Bayesian Reconstructions From Emission Tomography Data Using a Modified EM Algorithm

PETER J. GREEN

Abstract—**A new method of reconstruction from SPECT data is proposed, which builds on the EM approach to maximum likelihood reconstruction from emission tomography data, but aims instead at maximum posterior probability estimation, that takes account of prior belief about "smoothness" in the isotope concentration. A novel modification to the EM algorithm yields a practical method. The method is illustrated by an application to data from brain scans.**

I. Introduction

In single-photon emission computerized tomography (SPECT), the patient is injected with, or inhales, a pharmaceutical tagged with a radioactive isotope (a radiopharmaceutical). The pharmaceutical is chosen as being known to concentrate in the organ of interest, in a manner related to the phenomenon under study, for example metabolic rate or blood flow. Some of the photons emitted by the radiopharmaceutical are collected in a system of detectors located around the patient, and the objective is to form a three-dimensional reconstruction of the pattern of isotope concentration from the resulting array of photon counts.

A relatively inexpensive and convenient device for recording the emitted photons is the *gamma camera*, which is rotated about an axis through the patient, to collect a sequence of projections. The physical details of the gamma camera and its use in tomography are described in detail in Jaszczak, Coleman, and Lim [11] and in the excellent monograph by Larsson [16] and so will not be covered here.

Because of the random nature of radioactive disintegration, the tomographic data are noisy, and it is appropriate to regard reconstruction as a statistical estimation problem; since the views of the patient that are obtained are *indirect*—the coordinates with respect to which the data are collected are not those in which the reconstruction is required—this estimation is not entirely straightforward. The calculations required are far from trivial because of the amount of information involved: the minimum scale in practical use for 3-D tomography requires reconstruction of the isotope concentration on a $64 \times 64 \times 64$ grid, based on 64 projections, each consisting of a 64×64 array of counts.

The standard algorithms in use today are adapted from those successfully used in transmission tomography, and involve *filtered back-projection.* These are fast and inexpensive as they operate entirely linearly on the data. However, the resulting reconstructions do suffer from well-documented deficiencies. Attempts to overcome these deficiencies have chiefly centered on more careful modeling of the process generating the projection data, and that is the approach taken here. We also take the view that more fundamental than the algorithm for reconstruction is the *principle* underlying that algorithm: what is being estimated, what is the logical basis for that estimation, and what is the performance of the algorithm with respect to that basis?

We follow the paradigm of Bayesian image analysis, introduced by Besag [1], [2] and Geman and Geman [5]. This involves the construction of two probability models. The first describes the manner in which the detected counts are generated by the tomographic transformation and other physical circumstances from the true isotope concentration. In particular, this will account for the Poisson variation in the counts arising from the random nature of radioactive disintegration. The second model is a probability distribution on the space of true patterns of isotope concentration and provides a means of quantitatively coding prior information or beliefs about such patterns available before the data are collected. This aspect of the modeling is crucial in the context of emission tomographic data: it is indeed hard to imagine knowing *nothing* about the true pattern (although coding one's knowledge numerically might not be straightforward). Reconstruction of the isotope concentration will be accomplished by considering the posterior distribution that follows by Bayes' theorem from the two model components: thus the probability distribution of isotope patterns will be treated as a Bayes prior distribution.

If the reader finds this Bayesian paradigm unappealing, an alternative approach which yields exactly the same algorithms follows from the principle of penalized likelihood. Because of the high dimension of the "parameter" space (the space of true isotope patterns), it is natural to

Manuscript received February 22, 1989; revised August 8, 1989. This work was supported by the Science and Engineering Research Council.

The author was with the University of Durham; Durham, UK. He is now with the Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK.

IEEE Log Number 8931402.

treat the tomography problem statistically as (Poisson) nonparametric regression. It is well known that in such contexts the maximum likelihood principle is either inappropriate or ill-defined (see, for example, [8]), and a common way of modifying the principle so that it does give sensible and efficient estimates is by penalizing the (log-) likelihood so that implausible isotope patterns (for example, those that are too "rough") are disfavored, even though they may have high likelihood for the given data. The reconstructions that we aim to compute can thus be viewed either as posterior modes (*maximum a posteriori* estimates) or as maximum penalized likelihood estimates. In practice, we shall have to be content with approximately achieving this goal.

It will be clear in what follows that our approach to the SPECT problem makes considerable use of the highly original contributions to the subject by Shepp and Vardi [24] and Geman and McClure [6].

II. Modeling the Photon Counts

In this section, we introduce some basic notation and derive the first of our two probability models: that for the detected photon counts, *given* the isotope concentration.

From the outset, we shall suppose that the 3-D space over which the reconstruction is required, the *body space*, is finely subdivided into rectangular pixels or voxels. We are thus recognizing that spatially discrete data only allow a spatially discrete reconstruction, and that we cannot expect an arbitrarily fine spatial resolution in our result. This issue is explored from a theoretical point of view in Johnstone and Silverman [12]. There is nothing inherently natural about a rectangular pixelization, and there may be merit in other regular tessellations, e.g., using hexagonal prisms. There are thus a finite, although large, number of isotope concentrations to be estimated: we index these by the subscript s, and let x_s denote the unknown isotope concentration in pixel s, $s = 1, 2, \cdots, S$. The whole array $\{x_s\}$ will be denoted by x.

The data collected form a three-dimensional array of photon counts, but the gamma camera is not inherently either three-dimensional or discrete. After passing through a collimator, which only admits photons whose paths are nearly exactly perpendicular to the face of the camera, incoming photons strike a crystal where fluorescence occurs as a result of photomultiplication. The coordinates of the fluorescence are measured discretely, and the photons thus indirectly "binned" into a rectangular array. The gamma camera is then advanced around the patient's axis, another projection recorded, and so on. The resulting bins are thus indexed by angle and pixel within a projection. We shall use the single subscript t to index such bins; y_t denotes the recorded count within bin t, $t = 1, 2, \cdots, T$ and $y = \{y_t\}$ the complete collection of data. We will refer to t as varying in *projection space*.

With this notation, our modeling requirements are as follows:

1) the data model $p(y|x)$, and
2) the prior model $p(x)$.

Here, we are using an informal notation in which the symbol p is used to denote a generic probability density. The remainder of this section will be concerned with 1); consideration of 2) will be postponed until Section IV.

