda_j \,|\lambda_i : i \in N_j)$$

If λ_i is a neighbor of λ_j , then λ_j is required to be a neighbor of λ_i. Neighborhoods are referred to as 0^{th} order, 1^{st} order,.., N^{th} order. Figure 1 shows the 0^{th} order, 1^{st} order, and 2^{nd} order neighborhoods for the 2-D lattice and the 1^{st} order neighborhood for the 3-D lattice. By the Hammersly-Clifford theorem (1971) [2] a random field defined on a lattice is a Markov random field if and only if its distribution function corresponds to a Gibbs function. To define the form of a Gibbs function, we must first define a clique. A clique is either a single site or a set of sites such that each site in the clique is a neighbor of all other sites in the set. The clique types associated with each neighborhood are also shown in figure 1. For a rectangular 2-D lattice and a 1^{st} neighborhood shown in figure 1, the cliques are sets of sites consisting of a single site or two horizontally or vertically adjacent sites.

A Gibbs distribution is then a probability measure on the set of configurations $\{\lambda\}$ which has the form

$$f(\lambda) = \frac{1}{K} e^{\frac{-U(\lambda)}{\beta}}$$

where β is a constant, K is the normalizing constant (partition function), and $U(\lambda)$ is termed the energy function. The energy function has the form

$$U(\lambda) = \sum_{\alpha C} V_c(\lambda)$$

where C denotes the set of all cliques and $V_c(\lambda)$, termed a potential function, is a function on clique c.

Gibbs functions provide a powerful class of correlated priors. The appeal of the Gibbs prior is that it can be defined outside of the normalizing constant simply by defining a suitable pixel neighborhood and potential functions on the cliques associated with that neighborhood. The intent is to capture a desired property of the unknown image by a suitable choice of these. For Bayesian reconstruction, the normalizing constant, which depends on β, need not be calculated. The parameter β controls the degree to which the modes of the Gibbs prior stand out. As $\beta \rightarrow +\infty$ the integral of the prior in a small region about a mode tends toward the same value as that about any point so that the distribution tends to the uniform. As $\beta \rightarrow 0$ the prior becomes increasingly more pronounced about its modes. With a Gibbs prior, (1) becomes

$$\max_{\lambda} B(\lambda \,|y) = \log f(y \,|\lambda) - \frac{1}{\beta} \sum_{\alpha C} V_c(\lambda)$$

so that as $\beta \rightarrow +\infty$ Bayesian reconstruction is unaffected by the prior, which tends to the uniform, and reduces to maximum likelihood. In this work and in [4] and [5], acceptable values for β were obtained by trial and error. Furthur, as shown in the result section, β values from the interval $+\infty$ to some lower limit value can produce an improvement in the reconstruction. Smaller β values produce a degradation due to an over-influence by the prior. There is a need for statistical methods of determining optimal β values given suitably normalized potential functions and the data set.

Pixel configurations of lowest energy are of highest probability. It is common to choose energy functions which penalize configurations with neighboring pixels differing by large amounts. The Gaussian prior with a diagonal covariance matrix **H** and mean image **m** chosen by Levitan and Herman in [8] is also a particular case of a Gibbs prior with a 0^{th}-order neighborhood. A 0^{th}-order neighborhood has only cliques containing a single pixel. The corresponding energy function is $-\frac{\gamma}{2}(\lambda-\mathbf{m})^T\mathbf{H}(\lambda-\mathbf{m})$. For the work in this paper, we have chosen a 1^{st} order neighborhood (Fig. 1). In [1] , a 1^{st}-order neighborhood is considered unrealistic for most applications. However, in three dimensions a 1^{st}-order neighborhood results in 6 neighbors for every pixel versus 4 in 2 dimensions. As results presented here indicate, a 1^{st} order neighborhood in three dimension may be sufficient for many applications. In this paper we examine the following three potential functions. None of the resulting Gibbs priors impose any mean on the image. In each, the potential function on cliques containing a single site have been defined equal to zero.

