CHAPTER

21

Iterative Image Reconstruction

DAVID S. LALUSH* and MILES N. WERNICK†

**Department of Biomedical Engineering, North Carolina State University, Raleigh, North Carolina and University of North Carolina at Chapel Hill, North Carolina.*

†Departments of Electrical and Computer Engineering and Biomedical Engineering, Illinois Institute of Technology, Chicago, Illinois and Pridictek, LLC, Chicago, Illinois

I. INTRODUCTION

Traditionally, tomographic image reconstruction has been framed as a simple mathematical problem, namely that of inverting a discrete form of the Radon transform, as explained in Kinahan *et al.* (Chapter 20 in this volume). In this traditional approach, it is assumed that the data consist of line integrals of the object distribution, and no attempt is made to model explicitly the randomness of the gamma-ray counting process. This simplified version of the reconstruction problem is solved exactly by filtered backprojection (FBP), a method which allows emission tomography (ET) images to be computed very quickly. Unfortunately, because real ET data are not precisely described by the FBP model, the resulting images can exhibit significant inaccuracies. This is especially true in single-photon emission computed tomography (SPECT), where attenuation can cause severe artifacts if it is not suitably accounted for (see King *et al.*, Chapter 22 in this volume).

Avoiding the shortcomings of FBP requires that the reconstruction problem be framed in a way that more closely resembles reality. Rather than assuming a Radon model, modern reconstruction techniques use a more general linear model that can allow for a rich description of the blurring and attenuation mechanisms in the imaging process. Statistical reconstruction techniques in addition incorporate probabilistic models of the noise and, in the case of Bayesian methods, of the image itself.

The price of these enhancements is that the resulting mathematical problem is more difficult to solve than that of Radon transform inversion. Indeed, the solution generally cannot be written explicitly or, when it can be, the analytic form of the solution is impractical to compute. Thus, most reconstruction algorithms that attempt to incorporate an accurate imaging model are *iterative*, meaning that the estimated image is progressively refined in a repetitive calculation.

The principal trade-off between iterative techniques and FBP is one of accuracy versus efficiency. Iterative algorithms invariably require repeated calculations of projection and backprojection operations. Thus, they can require substantially greater computation time than FBP. Accurate modeling of physical effects in iterative algorithms can improve accuracy, but this added refinement can further compound the processing time. Initially, this disadvantage hampered the transition of iterative techniques from the

research lab to the clinic. However, iterative techniques are now in widespread clinical use, owing to improvements in computer power and the development of efficient modeling techniques and fast reconstruction algorithms.

An additional distinction between iterative methods and FBP is in the appearance of the images they produce; in particular, the noise texture and image detail in an FBP image can look significantly different than that in an iterative reconstruction. Therefore, physicians must take these distinctions into account when interpreting the images.

There is not yet a consensus that iterative reconstructions are always superior to FBP images or, at least, that the benefits of iterative reconstructions always justify the increased computational costs; therefore, the two approaches will probably continue to coexist for some time.

This chapter presents the general principles of iterative image reconstruction and is intended to provide the reader with a starting point for further study. We loosely classify iterative methods into a few major types and provide details on some prominent examples of these major categories, but the reader should bear in mind that there are many variations on each theme in addition to the examples given. This chapter is intended as a tutorial on the basic concepts and not as a comprehensive survey of the literature. Regretfully, some algorithms deserving of more extensive discussion do not receive it, and some algorithms deserving of mention have no doubt been overlooked.

We begin with a formulation of the tomography problem as a linear inverse problem and define the statistical characteristics of the measured data. We then describe the two main components of any iterative method: (1) the criterion for selecting the best image solution and (2) the algorithm for finding that solution.

II. TOMOGRAPHY AS A LINEAR INVERSE PROBLEM

A. Linear Model of the Imaging Process

Any ET reconstruction problem can be formulated as the following estimation problem:

> Find the object distribution **f**, given (1) a set of projection measurements **g**, (2) information (in the form of a matrix **H**) about the imaging system that produced the measurements, and, possibly, (3) a statistical description of the data and (4) a statistical description of the object (Fig. 1).

This problem statement applies equally to PET and SPECT and to all types of hardware configurations—for example, ring systems or dual rotating cameras in PET and parallel, fan, cone, pinhole, or coded-aperture collimation in SPECT.

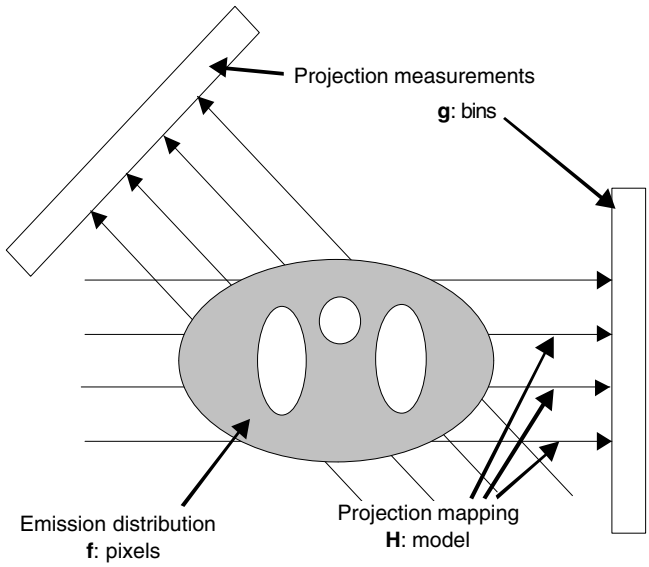

FIGURE 1 A general model of tomographic projection in which the measurements are given by weighted integrals of the emitting object distribution.

Indeed, if we assume that the imaging process is linear, the ET reconstruction problem is essentially similar to any linear inverse problem of the following form:

$$g_i = \int_{\mathbb{R}^D} d\mathbf{x} f(\mathbf{x}) h_i(\mathbf{x}), i = \mathbf{1}, \ldots, P \tag{1}$$

where $\mathbf{x}$ is a vector denoting spatial coordinates in the image domain, g_i represents the ith measurement, and $h_i(\mathbf{x})$ is the response of the ith measurement to a source at $\mathbf{x}$. In two-dimensional (2D) slice imaging, $D = 2$ and $\mathbf{x} = (x, y)$; in three-dimensional (3D) imaging, $D = 3$ and $\mathbf{x} = (x, y, z)$. The point-spread function (PSF) $h_i(\mathbf{x})$ can represent the effects of attenuation and all linear sources of blur. A good example of how various effects can be incorporated in the PSF can be found in Qi *et al.* (1996). For simplicity, we have suppressed additive contributions to the data, such as accidental coincidences (randoms) in PET.

What distinguishes tomography from other problems described by this linear model is that g_i are projections. Here, we use the term loosely to mean that the data are Radon-like, which implies a specific form for $h_i(\mathbf{r})$. In the ideal 2D Radon model on which FBP is based, $h_i(\mathbf{x}) = \delta(\mathbf{s}^T_i \mathbf{x} - t_i)$, in which case the measurements are simply line integrals of the form $g_i = \int_{L_i} f(\mathbf{x}) d\mathbf{x}$, where L_i is the line $\mathbf{s}^T_i \mathbf{x} = t_i$. Real ET data cannot be described by integrals over infinitely thin lines but instead must be thought of as integrals over finite strip- or cone-shaped regions of the object.

For computing purposes, we cannot represent the reconstructed image by a continuous-domain function; instead, we estimate a sampled version of the image, described in a discrete domain by column vector **f** (Fig. 2). Thus, each

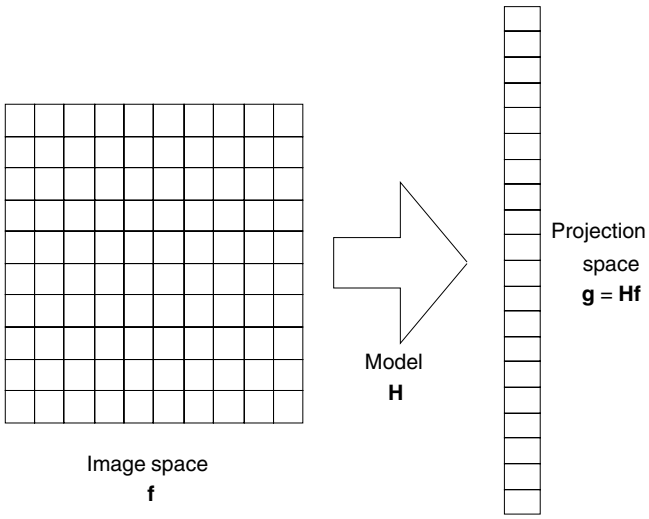

FIGURE 2 A discrete model of the projection process.

measurement in Eq. (1) can be approximated by the following system of linear equations:

$$g_i = \boldsymbol{h}_i^T \mathbf{f}, \quad i = 1, \ldots, P \tag{2}$$

which can be summarized by a single matrix equation as follows:

$$\mathbf{g} = \mathbf{Hf} \tag{3}$$

Here, $\mathbf{h}_i$ is the ith row of $\mathbf{H}$, and each element of $\mathbf{f}$, denoted by f_j, $j = 1, \ldots, N$, represents one pixel in the image space. For our purposes, the ordering of the pixels in $\mathbf{f}$ does not matter, but lexicographic ordering is usually assumed. In this general notation, $\mathbf{f}$ may represent either a 2D slice image or a 3D volume image, and complicated imaging systems can be readily represented within this approach by appropriate definition of $\mathbf{H}$.

We use the word *pixel* to refer to elements of the image domain, although it should be understood to encompass the term *voxel*, which is an element of a volume image. Each pixel is associated with a basis function $\phi_j(\mathbf{x})$, which transforms the continuous-domain function $f(\mathbf{x})$ into pixel values $f_j = \int_{\mathbb{R}^D} f(\mathbf{x})\phi_j(\mathbf{x})d\mathbf{x}$. The most commonly used pixel basis functions are ones that are constant within small, nonoverlapping rectangular (or cubic) regions arranged in a rectangular grid. In this case, the intensity of a pixel is intended to represent the expected number of emissions from within that pixel. However, there are benefits to using other types of basis functions, such as Gaussian basis functions (Lewitt, 1992) or finite-element models that adapt to the content of the image (Brankov *et al.*, 2004).

Note that the projection space is also discrete, with the projection data represented by the vector $\mathbf{g}$. The elements of $\mathbf{g}$ are referred to here as *projection bins* or simply *bins*, and every projection measurement is represented by one bin. Again, the ordering of the individual bins is not important. Bins are generally sampled uniformly, although they need not be.

In Eq. (3), $\mathbf{H}$ is a $P \times N$ matrix called the *system matrix*, which describes the imaging process, and can represent attenuation and any linear blurring mechanisms. Each element of $\mathbf{H}$ (denoted by h_{ij}) represents the mean contribution of pixel j in the object to bin i in the projections. It is in the specification of $\mathbf{H}$ that the model of the projection process can become as simple or as complex as we require because the intensity of a projection bin is a weighted sum of intensities of the image pixels. To represent the Radon case, the matrix elements are defined so that a projection bin receives contributions only from pixels that are intersected by a given line and the contributions of pixels that do not intersect the line are set to zero. The linear model can also represent a more realistic case wherein a projection bin receives contributions from many pixels, each weighted according to the relative sensitivity of the projection bin to each pixel. These contributions are affected by physical factors such as attenuation, detector response, and scatter and can be estimated from knowledge of the system design and measurement of the patient attenuation distribution.

B. Statistical Model of Event Counts

So far, we have dealt only with the average behavior of the imaging system and have neglected the variability inherent in the photon-counting process used in ET. In consideration of the randomness in the projection data, Eq. (3) should truly be written as:

$$E[\mathbf{g}] = \mathbf{Hf} \tag{4}$$

where $E[\cdot]$ denotes expected value.

1. Poisson Model

Photon emissions are known to obey the Poisson distribution, and photon detections also obey the Poisson distribution, provided that detector dead time can be neglected and that no correction factors have been applied to the data. In this case, the numbers of events detected in the projection bins are independent of one another. Thus, the probability law for $\mathbf{g}$ is given by:

$$p(\mathbf{g}; \mathbf{f}) = \prod_{i=1}^{P} \frac{\bar{g}_i^{\,g_i} \exp(-\bar{g}_i)}{g_i!} \tag{5}$$

where $\bar{g}_i$ is the ith element of $E[\mathbf{g}] = \mathbf{Hf}$:

$$\bar{g}_i = \sum_{j=1}^{N} h_{ij} f_j \tag{6}$$

Although it is not exact for real imaging systems, the Poisson model is a good description of raw ET data and is the most commonly used model in the ET field. However, other probability models are often used. For example, when a PET imaging system internally corrects for random coincidences

by subtracting an estimated randoms contribution, the statistics can be described by a shifted Poisson model (Yavus and Fessler, 1996). Gaussian models, which are described in the following section, are also often used because of their practical advantages.

2. Gaussian Model

In cases in which the mean number of events is reasonably high, the Poisson law in Eq. (5) can be approximated by the following Gaussian probability density function (PDF):

$$p(\mathbf{g};\mathbf{f}) = k\exp\left[-\frac{1}{2}\sum_{i=1}^{P}\frac{(g_i-\bar{g}_i)^2}{\bar{g}_i}\right] = k\exp\left[-\frac{1}{2}(\mathbf{g}-\mathbf{Hf})^T\mathbf{C}^{-1}(\mathbf{g}-\mathbf{Hf})\right] \tag{7}$$

where k is a normalizing constant, and $\mathbf{C} = \text{diag}\{\bar{g}_1, \ldots, \bar{g}_P\}$ is the covariance matrix of $\mathbf{g}$. This approximation, which models the Poisson distribution to second order, is reasonably accurate when the mean numbers of events $\bar{g}_i$ are 20 or greater (Kalbfleisch, 1985); at low counts, the Poisson distribution becomes asymmetric about its peak, whereas the Gaussian distribution is always symmetric. Note that negative values of the elements of $\mathbf{g}$ have a probability of zero in the Poisson law, whereas the Gaussian approximation permits negative values. Thus, Poisson-based algorithms often have built-in constraints of nonnegativity, whereas Gaussian-based algorithms require additional constraints to achieve nonnegativity.

C. Spatiotemporal (4D) Imaging Model

In writing Eq. (1), we suppressed the fact that the activity distribution in the patient is actually a function of time. This fact is not considered in a static ET study, in which all the counts measured during the imaging session are summed together and used to produce a single static image of the patient. In this case, $f(\mathbf{x})$ in Eq. (1) should be interpreted as the time average of the spatiotemporal activity distribution $f(\mathbf{x},t)$.

Whereas a static study is concerned only with the spatial distribution of the tracer, gated studies and dynamic studies also measure temporal variations of the tracer concentration, as explained in Wernick and Aarsvold (Chapter 2 in this volume). In an ET study, there are two types of temporal variations of interest: (1) fluctuations caused by physiological interactions of the tracer with the body and (2) cardiac motion, which helps assess whether the heart is functioning normally. Other temporal variations, including respiratory motion, voluntary patient motion, and the steady decline of activity associated with radioactive decay, are uninformative effects to be corrected for, if possible.

For time-sequence imaging, as in the case of dynamic or gated studies, the imaging model may be expressed as follows:

$$g_{ik} = \frac{1}{\tau_k}\int_{I_k} dt \int_{\mathbb{R}^{D+1}} d\mathbf{x}\, f(\mathbf{x},t)h_i(\mathbf{x},t) \qquad i = 1,\ldots,P,\ k = 1,\ldots,K \tag{8}$$

where I_k is the time interval of duration τ_k during which the kth frame of data is acquired. Equation (8), which describes the entire time sequence of data, can be written in discrete form as follows:

$$\tilde{\mathbf{g}} = \tilde{\mathbf{H}}\tilde{\mathbf{f}} \tag{9}$$

where the space-time system matrix $\tilde{\mathbf{H}}$ is given by:

$$\tilde{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_1 & & \mathbf{0} \\ & \mathbf{H}_2 & \\ \mathbf{0} & & \ddots \\ & & \mathbf{H}_K \end{pmatrix} \tag{10}$$

and the space-time data $\tilde{\mathbf{g}}$ and image $\tilde{\mathbf{f}}$ are concatenations of all the data and image frames, that is, $\tilde{\mathbf{g}} = (\mathbf{g}_1^T, \ldots, \mathbf{g}_K^T)^T$ and $\tilde{\mathbf{f}} = (\mathbf{f}_1^T, \ldots, \mathbf{f}_K^T)^T$. It can often be assumed that the system matrix is time-invariant, that is, $\mathbf{H}_1 = \mathbf{H}_2 = \ldots = \mathbf{H}_K$, in which case $\tilde{\mathbf{H}} = \mathbf{I}\otimes\mathbf{H}_1$, where $\otimes$ denotes the Kronecker product (Wernick *et al.*, 1999).