Assuming only that emissions from pixel s are completely random, with rate constant in time, and that "dead-time" effects in the counting can be neglected, the stream of photons from s arriving and detected at bin t forms a Poisson process homogeneous in time, and since each photon is detected in at most one bin and the emissions are independent, these Poisson processes are independent for all s and t. From elementary superposition properties of Poisson processes, therefore, the observable counts y are independent Poisson distributed random variables, with

$$y_t \sim \text{Poisson}\left(\sum_s a_{ts}x_s\right) \tag{1}$$

and this defines $p(y|x)$ to be

$$p(y|x) = \prod_t \frac{\exp\left(-\sum a_{ts}x_s\right)\left(\sum a_{ts}x_s\right)^{y_t}}{y_t!}. \tag{2}$$

It is important to note that this model involves no assumptions, other than those stated and the effects of the spatial discretization of body space. It is an unusual luxury for probability model based statistical analysis to be built on a model that has such strong justification. The only difficulties lie in determination of the coefficients, or weights, $\{a_{ts}\}$. The weight a_{ts} is simply, although somewhat circularly, defined to be the mean number of photons detected at t originating from a point $\boldsymbol{u}$, per unit isotope concentration at $\boldsymbol{u}$, integrated over $\boldsymbol{u}$ in the pixel s.

These weights depend on various factors: the geometry of the detection system, the activity of the isotope and exposure time, and the extent of attenuation and scattering between source and detector. It should be noted that this is a very much more complicated situation than arises in positron emission tomography (PET: [24]) where attenuation can be neglected. SPECT is much more important in practice because it is cheaper and uses standard apparatus. (In the U.K., SPECT is probably in use in at least 200 hospitals, compared to perhaps 10 using PET [C. J. Gibson, personal communication].) Numerical experiments on reconstruction algorithms have typically used simulated data, and so the weights a_{ts} have been assumed known. In analyzing real data, the weights must either be constructed with some care, or estimated from separate transmission tomography experiments on the same patient.

The weights used in the examples in this paper are based on a simple model for the physical circumstances in which the data are recorded. Scattering is neglected, and attenuation considered to be at a constant rate per unit distance within an idealized elliptical boundary representing the patient's body. The rate of emission declines appreciably during the period in which the data are collected, so a decay term is included. A more detailed description of the model for the weights is given in the next section.

The model 1), or equivalently 2), has much wider applicability than the SPECT problem that is the focus of this paper. With appropriate changes to the definitions of pixel s and detector bin t, the model applies to any of the image acquisition technologies that collect particles emitted by radioactive disintegration. In particular, it applies to other uses for the gamma camera, in collecting single projections or sequences of images, and it also applies to PET and even to modern low-intensity optical astronomy. Much of what follows in this paper is therefore also appropriate to other problems, but only SPECT will be considered when we discuss details of practical implementation.

III. Modeling the Weights

In this section, we consider modeling the coefficients $\{a_{ts}\}$ of (1) and (2). The aim is to use simple physical and geometrical arguments to derive values for these constants that are sufficiently accurate for reasonable reconstruction of x, without the need for auxiliary transmission experiments. The values will involve only known physical constants and easily measured dimensions, and do not depend on the data obtained during the scanning of the patient. In a later section, we consider a method for automatically revising these coefficient values following a preliminary analysis of the data.

Recall that a_{ts} represents the mean number of photons recorded at bin t, per unit concentration of the isotope at pixel s. It is therefore influenced by a number of factors as follows:

a) activity of isotope,
b) exposure time,
c) rate of radioactive decay,
d) rate of elimination of the radiopharmaceutical from the organ,
e) spatial correspondence between detector and pixel,
f) detector geometry,
g) absorption of photons,
h) photon scattering.

Note that d), g), and h) depend on the patient, and so in the absence of auxiliary data can at best be approximated. In the present paper, we shall neglect a), b), d), and h). The effect is to assume that each projection has the same exposure time, that elimination is negligible within the duration of the data acquisition, that scattering can be neglected, and finally means that our reconstruction has an arbitrary scale: some external calibration is needed to express the reconstructed x in physical units. It would not be difficult to remove some of these restrictions. The remaining factors are accounted for by assuming that each a_{ts} is the product of three factors:

i) the proportion of radioactivity that has not decayed by the time at which photons are collected in detector t,
ii) the solid angle of view of the center of pixel s into detector t, which is treated as a cylindrical tube of known length and radius, and
iii) the proportion of emissions that survive attenuation.

These considerations apply for three-dimensional reconstruction from a sequence of two-dimensional projections. Our numerical examples will address the usual smaller problem obtained by ignoring the third dimension: the spatial component parallel to the axis of rotation. We thus reconstruct each section or slice of the patient, perpendicular to this axis, from the one-dimensional projections corresponding to that slice. It is then necessary to account for the third dimension: for example, emissions outside a slice may be registered within that slice. The approximation made here is that successive slices are sufficiently similar that they can be assumed equal: x_s therefore denotes the concentration throughout a narrow prism parallel to the axis of rotation. The solid angle in ii) is then replaced by its integral over the third dimension.

Representing the attenuation process is the most difficult aspect of modeling the weights. In PET, attenuation is usually neglected, partly because the radiation is of a different energy where absorption is less, and partly because to first order its effect is eliminated by the geometry of the situation. The literature on SPECT contains many suggestions for the approximate elimination of the attenuation effect, including arithmetic or geometric averaging of opposing projections. We are taking the different approach of attempting to model the effect from geometrical considerations.

It is well known that attenuation operates multiplicatively, so that the proportion of photons surviving attenuation along a line L is of the form $\exp\left(-\int_L \mu(u)\, du\right)$; the linear attenuation rate μ varies markedly with the medium, and in particular has different values in bone, soft tissue, and air. The approximation we will make is that μ has a constant value $0.099\ \mathrm{cm}^{-1}$ (C. J. Gibson, personal communication) within the patient's body, and 0 elsewhere. Further, the body is approximated as a cylinder with elliptical cross section, parallel with the tomographic axis. The dimensions of the ellipse can be easily measured from the patient, or estimated from a trial reconstruction from the data, presuming that the body outline can be readily distinguished.

Calculation of our approximate weights thus reduces to derivation of some simple but tedious trigonometrical formulas, and substitution into these of a dozen or so physical constants and measurements. Precise details can be found in a computer program available from the author.

Notice that since the patient will not typically be perfectly centered on the tomographic axis, there are no symmetries among the $\{a_{ts}\}$ that can be exploited to save storage or computing time, even under the simple model described above.

There are few previous published treatments of these issues. Among most papers that have explicitly used the model 1), there seems to be either no detail given about the construction of the weights, or a side-stepping of the whole matter by using only artificial projection data generated either precisely from the model, or by simulating from the continuous Poisson model that underlies it. Attempts to model the weights realistically include the papers by Floyd, Jaszczak, and Coleman [4] and Veklerov, Llacer, and Hoffman [30], which both use Monte-Carlo methods to generate values for the weights. We believe that it is essential to attempt to model the weights properly, perhaps in the way suggested above and, further, that the weight values will in general need to be modified in light of the data, following diagnostics generated during reconstruction (see Section VIII).