(1) $V_1(\lambda_j;\lambda_i) = (\lambda_j - \lambda_i)^2$

(2) $V_2(\lambda_j;\lambda_i) = \dfrac{(\lambda_j-\lambda_i)^2}{\delta^2 + (\lambda_j - \lambda_i)^2}$

(3) $V_3(\lambda_i;\lambda_j) = \log\left(1 + \left(\dfrac{\lambda_j-\lambda_i}{\mu}\right)^2\right)$

The first is used by Geman and Geman in [4]. The second is equivalent to that used by Geman and McClure in [5]. The third potential function is a compromise between the first two. A plot of these three potential functions suitably normalized versus difference between the two pixels in a clique is shown in figure 2. This normalization simplifies the interpetation of the effects of the priors and enables some comparison for a given β value. The first potential function increasingly penalizes the separation between neighboring pixels. In addition, it does so at an increasing rate as the separation increases. To improve on this, we seek a potential function which penalizes separations within uniform

Figures and Graphs

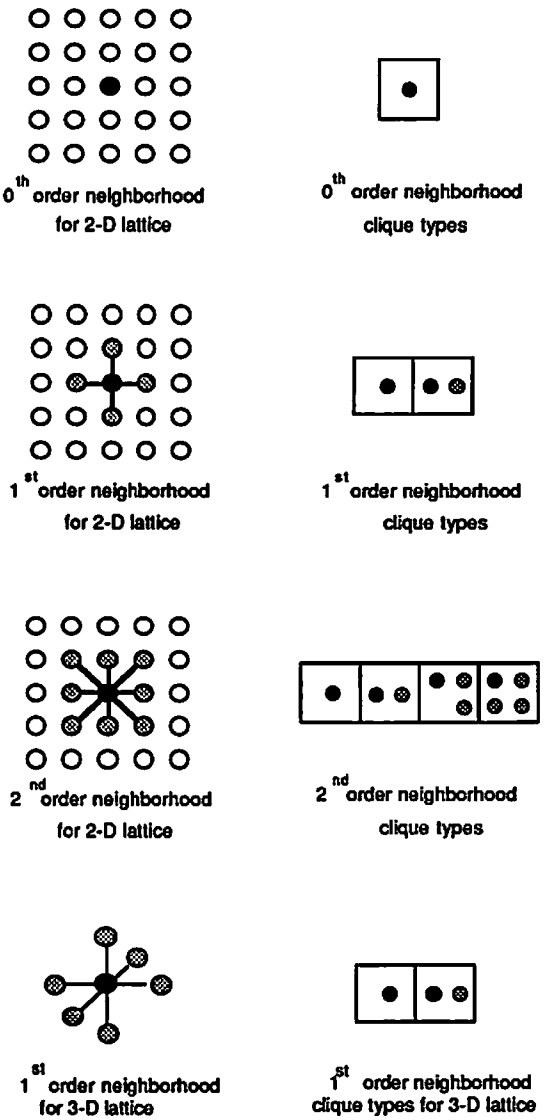

Figure 1

0^{th}, 1^{st}, and 2^{nd} order neighborhoods and clique types for 2-D lattices and 1^{st} order neighborhood and clique types for 3-D lattice

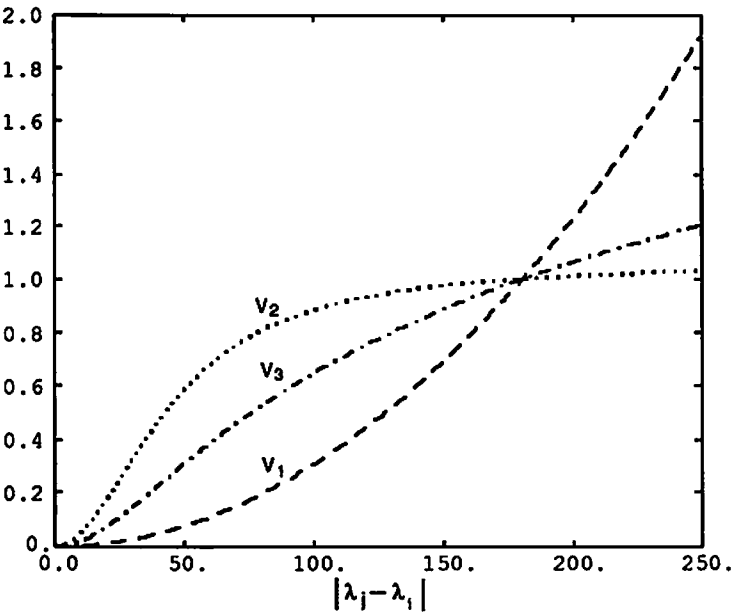

Figure 2

Plot of normalized potential functions vs. difference between two pixels in a clique.

regions without unduly penalizing the larger separations which we foresee occuring at the boundary between two different regions of the image. A good example of such a boundary is the region outside of the patient versus the region inside of the patient. The second prior encourages neighboring pixels to be of similar value until they have become separated through iterative reconstruction by a given threshold δ. At this point the prior allows furthur separation of their values at relatively small increase in the penalty. We found that reconstructions of single photn emission images from very low total count numbers benefitted from potential functions which increasingly penalize pixel separation over the full range of separation values.