III. COMPONENTS OF AN ITERATIVE RECONSTRUCTION METHOD

It is important to recognize that any method for image reconstruction is composed of two related but distinct components. The first component, which we call the *criterion*, is the statistical basis or governing principle for determining which image is to be considered as the best estimate of the true image. The second component, which we call the *algorithm*, is the computational technique used to find the solution specified by the criterion. In short, the criterion is a strategy, and the algorithm is a set of practical steps to implement that strategy.

IV. IMAGE RECONSTRUCTION CRITERIA

A. Constraint Satisfaction

A simple approach to image reconstruction is to view the problem as one of finding an image that satisfies the constraints dictated by the measured data and prior knowledge (such as the nonnegativity of pixel intensity values). This is the basis for a variety of algorithms known in image reconstruction as algebraic reconstruction techniques (ART) (Herman, 1980); in engineering, these methods are known as *vector-space projection methods* (Stark and Yang, 1998).

To illustrate this idea, let us first suppose that the data are uncorrupted by noise. In this case, each projection

measurement $g_i = \mathbf{h}_i^T\mathbf{f}$, $i = 1, \ldots, P$, defines a hyperplane in which the solution $\mathbf{f}$ must lie. In this sense, every hyperplane $g_i = \mathbf{h}_i^T\mathbf{f}$ is viewed as a set to which the solution is constrained to belong, and the solution must lie in the intersection of all these sets. If the number of measurements (equations) is greater than or equal to the number of pixels (unknowns), then the intersection of all the constraint sets may be a single point and the solution would be unique. A set of constraints that has a nonempty intersection is said to be *consistent*.

The essence of all constraint-satisfaction methods is that they aim to satisfy the known constraints. However, the problem is complicated when noise is present. In this case, each hyperplane (defined by a projection measurement) will be shifted by a random distance and the intersection of the hyperplanes (constraint sets) will be empty, unless the number of measurements is less than the number of pixels, in which case an infinite number of solutions may exist. The effects of noise can be alleviated by introducing constraints based on prior knowledge, usually representing image smoothness and non-negativity. Iterative algorithms for solving the constraint-satisfaction problem are described in Section VB.

The weakness of constraint-based methods is that they offer no mechanism for incorporating an explicit statistical model of the data. Thus, although these methods enjoyed considerable interest early in the development of ET, they have been largely supplanted by the maximum-likelihood and Bayesian methods described next.

B. Maximum-Likelihood Criterion

The maximum-likelihood (ML) criterion is a standard statistical-estimation criterion, proposed by R. A. Fisher (1921). In the ML criterion, it is presumed that the probability law $p(\mathbf{g};\mathbf{f})$ for the observation vector $\mathbf{g}$ is determined by some unknown deterministic parameter vector $\mathbf{f}$, which in our case is the object distribution we hope to reconstruct. In this context, $p(\mathbf{g};\mathbf{f})$ is called the likelihood function, which we denote by $L(\mathbf{f})$.

The ML criterion gives us a prescription for deciding which image, among all possible images, is the best estimate of the true object. The ML criterion can be stated simply as follows:

> **ML criterion:** Choose the reconstructed image $\hat{\mathbf{f}}$ to be the object function $\mathbf{f}$ for which the measured data would have had the greatest likelihood $p(\mathbf{g};\mathbf{f})$.

In this sense, the ML criterion seeks a solution that has the greatest statistical consistency with the observations. Symbolically, the ML strategy can be written as follows:

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{g};\mathbf{f}) \tag{11}$$

that is, choose the value of $\mathbf{f}$ for which $p(\mathbf{g};\mathbf{f})$ is greatest.

ML estimators have some desirable properties that justify their use in many situations (Van Trees, 1968). First, ML estimators are *asymptotically unbiased*, meaning that, as the number of observations becomes large, the estimates become unbiased (i.e., $E[\hat{\mathbf{f}}] \rightarrow \mathbf{f}$). Second, ML estimators are *asymptotically efficient*, meaning that (again, for large data records) they yield the minimum variance among unbiased estimators. In other words, ML estimators are less susceptible to noise than other unbiased estimators.

Unfortunately, although ET images reconstructed using the ML criterion have the least variance (image noise) among unbiased estimators, the variance is still unacceptably high. Therefore, we invariably choose to permit a certain amount of bias in the reconstructed image in exchange for reduced variance. This is accomplished by introducing spatial smoothing in the images, which reduces image noise at the cost of reduced fidelity in the mean. As we see later, smoothing can achieved *explicitly* (by Bayesian methods or lowpass filtering) or *implicitly* (by prematurely stopping an iterative ML algorithm before it actually reaches the ML solution).

C. Least-Squares and Weighted-Least-Squares Criteria

In statistical estimation problems in which the likelihood function is unknown, one can instead use the least-squares (LS) principle to determine the best solution. In the context of image reconstruction, the LS criterion can be stated as follows:

> **LS criterion:** Choose the value of $\mathbf{f}$ that, if observed through the system matrix $\mathbf{H}$, would yield projections $\mathbf{Hf}$ that are most similar to the observed projections $\mathbf{g}$ (in terms of Euclidean distance).

Thus, the LS solution also aims to maximize the consistency between the observed data and the reconstructed image. The LS estimation method can be expressed symbolically as follows:

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\mathbf{g} - \mathbf{Hf}\|^2 = \arg\min_{\mathbf{f}} \sum_{i=1}^{P} \left(g_i - \sum_{j=1}^{N} h_{ij} f_j \right)^2 \tag{12}$$

Equation (12) can be solved analytically to obtain the following closed-form solution:

$$\hat{\mathbf{f}} = \mathbf{H}^{+}\mathbf{g} \tag{13}$$

where $\mathbf{H}^{+} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{g}$ is the pseudoinverse of $\mathbf{H}$ (here we have assumed that $\mathbf{H}^T\mathbf{H}$ is invertible). This

closed-form solution is not often used in ET because of the large dimension of **H**; therefore iterative procedures are normally employed.

When we have knowledge that some of the projection data g_i have greater variance than others, we can weight each of the error terms in the summation in Eq. (12) differently. This approach, called weighted-least-squares (WLS) estimation, can be written as follows:

$$\begin{aligned}\hat{\mathbf{f}} &= \arg\min_{\mathbf{f}} (\mathbf{g}-\mathbf{Hf})^T \mathbf{D}(\mathbf{g}-\mathbf{Hf}) \\ &= \arg\min_{\mathbf{f}} \sum_{j=1}^{p} d_i \left(g_i - h_{ij} f_j \right)^2 \end{aligned} \tag{14}$$

where **D** is a diagonal matrix, with elements d_i on the diagonal. The weights d_i are usually chosen to be $(\mathrm{var}[g_i])^{-1}$. For ET data, which are Poisson-distributed, the variance equals the mean, so $d_i = \overline{g}_i^{-1}$. Like the LS solution, the WLS solution can be written in closed form as follows:

$$\hat{\mathbf{f}} = (\mathbf{H}^T\mathbf{D}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}^{-1}\mathbf{g} \tag{15}$$

but, again, iterative methods are usually used instead because of the large dimension of **H**. Another WLS approach is the iteratively reweighted LS method, in which **D** is estimated from projections of the current image estimate.

It is important to recognize that, although the LS and WLS criteria do not explicitly refer to any probability model for the data, they are actually equivalent to ML estimation under a Gaussian model. In fact, WLS reconstruction was one of the earliest ML methods to be applied to ET reconstruction (Huesman *et al.* 1977). To see the connection between WLS and ML, compare the WLS function in Eq. (14) to the Gaussian model of ET data in Eq. (7). These functions are equivalent if we choose $\mathbf{D} = \mathbf{C}^{-1}$. In addition, the LS function is identical to ML estimation under an assumption that the observations g_i have equal variance, which is a poor assumption in ET.

D. Shortcoming of Maximum-Likelihood, Least-Squares, and Weighted-Least-Squares Methods

The shortcoming of the aforementioned statistical methods, when applied to ET image reconstruction, is that they tend to produce images that are exceedingly noisy. This problem arises because these classical criteria aim solely to enforce maximal consistency between the reconstructed image and the measured data. Unfortunately, because the data are noisy, the image that is most consistent with these data also tends to be noisy.

This is particularly serious in ET reconstruction because ET imaging systems are lowpass in nature (i.e., they produce data that are blurred), whereas the noise is broadband (i.e., it contains components at all frequencies).[1] To see intuitively why this presents a problem, let us consider the LS criterion. According to Eq. (13), the LS solution is obtained by applying the pseudoinverse of **H** to the data. If **H** is a lowpass operator, then its pseudoinverse is a highpass operator; thus it tends to amplify noise in the image. Specifically, if we view noise in the observed data as an additive zero-mean contribution **n** (so that $\mathbf{g} = \mathbf{Hf} + \mathbf{n}$) the LS solution is:

$$\begin{aligned}\hat{\mathbf{f}} &= \mathbf{H}^{+}\mathbf{g} \\ &= \mathbf{H}^{+}(\mathbf{Hf}+\mathbf{n}) \\ &= \mathbf{f} + \mathbf{H}^{+}\mathbf{n}\end{aligned}$$

The first term in the solution is the correct answer, but the second term consists of noise that has been subjected to a highpass operator that amplifies high-frequency components. The result, $\hat{\mathbf{f}}$, is generally an extremely noisy image.

E. Bayesian Methods

The ML, LS, and WLS methods are referred to as *classical* estimation criteria, which refers to their assumption that **f** is unknown but deterministic (not random). Classical methods are based on the notion that the data alone should determine the solution and that no prejudice from the experimenter should influence the estimate. In contrast, Bayesian methods assume that the unknown quantity **f** is itself random and can therefore be described by a PDF $p(\mathbf{f})$ that is known in advance of data collection. This permits the experimenter to inject his or her own prior expectations about **f** into the process. For example, if we image a patient's brain, we can imagine this person's brain to be a sample drawn randomly from some hypothetical population of brains, defined by $p(\mathbf{f})$. This PDF, called the *prior* (because it reflects information known in advance), permits us to modify the reconstructed image to conform to our expectations. We know that a positron emission tomography (PET) brain image should look something like a brain and not a car, a heart, or anything else that is not a brain. Thus, we can postulate in advance (*a priori*) that images that look like cars or hearts should have probability zero and those that look like brains should have some positive probability.

Philosophically speaking, introducing our prior beliefs may not be appropriate in a hypothesis-driven science experiment because it biases the outcome toward our preconceived expectations. However, the Bayesian approach is a very helpful practical tool for image reconstruction, provided that the prior is reasonable.

Ideally, the prior PDF might precisely define our prior knowledge about the true image, such as that the true image is

[1] Here we use the terms *lowpass* and *broadband* loosely because they apply properly only to space-invariant systems and stationary noise processes, which can be described by block-circulant systems and covariance matrices. However, analogous concepts can be discussed in the context of singular value decomposition.

a brain. Unfortunately, such a belief can be difficult to express mathematically. Therefore, Bayesian reconstruction methods usually do not make such ambitious statements about the true image **f**. Instead, these methods usually aim simply to encourage the reconstructed image to be smooth, so as to suppress the effect of noise. Specifically, a low probability is assigned to image solutions that have lots of fine detail on the assumption that these features are probably due to noise. This assumption is based on the knowledge that, because of its blurring effect, the imaging system **H** strongly suppresses image detail; therefore, any such detail that persists in the reconstruction is likely to have arisen from noise.

1. Theory of Bayesian Estimation

The main conceptual shortcoming of the ML criterion is that it fails to consider the consequences of our choosing one image solution over another. Bayes' theory, on the other hand, begins by stating these consequences explicitly in the form of a quantity called the loss function, denoted by $\lambda(\mathbf{f},\hat{\mathbf{f}})$. In the image reconstruction problem, this loss function measures the extent to which the reconstructed image $\hat{\mathbf{f}}$ deviates from the true image **f**. The loss function typically used in ET image reconstruction is called a *hit-or-miss* loss function:

$$\lambda(\mathbf{f},\hat{\mathbf{f}}) = \begin{cases} 0 & \left|\mathbf{f}-\hat{\mathbf{f}}\right| < \delta \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

where δ is a positive constant, and $|\cdot|$ denotes L_1 norm, $|\mathbf{x}| = \Sigma_i |x_i|$. This loss function states that the reconstructed image $\hat{\mathbf{f}}$ is perfectly acceptable (zero loss) if it is sufficiently close to the true image **f** and that any less accurate result is equally unfavorable (unit loss).[2]

The aim of Bayesian methods is to find a criterion for choosing $\hat{\mathbf{f}}$ that will minimize the *average* loss that we experience when this criterion is applied. It is easy to show (Kay, 1993) that the minimization (on average) of Eq. (16) leads to the maximum *a posteriori* (MAP) criterion, which is stated as follows:

> **MAP criterion:** Choose the value of **f** that maximizes the posterior PDF, $p(\mathbf{f}|\mathbf{g})$.

Symbolically, the MAP criterion can be expressed as:

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f} \mid \mathbf{g}) \tag{17}$$

By using Bayes' law, given by:

$$p(\mathbf{f} \mid \mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})}{p(\mathbf{g})} \tag{18}$$

the MAP criterion can be written in a more useful form as follows:

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \frac{p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})}{p(\mathbf{g})} \tag{19}$$

By taking the logarithm of the quantity to be maximized, and omitting $p(\mathbf{g})$ (because it is not a function of **f**), the MAP criterion can be simplified to the following form:

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \left[\ln p(\mathbf{g} \mid \mathbf{f}) + \ln p(\mathbf{f})\right] \tag{20}$$

From Eq. (20), we see that the MAP criterion is similar to the ML criterion; however, it uses the logarithm of the prior to penalize image solutions that do not have the expected properties. Specifically, the maximization in Eq. (20) attempts to balance consistency with the data **g** (as expressed by the likelihood term) with conformance to our prior expectations (expressed by the prior term). In other words, it aims to produce an image that is reasonably consistent with the data while not being too noisy. Note that the ML criterion can be viewed as a limiting case of the MAP criterion as the prior $p(\mathbf{f})$ tends toward a uniform PDF, which implies that *a priori* we do not prefer any solution over any other.

An additional benefit of the MAP criterion is that the function to be maximized can be sharper than the likelihood function and thus can make iterative reconstruction algorithms more efficient. Figure 3 illustrates the effect of the prior on a relatively flat ML objective function. The sharpness of the prior about its peak determines the sharpness of the MAP criterion. The resulting MAP solution is pushed away from the ML solution and toward the peak of the prior. Obviously, if the peaking of the prior is extreme, then the resulting solution will depend primarily on the prior and the measured data will be ignored. This undesirable result can be avoided by using a weak prior, so

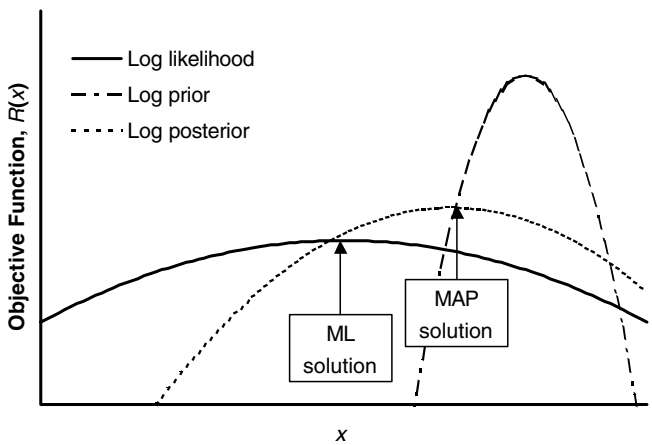

FIGURE 3 Comparison of objective functions in a simple 1D example. The MAP objective function (the log posterior) is sharper than the ML one (the likelihood), making the solution easier for an iterative algorithm to find.

[2]Other choices for the loss function lead to different reconstruction criteria. Notably, an L_2-error loss function implies a conditional mean, or minimum mean-square error, solution, whereas an L_1-error loss function leads to the median of the posterior as the solution.

that the solution captures the main properties of the ML solution while pushing the solution slightly in a direction that emphasizes smoothness.

2. Approaches to Defining the Prior

As we have seen, the MAP criterion is a simple one and not very different in form from the ML criterion. However, it can be difficult to express prior knowledge, by way of $p(\mathbf{f})$, in a mathematically meaningful and efficient way. One might be tempted, as we suggested earlier, to use the prior to express our knowledge that the true image should look like a brain, for example. However, this approach can be dangerous. For example, suppose that our prior model did not anticipate patients with a particular type of lesion. In this case, the reconstruction algorithm might view images containing this lesion as improbable and thus attempt to suppress the lesion in the reconstruction.