IV. Reconstruction Without Prior Information

There are many approaches to the reconstruction for SPECT that do not make use of any prior information about x, whether represented by a prior distribution $p(x)$ or in some other way. The standard method of filtered back-projection does not even use the model 2) for $p(y|x)$, but rather operates entirely linearly on the data to create an estimate of x of the form

$$\hat{x}_s = \sum_{\delta,\theta} b_{s\delta\theta} \sum_{\epsilon} f_{\delta\epsilon} y_{\epsilon\theta}$$

where we have written t as (ϵ, θ), separating the coordinates of pixel within a projection and angle. The coefficient $f_{\delta\epsilon}$ represents the filter applied to each projection, and $b_{s\delta\theta}$ the back-projection operator. Both of the coefficient arrays b and f enjoy a considerable number of symmetries. Construction of these coefficients does not involve acknowledging either the Poisson variation underlying the data, or attenuation: rather they are chosen so that the quality of the reconstruction is not much affected by ignoring such details. Since it takes such little account of the physical processes underlying emission tomography, it is remarkable how effective filtered back-projection is. The chief problems with the resulting reconstructions are that they tend to have both radial and ring-like artifacts, that they underestimate the isotope concentration in the middle of the body, and that they appear noisy (see, for example, [30]).

In pioneering work directed primarily at the PET problem, Shepp and Vardi [24] and Lange and Carson [31] independently proposed using model 1) explicitly. They suggested reconstructing x using the principle of maximum likelihood estimation (MLE), and developed an EM algorithm for iterative approximation of the MLE solution: that $\hat{x} = \hat{x}(y)$ maximizing $p(y|x)$ as given in (2), subject to the physically-necessary positivity condition $x_s \geq 0$ for all s. This approach was further developed by Vardi, Shepp, and Kaufman [28]. There has been some ambiguity about the uniqueness of the MLE. Of course if T, the number of bins, is less than S, the number of pixels, the solution $\hat{x}$ cannot be unique; in practice it will typically not be unique even when T is somewhat more than S. The EM algorithm still converges to *a* maximum of (2); the set of maxima is convex. Choice of the initial estimate affects the final estimate and "is somewhat akin to a choice of a Bayes prior" [28]. However, Shepp and Vanderbei [23] argue that, with photon counts typical of real emission tomography experiments, the maximum is rendered essentially unique by the positivity condition. The EM algorithm will be discussed in more detail in Section VI: it underlies our own proposal for reconstruction.

It is common experience that MLE reconstructions have the unattractive feature of being very noisy in appearance. (In fact, the MLE is never achieved by the EM approach, as it converges so slowly, but at the point where the EM iterations are abandoned as changing the current estimate of x so little, the noisiness in x is typically *increasing*.) Vardi, Shepp, and Kaufman suggest that "either a slightly smoothed version of the MLE or, alternatively, an EM reconstruction that starts . . . uniform . . . and is run for a limited number of iterations . . . gives very good reconstructions." We find the latter alternative unappealing: developing the former, but based on maximizing the posterior probability, is the main point of this paper.

The idea of stopping the EM iteration at a point determined by a statistical significance test is explored by Veklerov and Llacer [29]. The recent literature has contained a number of ideas for improving the reconstruction yielded by the EM approach, including papers by Snyder and Miller [26], Lange, Bahn, and Little [15], Levitan and Herman [17], Liang and Hart [19], Tanaka [27], and Silverman, Jones, Wilson, and Nychka [25]. Connections between some of these and our proposed approach are drawn in Section VI.

Other methods based on (1) are discussed by Vardi, Shepp, and Kaufman, including least-squares estimation, Stein type estimates, and moment estimates including filtered back-projection. We do not discuss these further here.

Statisticians recognize (2) as a simple example of a *generalized linear model* [20], determined by the Poisson error model and identity link function. Such models are conventionally fitted by an iterative scheme known as the method of scoring, a modification of the Newton–Raphson procedure, but of course the scale of the present problem renders such matrix methods impractical. Further, such an approach, in common with some of the others mentioned, needs nontrivial modification to allow imposition of the positivity constraints.

V. Modeling Prior Information

Selection of a prior distribution $p(x)$ for the isotope concentration x should properly involve consideration of the physician's expectations regarding the spatial distribution of isotope concentration in the organs of interest, and indeed the shapes and sizes of those organs. Specification of such information in probability form seems a challenging task, involving modeling at a very high level. Fortunately, it is now often considered in Bayesian image analysis that long-range, high-level structures in the prior may not be important: the posterior $p(x|y)$ is chiefly sensitive only to local properties of $p(x)$. Put another way, suppose that we have two candidate prior distributions, $p(x)$ and $\tilde{p}(x)$, that are very different, yet the local conditional probability $p(x_A | x_{\nu A})$ is similar to $\tilde{p}(x_A | x_{\nu A})$ whenever A is a small set of pixels and νA a set of pixels near to A. Then $p(x | y)$ will be close to $\tilde{p}(x | y)$ in most important respects. This statement is deliberately made imprecisely here, but can be made more rigorous (for example, see Green, in discussion of Besag [2]).

We can therefore concentrate on local properties only. Following Geman and McClure [6], [7], we will in fact only allow pairwise interactions, among pairs of pixels that are neighbors. Specifically, we suppose that

$$p(x) \propto \exp\left(-\beta V(x)\right) = \exp\left(-\beta \sum_{s,r} w_{sr} \phi\left(\frac{x_s - x_r}{\delta}\right)\right) \tag{3}$$

where β and δ are parameters, w_{sr} is a weight coding the strength of neighborliness between pixels s and r: $w_{sr} = 1$ if s and r are orthogonal nearest neighbors, $\sqrt{1/2}$ for diagonal neighbors, and 0 otherwise. The function ϕ is nonnegative and symmetric about 0, and monotonically increasing for positive values of its argument. For example, Geman and McClure choose $\phi(u) = (1 + u^{-2})^{-1}$, and we will use $\phi(u) = c_1 \log\cosh(c_2 u)$ where c_i are chosen to match Geman and McClure's function by taking $c_1 = 27/128$ and $c_2 = 16/3\sqrt{3}$; this matching is derived by arranging that max ϕ' and $\phi''(0)$ coincide for the two functions: they are then very close for all u with $|u| < 1$.

The distribution (3) is in fact improper if the range of values of $\{x_s\}$ is unbounded, but this need not concern us: it nevertheless gives rise to a proper posterior distribution.