4. The Application to Emission Tomography

Reconstructing the distribution of a radioactively tagged compound in a patient is important in assessing organ function and in assessing the efficiency of chemotherapeutic drugs. Upon ingestion, a radiopharmaceutical either rapidly or gradually accumulates in various regions of the body. As the radioactive atoms in the compound decay, γ-rays are emitted. Some of these γ-rays exit the body and are recorded by the imaging system. A γ-ray imaging system records each detected γ-ray into one of several hundred thousand bins. The number of counts in each bin comprise the data. Corresponding to each bin is a relatively small 3-D volume from which it can be said, with high probability, the emission originated. However, these 3-D volumes overlapp one other so that any point in the source space may belong to 50 or more bins. Typically, several million γ-rays will be detected and registered into several hundred thousand bins. The number of counts in these bins can be shown to be conditionally independent Poisson random variables [8]. To determine the distribution of radiopharmaceutical, we divide the 3-D space viewed by the imaging system into small volumes called pixels. By reconstructing the Poisson mean of each source pixel, we characterize the distribution of isotopically tagged compound in the patient.

An emission imaging system has a complete/incomplete data formulation in terms of the observed but incomplete data $\mathbf{y}$ and complete data set $\{\mathbf{x}_{ij}\}$ where $\mathbf{y}_i$ represents the number of counts in bin i and $\mathbf{x}_{ij}$ represents the number of γ-ray emissions from source pixel j detected at bin i [16]. Clearly, $\mathbf{y}_i = \sum_i \mathbf{x}_{ij}$. The complete data $\mathbf{x}_{ij}$ are well modeled as independent Poisson random variables with means $\mathbf{P}_{ij}\lambda_j$ where λ_j is the mean of the total number of γ-ray emissions from source pixel j and $\mathbf{P}_{ij}$ is the probability that a γ-ray emitted from source pixel j will be detected at camera bin i. Thus,

$$f(\mathbf{x}\,|\lambda) = \prod_i \prod_j e^{-(\mathbf{P}_{ij}\lambda_j)} \frac{(\mathbf{P}_{ij}\lambda_j)^{\mathbf{x}_{ij}}}{(\mathbf{x}_{ij})!}$$

The logarithm of $f(\mathbf{x}\,|\lambda)$ is a linear function of the $\mathbf{x}_{ij}$'s plus terms independent of λ. Therefore, for the E-step (6) we only need to compute $E\{\mathbf{x}_{ij}\,|\mathbf{y}\lambda^k\}$. The variables $\{\mathbf{x}_{iq}: q = 1,..,N\}$ are independent and Poisson with means $\mathbf{P}_{iq}\lambda_q$. In addition, $\sum_q \mathbf{x}_{iq} = \mathbf{y}_i$ so that

$$f(\mathbf{x}_{i1}..\mathbf{x}_{iN}\,|\mathbf{y};\lambda^k) = f(\mathbf{x}_{i1}..\mathbf{x}_{iN}\,|\mathbf{y}_i;\lambda^k)$$

The joint distribution of a set of independent Poisson variables z conditioned on their sum is a multinomial distribution with probability in each class j equal to $\frac{E\{\mathbf{z}_j\}}{\sum E\{\mathbf{z}_j\}}$ [13]. The expected value of a multinomial variable is equal to the probability in its class times the total number of trials. Therefore, the joint distribution of $\{\mathbf{x}_{iq}: q=1,..N\}$ conditioned on $\sum_q \mathbf{x}_{iq}=\mathbf{y}_i$ is

$$f(\mathbf{x}_{i1}..\mathbf{x}_{iN}\,|\mathbf{y}_i;\lambda^k) = \frac{(\mathbf{y}_i)!}{(\mathbf{x}_{i1})!\cdots(\mathbf{x}_{iN})!}(\mu_{i1})^{\mathbf{x}_{i1}}\cdots(\mu_{iN})^{\mathbf{x}_{iN}} \qquad \text{where} \quad \mu_{ij} = \frac{\mathbf{P}_{ij}\lambda_j}{\sum_q \mathbf{P}_{iq}\lambda_q}$$

$E\{\mathbf{x}_{ij}\,|\mathbf{y}\lambda^k\}$, which is the mean of class j of the multinomial distribution is equal to $\mathbf{y}_i * \frac{\mathbf{P}_{ij}\lambda_j^k}{\sum_q \mathbf{P}_{iq}\lambda_q^k}$.