Owing to the difficulties associated with highly specific priors, the image reconstruction field has favored the use of priors that are simple and generic. The most prevalent approach in ET reconstruction is to use the prior simply to specify the belief that smooth images are more plausible than noisy ones. Such a prior makes a relatively bland statement about the true object distribution and can be applied generally because it does not aim to specify whether the object is a brain or a heart. The principal disadvantage of this approach is that legitimate image features can be obliterated along with the noise.

3. The Gibbs Distribution

A smooth image is one in which it is highly probable that neighboring pixels have similar intensity values. A prior that encourages this property attempts to suppress sharp transitions between pixels on the basis that such features are probably due to noise. By viewing the true pixel intensity values as random variables, a smooth image can be described as one in which intensity values of neighboring pixels are highly correlated while distant pixels are less so.

A simple mathematical model having this property is the Markov random field, which can be described by the Gibbs PDF (Geman and Geman, 1984):

$$p(\mathbf{f}) = \frac{1}{Z}\exp[-\beta U(\mathbf{f})] \tag{21}$$

where Z is a normalizing constant called the *partition function*; β is a scalar weighting parameter that determines the peaking of the distribution about its maxima; and $U(\mathbf{f})$ is the energy function. The energy function is a weighted sum of *potential functions*, which are functions of small sets of neighboring pixels, called *cliques*, denoted by S_c ($c = 1, \ldots, C$):

$$U(\mathbf{f}) = \sum_{c=1}^{C} \sum_{i,j,k\cdots \in S_c} w_{ijk\cdots} V_c(f_i, f_j, f_k, \cdots) \tag{22}$$

In Eq. (22), pixels indexed by $i,j,k, \ldots$, are elements of the same clique. This general model encompasses a number of priors proposed for ET reconstruction, including Gaussian (Levitan and Herman, 1987) and entropy (Liang *et al.*, 1989) priors. Cliques may have any number of pixels. For example, the total intensity in the image can be constrained by using a clique that includes all the pixels in the image and an energy function based on the sum of the intensities of all pixels in the clique. It is permissible to mix clique sizes and energy functions in the Gibbs model, and a particular pixel may be a member of more than one clique.

In most ET reconstruction applications, there is a clique for each pixel i, and each clique consists of pixels that are nearest to pixel i. Methods using two-pixel cliques (e.g., Lee *et al.*, 1995) generally use potential functions that are related to the difference in intensity between the two pixels, and the potential functions do not vary across the image. As a result, Eq. (22) is often simplified to:

$$U(\mathbf{f}) = \sum_{i=1}^{N} \sum_{j\in S_i, i<j} w_{ij} V_{ij}(f_i - f_j) \tag{23}$$

Although there are many possibilities for the clique structure and clique weights, usually cliques consist of local neighborhoods and the weights are usually determined by the inverse of the distance between the centers of the two pixels in the clique. The principal distinction between methods usually lies in the choice of potential function, which can strongly determine the smoothness properties of the MAP solution (Lalush and Tsui, 1992, 1993).

Clique potentials are generally bowl-shaped functions, as shown in Figure 4. The clique potential is zero at $r = f_i - f_j = 0$ and increases with increasing difference between the neighboring pixels. Thus, large intensity differences increase the energy function $U(\mathbf{f})$, which reduces the prior probability of the image. Clique energy functions fall roughly into two categories, based on how the behavior of the function for large r. Quadratic and higher-order functions (Fessler, 1994) apply increasing smoothing power as pixel differences increase. Linearly increasing functions (Green, 1990; Hebert and Leahy, 1992; Lalush and Tsui, 1993; Mumcuoglu *et al.*, 1994) tend toward a linear function of r for large r. Priors that increase more slowly than a quadratic are sometimes called *edge-preserving* priors because they can have a selective effect on smoothing, removing small differences due to noise while retaining edge sharpness in the final result. Linearly increasing functions may be convex or nonconvex. Because exhaustive comparisons are difficult, it has not been established that any one type of clique potential is always preferred.

4. The Prior as a Penalty Term

Using the Gibbs prior of Eq. (21), the log posterior PDF becomes:

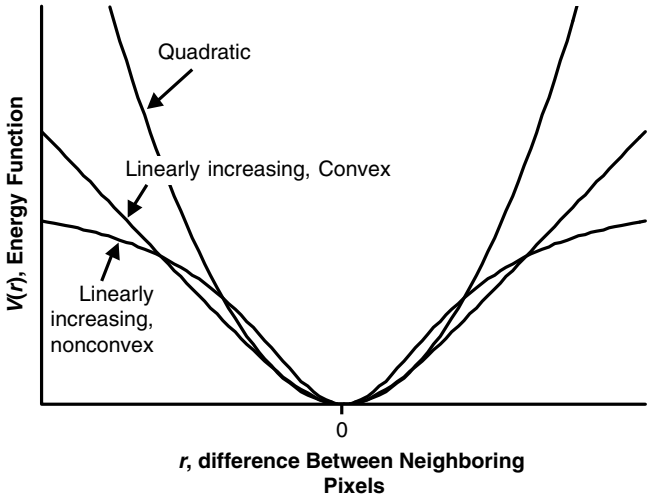

FIGURE 4 Potential functions used in Gibbs priors.

$$\ln p(\mathbf{f} \mid \mathbf{g}) = \ln L(\mathbf{f}) - \beta U(\mathbf{f}) \tag{24}$$

where the likelihood function $L(\mathbf{f})$ may have either the Poisson or Gaussian form described earlier. Noisy images can produce large values of the likelihood function, but they will be penalized by the prior term $\beta U(\mathbf{f})$ and thus will tend not to be chosen when maximizing $\ln p(\mathbf{f}|\mathbf{g})$. In this way, a balance is struck between the requirements of the measured data, through the likelihood, and the requirements of the prior. This balance is governed by the weighting parameter β. If β is set to zero, the MAP solution is simply the ML solution; as β becomes large, the prior term dominates the maximization. Thus, β, which is called a *hyperparameter* of the optimization problem, is the most significant determinant of the degree of smoothing present in the MAP solution. Usually, β is set by the user based on experience and desired results, but there are several methods for automatically determining it (Higdon *et al.*, 1997; Johnson *et al.*, 1991; Zhou *et al.*, 1995).

The form of the prior determines whether Eq. (24) is unimodal or whether it may have disconnected local maxima. For priors with the form of Eq. (23), the clique potential function must be strictly convex in order to ensure that the objective function is unimodal (Lange, 1990). Thus, only convex priors (Fessler, 1994; Green, 1990; Mumcuoglu *et al.*, 1994) can guarantee convergence to a global maximum of the log posterior function $\ln p(\mathbf{f};\mathbf{g})$. Nonconvex priors (Hebert and Leahy, 1992; Lalush and Tsui, 1993) have been proposed and shown to have useful properties as well, despite the fact that they may converge only to local maxima.

The MAP method is also sometimes known as the *penalized ML method* because the prior term can be viewed as a penalty on solutions that have undesirable properties. By using this terminology one acknowledges that the chosen prior may not truly describe the PDF of the image, but may be only a practical device used to discourage unwanted image characteristics.

It should also be noted that the penalization approach can be extended to LS methods. For example, one can add a penalty term to the WLS function to obtain a method called penalized WLS (PWLS) (Fessler, 1994; Kay, 1993), which is equivalent to MAP reconstruction with a Gaussian likelihood function.

5. Anatomical Priors

Simple smoothing priors based on the Gibbs distribution are the predominant approach for MAP reconstruction. However, another approach involving what are called *anatomical priors* has also been widely studied (Chiao *et al.*, 1994; Gindi *et al.*, 1991; Leahy and Yan, 1991; Ouyang *et al.*, 1994; Brankov, Yang, Leahy, *et al.*, 2002). A major shortcoming of the basic Gibbs prior is that it penalizes all kinds of abrupt intensity variations in an effort to suppress noise. Unfortunately, in doing so it may also suppress legitimate image boundaries at the edges of anatomical features. Methods based on anatomical priors aim to identify valid image boundaries by using information gained from a second image of the patient obtained by an imaging modality such as magnetic resonance (MR) imaging. These methods have not gained wide acceptance, largely because of the difficulty of reconciling the different kinds of information that appear in the MR and ET images. For example, an anatomical boundary in an MR image may not have a corresponding functional boundary in the ET images, so imposing such a boundary would create a false image feature.

F. Criteria for Reconstruction of Image Sequences: 4D Reconstruction

Traditionally, image sequences have been reconstructed by applying the methods described thus far to each image frame, one by one. This approach misses an opportunity to reduce the effect of noise by encouraging smoothness between image frames. Between-frame smoothing can be used to great advantage in ET reconstruction because ET data usually vary slowly in time, whereas the noise is uncorrelated from frame to frame. In electrical engineering, image-recovery methods that exploit commonalities between images are called *multichannel methods*; a summary of these methods is given (Galatsanos *et al.*, 2000). The application of multichannel methods to ET has received considerable attention recently and is referred to as *spatiotemporal* or *4D reconstruction*.

1. Dynamic Imaging

Dynamic imaging, which may be achieved by PET or SPECT (see Wernick and Aarsvold, Chapter 2 in this volume), is used to identify temporal variations in the radiotracer concentration that reveal information about organ

physiology. In dynamic imaging, the organ of interest is either reasonably stationary or is simply treated as such. In cardiac imaging, the result of this assumption is a motionless (and motion-blurred) image of the heart that does not depict cardiac motion but instead focuses on capturing important information about the kinetics of the tracer.

In dynamic imaging, between-frame smoothness can be achieved by enforcing smoothness in the reconstructed images along the time dimension of the image sequence. This is illustrated in Figure 5a, where, for simplicity, each image frame is depicted as a 2D slice. The connected pixels are those that are assumed to be correlated and may be grouped in the same clique in a Gibbs prior. The temporal variations in dynamic image sequences are usually very slowly varying, and the noise level can be exceedingly high; therefore, dynamic imaging can benefit greatly from temporal smoothing.

Many 4D reconstruction methods for dynamic imaging are based on representations of temporal variations in the image sequence using smooth time functions. For example, principal component (or Karhunen-Loève) basis functions were used in (Kao *et al.*, 1997; Wernick *et al.*, 1999). These basis functions, which are tailored to the specific data set, are designed to represent the data compactly (and thus yield a fast reconstruction) and to isolate and reject orthogonal noise components. Spline functions have also been used to model the variations in dynamic image sequences within the reconstruction algorithm (Nichols *et al.*, 1999; Reutter *et al.*, 2000). Many methods that estimate kinetic parameters from projection data assume exponential functions as the basis functions because these are solutions to standard kinetic models (See Morris *et al.*, Chapter 23 in this volume; Coxson *et al.*, 1990; Zeng *et al.*, 1995; Reutter *et al.*, 1998; Hebber *et al.*, 1997; Limber *et al.*, 1995). Several recent 4D reconstruction methods instead use various smoothing constraints on the temporal variations in the image sequence (Farncombe *et al.*, 1999, 2001), such as regional monotonicity of the time functions.

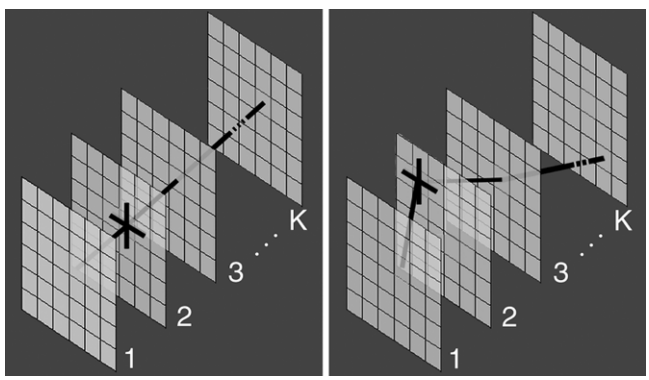

FIGURE 5 Smoothing used in spatiotemporal reconstruction (a) without motion compensation and (b) with motion compensation.

Dynamic SPECT has become a prominent research topic in recent years because traditional SPECT acquisitions have an important practical problem to overcome. For tracers with rapid kinetic properties, the tracer distribution in the body changes during the time required for the gamma camera to make a complete revolution about the patient. Thus, a complete set of projections is not available at any given instant of time. Therefore, images reconstructed by conventional means exhibit serious artifacts, and 4D reconstruction is essential.

2. Gated Imaging

In gated cardiac imaging, the dynamics of the tracer distribution are usually disregarded in favor of capturing cardiac motion. Whereas a dynamic study produces a long time sequence (like a movie), gated imaging produces a short, looping image sequence that represents a composite representation of a single heartbeat that summarizes data acquired over a large number of cardiac cycles.

In gated imaging, it is critical that cardiac motion be accounted for in the smoothing criterion. It is possible to accomplish this by viewing the heart's motion as a source of intensity fluctuations in each pixel (Narayanan *et al.*, 1999). However, a more appealing approach is to ensure that smoothing is performed between pixels in different frames that contain roughly equivalent tissue. The reasoning is that, if the myocardium (heart wall) were to occupy a given pixel in one frame but not in the next, the smoothing method would wrongly average (blur together) two distinct portions of the object. This will reduce the effect of noise, but will also distort the appearance of the object and its motion.

One approach uses the intensity-selective nature of certain Gibbs priors to enforce smoothing only among spatiotemporal neighboring pixels that share similar intensity (Lalush and Tsui, 1996, 1998a). This method does not require an explicit motion estimate and yet is able to resist the distortion.

Other 4D reconstruction methods for gated imaging bend the trajectories followed by the smoothing so as to follow the motion of the heart, as illustrated in Figure 5b. These techniques incorporate a motion estimation algorithm to determine the motion trajectories. Motion estimation has a long history of developments in the image-processing literature (Tekalp, 1995), but its application in gated ET imaging is relatively recent. Optical flow (Klein *et al.*, 1997) and deformable mesh modeling (Brankov, Yang, Wernick, *et al.*, 2002) have been studied, as well as simultaneous iterative methods of motion estimation and reconstruction (Mair *et al.*, 2002). Once motion is determined, smoothness can be enforced through the use of Gibbs priors, as demonstrated in (Lalush and Tsui, 1996, 1998a). Smoothness can also be enforced by using postreconstruction temporal filters (Brankov, Yang, Wernick, *et al.*, 2002), an idea that was first proposed (without motion estimation) by (King and Miller 1985).

The simultaneous reconstruction of gated and dynamic cardiac images has recently been proposed, allowing both wall motion and tracer dynamics to be assessed (Feng *et al.*, 2003). This approach, which has been termed *5D reconstruction*, aims to gather more complete and accurate information about the heart from the available data.

V. ITERATIVE RECONSTRUCTION ALGORITHMS

Earlier we explained that an iterative image reconstruction method consists of a criterion for selecting the best image, combined with an algorithm for computing that image. Many different iterative approaches to solving the ET reconstruction problem have been studied, but they share some common traits. In the following section, we outline the common features of most iterative reconstruction algorithms and discuss some of their general properties.

A. General Structure of Iterative Algorithms

Most iterative reconstruction algorithms fit the general model shown in Figure 6. The process is begun with some initial estimate $\hat{\mathbf{f}}^{(0)}$ of the pixel intensity values in the image. A projection step is applied to the current image estimate $\hat{\mathbf{f}}^{(n)}$, which yields a set of projection values $\hat{\mathbf{g}}^{(n)}$ that would be expected if $\hat{\mathbf{f}}^{(n)}$ were the true image. The predicted projections $\hat{\mathbf{g}}^{(n)}$ are then compared with the actual measured data $\mathbf{g}$ to create a set of projection-space error values $\mathbf{e}_g$. These are mapped back to the image space through a backprojection operation to produce image-space error values $\mathbf{e}_f$ that are used to update the image estimate, which becomes the new estimate $\hat{\mathbf{f}}^{(n+1)}$. This process is repeated again and again until the iteration stops automatically or is terminated by the user. Each of these repetitions is called an *iteration*. At the conclusion of the process, the current image estimate is considered to be the final solution.