The log cosh function generates quite a flexible family of prior distributions: in particular, if β and $\delta \to \infty$ in such a way that $\beta\delta^{-2} \to \kappa$, then $\beta\phi(u/\delta) \to \kappa u^2$ for all u, while if β and $\delta \to 0$ such that $\beta\delta^{-1} \to \kappa$, then $\beta\phi(u/\delta) \to \kappa|u|$. The former case gives a Gaussian prior, the latter corresponds with a proposal due to J. Besag (personal communication). The log cosh function may be closely approximated by a curve that is piecewise quadratic and linear for $|u| < 1$ and ≥ 1, respectively; there may be computational advantages in such an approximation. An important property of the function and this approximation, not shared by Geman and McClure's function, is convexity of $\log p(x)$ and, hence, $\log p(x | y)$, with obvious numerical advantages in optimization. However, which function is more appropriate depends on the typical pattern of isotope concentrations in the organ in question.

For any ϕ, the form of (3) is such as to treat equal differences in neighboring x values equally at all points of the scale: it may be more appropriate first to change the x scale by applying a logarithmic or power transformation, but we do not consider this here. As in other applications of Bayesian image analysis, there is potential for employing a richer class of priors, for example, incorporating edge sites [5], or allowing β and δ to vary stochastically across the scene ([2] and discussion).

VI. EM Algorithm for Bayesian Reconstruction

Our aim is to choose $x = x(y)$ to maximize the posterior probability

$$\begin{aligned}\log p(x|y) &= \log p(y|x) + \log p(x) + \text{constant}\\ &= \sum_t \left(y_t \log\left(\sum_s a_{ts}x_s\right) - \sum_s a_{ts}x_s\right)\\ &\quad - \beta V(x) + \text{constant}\end{aligned} \tag{4}$$

where $V(x)$ is given by (3). Since this objective function is not quadratic, some iterative algorithm will be needed, and the large size of the problem means that conventional Newton or gradient methods are not practicable. We will use the EM algorithm instead [3]. The EM algorithm is a general approach for maximizing a likelihood or posterior distribution when some of the data are "missing" in some sense, and observation of that missing data would have greatly simplified the estimation of parameters. In the present case, data are "missing" not because of any censoring or misrecording, but because of the superposition of the Poisson streams of photons. This means that we cannot observe the more basic data $\{z_{ts}\}$, the number of photons recorded at bin t *emitted from pixel s*. Equivalently, what is missing is a label on each recorded photon stating its source in body space. The observed data are $y_t = \Sigma_s z_{ts}$.

Let us begin by reviewing the application of the EM algorithm to reconstructing x without using the prior term, that is, by maximizing the log-likelihood $p(y|x)$ alone. The result will be the maximum likelihood reconstruction $\tilde{x}$. If the z_{ts} *were* observed, then $\tilde{x}$ would be very easily calculated. Since the Poisson model 1) holds good for the disaggregated data:

$$z_{ts} \sim \text{Poisson}\ (a_{ts}x_s), \text{ independently,} \tag{5}$$

we immediately find

$$\tilde{x}_s = \frac{\sum_t z_{ts}}{\sum_t a_{ts}}. \tag{6}$$

The $\{z_{ts}\}$ can be estimated using another result from elementary probability theory, so that

$$z_{ts} = \frac{y_t a_{ts} x_s}{\sum_{s'} a_{ts'} x_{s'}} \tag{7}$$

which is just its conditional expectation given y. Since the Poisson distributions (5) form a linear exponential family, the $\{z_{ts}\}$ are complete-data sufficient statistics [3], so that the EM algorithm for estimating $\tilde{x}$ from y is obtained by combining (6) and (7) to give

$$\tilde{x}_s^{\text{new}} = \frac{\tilde{x}_s^{\text{old}} \sum_t y_t a_{ts} \Big/ \left(\sum_{s'} a_{ts'} \tilde{x}_{s'}^{\text{old}}\right)}{\sum_t a_{ts}}$$

as an iteration producing the updated estimate $\hat{x}^{\text{new}}$ from $\hat{x}^{\text{old}}$. This iteration is repeated until apparent convergence. This derivation is essentially that of Shepp and Vardi [24], and we have already discussed in Section V some of the shortcomings of the resulting reconstruction. One rather attractive feature of the method, however, is that the positivity constraints are automatically satisfied, providing the initial estimate is entirely positive.

It is interesting to note that the EM algorithm for this problem can be seen as an iterative forward/backward projection technique. Calculation of the fitted values $\Sigma_s a_{ts}x_s$ corresponds to forward projection, and their use in (6) and (7) to reestimate x_s is backward projection.

Our present objective is to use a similar approach to construct the maximum posterior probability reconstruction $\hat{x}$. As noted by Dempster, Laird, and Rubin, the EM algorithm can still be appropriate: it will increase the posterior probability at each stage, and convergence can usually be established if $p(x \mid y)$ is convex in x.

There is no change to what is treated as the missing data, so that the E step given by (7) remains the same. The M step should now maximize $\log p(x \mid z)$, the log posterior probability of x given the missing data z. This is as follows:

$$\sum_{t,s} \left(z_{ts} \log \left(a_{ts}x_s\right) - a_{ts}x_s\right) - \beta V(x) + \text{constant}$$

and so, after differentiating, we have to solve

$$\frac{\sum z_{ts}}{x_s} - \sum a_{ts} - \beta \frac{\partial}{\partial x_s} V(x) = 0. \tag{8}$$

In contrast to the nonBayesian case where $\beta = 0$, direct solution of this is completely impractical, except in the degenerate case where the pixels are independent under the prior. Instead, we will use a "one-step-late" (OSL) approximation, as proposed by Green [10] to facilitate an EM-type approach to quite general maximum penalized likelihood or posterior probability problems. We evaluate the partial derivatives of $V(x)$ in (8) at the *current* estimate $\hat{x}^{\text{old}}$, and thus have the simple updating equations

$$\hat{x}_s^{\text{new}} = \frac{\sum z_{ts}}{\sum a_{ts} + \beta \dfrac{\partial}{\partial x_s} V(x)\Big|_{\hat{x}^{\text{old}}}}. \tag{9}$$

Solving these involves only trivially more computing effort than when $\beta = 0$, so it is entirely practicable.

We cannot give a proof that the OSL algorithm converges in general, as this will depend on the form of the function $V(x)$ and the value of β. However, empirical evidence suggests that it usually converges when V is as defined in Section V and β is not too large. The algorithm does at least have a fixed point justification: if it converges, to $\hat{x}$ say, then $\hat{x}$ is a solution to (8), and hence also to $(\partial/\partial x) \log p(x \mid y) = 0$.