The E-step can now be formed as

$$E_x\{\log f(\mathbf{x}\,|\lambda)\,|\mathbf{y}\lambda^k\} = \sum_j \left(-\mathbf{a}_j\lambda_j + \mathbf{b}_j^k \log\lambda_j\right) + \text{terms independent of } \lambda \qquad (8)$$

$$\text{where} \quad \mathbf{a}_j = \sum_i \mathbf{P}_{ij} \qquad \text{and} \qquad \mathbf{b}_j^k = \sum_i \frac{\mathbf{y}_i \mathbf{P}_{ij}\lambda_j^k}{\sum_q \mathbf{P}_{iq}\lambda_q^k}$$

If the $\mathbf{P}_{ij}$'s have been normalized as in [16], then $\mathbf{a}_j = 1$ for all j. In forming $Q(\lambda\,|\lambda^k)$ of (4) we can omit the terms independent of λ since these do not affect the M-step (7). For the generalized EM approach, the M-step is to find a λ^{k+1} such that $Q(\lambda^{k+1}\,|\lambda^k) > Q(\lambda^k\,|\lambda^k)$ where from (4) and (8)

$$Q(\lambda\,|\lambda^k) = \sum_j \left(-\mathbf{a}_j\lambda_j + \mathbf{b}_j^k \log\lambda_j\right) - \sum_{\alpha \in C} \frac{V_c(\lambda)}{\beta} \qquad (9)$$

The algorithm we present here is a generalized EM algorithm and it is explained in the section to follow. It is formulated for the class of correlated priors represented by Gibbs functions and its implementation closely follows that of the EM likelihood algorithm. The algorithm performs a coordinate gradient ascent of $Q(\lambda\,|\lambda^k)$ with an initial step-size which resembles the EM likelihood algorithm for emission tomography [16]. We note that this algorithm updates pixels sequentially and that updated values are used to update the pixels that follow. Therefore, in step 2a below, the superscript has been omitted from λ_i in $V(\lambda_i;\lambda_j^k)$ because the neighboring pixels λ_i may consist of both updated and un-updated pixels. At the k^{th} iteration, perform the following steps:

(1) For all image pixels, compute the usual EM likelihood algorithm [16] updated variables $\lambda_j^{EM} = \dfrac{\mathbf{b}_j^k}{\mathbf{a}_j}$ where $\mathbf{a}_j$ and $\mathbf{b}_j^k$ are defined as in (8).

(2) To update the image, visit pixel sites sequentially. When visiting a pixel λ_j, do (2a)-(2d)

(2a) compute C_1 and C_2 where

$$C_1 = \mathbf{a}_j\left(-\lambda_j^k + \lambda_j^{EM}\log\lambda_j^k\right) - \sum_{i\in N_j}\frac{V(\lambda_i;\lambda_j^k)}{\beta}$$

$$C_2 = \frac{1}{\beta}\sum_{i\in N_j}\frac{\partial}{\partial\lambda_j^k}V(\lambda_i;\lambda_j^k)$$

(2b) Set $\alpha = 1$.

Compute $\lambda_j^{k+1} = \lambda_j^{EM} - \dfrac{C_2\lambda_j^k}{\mathbf{a}_j}$

If $\lambda_j^{k+1} > 0$ go to (2d)

If $\lambda_j^{k+1} \le 0$ compute $\alpha = \dfrac{.5}{1 - \dfrac{\lambda_j^{EM}}{\lambda_j^k} + \dfrac{C_2}{\mathbf{a}_j}}$.

(2c) Compute $\lambda_j^{k+1} = (1-\alpha)\lambda_j^k + \alpha\left\{\lambda_j^{EM} - \dfrac{C_2\lambda_j^k}{\mathbf{a}_j}\right\}$

(2d) Check if $\mathbf{a}_j\left(-\lambda_j^{k+1} + \lambda_j^{EM}\log\lambda_j^{k+1}\right) - \sum_{i\in N_j}\dfrac{V(\lambda_i;\lambda_j^{k+1})}{\beta} \ge C_1$

If yes, update pixel j to λ_j^{k+1} and visit the next pixel. If no, divide α by 2 and return to step (2c).

To further clarify the steps, when visiting a pixel j, step 2a computes two values which are functions of the data and of the updated and un-updated pixels in the neighborhood of pixel j. In step 2b, if the step-size $\alpha = 1$ results in a negative X_j^{k+1}, half the step-size α which gives $X_j^{k+1} = 0$ is computed. Step 2c implements a coordinate ascent [11] step for the specified step-size, and step 2d ensures that the step-size α has resulted in an increase in $Q(\lambda \,|\lambda^k)$. If $Q(\lambda \,|\lambda^k)$ has not been increased, step [2d] cuts the step-size in half. As $\beta \rightarrow +\infty$, $C_2 \rightarrow 0$ and the pixels are updated by setting them equal to the EM likelihood updated pixel values. Step 2d , which would then represent a check to ensure the likelihood function has been increased, is always satisfied so that steps 2a, 2c, and 2d are no longer necessary. As $\beta \rightarrow +\infty$, this algorithm thus reduces to the EM likelihood algorithm.