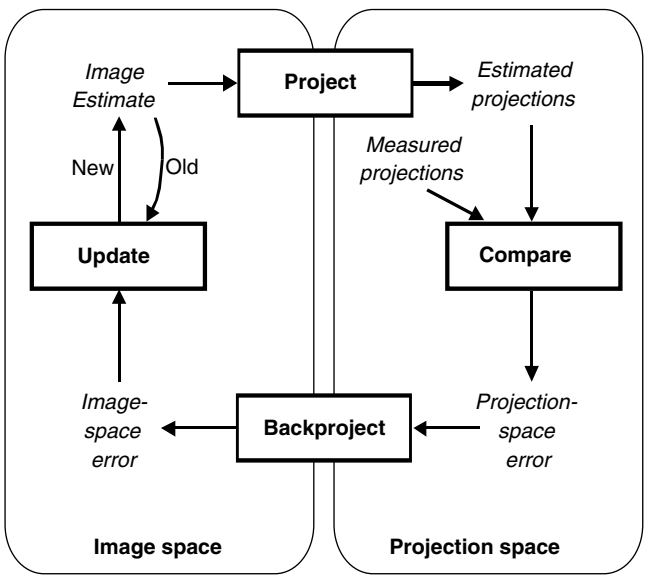

FIGURE 6 Flowchart of a generic iterative reconstruction algorithm.

A critical but often-overlooked topic in image reconstruction is the practical computational choices involved in implementing the forward and backprojection steps involved in all iterative algorithms. In many cases, this issue can preclude the use of certain algorithm types because they are simply impractical to compute. Various approaches to defining the projection and backprojection operations are described in detail in Section VIIIA.

Note that we have not specified the details of the projection, comparison, backprojection, and update steps. It is principally in these steps that individual reconstruction algorithms differ. Note that direct reconstruction methods such as FBP use only the backprojection portion of the loop, so that there is no feedback about whether the image estimate, when projected, is consistent with the measured data. The power of iterative methods lies in the use of this feedback loop to refine the reconstructed image.

B. Constraint-Satisfaction Algorithms

Let us begin by reviewing the earliest iterative algorithms, which are based on the constraint satisfaction strategy introduced in Section IVA. The earliest algorithm of this kind, developed in 1937 by Kaczmarz (Rosenfeld and Kak, 1982; Kaczmarz, 1937), was designed for the solution of systems of linear equations. In the ET field, constraint-satisfaction methods are best known generally as ART (Gordon *et al.*, 1970; Herman, 1980) and come in many varieties. Constraint-satisfaction methods have gained significant popularity in the electrical engineering community, where they are known as *vector-space projection methods* (Stark and Yang, 1998). Currently constraint-based methods are less prominent in the ET field than statistical methods, but they provide useful insights into the issues that underlie the reconstruction problem. These methods remain popular as solutions to other inverse problems, such as retrieval of Fourier phase information from magnitude-only data, and are worth understanding.

A brief summary of a general class of methods, called *projections onto convex sets* (POCS), is given in Section VIIIB. The following methods are all variations on this ap-proach, in which all the known information (data and prior knowledge) are based on constraints. The idea of POCS methods is to specify the solution as being a point in the solution space that satisfies all the constraints and thus lies in the intersection of all the sets that describe these constraints.

1. The Kaczmarz Method/Algebraic Reconstruction Technique

Our imaging model, $\mathbf{g} = \mathbf{Hf}$, can be considered to be a set of simultaneous equations, one for each projection bin. Each linear equation in the imaging model, $g_i = \mathbf{h}_i^T\mathbf{f}$, defines a hyperplane in the vector space in which $\mathbf{f}$ is defined. Therefore, assuming that this set of simultaneous equations is consistent (which occurs when there is no noise), the solution is any point that lies in the intersection of all the hyperplanes. This point can be determined by a process in which, starting from an initial estimate $\hat{\mathbf{f}}^{(0)}$, the vector is repeatedly projected onto all the hyperplanes (here, the word *projection* is used in the linear algebra sense, not the tomography sense).

The operation of projecting a point onto a hyperplane is a simple one, given by:

$$\hat{f}_i^{(n+1)} = \hat{f}_i^{(n)} + h_{ji}\frac{\left(g_j - \sum_k h_{jk}\hat{f}_k^{(n)}\right)}{\sum_k h_{jk}^2} \tag{25}$$

Figure 7 shows a simple example of how the algorithm proceeds when there are only two pixels (two dimensions) and two measurements (two constraint sets).

A nonnegativity constraint set can also be introduced into the procedure. The operation of projecting onto this set is simple. In each iteration, any pixel having a negative value is set to zero:

$$\hat{f}_i^{(n+1)} = \begin{cases} \hat{f}_i^{(n)} & \text{if } \hat{f}_i^{(n)} \geq 0 \\ 0 & \text{if } \hat{f}_i^{(n)} < 0 \end{cases} \tag{26}$$

Similarly, pixels outside a particular image region S can be constrained to be zero by simply setting pixels outside S to zero in each iteration:

$$\hat{f}_i^{(n+1)} = \begin{cases} \hat{f}_i^{(n)} & \text{if } \hat{f}_i^{(n)} \in S \\ 0 & \text{if } \hat{f}_i^{(n)} \notin S \end{cases} \tag{27}$$

Various smoothing constraint sets may also be defined, as explained in Stark and Yang (1998).

The convergence rate of the Kaczmarz/ART method may be dependent on the orthogonality of the successive equations. Note in Figure 8a that when the equations are far from orthogonal, the process may require many iterations to reach the intersection. On the other hand, when the equations are more nearly orthogonal, the process requires relatively few iterations (Fig. 8b). It is important to note that when there is no unique solution the intersection of all the hyperplanes is empty, and so the iterative process oscillates among several solutions (Fig. 8c). This behavior can be avoided by underrelaxation, that is, by stopping short of each constraint set in each projection step. If done properly, this can lead to a LS solution to the problem (Censor *et al.*, 1983), which is a compromise solution lying in the middle of the triangular region in Figure 8c.

One disadvantage of ART is that it updates based on one measurement at a time. This requires a matrix-based ap-

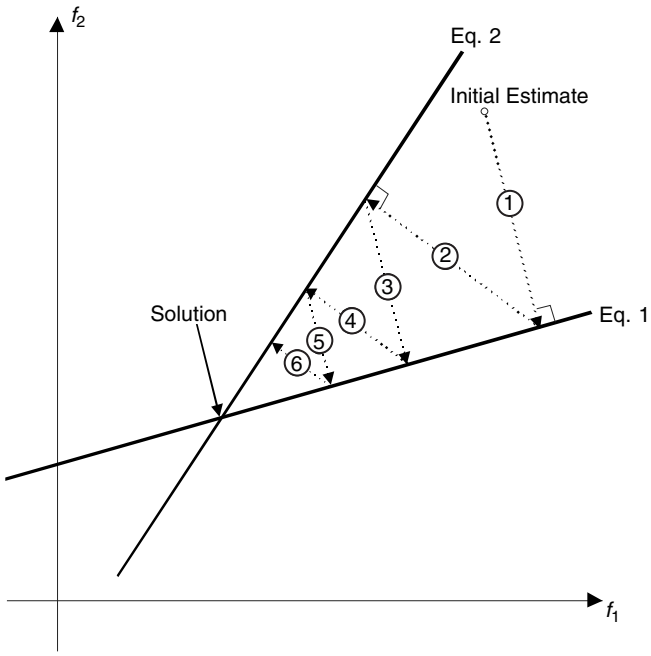

FIGURE 7 A simple example of the Kaczmarz procedure. The current estimate is projected successively onto each line by finding the point on each line that lies closest to the current estimate.

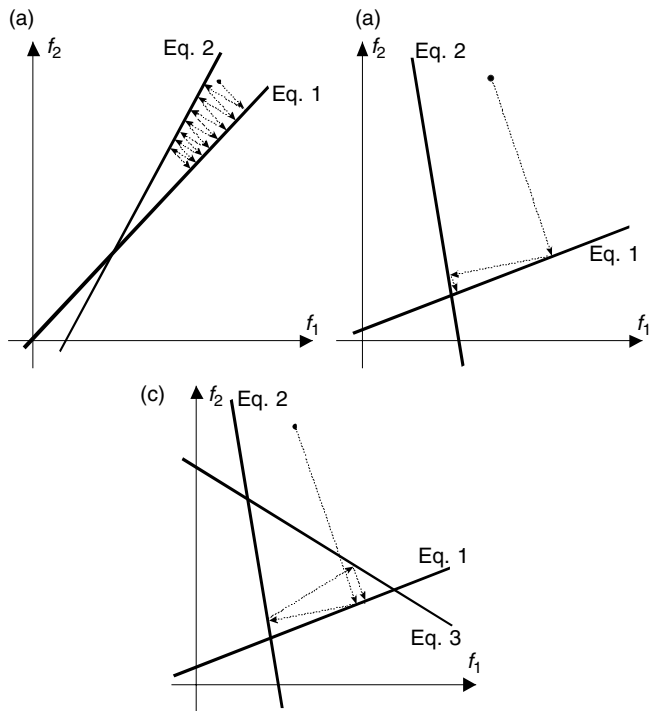

FIGURE 8 Illustrations of the Kaczmarz method in different situations, showing how the convergence of the method may be affected by the arrangement of the hyperplanes describing the linear equations.

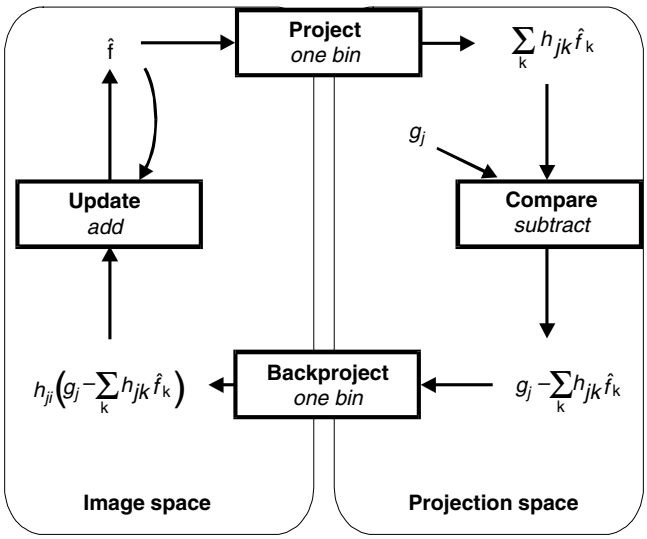

FIGURE 9 The algebraic reconstruction technique in the form of the general iterative model.

proach to the model so that individual elements of **H** can readily be accessed. This makes ART practical only for slice-by-slice reconstruction of parallel projection data.

2. Variations on ART

Several variations on ART have been proposed, and all fit the general iterative model of Figure 6, as shown in Figure 9. Multiplicative ART (MART) (Gordon *et al.*, 1970) uses a multiplicative error and update:

$$\hat{f}_i^{(n+1)} = \hat{f}_i^{(n)} \frac{g_j}{\sum_k h_{jk}\hat{f}_k^{(n)}} \text{ if } h_{ji} > 0 \tag{28}$$

It has been shown that MART can be derived in a POCS framework (Mailloux *et al.*, 1993), but that its performance depends on the initial estimate. MART features automatic enforcement of nonnegativity: If the initial pixel intensities are nonnegative, then all iterated image estimates remain nonnegative. It is also simple to constrain pixels outside a region S to be zero by initializing them to zero.

Another variation of ART is the simultaneous iterative reconstruction technique (SIRT)[13], in which the update is performed for one pixel at a time using all equations that contribute to that pixel:

$$\hat{f}_i^{(n+1)} = \hat{f}_i^{(n)} + t\sum_j h_{ji} \frac{\left(g_j - \sum_k h_{jk}\hat{f}_k^{(n)}\right)}{\sum_k h_{jk}^2} \text{ if } > 0 \tag{29}$$

The SIRT method is a parallel POCS algorithm, in which the projections are all performed simultaneously and then averaged. In SIRT, the relaxation parameter t, which may vary as the iterations progress, controls the convergence characteristics. An extension of SIRT, which has all pixels updated at the same time, allows the use of a projector-based model and removes the requirement for storing the matrix **H**.

In the following sections, we describe the most predominant iterative algorithms, WLS, ML, and MAP, which are based on statistical estimation criteria.

C. The Maximum-Likelihood Expectation-Maximization Algorithm

For several years, the leading iterative reconstruction algorithm for PET and SPECT has been the maximum-likelihood expectation-maximization (ML-EM) algorithm and its variations. The ML-EM algorithm was first proposed formally in 1977 as the solution to incomplete data problems in statistics (Dempster *et al.*, 1977) and has since found application in a wide range of statistical applications. Strictly speaking, the ML-EM algorithm presented in (Dempster *et al.*, 1977) is not an algorithm at all but rather a general prescription for developing algorithms that can be applied to a broad range of specific ML estimation problems (McLachlan and Krishnan, 1997).

Using this general prescription, an ML-EM algorithm for ET reconstruction was demonstrated in 1984 (Lange and Carson, 1984; Shepp and Vardi, 1982); however, the resulting iterative formula had been derived by a different approach in the 1970s (Lucy, 1974; Richardson, 1972) and was already known to the field of astronomy as the Richardson-Lucy algorithm. When applied to the ET reconstruction problem (or indeed to any linear inversion problem with Poisson noise), the ML-EM framework yields the following simple iterative equation, which is easy to implement and understand:

$$\hat{f}_j^{(n+1)} = \frac{\hat{f}_j^{(n)}}{\sum_{i'} h_{i'j}} \sum_i h_{ij} \frac{g_i}{\sum_k h_{ik}\hat{f}_k^{(n)}} \tag{30}$$

Further explanation of the ML-EM algorithm, along with a derivation of the well-known iterative expression in Eq. (30), is provided in Section VIII C.

1. Properties of ML-EM

The ML-EM algorithm in Eq. (30) has a very simple form that fits our general model of an iterative algorithm (see Fig. 10). In fact, it is strikingly similar to MART (Eq. 28), the primary difference being that all pixels are updated simultaneously in the ML-EM algorithm. Its convergence behavior is consistent and predictable. Because the error and updates are multiplicative, ML-EM automatically imposes a nonnegativity constraint and allows for selected pixels to be preset to zero. The algorithm is simple to implement by computer using a projector-based model and eventually leads to a constrained ML solution.

The ML-EM algorithm for ET reconstruction has two main shortcomings. First, the convergence of the algorithm

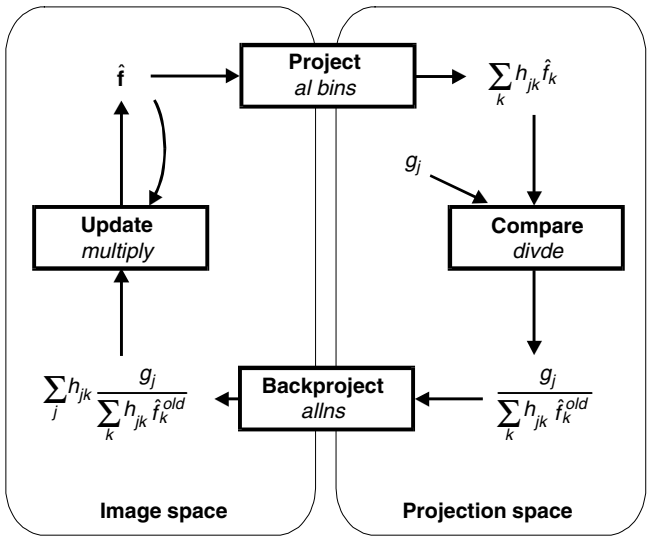

FIGURE 10 The maximum-likelihood expectation-maximization algorithm in the form of the general iterative model.

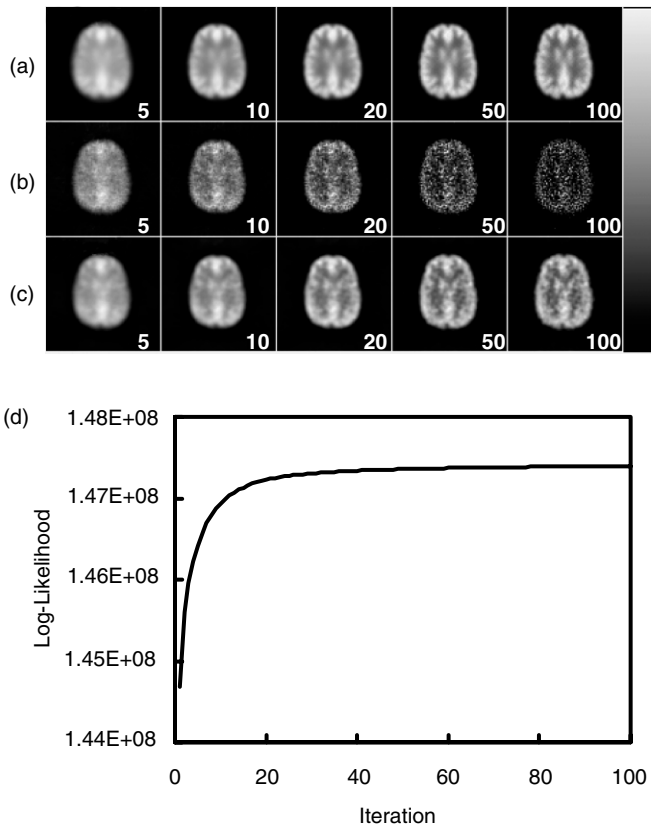

FIGURE 11 Convergence properties of the ML-EM algorithm. The images show the progression of iterated image estimates from simulated brain SPECT data. The algorithm modeled nonuniform attenuation in the matrix **H**. The cases shown are (a) noise-free, (b) noisy, and (c) the noisy reconstructions followed by a 3D Butterworth lowpass filter of order 4 and cut-off 0.2 cycles per pixel. The numbers in each cell indicate the iteration number. (d) The graph plots the log-likelihood function after each iteration up to 100 for the noisy data set.

is slow. A usable solution may require 30–50 iterations. Because each iteration requires one forward projection and one backprojection, the ML-EM algorithm can be expected to require one to two orders of magnitude more processing time than FBP. However, one should keep in mind that the ML-EM algorithm can perform much better than FBP because of its ability to solve the generic linear model and thus compensate for nonuniform attenuation, and so forth. Nevertheless, the required computation time initially hampered acceptance of the ML-EM method in clinical use.