The rates of convergence of the EM and OSL algorithms have been discussed by Green [10]. Asymptotically, the error $\hat{x}^{\text{new}} - \hat{x}$ declines geometrically with the iteration number for the OSL algorithm, at a rate given by the spectral radius of $(B + C)^{-1}(C - \beta K)$. Here B is the observed information matrix, C is the expectation of the additional information in the missing data, given the observed data, and K is the matrix of second-order partial derivatives of $V(x)$; all are evaluated at $x = \hat{x}$. The convergence rate for the (impracticable) EM method is similarly determined by $(B + C + \beta K)^{-1}C$, and is slightly slower, at least for small β.

Elementary calculations show that for the Poisson model 1), B has entries

$$B_{rs} = \sum_t \frac{y_t}{\hat{\mu}_t^2} a_{tr}a_{ts}$$

where $\hat{\mu}_t = \Sigma_s a_{ts}\hat{x}_s$, and $(B + C)$ is diagonal with

$$(B + C)_{ss} = \hat{x}_s^{-1} \sum_t \frac{y_t a_{ts}}{\hat{\mu}_t}.$$

It is readily established that B and C are positive definite, so when $\beta = 0$, the required spectral radius is $1 - \lambda_1$ where λ_1 is the smallest eigenvalue of $(B + C)^{-1}B$, or equivalently of the symmetric matrix $(B + C)^{-1/2}B(B + C)^{-1/2}$. Finding the spectrum of this matrix seems far from straightforward, but of course λ_1 can be characterized by

$$\lambda_1^2 = \inf_{v \neq 0} \frac{v^T B(B + C)^{-1} Bv}{v^T(B + C)v}.$$

Thus, upper bounds on λ_1 can be constructed by evaluating the right-hand side of this expression for various vectors v. This calculation is straightforward, given the forms for B and $(B + C)$ quoted above. The bound can be expressed in the form

$$\frac{\sum_r c_r \left(\sum_t w_{rt}\left(\sum_s a_{ts}v_s\right)\Big/\left(\sum_s a_{ts}\hat{x}_s\right)\right)^2}{\sum_r c_r(v_r/\hat{x}_r)^2}$$

where $c_r = \hat{x}_r \Sigma_t y_t a_{tr}/\hat{\mu}_t$ are positive numbers, and $w_{rt} = (y_t a_{tr}/\hat{\mu}_t)/\Sigma_{t'}(y_{t'} a_{t'r}/\hat{\mu}_{t'})$ are positive weights summing to 1. Thus, it is evident that the bound will be small when $\|\Sigma_s a_{ts}v_s\|/\|\Sigma_s a_{ts}\hat{x}_s\|$ is much smaller than $\|v\|/\|\hat{x}\|$, and this will tend to occur when elements of v alternate abruptly in sign. Calculations based on the circumstances of the data discussed in Section IX showed that a particularly small bound on λ_1^2 was obtained when $v_s = x_s^3 \epsilon_s$ where ϵ_s consisted of a checkerboard of alternating ± 1's. This led to a lower bound on the spectral radius of 0.99938, confirming the extremely slow rate of convergence observed in practice.

Similar calculations were performed for the OSL method using the log cosh prior distribution from Section V, and a range of positive β and δ. The resulting lower bounds are given in Table I, and support empirical evidence that convergence is appreciably faster after the introduction of the prior.

TABLE I
Lower Bounds on Convergence Rate $1 - \lambda_1$

β	δ	bound
0.0	-	0.99938
80	1000	0.94341
0.3	50	0.93638
0.2	70	0.97586
0.2	50	0.95758
0.2	30	0.90772
0.1	50	0.97878
0.0075	5	0.96562

Other approaches to maximizing the posterior probability (4) have appeared in recent literature. Although our formulation of the problem is almost identical to that of Geman and McClure [6], [7], we have followed a different approach to its solution. They have proposed a number of stochastic algorithms for reconstructing x. The form of the posterior distribution (4) is such as to give the *posterior local characteristics* $p(x_s | y, \{x_r, r \neq s\})$ a particularly simple form. This allows them to simulate a Markov chain on the x space whose equilibrium distribution is $p(x|y)$, and so can be used to estimate $E(x|y)$. Similarly, annealing can be used to iterate towards the x maximizing $p(x|y)$, as in Geman and Geman [5]. Practical considerations have led to compromises employing deterministic neighbors of these algorithms, including Iterated Conditional Modes [2] and gradient descent. One advantage of the stochastic algorithms is that they can be adapted to yield estimates of β, with the rather impressive results to be seen in Geman and McClure [7].

Adaptation of the EM approach to the Bayesian reconstruction problem has been previously suggested by several authors. Liang and Hart [19] restrict attention to the use of priors for which the EM updating equations (8) have an explicit solution, after some approximation. Levitan and Herman [17] suppose that the prior is Gaussian and indeed, in presenting an explicit algorithm, assume that the pixels are independent; Lange, Bahn, and Little [15] also use various prior distributions in which pixels are independent. While this restriction gives rise to explicitly soluble EM updating equations, we do not believe that it can model prior belief about x adequately, and thus cannot satisfactorily "smooth" the data.

Silverman, Jones, Wilson, and Nychka [25] modify the EM algorithm for the unpenalized problem by including an additional smoothing step in each cycle. This is intuitively appealing, but it seems difficult to give it an objective justification. They are able to relate their approach approximately to a penalized likelihood method that uses a certain penalty which is quadratic in the square roots of x; for more information on this connection, see [10].

In none of the papers cited above, with the exception of those by Geman and McClure, is there any analysis of real tomographic data.

VII. Improving the One-Step-Late EM Algorithm

The derivation of the updating algorithm defined in (9) is somewhat naive from a numerical-analytic point of view. Surprisingly, we shall see in the examples in Section IX that it can give quite acceptable performance, but it is worth considering some simple methods for improving it. Further work is needed in this area: it is the subject of current research, and will appear elsewhere.

As we stated in Section VI, the one-step-late approximation to the EM algorithm converges *more* quickly than the true EM iteration, when the prior parameter β is sufficiently small. When this is not the case, one of the only practical ways of achieving the EM algorithm is by an inner iteration that updates the calculation of the derivatives of V, but not the missing data z_{ts}. Specifically, given $\hat{x}^{\text{old}}$, let $z_{ts} = y_t a_{ts} \hat{x}_s^{\text{old}} / \Sigma_{s'} a_{ts'} \hat{x}_{s'}^{\text{old}}$, and set $\hat{x}^{\text{new},0} = \hat{x}^{\text{old}}$. Then for $m = 1, 2, \cdots$, iterate on

$$\hat{x}_s^{\text{new},m} = \frac{\sum z_{ts}}{\sum a_{ts} + \beta \dfrac{\partial}{\partial x_s} V(x) \big|_{\hat{x}^{\text{new},m-1}}} \tag{10}$$

until convergence, then set $\hat{x}^{\text{old}} = \hat{x}^{\text{new},m}$, and recalculate z_{ts}, and repeat. Some limited experience suggests that this modification does not speed convergence to $\hat{x}$; it seems better to spend the extra computing effort on another one-step-late update.