Let us consider why this is a generalized EM algorithm. If step 2d ensures a monotonic increase of $Q(\lambda \,|\lambda^k)$ and step 2c is guaranteed to arrive at a X_j^{k+1} satisfying step 2d, then at the conclusion of the M-step $Q(\lambda^{k+1} \,|\lambda^k) > Q(\lambda^k \,|\lambda^k)$ and the algorithm is a generalized EM algorithm. Let the potential function on cliques containing a single pixel be set to zero. For a 1st order neighborhood, specify a potential function $V(\lambda_i;\lambda_j)$ evaluated on all cliques containing two pixels. The energy function has the form $\frac{1}{\beta}\sum_{ij \in C} V(\lambda_i;\lambda_j)$. Let N_j denote the set of indices of pixels which are neighbors of pixel j. In order to show that (9) is increased at each stage, there are only two cases we must consider: updating a pixel whose neighbors have not been updated in the present iteration, and updating a pixel for which one or more of the neighbors have been updated. Without loss of generality, let us examine steps 2a-2d for two pixels λ_i and λ_j which are neighbors by writing $Q(\lambda \,|\lambda^k)$ (9) explicitly in terms of λ_i and λ_j .

$$Q(\lambda_i;\lambda_j;\lambda_l: l \neq i,j \,|\lambda^k) = -\mathbf{a}_i\lambda_i + \mathbf{b}_i^k \log\lambda_i - \sum_{\substack{l \in N_i \\ l \neq j}} \frac{V(\lambda_i;\lambda_l)}{\beta} - \frac{V(\lambda_i;\lambda_j)}{\beta} - \mathbf{a}_j\lambda_j + \mathbf{b}_j^k \log\lambda_j$$

$$- \sum_{\substack{l \in N_j \\ l \neq i}} \frac{V(\lambda_j;\lambda_l)}{\beta} + \sum_{l \neq i,j} \left(-\mathbf{a}_l\lambda_l + \mathbf{b}_l^k \log\lambda_l\right) - \sum_{\substack{pq \in C \\ p;q \neq i \text{ or } j}} \frac{V(\lambda_p;\lambda_q)}{\beta} \qquad (10)$$

Let us first visit pixel i whose neighbors have not been updated. Step 2d ensures an increase in the sum of terms 1, 2, 3, and 4 in (10), with no effect on the other terms. Therefore, X_i^{k+1} satisfies

$$Q(X_i^{k+1};X_j^k X_l^k \; l \neq i,j \,|\lambda^k) > Q(X_i^k X_j^k X_l^k \; l \neq i,j \,|\lambda^k)$$

For the second case, let us then visit and update pixel j a neighbor of pixel i , step 2d

furthur guarantees an increase in the sum of terms 4, 5, 6, and 7 in (10), with no effect on the other terms. Therefore, λ_j^{k+1} satisfies

$$Q(\lambda_i^{k+1};\lambda_j^{k+1};\lambda_l^k\ l\neq i,j\ |\lambda^k) > Q(\lambda_i^{k+1};\lambda_j^k;\lambda_l^k\ l\neq i,j\ |\lambda^k) \tag{11}$$

Each pixel that is updated according to step 2d results in an increase in $Q(\lambda\,|\lambda^k)$ and from section 2, $B(\lambda\,|\mathbf{y})$ is increased.

It remains to show that step 2c will result in an updated pixel satisfying step 2d. Step 2c is a coordinate gradient ascent [11] of $Q(\lambda\,|\lambda^k)$ initialized at λ^k . At each step of a coordinate ascent algorithm, only changes to a single element λ_j are allowed such that a monotonic increase in $Q(\lambda\,|\lambda^k)$ is achieved. Each element is addressed in some prescribed order so that for bounded functions ultimate convergence to a stationary value is assured. Let us take the case above, where λ_i^k has been updated to λ_i^{k+1} and we wish to update λ_j^k. A coordinate gradient ascent of $Q(\lambda_i^{k+1};\lambda_j^k;\lambda_l^k\, l\neq i,j\ |\lambda^k)$ takes the form

$$\lambda_j^{k+1} = \lambda_j^k + \mathbf{m}_j \frac{\partial}{\partial \lambda_j^k} Q(\lambda_i^{k+1};\lambda_j^k;\lambda_l^k\ l\neq i,j\ |\lambda^k) \tag{12}$$

where $\mathbf{m}_j$ is any positive value. From (10)