The second shortcoming of the ML-EM algorithm is that the ML criterion on which it is based yields very noisy reconstructed images, as explained earlier. Thus, as the ML-EM iterations proceed, and the algorithm approaches the ML solution, the variance of the image estimate, which is manifested as noise, increases.

In practice, the ML-EM algorithm yields good results if the iterative procedure is stopped prematurely, and the results may, in addition, benefit from application of a postreconstruction lowpass filter, which is a common approach in clinical applications. Several approaches for deciding when to stop the iterations were proposed after the introduction of the ML-EM algorithm (e.g., Llacer and Veklerov, 1989). An alternative approach is the method of sieves (Snyder and Miller, 1985), which involves smoothing within each iteration to constrain the solution.

The behavior of the ML-EM algorithm with iterations is demonstrated in the example in Figure 11, which shows brain SPECT data realistically simulated using the Zubal brain phantom (Zubal *et al.*, 1994). ML-EM has been shown consistently to cause low spatial frequencies to appear first and then gradually to develop higher spatial frequencies as the iterations progress (Wilson *et al.*, 1994). Thus, as illustrated in the noise-free results of Figure 11a, early stoppage of ML-EM iterations amounts to an implicit form of smoothing of the reconstructed image. Figure 11b shows the image noise increasing as the estimate approaches the ML solution, but the postfiltered results (Fig. 11c) show little change after 50 iterations because the filter removes the frequency components that emerge in later iterations. The graph in Figure 11d shows the progression of the log-likelihood as a function of iterations, indicating that, although the log-likelihood appears to plateau early in the iterative process, the image estimates continue to change. The log-likelihood function is generally not a good measure of image quality for the same reason that the ML criterion is not an ideal reconstruction criterion, and it should be noted that images having nearly the same log-likelihood value can appear very different.

2. Variations on ML-EM

To address the problem of slow convergence, several methods for accelerating the ML-EM algorithm have been

proposed. These include methods for extrapolating and increasing the magnitude of the change made at each iteration (Kaufman, 1987; Lange *et al.*, 1987; Lewitt and Muehllehner, 1986; Rajeevan *et al.*, 1992; Tanaka, 1987), using grids of different sizes to reduce processing time (Pan and Yagle, 1991; Ranganath *et al.*, 1988), and increasing the number of updates by using only part of the data in each update (Browne and De Pierro, 1996; Byrne, 1996; Hudson and Larkin, 1994). Most of these methods improve the initial convergence of the algorithm; that is, they arrive at the log-likelihood plateau in fewer iterations. However, after reaching the plateau, the convergence rate is generally not increased greatly. An exception to this is the ordered-subsets EM (OS-EM) algorithm, which we discuss later.

The great success of the OS-EM algorithm in speeding up the reconstruction process, along with improvements in computer power, has lessened the practical need for additional acceleration strategies. However, there are other strategies worth noting. For example, the space-alternating generalized EM (SAGE) algorithm (Fessler and Hero, 1994) improves the convergence rate by updating each pixel individually and using a matrix-based projection model. SAGE is based on an alternate approach to the complete-data space, using instead a series of different complete-data spaces. The resulting algorithm converges to the ML estimate, obtains a usable result in fewer than 20 iterations, and can easily incorporate smoothing constraints. However, SAGE is inefficient for large 3D problems where projector-based models (Section VIIIA) must be used.

D. Least-Squares and Weighted-Least Squares Algorithms

The WLS criterion presented in Section IVC results in a quadratic function to be optimized. Optimization problems involving quadratic objective functions have been widely studied; therefore, there are many tools available for WLS reconstruction. The LS criterion is a special case of WLS, so the following discussion can be applied to LS reconstruction by choosing the weighting matrix to be the identity matrix.

1. General Structure of Methods

Many algorithms can be used to solve the WLS optimization problem in Eq. (14). All of them follow the same basic additive update formula:

$$\hat{\mathbf{f}}^{(n+1)} = \hat{\mathbf{f}}^{(n)} + t\,\Delta\mathbf{f}^{(n)} \tag{31}$$

where n is the iteration number; the scalar t is referred to as the *step size*; and the vector $\Delta\mathbf{f}^{(n)}$, which has the same dimensions as the image, is called the *step direction*. If we consider the multidimensional solution space, wherein each potential solution image is a point, then we can consider the iterative update as a movement through the solution space from $\hat{\mathbf{f}}^{(n)}$ to $\hat{\mathbf{f}}^{(n+1)}$.

2. Optimal Step Size

Algorithms for solving the WLS problem differ primarily in the manner in which the step direction is determined. For an arbitrary step direction, it is relatively simple to determine the step size that results in the greatest decrease in the weighted squared error. Let $J(\mathbf{f}) = (\mathbf{g} - \mathbf{Hf})^T\,\mathbf{D}(\mathbf{g} - \mathbf{Hf})$ denote the WLS objective function. Then we can optimize the step size by setting:

$$\frac{\partial J(\mathbf{f})}{\partial t}\Big|_{\mathbf{f}^{(n)}+t\Delta\mathbf{f}^{(n)}} = 0$$

and solving for t to obtain:

$$t = \frac{\Delta\mathbf{f}^{(n)^T}\mathbf{H}^T\mathbf{D}\left(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(n)}\right)}{\Delta\mathbf{f}^{(n)^T}\mathbf{H}^T\mathbf{D}\mathbf{H}\Delta\mathbf{f}^{(n)}} \tag{32}$$

3. Steepest Descent

The steepest descent (SD) (or gradient descent) algorithm uses the gradient of the objective function as the step direction. In other words, the algorithm proceeds in each step in the direction in which the objective function decreases most quickly:

$$\Delta\mathbf{f} = -\frac{\partial J(\mathbf{f})}{\partial \mathbf{f}}\Big|_{\hat{\mathbf{f}}^{(n)}} = -\mathbf{H}^T\mathbf{D}\left(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(n)}\right) \triangleq \mathbf{p} \tag{33}$$

To implement the algorithm efficiently requires a variation on the model of Figure 6, as shown in Figure 12. In this case, the projection space error $\mathbf{e} = (\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(n)})$ is updated incrementally along with the new image estimate and the next step direction. The initial gradient must be computed with an additional backprojection operation before the first iteration. Like the model in Figure 6, only a single backprojection and a single projection operation are required at each iteration. One iteration of SD takes about the same time as one iteration of ML-EM.

Although intuitive and relatively simple to implement, SD algorithms generally require more iterations than conjugate gradient algorithms to reach a usable solution. It is possible to enforce a nonnegativity constraint by setting negative pixels to zero after each iteration, but this requires recomputing the error vector $\mathbf{e}$ at the beginning of each iteration with an additional projection operation.

4. Conjugate Gradient

The conjugate gradient (CG) algorithm is more efficient than SD (Press *et al.*, 1988) and has been applied effectively to reconstruction in ET (Huesman *et al.*, 1977; Kaufman, 1993; Tsui *et al.*, 1991). In CG, all step directions are chosen to be conjugate to one another:

$$\left(\mathbf{H}\Delta\mathbf{f}^{(i)}\right)^T\mathbf{D}\left(\mathbf{H}\Delta\mathbf{f}^{(j)}\right) = 0 \;\text{ if } i \neq j \tag{34}$$

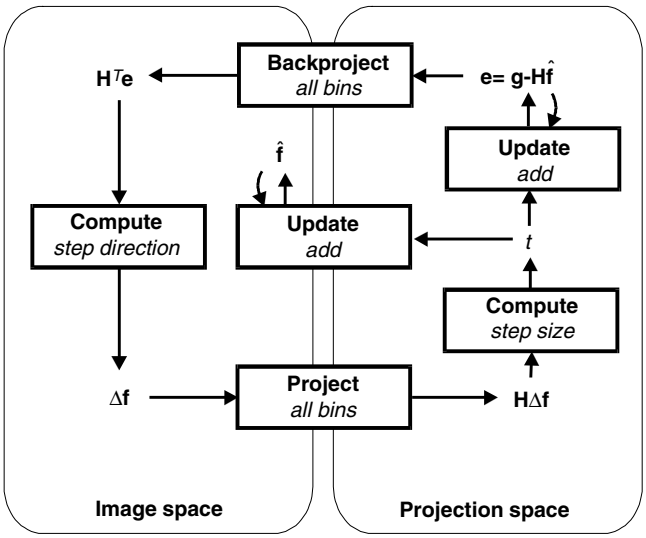

FIGURE 12 Weighted least-squares algorithms in a modified form of the general iterative model. The steepest descent, conjugate gradient, and coordinate descent forms differ only in how the step directions are computed.

where i and j are iteration numbers. Thus, minimization along a step direction does not interact with minimizations performed with respect to any previous step directions. In other words, each new step does not spoil the work of previous steps. For this reason, CG is guaranteed to find the minimum of a quadratic objective in N iterations, where N is the number of pixels in the image. However, a small 64×64 image will require 4096 iterations to reach the minimum! Fortunately, a usable image estimate is generally obtained in far fewer iterations.

The following relation, called the Fletcher-Reeves variant (Press *et al.*, 1988), can be used to compute conjugate gradient step directions:

$$\Delta\mathbf{f}^{(n+1)} = \mathbf{p}^{(n)} + \frac{\mathbf{p}^{(n)^T}\mathbf{p}^{(n)}}{\mathbf{p}^{(n-1)^T}\mathbf{p}^{(n-1)}}\Delta\mathbf{f}^{(n)} \tag{35}$$

where the superscript n is the iteration number and the gradient $\mathbf{p}$ is computed as in Eq. (33). Here, the step direction depends on the previous step direction as well as the gradient directions at the current and previous iterations. This results in a slightly more complex computation of the step direction as compared to steepest descent, but a more efficient minimization of the weighted squared error. Enforcing a nonnegativity constraint is problematic, however, because manipulating negative pixels after one iteration effectively changes the step direction and ruins the delicate sequence of conjugate step directions generated. The implementation of CG can be done in the same form as Figure 12, substituting Eq. (35) for the step direction computation.

5. Coordinate Descent

A relatively simple approach to generating step directions is to update one pixel at a time. If a matrix-based model can be used (i.e., if the matrix $\mathbf{H}$ can be computed and stored), then a coordinate descent approach can be extremely efficient. If not, then coordinate descent is very inefficient. Coordinate descent has the added advantage of permitting the simple application of a nonnegativity constraint: If a pixel goes negative at one update, it is simply thresholded to zero with no real penalty to the algorithm. For coordinate descent, $\Delta\mathbf{f}$ is simply a vector with a 1 in the position of the pixel to be updated and zeros elsewhere. The optimal step size is again computed using Eq. (32), but it is only necessary to apply the elements of $\mathbf{H}$ that operate on the pixel in question, hence the need to be able to look up individual elements of $\mathbf{H}$. Coordinate descent has been applied successfully to ET reconstruction (Fessler, 1994) with the addition of a smoothing penalty, a method that is discussed later.

6. Properties of WLS Algorithms

The WLS-CG algorithm is not much faster than ML-EM, unless a transformation of the image space is first performed. This transformation matrix, called a *preconditioner* or *relaxation matrix* (Golub and Van Loan, 1989), seeks to equalize the curvature of the objective function along all its axes, making the algorithm more efficient in finding the minimum. The preconditioning process involves creating a transformed image space $\mathbf{v}$ as follows:

$$\mathbf{v} = \mathbf{M}\mathbf{f},\ \mathbf{M} = [\text{diag}\,(\mathbf{H}^T\mathbf{D}\mathbf{H})]^{-1/2} \tag{36}$$

The algorithm then solves for the gradients and step directions in the $\mathbf{v}$-space, converts the step direction back to the $\mathbf{f}$-space, and solves as usual for step size. The result (Lalush and Tsui, 1995) is that the computation of the current gradient $\mathbf{p}$ becomes:

$$\mathbf{p}^{(n)} = -\mathbf{M}^{-1}\mathbf{H}^T\mathbf{D}\left(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(n)}\right) \tag{37}$$

and the step direction in the $\mathbf{f}$-space is computed directly for preconditioned CG as:

$$\Delta\mathbf{f}^{(n+1)} = \mathbf{M}^{-1}\mathbf{p}^{(n)} + \frac{\mathbf{p}^{(n)^T}\mathbf{p}^{(n)}}{\mathbf{p}^{(n-1)^T}\mathbf{p}^{(n-1)}}\Delta\mathbf{f}^{(n)} \tag{38}$$

The preconditioned WLS-CG algorithm has been claimed to converge up to 10 times faster than ML-EM (Tsui *et al.*, 1991). In general, from 8 to 15 iterations are required to reach a usable solution. Iterated image comparisons are shown in Figure 13. These show that, without preconditioning, convergence is still rather slow. With preconditioning, convergence is rapid, although rather significant changes occur from one iteration to the next at early iterations. Also, the spatial frequency recovery is not as gradual or predictable as in ML-EM. As in ML-EM, image noise increases with further iterations as the solution gets

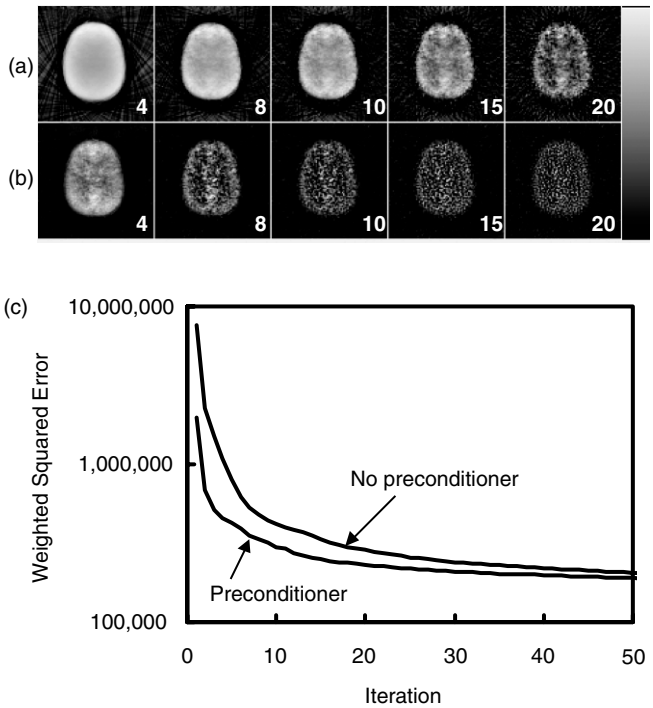

FIGURE 13 The convergence of the weighted least-squares conjugate gradient algorithm (a) without and (b) with a preconditioner. The images show the iterated image estimates from simulated brain SPECT data at the iteration numbers indicated. (c) The graph shows the weighted squared error as a function of iteration.

closer to the Gaussian ML solution. However, image noise with preconditioned WLS-CG tends to be somewhat higher than that obtained from ML-EM. This can be attributed to the lack of a nonnegativity constraint; when a pixel is affected by noise and tends to become negative, a nearby pixel may seek to cancel out this effect and increase in a positive direction. If pixel values are constrained from becoming negative, as in ML-EM, then large positive variations tend to be suppressed as well.

7. Other Poisson-Based ML Algorithms

Some of the optimization principles applied to the quadratic WLS objective have been applied to maximizing the Poisson likelihood function. One reason for this is to more accurately model the statistics in projection bins in which few events are acquired and the Gaussian approximation is less accurate. It is possible to derive a CG algorithm based on the nonquadratic Poisson likelihood (Mumcuoglu *et al.*, 1994). It is important to note that, with a nonquadratic objective, the conjugate gradient method no longer has the guarantee of convergence in N iterations, although this is of little practical use in the case of image reconstruction. Also, although the use of a Poisson objective constrains the projections of the image estimates to be nonnegative, individual pixel values may still become negative; thus, there is no inherent nonnegativity constraint on the pixels.