Generally, the EM method is notorious for generating algorithms that are slow to converge. At a typical stage, $\hat{x}^{\text{new}}$ is a poor approximation to $\hat{x}$. However, if successive iterations tend to lie in approximately the same direction in the space of all x it may be that $\hat{x}^{\text{new}}$ is a good guide to the direction in which $\hat{x}$ lies from $\hat{x}^{\text{old}}$, so that it is worth searching along the line

$$\hat{x}^{\theta} = \hat{x}^{\text{old}} + \theta(\hat{x}^{\text{new}} - \hat{x}^{\text{old}}) \tag{11}$$

and optimizing by maximizing $p(\hat{x}^{\theta}|y)$ over θ. This search could be implemented by a bisection method, or by quadratic extrapolation, etc. Such modifications to the EM procedure for MLE in PET are considered in much detail by Kaufman [13], whose experience, with which ours agrees, is that such modifications are well worthwhile. Such searches will usually involve examining $\hat{x}^{\theta}$ for some θ outside the interval $[0, 1]$, and if so, care must be taken that the elements of $\hat{x}^{\theta}$ remain nonnegative. The range of valid θ is easily determined by inspecting (11). This matter has also been investigated by Lange, Bahn, and Little [15]; see also the references therein. The simpler option of simply scaling up the EM step by a fixed multiplier θ has been proposed by Lewitt and Muehllehner [18]. Algorithms like (11) have also been proposed for an EM solution to the problem of estimating finite mixture distributions; see Peters and Walker [22].

VIII. Diagnostics, and Improving the Model

A further advantage of our approach of treating reconstruction as a statistical estimation problem is that a whole battery of diagnostic techniques are immediately available to investigate the adequacy of the estimated reconstruction, and thereby of the model underlying it. The log-likelihood ratio statistic or *deviance* [20] is defined for the present problem as

$$D = 2\sum_t \left(y_t \log \frac{y_t}{\sum a_{ts}\tilde{x}_s} - y_t + \sum a_{ts}\tilde{x}_s \right).$$

In regular estimation problems, with x estimated by maximum likelihood, this would have a χ^2 distribution on $T - S$ degrees of freedom under the null hypothesis that the model 1) was correct. Here the standard distribution theory does not apply, but we would nevertheless expect that the realized value of D should be substantially less than T, the number of counts, if the model were satisfactory. We recommend that both the deviance D and the log-posterior $p(x \mid y)$ be monitored as iteration proceeds.

More specific diagnostics can be obtained by examining residuals. Since $y_t \sim \text{Poisson}(\mu_t)$ where $\mu_t = \Sigma a_{ts}x_s$, y_t has both mean and variance equal to μ_t, and therefore $(y_t - \mu_t)/\sqrt{(\mu_t)}$ is standardized. Of course, μ is unknown, but is estimated by $\hat{\mu}_t = \Sigma a_{ts}\hat{x}_s$, and the quantities $(y_t - \hat{\mu}_t)/\sqrt{(\hat{\mu}_t)}$ are *residuals* that approximately have mean 0 and variance 1. Further, apart from small correlations induced by estimation, they should only reflect the Poisson variation in the data, and therefore appear random and unstructured. Any pattern reflects some failure of the model 1). Since the Poisson assumption is difficult to criticize, pattern in residuals is pointing to deficiency in the model for the weights $\{a_{ts}\}$. It would be surprising if such deficiencies did not exist.

When a clear pattern is seen, it can be used to modify the model for the weights. Suppose the weights *should* be $c_t a_{ts}$ where a_{ts} are as already determined. Then $y_t \sim \text{Poisson}(c_t \Sigma a_{ts}x_s)$, and if x were known, we would estimate c_t by $y_t/\Sigma a_{ts}x_s$. In fact, we have only an estimate of x, but assuming that the c_t vary smoothly in projection space (otherwise a pattern in the residuals would not have been evident) we can estimate c_t by $(Sy)_t/(S\hat{\mu})_t$ where S is a suitable smoother in projection space, and $\hat{\mu}$ is the result of a preliminary reconstruction $\hat{x}$. This idea works well in practice; see Section IX.

IX. Implementation and Examples

The OSL method has been implemented on a Sun 3/160 workstation. The algorithm is straightforward and the program has no special features of note except perhaps regarding the representation of body space and the weight matrix (a_{ts}).

The method as described makes no assumption about the pixelization of body space: for the SPECT problem, we prefer to retain the conventional rectangular grid, rather than one of the circular grids advanced by Kearfott [14], Kaufman [13], or Silverman, *et al.* [25]. Unlike the PET problem, SPECT offers no symmetries to be exploited, as the attenuation pattern is determined by the patient's body. Further, the prior distribution $p(x)$ cannot be specified in a spatially stationary way when the grid is not regular, so there is a risk of artifacts being introduced.

In our analysis of head sections we generally reconstruct on a 48 × 48 grid of 0.55 cm squares from 64 projections each of 52 counts. With these dimensions, only about 10 percent of the weights $\{a_{ts}\}$ are nonzero, and so a sparse matrix representation is used: the weights are stored columnwise together with pointers to the detectors t concerned. This choice of representation has implications for the order of looping over s and t in the OSL update; see also [13, sect. II].

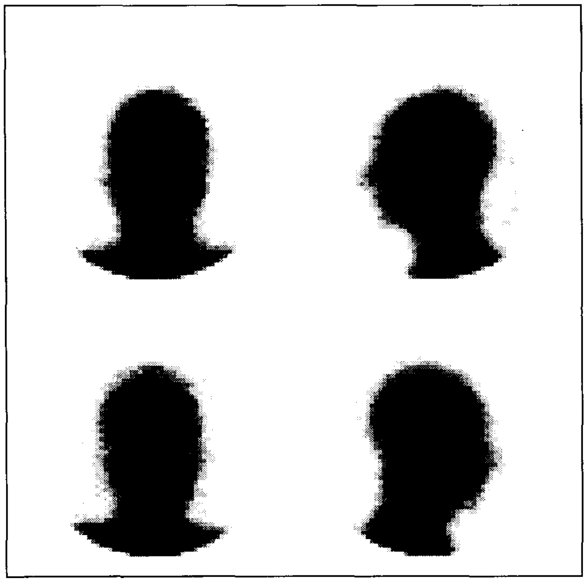

Fig. 1. Four of the 64 projections of raw data, each 64 × 64.