$$\frac{\partial}{\partial \lambda_j^k} Q(\lambda_i^{k+1};\lambda_j^k;\lambda_l^k l\neq i,j\ |\lambda^k) = -\mathbf{a}_j + \frac{\mathbf{b}_j^k}{\lambda_j^k} - C_2 = \mathbf{a}_j\left(-1 + \frac{\lambda_j^{EM}}{\lambda_j^k} - \frac{C_2}{\mathbf{a}_j}\right) \tag{13}$$

where C_2 is as defined in step [2a]. Let $\mathbf{m}_j = \dfrac{\alpha\lambda_j^k}{\mathbf{a}_j}$. Since λ^k is constrained to be positive, $0<\alpha\leq 1$, and $\mathbf{a}_j>0$, m_j is positive. Substituting $\mathbf{m}_j$ and (13) into (12) gives step 2c. The step direction follows the j^{th} coordinate directional derivative of $Q(\lambda_i^{k+1};\lambda_j^k;\lambda_l^k\ l\neq i,j\ |\lambda^k)$ and the initial step-size is chosen to mimic the EM likelihood algorithm. An increase in $Q(\lambda_i^{k+1};\lambda_j^k;\lambda_l^k\ l\neq i,j\ |\lambda^k)$ is possible with a sufficiently small step size if that derivative is non-zero. If the derivative is zero, step 2d is satisfied immediately. Steps sizes which do not satisfy step 2d are cut in half so that the algorithm quickly arrives at a step size increasing $Q(\lambda\,|\lambda^k)$.

This GEM algorithm monotonically increases $Q(\lambda\,|\lambda^k)$ and terminates at a point λ^k for which

$$[\nabla Q(\lambda^k\,|\lambda^k)]_j \longrightarrow \begin{Bmatrix} = 0 \text{ if } \lambda_j^k > 0 \\ < 0 \text{ if } \lambda_j^k = 0 \end{Bmatrix} \tag{14}$$

for all j; i.e. at a point where the directional derivative in all feasible directions is less than or equal to zero. From (4)

$$\nabla B(\lambda \,|y) = \nabla Q(\lambda \,|\lambda^k) - \nabla E_x\{\log f(x \,|y\lambda) \,|y\lambda^k\}$$

Since λ^k maximizes $E_x\{\log f(x \,|y\lambda) \,|y\lambda^k\}$ (5), $\nabla E_x\{\log f(x \,|y\lambda^k) \,|y\lambda^k\} = 0$ [19]. This holds regardless of whether or not λ^k lies on a boundary. Therefore $\nabla B(\lambda^k \,|y) = \nabla Q(\lambda^k \,|\lambda^k)$ and the algorithm terminates at a point λ^k for which the directional derivative of $B(\lambda^k \,|y)$ in any feasible direction is less than or equal to zero, i.e. $\{\lambda^k\}$ converges to a stationary point.

5. Results

The simulations we present model a 3-D single photon emission imaging system consisting of a parallel collimated gamma camera with 48^2 pixels viewing a 3-D volume with data collected from 48 different equispaced angles. Perfect collimation is assumed. A 3-D source space consisting of 48^3 pixels is reconstructed. The Gibbs distribution potential functions were defined as in section 3 for a 1^{st} order neighborhood consisting of the 6 nearest neighbors. The neighbors of an interior image pixel consist of the pixels above, below, and on all 4 sides totaling 6 neighbors. The missing neighbors of pixels located on the side boundaries are assumed zero. A free boundary [4] is used for pixels on the top and bottom planes of the 3-D reconstruction space. These pixels have fewer neighbors. Pixels on the bottom layer of the 3-D reconstruction space have no neighbor below while pixels on the top layer have no neighbor above.

Figure 3 shows the 3-D computer generated phantom used in our simulation study. A total mean of 2 million counts were generated from this 3-D phantom. Figures 4, 5, 6, and 7 provide a visual comparison of the reconstructions from the EM likelihood and GEM Bayesian algorithm with each of the three priors discussed in section 3. The reconstructions displayed in these figures are the result of 50 iterations of each specified algorithm. With more iterations, the EM likelihood reconstruction worsened furthur while the GEM reconstructions remained unchanged. For these reconstructions β=1 and the three potential functions were normalized as shown in figure 2. Figure 4, the EM likelihood reconstruction, shows the excessive non-smoothness reported by many authors [5],[7],[9],[17],[18]. Figures 5, 6, and 7 show that all three Gibbs priors produced a considerable visible improvement in the reconstruction.