The biggest problem created by applying SD or CG to the Poisson-based objective is that there is not an explicit solution for the optimal step size as in Eq. (32). Thus, either a heuristic for step size must be used, or a 1D search method (Mumcuoglu *et al.*, 1994) must be employed to find the optimal step size. The former makes the algorithm difficult to use because the convergence will be data-dependent and dependent on the heuristic used. The 1D search is effective and does not significantly impact processing time, as shown in Mumcuoglu *et al.* (1994), but adds to the complexity of programming. Furthermore, preconditioning is required to achieve convergence rates comparable to those from WLS-CG.

E. Maximum A Posteriori Reconstruction Algorithms

As explained earlier, the ML solution is generally too noisy to be useful, so something must be done to control noise. In current clinical applications, the usual approach is to stop the iterations early and/or apply a linear filter. However, the MAP approach offers a more flexible and principled method of encouraging desirable properties in the reconstructed image.

1. MAP Algorithms

Owing to the similarities between the MAP and ML objective functions, most of the algorithms used for ML reconstruction have MAP counterparts. However, the prior term can often create complications that require approximations or additional steps. In other cases, the algorithm can be applied without difficulty, but only for certain forms of the prior.

The analog of the ML-EM algorithm is the following MAP-EM algorithm (Green, 1990):

$$\hat{f}_j^{(n+1)} = \frac{\hat{f}_j^{(n)}}{\sum_{i'} h_{i'j} + \beta \dfrac{\partial U(\mathbf{f})}{\partial \mathbf{f}}} \sum_i h_{ij} \frac{g_i}{\sum_k h_{ik} + \hat{f}_k^{(n)}} \tag{39}$$

This algorithm deviates from the ML-EM algorithm principally in that there is a prior term in the denominator. This term is problematic because it is supposed to be evaluated using the next estimate, $\hat{f}_j^{(n+1)}$, which is not yet available. MAP-EM algorithms differ in the approach they use to address this problem. The simplest approach is to evaluate the derivative term at the previous image estimate, a method called the one step late (OSL) procedure (Green, 1990). The MAP-EM OSL algorithm has been shown to converge to the MAP solution for only certain forms of the prior (Lange, 1990). Another type of approach, called a generalized EM (GEM) MAP algorithm (Hebert and Gopal, 1992), sequentially updates pixels and verifies that each update increases the posterior density, thus ensuring convergence to a maximum of the posterior density.

Several proposed algorithms have been based on the Gaussian likelihood function. The PWLS algorithm (Fessler 1994) employs a coordinate descent approach and uses a quadratic prior to solve explicitly for the optimal step size t. The algorithm is very efficient, usually requiring approximately 10–15 iterations, but requires a matrix-based approach and restricts priors to only quadratic forms. More general priors and projector-based models can be accommodated by a MAP conjugate gradient (MAP-CG) algorithm (Lalush and Tsui, 1995), which is analogous to the WLS-CG algorithm. This algorithm also requires approximately 10–15 iterations, but adds some programming complexity with the need to perform a local linear fit of the prior term in the step size calculation and has no nonnegativity constraint. It is also possible to use an EM algorithm with a Gaussian likelihood and a Gaussian prior (Levitan and Herman, 1987), but the resulting algorithm requires more iterations than PWLS or MAP-CG.

A Poisson-based CG algorithm, termed *preconditioned conjugate gradient* (PCG) (Mumcuoglu *et al.*, 1994), is the MAP analog of the Poisson-based CG algorithm. It requires line searching to optimize the step size, but its convergence rate is competitive with PWLS and MAP-CG and it does not have the convergence problems of the OSL method.

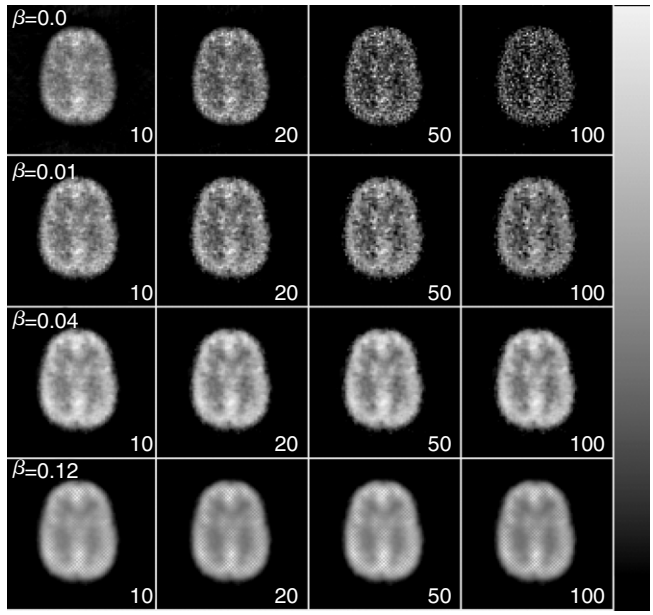

FIGURE 14 Iterated image estimates for the maximum *a posteriori* expectation-maximization one-step-late algorithm from noisy simulated brain SPECT data. Each row shows results for a different value of the weighting parameter, β, which determines the relative weight placed on the smoothing prior.

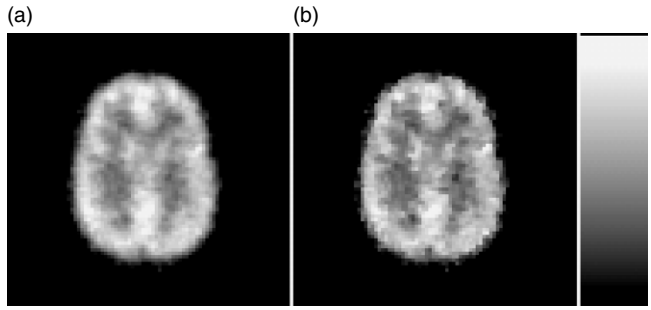

FIGURE 15 Results of using different types of potential functions with MAP-EM, zoomed to show detail. (a) Using a quadratic potential. (b) Using a linearly increasing, convex potential set to discourage small pixel differences.

2. Properties of MAP Reconstructions

MAP methods alleviate the principal problems associated with ML algorithms. First, MAP reconstructions are smoother than their ML counterparts. Second, iterated MAP estimates tend to reach a point at which they change very little with further iterations, indicating approximate convergence (this behavior can be controlled by the weighting parameter β) (Lalush and Tsui, 1992). Figure 14 shows an example of the convergence behavior of MAP-EM OSL reconstructions with various values of β compared to ML-EM reconstruction. For moderate values of β, MAP reconstructions reach a point of effective convergence and are clearly smoother than their ML counterparts. As the value of β is decreased, the degree of smoothing is reduced because the criterion places less weight on the smoothing requirement of the prior. As β is increased, smoothing increases, but potentially important features in the image begin to degrade. Setting the weighting parameter is critical because excessive β values invariably result in a loss of contrast and detail and insufficient β values produce images that are too noisy.

Different types of priors produce different smoothing characteristics, as shown in Figure 15. Quadratic priors tend to result in smoothing that is qualitatively similar to that obtained by space-variant linear filtering (Fessler and Rogers, 1996). Linearly increasing priors (Green, 1990; Hebert and Leahy, 1992; Lalush and Tsui, 1992, 1993; Mumcuoglu *et al.*, 1994) usually have an adjustable *scaling parameter* that sets the degree to which small pixel differences are discouraged. If strong smoothing of small pixel differences is applied, the result is a piecewise-continuous image with a number of regions of uniform intensity. This can result in spurious regions due to noise and false boundaries, which can be distracting to the viewer. Also, small and low-contrast image features may be lost. If the scaling parameter is set to allow small differences between neighboring pixels, the result can smooth noise while preserving a degree of sharpness in object boundaries.

Although MAP reconstruction successfully smoothes noise and improves convergence, it also has several disadvantages. First, success can be highly dependent on the choice of parameters. Unfortunately, it is not efficient to use

a trial-and-error approach to parameter selection, as might be used with postreconstruction filters, because an entire iterative reconstruction must be performed to assess the result from one set of parameters. Second, excessive smoothing using Gibbs priors can result in a loss of image features or, in some cases, the creation of spurious features. This is a manifestation of the fact that the MAP estimator adds some bias to the ML problem in exchange for reduced noise variance. Finally, MAP algorithms apply smoothing properties that are somewhat different from the traditional Fourier-domain filters used in nuclear medicine, and so they may be difficult for physicians to interpret initially.

F. Subset-Based Reconstruction Algorithms

As noted earlier, the ML-EM algorithm has many desirable properties, but it suffers from slow convergence, a problem that has been addressed very successfully by subset methods (also called *block-iterative* or *row-action* methods). These techniques break up the full set of projection data into a series of mutually exclusive subsets and apply the reconstruction algorithm to each subset sequentially. Each pass through the data set is used to effect a greater number of iterative refinements, which results in significant acceleration compared with ML-EM, usually on the order of the number of subsets.

1. The OS-EM Algorithm

The OS-EM algorithm (Hudson and Larkin, 1994) is a simple modification of the ML-EM algorithm, given by:

$$\hat{f}_j^{new} = \frac{\hat{f}_j^{old}}{\sum_{i' \in S_n} h_{i'j}} \sum_{i \in S_n} h_{ij} \frac{g_i}{\sum_{i=1}^{N} h_{il} \hat{f}_l^{old}} \tag{40}$$

where the backprojections are performed only for the projection bins belonging to the subset S_n. At each update, a different subset of the projection data is used. Generally, one update in this algorithm is called a *subiteration* and one pass through all of the subsets is referred to as an *iteration*. Thus, the processing time for one iteration of OS-EM is comparable to that of one iteration of ML-EM.

The organization of the subsets is important to the performance of the algorithm. In addition, mathematical difficulties can result if any subset does not contain some contribution from every pixel in the field of view; in this case, the first summation in the denominator of Eq. (40) is zero. This is an important consideration in nonparallel projection geometries.

When reduced to a single subset encompassing all the projection data, the OS-EM algorithm reduces to the ML-EM algorithm. At the other extreme, one could define each subset to include a single projection bin, resulting in an algorithm similar to MART; however, this would preclude the use of projector-based models of **H**.

Usually, subsets are organized in groups of projection bins associated with one projection view or camera position, which is convenient for projector-based models. Figure 16 illustrates the typical procedure for a parallel projection model. First, a number of projection views are used to compute the update for the first subset. Then, more views are used to compute the next update, and so on. After all subsets have been used, the process begins again with the first subset.

Usually, the members of a subset are chosen to have maximum angular distance between them. For example, when creating 16 subsets from data with 128 projection views over a 360° arc, each subset will contain all the bins from eight views spaced at angular intervals of 45°. The sequence of updating subsets usually aims to maximize the angular spacing between successive subsets. In our example, the first subset may be chosen at cardinal directions, the second at 22.5° from the first, and the third subset offset one angular increment from the first (approximately 2.8°).

These are typical and intuitive approaches to organizing the subsets; however, evidence suggests that results are not

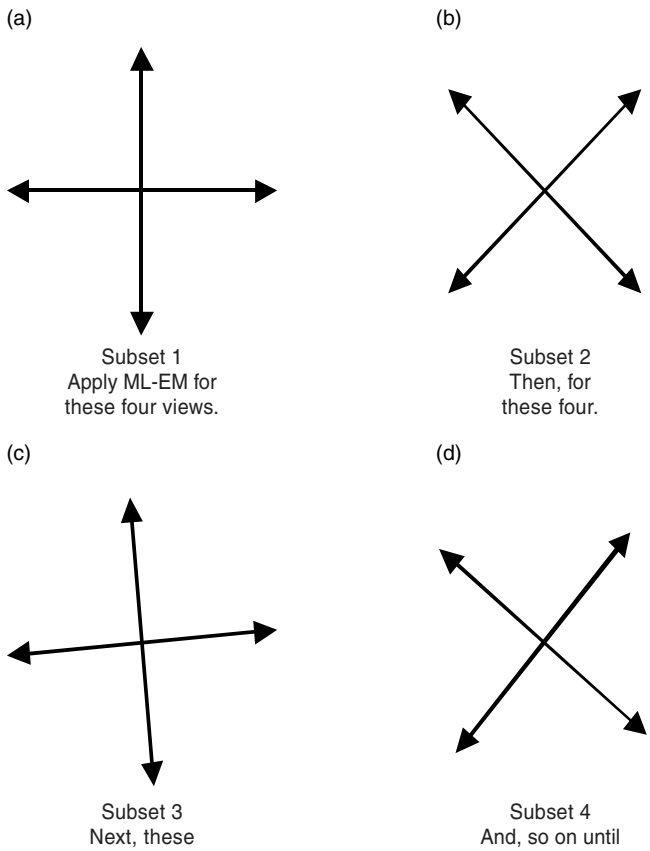

FIGURE 16 The process of using subset updates in a reconstruction algorithm. (a) Subset 1: Apply ML-EM for these four views. (b) Subset 2: Then apply the algorithm for these four. (c) Subset 3: Then apply the algorithm for these four. (d) Subset 4: Continue until all data are used.

very sensitive to the particular views chosen for a subset or to the ordering of the subsets, although suggestions for balancing the subsets are given in Hudson and Larkin (1994). On the other hand, the *number* of subsets, or alternately the number of views per subset, governs the degree of acceleration over ML-EM. A rule of thumb states that OS-EM at n iterations reaches roughly the same point of convergence as ML-EM at (Number of Subsets) $\times$ n iterations. The noise increases that much more quickly also, so the algorithm must be stopped and smoothing applied, just as in ML-EM.

2. Properties of OS-EM

Figure 17 shows examples of iterated image estimates using OS-EM for various numbers of subsets. Sustantial acceleration is achieved with little impact on the resolution and noise properties of the images. OS-EM appears to have nearly all the desirable properties of ML-EM, but with significant speed improvement. It generally requires fewer than seven iterations, compared to 8–12 for the fastest WLS-based algorithms. It has a built-in nonnegativity constraint and shares the property of ML-EM that low spatial frequencies converge first, with higher spatial frequencies improving with further iterations (Lalush and Tsui, 1998b). Its convergence properties are therefore very predictable and reproducible. The OS-EM algorithm is simple to implement, although there is the small added complexity of having to choose the number of subsets.

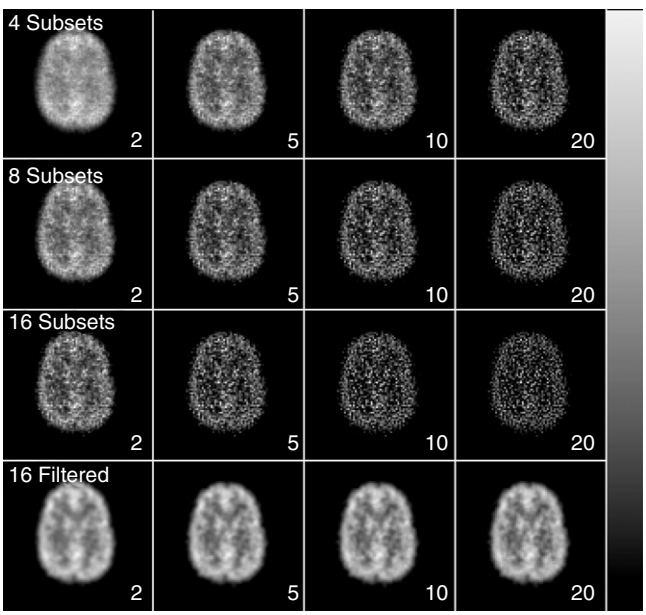

FIGURE 17 Iterated image estimates using the ordered-subsets expectation-maximization algorithm for simulated brain SPECT data. Iteration numbers are indicated in each cell. The original data had 128 views over a 360° arc and was broken up into 4, 8, and 16 subsets. The last row shows the results of filtering the 16-subset iterated estimates with a 3D Butterworth lowpass filter of order 4 and cut-off 0.2 cycles per pixel.

With all of its similarities to ML-EM, it is important to note that OS-EM is not really an EM algorithm and, at this writing, has no general proof of convergence. Experience with the algorithm has been generally good, and it has been shown, at least in some cases, to give results nearly identical to ML-EM in the mean (Lalush and Tsui, 2000). The principal cost of using the subset method is an increase in image noise for the same level of bias as compared to ML-EM (Lalush and Tsui, 2000). In other words, as more subsets are used in an effort to increase acceleration, image noise becomes worse. Thus, users should be wary of using a large number of subsets; modest acceleration of 8–10 times is possible with very little increase in noise.