We will illustrate the OSL method applied to data kindly made available by Dr. C. J. Gibson of the Northern Regional Medical Physics Department, Durham. Fig. 1 shows 4 of the 64 projections. We will concentrate on a horizontal slice, obtained by aggregating the 29th and 30th rows from the top of these pictures. The relevant data are shown in Fig. 2 as a sinogram, the horizontal axis representing angle of view and the vertical one linear displacement: note that only data from the central 52 detectors are being used, to reduce the computing time. There are a total of 219123 photons in the data displayed in this image. The EM algorithm was run for 16 complete sweeps starting from a "flat" image: and using no prior term. The result was the trial reconstruction shown in Fig. 3. At this point, residuals were calculated, as defined in Section VIII; these are displayed in Fig. 4 and reveal a pronounced pattern. Discussions with Dr. Gibson suggested that this might be caused by absorption of photons within the couch supporting the patient. The weights $\{a_{ts}\}$ were modified as described in Section VIII, and the algorithm ran for a further 128 iterations, using the log cosh function, with $\beta = 0.2$ and $\delta = 50.0$. At this point further changes were imperceptible. The resulting reconstruction is shown in Fig. 5; the corresponding residual pattern is in Fig. 6, and shows little cause for concern. The deviance associated with this reconstruction is 2275. In Fig. 7, we compare four different reconstructions, one (top left) being a filtered back-projection reconstruction supplied with the data. The other three were obtained in the same

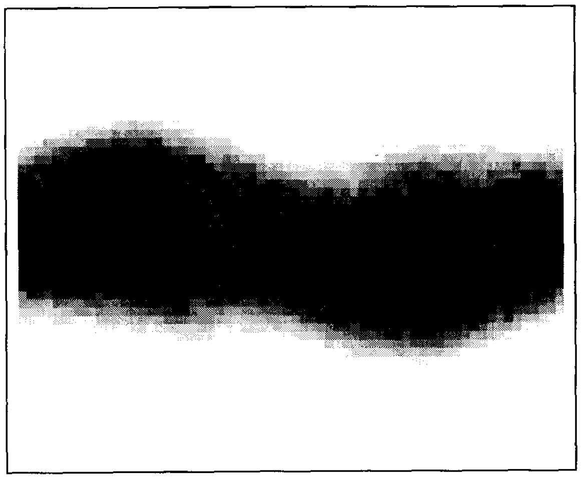

Fig. 2. Sinogram slices 29 and 30 from Fig. 1.

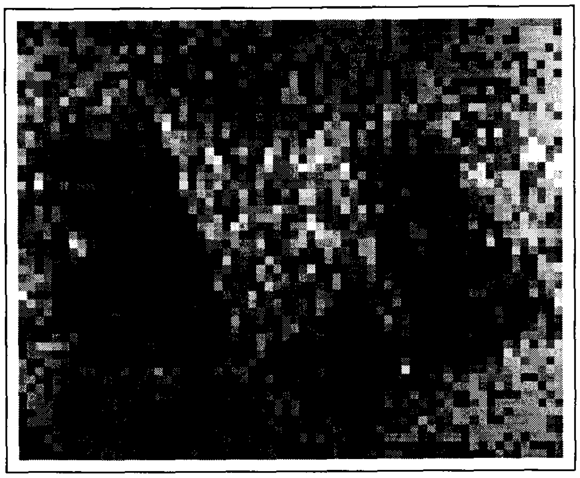

Fig. 4. Residuals corresponding to Fig. 3.

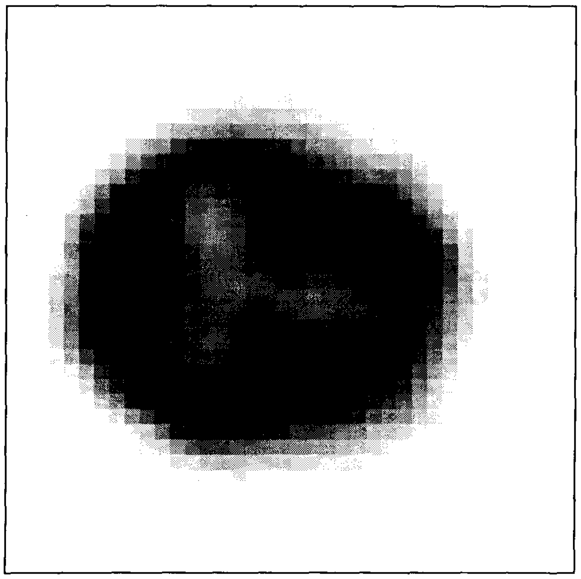

Fig. 3. Trial reconstruction after 16 EM iterations.

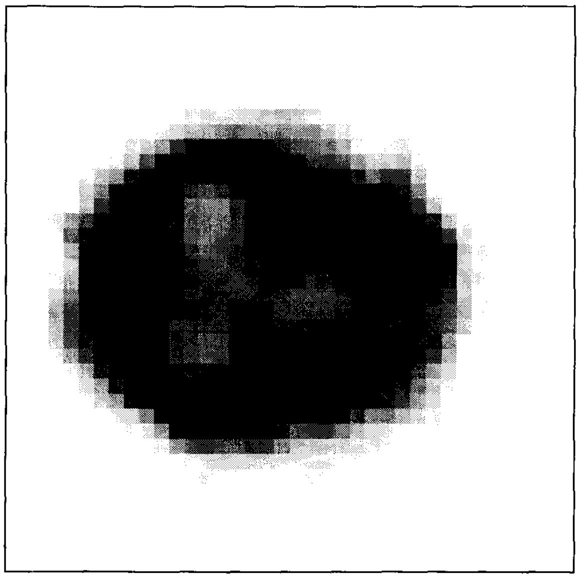

Fig. 5. An OSL reconstruction after 16 + 128 iterations.

manner as Fig. 5, but with varying choices for β and δ, namely: (top right) $\beta = 0$, approximating the ML reconstruction; (bottom left) $\beta = 0.0075$, $\delta = 5$, close to an "absolute value" prior; and (bottom right) $\beta = 80$, $\delta = 1000$, close to a Gaussian prior. The deviances for these three were 2253, 2266, and 2290, respectively. Note how the characteristics of the reconstruction change with the specification of prior information. None of these is considered to be satisfactory. The first is noisy and has spurious radial artifacts. The ML reconstruction is noisy, and this would get worse had the iteration been continued. The next is too discrete in the intensity scale, and the fourth has over-smoothed transitions in isotope concentrations.

In all of these examples, the resolution attainable in the reconstruction is limited by the coarseness of the grid on which the data are collected: as can be seen from Fig. 2, the maximum diameter of the head is about 30 pixels (each 0.625 cm across). Fine detail is inevitably obscured. Some of the remaining lack of clarity in the reconstruction can probably be attributed to the neglect of photon scattering in our model for the coefficients, and this could be remedied without affecting the proposed method. Note that in Figs. 3, 5, and 7 the images have been cropped to remove uninteresting peripheral background.