Figure 8 shows the L_2 norm error between the true source image and the reconstruction for 100 iterations of the EM likelihood algorithm and for 100 iterations of the the GEM algorithm using the three different Gibbs priors in section 3. The GEM algorithm initially reduced the L_2 norm error as quickly as the EM likelihood algorithm. The EM likelihood algorithm characteristically iterated away from the true source image after a number of iterations while the GEM algorithm continued to reduce the L_2 norm error until convergence.

Figure 9 shows the Bayes (likelihood) value at each iteration. This shows a monotonic increase in these functions for the EM likelihood algorithm the GEM bayesian algorithm. Figure 10 examines a range of β values. Here, the L_2 norm error between the true source image and the reconstruction for 100 iterations of the GEM algorithm using the first potential function V_1 with different values of β is shown. According to this criterion, we found an improvement in the reconstruction for all values of $\beta \geq 1.0$.

These simulations were run in Fortran code on a Sun 3/110 workstation with a floating point accelerator. Some improvement in the speed was achieved by setting the parameter space to $\lambda_j \geq \phi > 0$. where ϕ was chosen as some small value such as $\phi = .0001$. If λ_j equaled ϕ and the the directional derivative was negative along that coordinate, λ_j was left equal to ϕ and the next pixel was visited. This reduced some of the time spent computing small steps for small pixel values converging towards zero, a task particularily time consuming as the algorithm converges. The EM likelihood algorithm required 48 sec per 3-D iteration while the GEM algorithm averaged 63 sec per 3-D iteration. The 3-D forward/back projections required the majority of the CPU time (47 sec per iteration) while pixel updating filled the remaining seconds.

6. Conclusion

The GEM Bayesian algorithm we have presented can be used with any locally correlated priors in the form of Gibbs functions and has the desireable theoretical convergence of its generalized EM formulation. For 3-D images, such as those encountered in emission tomography, this algorithm can provide an improvement over maximum likelihood reconstruction at a nominal computational cost. This work shows that some improvement can be achieved for a wide range of β values, but statistical methods for optimizing the choice of this parameter are still needed.

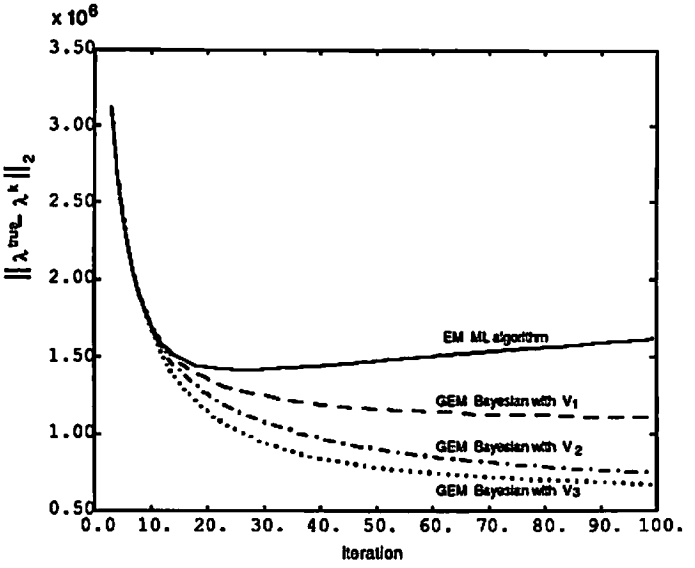

Figure 8

Plot of the L_2 difference between the true 3-D image and the reconstruction for 100 EM ML and GEM Bayesian iterations.

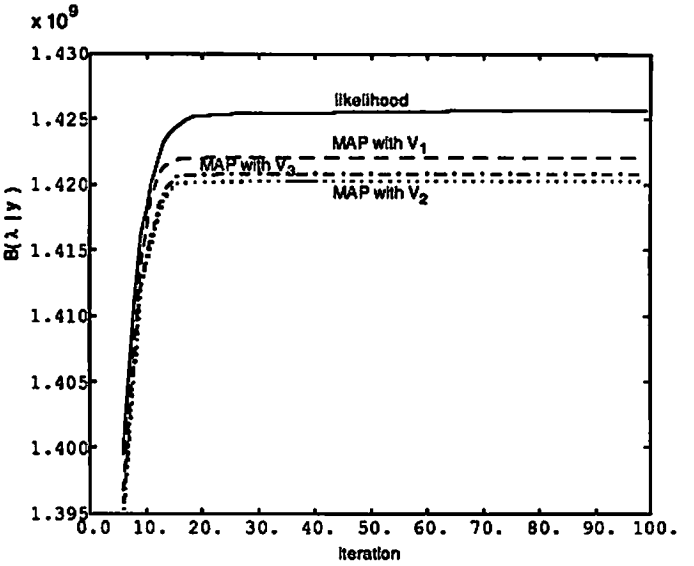

Figure 9

Plot of the likelihood and Bayesian functions for 100 EM and GEM iterations with $\beta = 1.0$