3. Related Algorithms

Several other algorithms use the same subset concept as OS-EM to achieve usable results in a small number of iterations. The rescaled block-iterative (RBI) EM algorithm (Byrne, 1996) can be shown to converge to a consistent solution, provided one exists. This is a minor theoretical advantage, because most noisy data are bound to be inconsistent. In practical application, RBI-EM behaves almost identically to OS-EM, except that it tends to run approximately one-half as fast for the same number of subsets (Lalush and Tsui, 2000). The row-action maximum-likelihood algorithm (RAMLA) (Browne and De Pierro, 1996), developed independently, is almost identical to OS-EM, with the exception of a step-size adjustment factor. If this factor is decreased properly at successive iterations, then RAMLA will converge to the ML solution. Because the ML solution is not desirable due to its noise properties, this is again mostly a theoretical advantage. Several MAP analogs to the subset approach have also been developed (Hudson and Larkin, 1994; Lalush and Tsui, 1998a). For the most part, these cannot be proven to converge to the MAP estimate, but do appear to achieve the properties of MAP-EM, only faster.

G. Iterative Filtered Backprojection Algorithms

Many algorithms use an additive update that includes a filtered backprojection operation in the backprojection step. Although most of these iterative filtered backprojection (IFBP) methods were developed from intuitive arguments (i.e., without any specific governing criterion), they can be analyzed using some of the same principles that apply to WLS algorithms.

1. General IFBP Model

A general model for IFBP algorithms uses an additive update, as in Eq. (32), with the step direction determined by:

$$\Delta\mathbf{f}^{(n)} = \mathbf{B}\left(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(n-1)}\right) \tag{41}$$

where the backprojection matrix $\mathbf{B}$ involves some use of a ramp filter or filtered backprojection. Contrast this with

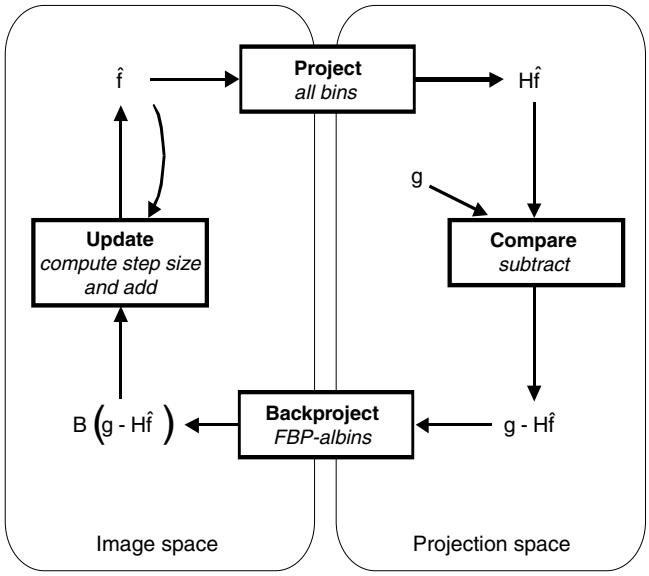

FIGURE 18 Iterative filtered backprojection algorithms in the form of the general iterative model.

the ML-EM and WLS algorithms, in which the backprojection step involves only the application of $\mathbf{H}^T$, which would be a conventional Radon backprojection in the case in which $\mathbf{H} = \mathbf{H}_{\text{Radon}}$. Different IFBP algorithms use different forms of the backprojection operation $\mathbf{B}$ and use different rules for determining step size. Figure 18 illustrates how IFBP algorithms follow the general iterative model.

2. Specific Algorithms

An early and widely used IFBP algorithm is the iterative Chang reconstruction method (Chang, 1979). This method was developed to facilitate reconstruction with correction for uniform or nonuniform attenuation in SPECT. The Chang algorithm uses $\mathbf{B} = \mathbf{W}\mathbf{B}_{\text{FBP}}$, where $\mathbf{W}$ is a diagonal weighting matrix having the same dimension as the image and composed of factors based on the average attenuation experienced by each pixel over the acquired arc, and $\mathbf{B}_{\text{FBP}}$ is the filtered backprojection operation. The step size t is always 1. This leads the Chang algorithm to be potentially unstable or oscillatory as iterations progress (Lalush and Tsui, 1994), a behavior that depends on the data. Still, if the attenuation distribution is relatively uniform, Chang's method can give reasonably accurate results in the first or second iteration.

The It-W iterative reconstruction algorithm (Wallis and Miller, 1993) performs the backprojection step by first applying a ramp filter to the projection-space error and then backprojecting with $\mathbf{H}^T$. The result is then weighted by the square of the Chang weighting matrix, $\mathbf{W}^2$. The step size in this case is applied to maintain a constraint on the total intensity in the reconstructed image, which should help prevent instability. The algorithm is fast (Wallis *et al.*, 1998), but it is complex to program and requires a number of adjustments to perform well.

Approximate SD and CG IFBP algorithms have also been developed (Lalush and Tsui, 1994). These methods are based on minimizing an unusual squared-error function whose gradient is approximately the step direction given in Eq. (41). Because the step size is optimized in this case, the algorithm is guaranteed to be stable. When the backprojection operator from the Chang algorithm is used in this context, the result is a stable version of the Chang algorithm. However, evidence does not indicate that IFBP-CG outperforms WLS-CG in any way.

Many other IFBP algorithms have been proposed (Liang, 1993, Maze *et al.*, 1993; Pan *et al.*, 1993; Walters *et al.*, 1981; Xu *et al.*, 1993). These use various methods for computing $\mathbf{B}$ and the step size. For the most part, they are either potentially unstable or are stable but less efficient than methods with optimal step size. As with the other iterative algorithms, all IFBP algorithms suffer from increasing image noise as iterations proceed, so they are generally stopped after a certain number of iterations and the results are filtered.

VI. EVALUATION OF IMAGE QUALITY

In this chapter we have reviewed a great many approaches to image reconstruction, which begs the question: Which method is best? Unfortunately, the answer is not a simple one; each method has its strengths and weaknesses, and algorithms may perform differently in different applications. Several of the mathematical tools of image evaluation, such as receiver operating characteristic (ROC) analysis, which are explained in the last portions of Kao *et al.* (Chapter 6 in this volume), should be studied in conjunction with the following discussion.

It is now generally accepted that the true test of an imaging system or image-reconstruction algorithm is its ability to produce images that help achieve a desired goal (Barrett, 1990; Barrett *et al.*, 1995; Metz, 1978, 1986, 1989). This approach to image evaluation is called *task-based assessment*. For example, if the images are to be used by physicians to diagnose disease, then an algorithm can be considered to perform well if it produces images that lead to accurate diagnoses. On the other hand, if the reconstruction algorithm is to be used for functional brain mapping, in which statistical methods are applied to detect regions of neuronal activation, then a "good" algorithm is one that causes the statistical detection methods to produce accurate results. Finally, if the task is quantitative, such as the measurement of tracer concentration in a given image region, then the reconstruction algorithm should be judged by its quantitative accuracy.

A sound test of image quality for performing a quantitative task can consist of the use of a physical phantom in which the true value of the quantity to be estimated (such as tracer concentration) is known and can be compared to the value reflected in the reconstructed image. However, it is sometimes difficult to make such a study truly representative of the clinical task, and so realistic simulation methods may be used instead. Ideal testing for a visual task, such as diagnosis by a physician, usually requires a human-observer study (Metz, 1978, 1986, 1989), in which trained observers view images for which the true state is known, permitting the diagnostic accuracy of the observers to be assessed. This generally requires a large number of realistic simulations or patient images with confirmed diagnosis.

The effort required to conduct these ideal performance tests can be substantial; therefore, it may not be practical in early stages of algorithm development to conduct exhaustive tests of this kind. Initially, simpler tests are often used to narrow the search for the "best" algorithm and imaging methodology. These tests include measures of statistical estimation performance (e.g., bias and variance), measures of effective spatial resolution, and numerical observers that act as surrogates for human observers.

A. Bias and Variance

Because ET data are random, an image reconstructed from these data is also random—every time an object is scanned, a slightly different image will be obtained. Borrowing from statistical-estimation theory, image-reconstruction algorithms are often evaluated in terms of the mean and variance of the average intensity within a region of interest. Ideally, we hope that the mean reconstructed image will be identical to the true image (or at least identical to a discrete pixel representation of the image). Our success in achieving this objective can be quantified by using a quantity called *bias*, which is the difference between the mean reconstructed image and the true image, $\mathbf{b} = E[\hat{\mathbf{f}}] - \mathbf{f}$. Furthermore, we want the variance of the reconstructed image to be identically zero, meaning that the reconstruction method produces exactly the same image from every noisy set of projections of the same object.

Of course, it is impossible to achieve both of these goals simultaneously. In fact, effort toward one goal tends to work against attainment of the other. This notion, which is a common theme throughout the statistical-estimation field, is commonly called the *bias-variance trade-off*. It is possible to achieve low variance by smoothing the reconstructed image to the point at which it has no visible detail. In this case, the cost of low variance is high bias; the image is inaccurate in the mean. Conversely, it is also possible to achieve low bias by using no smoothing at all (e.g., by running the ML-EM algorithm to convergence); however, the resulting image will exhibit high variance (and will appear very noisy).

In image reconstruction, the degree of enforced smoothness controls the bias-variance trade-off. Smoothness may be controlled implicitly (by the number of iterations of the ML-EM algorithm) or explicitly by adjusting the parameter β in MAP-EM. Either way, a curve in bias-variance space can be generated by sweeping through a range of values for the governing parameter. Figure 11a shows how the mean of the ML-EM estimate improves as the iterations progress. Figure 11b shows that, at the same time, the variance (noise level) of the image worsens. Therefore, we must always settle for the best compromise between these two competing factors.

Bias and variance can be computed in two ways. The obvious way is to generate many noise realizations of a reconstructed image and empirically calculate the mean and variance of these results (Wilson *et al.*, 1994). Owing to the computational burden of this approach, it is often preferred to compute these statistics directly when possible (Barrett *et al.*, 1994, Fessler, 1996, Qi and Leahy, 1999, Qi and Huesman, 2001, Xing and Gindy, 2002, and Jinyi, 2003).

B. Effective Spatial Resolution

Due to the principle of superposition, a linear imaging system (which includes a reconstruction algorithm in the case of ET) is characterized entirely by its PSF, that is, the image of a point object. Thus, the PSF provides a complete description of the spatial resolution produced by the system or algorithm in the absence of noise. However, when the reconstruction algorithm is nonlinear, the PSF does not have such a clear meaning because the image of a point source depends not only on the imaging system and reconstruction method, but also on the object itself. Nevertheless, the effective PSF is a helpful way to shed light on the spatial resolution properties of the reconstruction algorithm. Two methods that have been proposed for quantifying resolution are the effective local Gaussian (Liow and Strother, 1993) resolution and the local impulse response (Stamos *et al.*, 1988 and Fessler and Rogers, 1996).

The use of spatial resolution as a metric implies that the task requires the visibility of small details, which is not always the case. Further, it does not consider the local power spectrum of noise and how noise may interfere with the task to be performed. For this reason, some authors compute a local noise power spectrum and compare this to the system frequency response derived from the PSF (Wilson *et al.*, 1994; Lalush and Tsui, 1998b) to get a more complete picture of the noise-resolution trade-off.

C. Numerical Observers

If an image is to be assessed visually by a physician, it is widely accepted that the most conclusive test of a reconstruction algorithm is a human-observer study. When the image is to be used to determine the presence or absence of

a lesion, a study can be designed to determine whether human observers (e.g., physicians) can make this decision reliably. However, human-observer studies are tedious and time-consuming; therefore, they are usually reserved for the final testing of algorithms. During the early development phases, it is helpful to have a computer algorithm, called a *numerical observer*, that can mimic human-observer performance and thus stand in as a surrogate for human observers. By far the most popular numerical observer is the channelized Hotelling observer (CHO) (Barrett *et al.*, 1993; Hotelling, 1931; Burgess *et al.*, 1997; Myers and Barrett, 1987; Myers *et al.*, 1990; Hutton and Strickland, 1997; Wollenweber *et al.*, 1999; Narayanan *et al.*, 2002).

The CHO uses the generalized likelihood-ratio test (GLRT) from statistical detection theory (Kay, 1993) as the lesion detector. The GLRT for the problem of an additive signal in Gaussian signal-independent noise is sometimes called the Hotelling detector. It so happens that the GLRT for this problem is a linear detector, which makes it particularly easy to implement. In the CHO, the features provided to the Hotelling detector are obtained by applying various bandpass filters to the image to coarsely simulate the receptive fields of the human visual system. The features so computed are called *channels*. Using this model of the human observer, it is possible to roughly predict how well humans will perform in detecting lesions from a class of images. It is possible to order reconstruction algorithms in terms of performance by comparing the lesion-detection accuracy predicted for each algorithm.

VII. SUMMARY

Iterative image reconstruction algorithms share a number of common traits. Most are general enough to address any tomographic problem that can be expressed by a linear relationship between the image pixels and the projection bins. Most fit reasonably well into a general model that involves repeating the processes of projecting an image estimate, comparing the estimated projections to the measured data to compute some form of error, backprojecting the error, and using the error to update the image estimate. The principle of feedback works in favor of the iterative algorithm, so that as iterations continue the image estimate gradually approaches some desired result. Different algorithms can be compared on the basis of a number of properties, but those that largely govern the acceptance of an algorithm are its stability, accuracy, speed, and ease of use or implementation.

Classical iterative algorithms, such as ART, are based on the Kaczmarz method of solving systems of linear equations. Although not widely used in nuclear medicine, they form an important basis for understanding the statistical algorithms that were developed later. Statistical algorithms have two basic parts: a statistical criterion (the basis for determining which image among the many possible is to be the solution) and a numerical algorithm (the method for finding the solution prescribed by the criterion). The most successful early statistical algorithm, ML-EM, is based on combining a Poisson statistical model with the EM algorithm. The positive traits of ML-EM spawned a number of variations, mostly seeking to improve on its slow convergence rate. Gaussian statistical models, related to WLS criteria, result in quadratic objective functions. These can efficiently use many established numerical algorithms, such as CG or coordinate descent, and result in faster reconstruction algorithms. Noise remains a problem, however.

IFBP algorithms, such as Chang's method, were developed intuitively, but are related to WLS algorithms. Because of the nontheoretical origins, some of these algorithms can be unstable. MAP criteria can be used to combat the noise sensitivity in ML solutions. MAP algorithms impose smoothness constraints, usually through establishing relationships between neighboring pixels and can therefore converge to usable solutions. The properties of smoothing are highly dependent on the choice of the smoothing distribution, and can lead to undesirable results if not carefully controlled. Ordered-subset approaches, such as OS-EM, offer all the positive features of ML-EM in far fewer iterations and have therefore come to dominate the field in recent years.

VIII. APPENDICES

A. Modeling the Projection of a Pixel

An often-overlooked topic in image reconstruction is the practical issue of computing forward projections of the current image estimate and backprojections of the predicted projections, which are integral steps in all the iterative algorithms we have discussed. Several different approaches for projecting a pixel were provided in the Donner algorithms library (Huesman *et al.*, 1977). These methods, which represent different trade-offs between efficiency, are illustrated in Figure 19. Among these are methods that model the activity in a pixel as being distributed in different ways, often referred to as *pixel-driven methods*. The simplest, most efficient, and least accurate is the point-projection model (Fig. 19a), wherein the intensity in a pixel is assumed to be contained in a point source at the center of the pixel. In projecting this model, it is only necessary to compute the bin location of the projection of the point and then to deposit all the intensity of that pixel in a single bin.