Further work is needed to fine-tune the approach to various practical circumstances, but our present conclusion is that, for these data, when $\beta = 0.2$ and $\delta = 50.0$ in the log cosh prior distribution, the OSL method has successfully yielded a "cleaner," more interpretable, reconstruc-

Fig. 6. Residuals corresponding to Fig. 5.

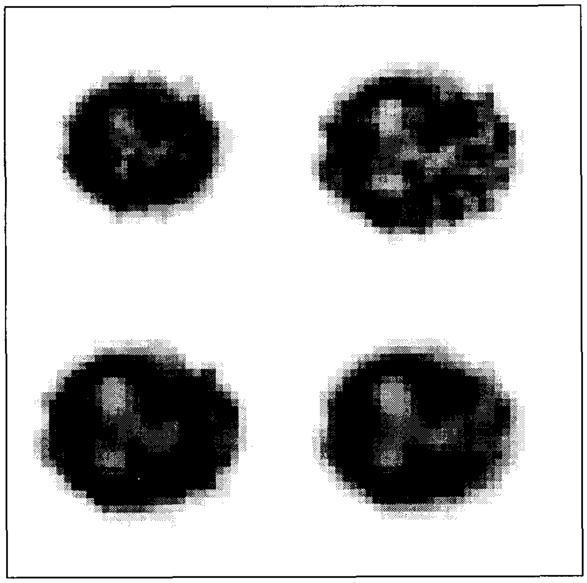

Fig. 7. Four unsatisfactory reconstructions.

tion. It is smoother generally, yet retains abrupt gradients in the estimated isotope concentrations where these are suggested strongly by the data.

Acknowledgment

The author's thanks are due to L. Shepp, Y. Vardi, L. Kaufman, D. Nychka, and J. Besag for invaluable discussions, and to C. Gibson for providing the data and much practical advice.

References

[1] J. Besag, "Discussion of paper by P. Switzer," *Bull. Int. Stat. Inst.*, vol. L-3, pp. 422-425, 1983.

[2] ——, "On the statistical analysis of dirty pictures (with discussion)," *J. Roy. Statist. Soc., B*, vol. 48, pp. 259-302, 1986.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Statist. Soc, B*, vol. 39, pp. 1-38, 1977.

[4] C. E. Floyd, R. J. Jaszczak, and R. E. Coleman, "Inverse Monte-Carlo: A unified reconstruction algorithm for SPECT," *IEEE Trans. Nucl. Sci.*, vol. NS-32, pp. 779-785, 1985.

[5] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721-741, 1984.

[6] S. Geman and D. E. McClure, "Bayesian image analysis: An application to single photon emission tomography," in *Proc. Amer. Statist. Assoc., Stat. Comp.*, 1985, sect., 12-18.

[7] ——, "Statistical methods for tomographic image reconstruction," ISI Tokyo session. *Bull. Int. Stat. Inst.*, vol. LII-4, pp. 5-21, 1987.

[8] I. J. Good and R. A. Gaskins, "Non-parametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255-277.

[9] P. J. Green, "Statistical methods for spatial image analysis," Invited discussion contribution, ISI Tokyo session. *Bull. Int. Stat. Inst.*, vol. LII-4, pp. 43-44, 1987.

[10] ——, "On use of the EM algorithm for penalized likelihood estimation," *J. Roy. Statist. Soc., B*, 1990.

[11] R. J. Jaszczak, R. E. Coleman, and C. B. Lim, "SPECT: Single photon emission computed tomography," *IEEE Trans. Nucl. Sci.*, vol. NS-27, pp. 1137-1153, 1980.

[12] I. M. Johnstone and B. W. Silverman, "Speed of estimation in positron emission tomography," *Ann. Statist.*, to be published.

[13] L. Kaufman, "Implementing and accelerating the EM algorithm for positron emission tomography," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 37-51, 1987.

[14] K. J. Kearfott, "Comment on paper by Vardi, Shepp and Kaufman," *J. Amer. Statist. Assoc.*, vol. 80, pp. 26-28, 1985.

[15] K. Lange, M. Bahn, and R. Little, "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 106-114, 1987.

[16] S. A. Larsson, "Gamma camera emission tomography," *Acta Radiologica*, Supplementum 363, Stockholm.

[17] E. Levitan and G. T. Herman, "A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 185-192.

[18] R. M. Lewitt and G. Muehllehner, "Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation," *IEEE Trans. Med. Imaging*, vol. MI-5, pp. 16-22, 1986.

[19] Z. Liang and H. Hart, "Bayesian image processing of data from constrained source distributions: I and II," *Bull. Math. Biol.*, vol. 49, pp. 51-91.

[20] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *J. Roy. Statist. Soc., A*, vol. 135, pp. 370-384.

[21] D. Nychka, "Some properties of adding a smoothing step to the EM algorithm," *Statist. Probabil. Lett.*, vol. 7, 1990.

[22] B. C. Peters and H. F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions," *SIAM J. Appl. Math.*, vol. 35, pp. 362-378, 1978.

[23] L. A. Shepp and R. J. Vanderbei, "New insights into emission tomography via linear programming," presented at NATO Meet. on Formation, Handling and Evaluation of Medical Images, Portugal, Sept. 1988.

[24] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction in positron emission tomography," *IEEE Trans. Med. Imaging*, vol. MI-1, pp. 113-122, 1982.

[25] B. W. Silverman, M. C. Jones, J. D. Wilson, and D. W. Nychka, "A smoothed EM approach to indirect estimation problems with particular reference to stereology and emission tomography," *J. Roy. Statist. Soc., B*, 1990.

[26] D. L. Snyder and M. I. Miller, "The use of sieves to stabilize images produced with the EM algorithm for emission tomography," *IEEE Trans. Nucl. Sci.*, vol. NS-32, pp. 3864-3872, 1985.

[27] E. Tanaka, "A fast reconstruction algorithm for stationary positron emission tomography based on a modified EM algorithm," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 98-105, 1987.

[28] Y. Vardi, L. A. Shepp, and L. Kaufman, "A statistical model for positron emission tomography (with discussion)," *J. Amer. Statist. Assoc.*, vol. 80, pp. 8-37, 1985.

[29] E. Veklerov and J. Llacer, "Stopping rule for the MLE algorithm based on statistical hypothesis testing," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 313-319, 1987.

[30] E. Veklerov, J. Llacer, and E. J. Hoffman, "MLE reconstruction of a brain phantom using a Monte-Carlo transition matrix and a statistical stopping rule," *IEEE Trans. Nucl. Sci.*, vol. NS-35, pp. 603-607, 1988.

[31] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," J. Comput. Assist. Tomog., vol. 8, pp. 306-316, 1984.