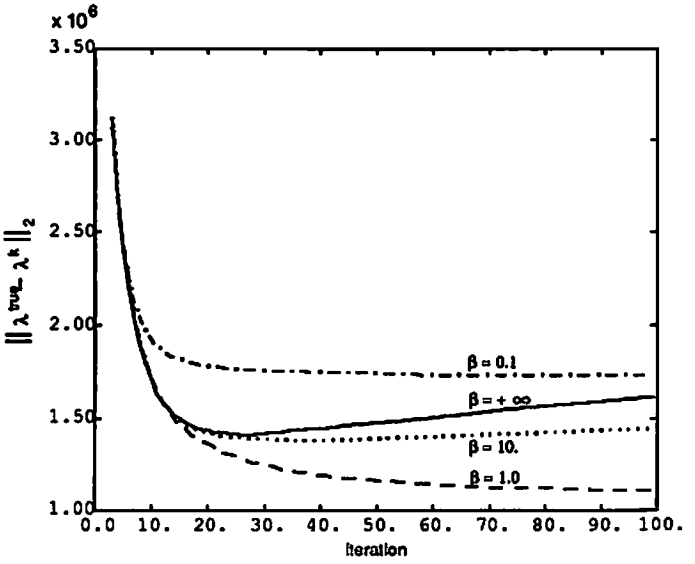

Figure 10.

Plot of the L_2 difference between the true 3-D image and the reconstruction for 100 GEM iterations with V_1 and different β values.

References

[1] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Royal Statist. Soc.*, no. 3, 1986.

[2] J. Besag, "Statistical Interaction and the Statistical Analysis of Lattice Systems," *J. Royal Statist. Soc.*, series B, vol. 34, 1972.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statist. Soc.*, series B, vol. 39, 1977.

[4] S. Geman, D. Geman "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. PAMI*, no. 6, Nov. 1984.

[5] S. Geman, D. McClure, "Bayesian Image Analysis: an Application to Single Photon Emission Tomography," *Proc. Statist. Comput. Sect.*, Amer. Statist. Assoc., 1985

[6] H. Hart, Z. Liang, "Bayesian Processing in Two Dimensions," *IEEE Trans. Med. Imaging*, no. 3, Sept. 1987.

[7] T.J. Hebert, R. Leahy, M. Singh, "Fast MLE for SPECT using an Intermediate Polar Representation and a Stopping Criterion," *IEEE Trans. Nucl. Sci.*, Feb., 1988.

[8] T.J. Hebert, R. Leahy, M. Singh, "ML Reconstruction for a Prototype Electronically Collimated Single Photon Emission System," SPIE Med. Imag. Conf., Feb. 1987.

[9] E. Levitan, G.T. Herman, "A Maximum A Posteriori Probability Expectation Maximization Algorithm for Image Reconstruction in Emission Tomography," *IEEE Trans. Med. Imaging*, no. 3, Sept. 1987.

[10] Z. Liang, H. Hart, "Bayesian Image Processing of Data from Constrained Source Distributions-Non-Valued, Uncorrelated and Correlated Constraints," *Bulletin of Mathematical Biology*, vol. 49, 1987.

[11] D. Luenberger, *Linear and Non-Linear Programming*, Addison-Wesley Publishing Co., Readings Mass., Second Edition, 1984.

[12] M. Miller, D. Snyder, S. Moore, "An Evaluation of the Use of Sieves for Producing Estimates of Radioactivity Distributions with the EM Algorithm for PET," *IEEE Trans. Nuc. Sci.*, vol. NS-33, 1986.

[13] C. Rao, *Advanced Statistical Methods in Biometric Research*, Hafner Press, 2nd Reprint 1974.

[14] A. Rosenfeld, A. Kak, *Digital Picture Processing I & II*, Academic Press, 1982.

[15] R. Serfling, *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, 1980.

[16] L. Shepp, Y. Vardi, "Maximum Likelihood Reconstruction for Emission Tomography," *IEEE Trans. Med. Imaging*, no. 2, Oct. 1982.

[17] D. Snyder, M. Miller, "The Use of Sieves to Stabilize Images Produced with the EM Algorithm for Emission Tomography," *IEEE Trans. Nuc. Sci.*, vol. NS-32, 1985.

[18] E. Veklerov, J. Lacer, "Stopping Rule for the MLE Algorithm Based on Statistical Hypothesis Testing," *IEEE Trans. Med. Imaging*, No. 4, Dec. 1987.

[19] C.F.J. Wu, "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, vol. 11, 1983.