Slightly more complex is the convex-disk model in Figure 19b. In this case, the intensity in a pixel is assumed

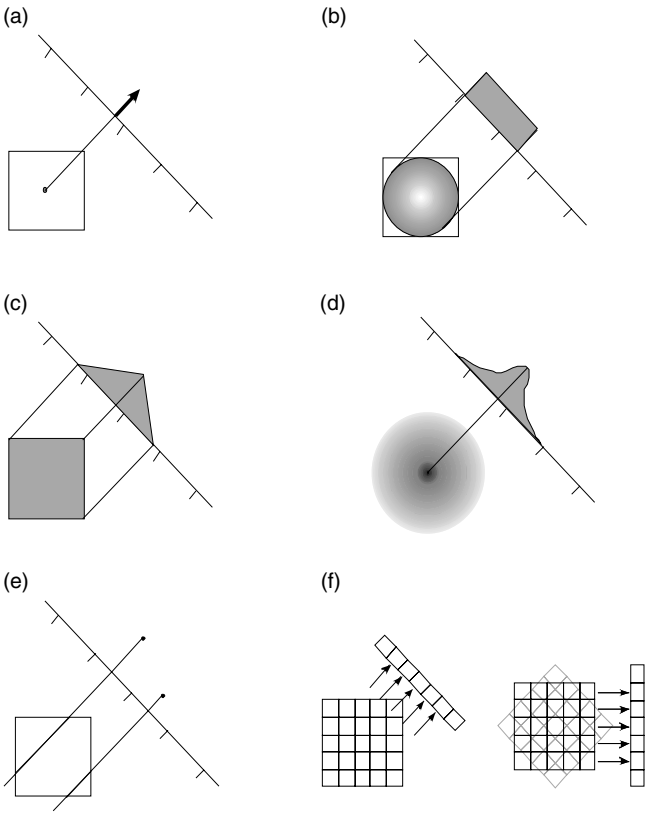

FIGURE 19 Approaches to modeling the projection of a pixel into a set of discrete bins. (a) Point-projection model. (b) Convex-disk model. (c) Area-weighted model. (d) Gaussian blobs. (e) Line-length approach. (f) Rotation-based projection model.

to be distributed in a disk fitting within the pixel. The disk intensity is not uniform but rather is lower in the center so that, when the disk is projected with a parallel-beam model, its projection becomes a rectangle, independent of the angle of projection. Because the projection is uniform, it is a simple matter to determine where the boundaries of the projection lie relative to the bin boundaries and to weight the pixel contributions to each bin accordingly. Note that whereas the point-projection method can model virtually any geometry, the convex-disk method is limited to parallel projection.

The most accurate example of pixel-driven methods is the area-weighted model (Fig. 19c). Here the pixel intensity is distributed uniformly over the space the pixel occupies. The projection of the pixel is angle-dependent, so the computation is more complex than in the previous cases. The contribution of a particular pixel to a particular bin is computed as the integral of the pixel projection that lies within the boundaries of the bin. In parallel projection, this is equal to the area of the intersection of a projection strip cast out from the projection bin with the space occupied by the pixel.

An alternative approach to modeling pixels is the use of Gaussian blobs (Fig. 19d). Because the parallel projection of a 2D Gaussian shape is a 1D Gaussian, it is a simple matter to compute the projection location of the center of the blob and then to distribute the projected Gaussian among the regions of the individual bins. The use of blobs also leads to a natural smoothing effect in the reconstructed image.

The pixel-driven methods are contrasted with ray-driven methods, which compute projections from the point of view of the projection bins. The line-length approach (Fig. 19e) was popular in early implementations because it computes projections with little storage requirement and also leads to a natural computation of attenuation factors. Here, a ray is cast out from the center of a given projection bin in the direction of projection. Every pixel that is intersected by the ray is assigned a weight proportional to the length of the ray in the region bounded by the pixel. To improve accuracy, it is possible to cast multiple rays out from each projection bin, but this makes the method much less computationally efficient.

Finally, an efficient approximation to the area-weighted model is the rotation-based projection model (Fig. 19f). In this case, we do not compute the projection of a given pixel onto the rotating bin array. Rather, the bin array remains fixed and the image is rotated by interpolation. Once in the rotated image frame of reference, the projection is always taken in the same direction. Backprojection can be accomplished by reversing the process. Several rotation-based approaches have been proposed (DiBella *et al.*, 1996; Wallis and Miller, 1997), the chief differences among them involving how the interpolation required for image rotation is accomplished.

For 3D imaging geometries, some of the methods described here have natural extensions. The area-weighted methods extend to volume-weighted methods. In converging-beam imaging, however, the computation of volume intersections is rather complex. It is possible to take advantage of rotation-based methods so that the volume intersections need only be computed in the rotated image frame of reference. A further extension results in a warping of the image space to convert converging projections into parallel projections (Zeng *et al.*, 1994).

Another important consideration in the modeling of projection is whether to explicitly compute and store the matrix $\mathbf{H}$ or to use a projection operator. In the first case, called a matrix-based model, the size of the matrix can result in enormous storage requirements if anything beyond a simple 1D projection of a 2D object is modeled. Also, the computation of the elements of the matrix can be time consuming except for simple cases. Furthermore, if a system has variable geometry, such as a SPECT system in which the radius of rotation varies from study to study, it becomes necessary to compute the matrix anew for each study per-

formed. An alternative is the more common projector-based approach. In this case, a subroutine is written to take an image and compute its projections without explicitly computing the individual elements of the matrix **H**. Thus, the output of the subroutine is **Hf**, even though the elements of **H** are never found. The projector-based approach is more flexible and has a smaller requirement for storage than the matrix-based approach, but it may require more computation time.

B. Projections onto Convex Sets

Let us suppose that the image **f** is unknown, but that we have knowledge about **f** in the form of constraint sets, C_i, $i = 1, \ldots, M$. In image reconstruction, most important constraints are those imposed by the data, but other prior information may also be represented as constraint sets. In the method of POCS, all the constraint sets are assumed to be convex, meaning that the line segment connecting any two points in the set also lies completely within the set.

In this formulation, the problem is to find an image vector **f** that satisfies all the constraints. In other words, the problem is to find a vector **f** that lies in the intersection of all the constraints, $\mathbf{f} \in \cap_{i=1}^{M} C_i$. Provided that the intersection is not empty and that all the sets are convex, the sequential POCS algorithm is guaranteed to converge to a point in the intersection. This is the basis for the standard ART algorithm.

The basic operation in the POCS algorithm is the projection (here, the term is used in the linear algebra sense, not in the tomography sense). Recall that the projection of a point **x** onto a set S is the point in S that is closest to **x**.

The sequential POCS algorithm makes use of projections in the following way. Starting from any initial estimate for the image vector **f**, the vector is projected onto each and every constraint set in sequence. The result of this sequence of steps becomes the new estimate, and the process is repeated until convergence. Each iteration can be expressed as $\hat{\mathbf{f}}^{(n+1)} = P_1 P_2 \ldots P_M \hat{\mathbf{f}}^{(n)}$, where P_i is the operation that projects any vector onto the set C_i.

A special case of the POCS algorithm is when all the constraint sets are hyperplanes, in which case the algorithm is known as the Kaczmarz method or ART.

C. Details of the Maximum-Likelihood Expectation-Maximization Algorithm

The ML-EM algorithm and its variants have dominated the field of iterative ET image reconstruction for many years; therefore, it is worth looking more closely at this widely used technique.

The EM algorithm, as proposed in Dempster *et al.* (1977), is a general prescription for developing iterative procedures for solving all sorts of ML estimation problems. The specific iterative formula in Lange and Carson (1984), the ML-EM algorithm for image reconstruction, is simply one application of this general problem-solving approach.

Let us begin by outlining the principles of the general EM methodology. There are many problems in statistics in which the ML solution is difficult to find, but would be easy to find if we had access to some additional data, which are sometimes called *missing data*. In some situations, these extra data are measurements that are literally missing (e.g., the experimenter failed to record them during data collection). In other situations, the notion of missing data is simply an abstraction that leads to a convenient solution to the problem. In these cases, the missing data are measurements that would be extremely helpful to have, but are not actually measurable. (ET reconstruction is an example of this latter use of the EM algorithm.) When applying the EM approach, the observed data that we actually have are called the *incomplete data*, denoted by **g**, and the full complement of data we wish we had are called the *complete data*, denoted by **s**.

In ET, the reason our photon count data are incomplete is that we do not know exactly where the photons collected in each projection bin came from. The process of tomographic projection mixes together the photons emitted from the object **f** into the projection bins g_j in a linear way described by the system matrix **H**. The task of image reconstruction is, in a sense, to unmix the photon counts by inverting the effect of **H**. Viewed in this way, our data would be complete if we knew not only how many counts were measured in projection bin j, but also how many of these came from each object pixel k. Thus, we can define an element of the complete data as a quantity s_{jk}, which denotes the number of photons emitted from pixel k that were detected in projection bin j. Of course, if we actually had access to data of this kind, the reconstruction problem would be trivial. The reason for postulating these unrealistic data is to allow us to use the EM mechanism for obtaining the ML solution.

1. Statement of the EM Algorithm

The EM algorithm consists of two alternating steps, which are repeated until convergence. These steps are called the *expectation step* (E-step) and the *maximization step* (M-step) and are defined as follows.

E-step Based on the current estimate $\hat{\mathbf{f}}^{(n)}$, compute:
$Q(\mathbf{f};\hat{\mathbf{f}}^{(n)}) = E[\ln p(\mathbf{s};\mathbf{f}) \mid \mathbf{g};\hat{\mathbf{f}}^{(n)}]$

M-step Choose the next estimate $\hat{\mathbf{f}}^{(n+1)}$ to maximize $Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})$:

$$\hat{\mathbf{f}}^{(n+1)} = \arg\max_{\mathbf{f}} Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})$$

The essence of the algorithm is as follows. If we had access to the complete data, the ML problem would be easy

to solve; however, in the absence of the complete data, the log-likelihood function for the complete data cannot even be computed. Thus, in each iteration, the average of this log-likelihood function is computed in the E-step and maximized in the M-step. The following derivation shows that this procedure increases the likelihood $p(\mathbf{g};\mathbf{f})$ in each iteration (so long as that is possible).

2. EM Algorithm Produces Nondecreasing Likelihood

Recall that the aim of ML reconstruction is to find the image $\mathbf{f}$ that maximizes the likelihood function $p(\mathbf{g};\mathbf{f})$ or, equivalently, the logarithm of the likelihood (or log-likelihood) $l(\mathbf{f}) = \ln p(\mathbf{g};\mathbf{f})$.[3] In the context of the EM algorithm, we call this the incomplete-data log-likelihood.

To begin, let us relate the log-likelihood of the incomplete (observed) data $\mathbf{g}$ to that of the idealized complete data $\mathbf{s}$. From basic probability theory:

$$p(\mathbf{s};\mathbf{f}) = p(\mathbf{s}\,|\,\mathbf{g};\mathbf{f})\,p(\mathbf{g};\mathbf{f}) \tag{42}$$

Taking the logarithm of both sides of Eq. (19) and rearranging, one obtains:

$$\begin{aligned} l(\mathbf{f}) &= \ln p(\mathbf{g};\mathbf{f}) \\ &= \ln p(\mathbf{s};\mathbf{f}) - \ln p(\mathbf{s}\,|\,\mathbf{g};\mathbf{f}) \end{aligned} \tag{43}$$

Taking the expected value of both sides of Eq. (43) with respect to $p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)})$ yields:

$$\begin{aligned} l(\mathbf{f}) &= E[\ln p(\mathbf{s};\mathbf{f})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}] - E[\ln p(\mathbf{s}\,|\,\mathbf{g};\mathbf{f})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}] \\ &= Q(\mathbf{f};\hat{\mathbf{f}}^{(n)}) - R(\mathbf{f};\hat{\mathbf{f}}^{(n)}) \end{aligned} \tag{44}$$

where $Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})$ and $R(\mathbf{f};\hat{\mathbf{f}}^{(n)})$ are the two expectations in the first line of Eq. (44).

Now let us compare the value of the incomplete-data log-likelihood $l(\mathbf{f})$ for two successive iterations, n and $n + 1$:

$$\begin{aligned} l(\hat{\mathbf{f}}^{(n+1)}) - l(\hat{\mathbf{f}}^{(n)}) = &\left[Q(\hat{\mathbf{f}}^{(n+1)};\hat{\mathbf{f}}^{(n)}) - Q(\hat{\mathbf{f}}^{(n)};\hat{\mathbf{f}}^{(n)})\right] \\ &+ \left[R(\hat{\mathbf{f}}^{(n)};\hat{\mathbf{f}}^{(n)}) - R(\hat{\mathbf{f}}^{(n+1)};\hat{\mathbf{f}}^{(n)})\right] \end{aligned} \tag{45}$$

We want to show that this difference is nonnegative, thus demonstrating that each iteration increases the likelihood when possible and, at worst, leaves it unchanged. The first term in square brackets in Eq. (45) is clearly nonnegative, because the M-step maximizes $Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})$ and, thus, $Q(\hat{\mathbf{f}}^{(n+1)};\hat{\mathbf{f}}^{(n)}) \geq Q(\hat{\mathbf{f}}^{(n)};\hat{\mathbf{f}}^{(n)})$. The second term is also nonnegative, which can be seen from the following.

The following identity holds for any real number t: $\ln t \leq t - 1$. Thus, we can write:

$$E[\ln \pi(\mathbf{s})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}] \leq E[\pi(\mathbf{s})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}] - 1 \tag{46}$$

where

$$\pi(\mathbf{s}) = \frac{p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n+1)})}{p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)})} \tag{47}$$

Evaluating the right-hand side of Eq. (46), we obtain:

$$\begin{aligned} E[\pi(\mathbf{s})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}] - 1 &= E\left[\frac{p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n+1)})}{p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)})}\Big|\mathbf{g};\hat{\mathbf{f}}^{(n)}\right] - 1 \\ &= \int \frac{p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n+1)})}{p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)})}\,p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)})\,d\mathbf{s} - 1 \\ &= 0 \end{aligned} \tag{48}$$

From Eqs. (46)–(48), we obtain:

$$\begin{aligned} E\left[\ln \pi(\mathbf{s})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}\right] &\leq 0 \\ E\left[\ln p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n+1)})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}\right] - E\left[\ln p(\mathbf{s}\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)})\,|\,\mathbf{g};\hat{\mathbf{f}}^{(n)}\right] &\leq 0 \\ R(\hat{\mathbf{f}}^{(n+1)};\hat{\mathbf{f}}^{(n)}) - R(\hat{\mathbf{f}}^{(n)};\hat{\mathbf{f}}^{(n)}) &\leq 0 \end{aligned} \tag{49}$$

Therefore, the second difference term in Eq. (45) is nonnegative, and we have established that:

$$l(\hat{\mathbf{f}}^{(n+1)}) \geq l(\hat{\mathbf{f}}^{(n)}) \tag{50}$$

In a variation of the EM algorithm, called the GEM *algorithm*, the M-step is replaced with a weaker step in which the expectation is not maximized. Instead, the GEM algorithm aims only to achieve $Q(\hat{\mathbf{f}}^{(n+1)};\hat{\mathbf{f}}^{(n)}) \geq Q(\hat{\mathbf{f}}^{(n)};\hat{\mathbf{f}}^{(n)})$.

3. EM Algorithm for ET Image Reconstruction

Now let us derive Eq. (30), which is the iterative formula for ML-EM image reconstruction in ET.[4] Recall that, in ET, a useful definition of the complete data is s_{im}, which is the (random) number of photons emitted from within pixel m and detected in projection bin i. Of course, this is an unmeasurable quantity, but it allows us conveniently to use the EM framework to solve the ML reconstruction problem. The complete data can be related to the observed projection data $\mathbf{g}$ and the image $\mathbf{f}$ as follows:

$$g_i = \sum_m s_{im} \tag{51}$$

$$E[s_{im}] = h_{im} f_k \tag{52}$$

The E-step of the EM algorithm requires an expression for the complete-data log-likelihood $\ln p(\mathbf{s};\mathbf{f})$. In ET, the counts s_{im} are independent Poisson-distributed random variables; therefore,

$$p(\mathbf{s};\mathbf{f}) = \prod_i \prod_m \frac{E[s_{im}]^{s_{im}} e^{-E[s_{im}]}}{s_{im}!} \tag{53}$$

and the log-likelihood is (using Eq. (52)):

$$\ln\, p(\mathbf{s};\mathbf{f}) = \sum_i \sum_m [s_{im} \ln (h_{im} f_m) - h_{im} f_m - \ln(s_{im}!)] \tag{54}$$

[3] This derivation is based partly on the derivation in McLachland and Krishnan (1997).

[4] This derivation is based on the one given in Lange and Carson (1984).

Now we are ready to compute the E-step of the EM algorithm.

E-step

Using Eq. (54), the E-step is computed as follows:

$$\begin{aligned} Q(\mathbf{f};\hat{\mathbf{f}}^{(n)}) &= E[\ln p(\mathbf{s};\mathbf{f}) \mid \mathbf{g};\hat{\mathbf{f}}^{(n)}] \\ &= \sum_i \sum_m \left\{ E[s_{im} \mid \mathbf{g};\hat{\mathbf{f}}^{(n)}] \ln (h_{im} f_m) - h_{im} f_m - E[\ln (s_{im}!)] \right\} \end{aligned} \quad (55)$$

The conditional mean of s_{im} in Eq. (55) is given by:

$$E[s_{im} \mid \mathbf{g};\hat{\mathbf{f}}^{(n)}] = \frac{h_{im}\hat{f}_m^{(n)}}{\sum_k h_{ik}\hat{f}_k^{(n)}} g_i \triangleq p_{im} \quad (56)$$

which is simply the fraction of the detected counts in projection bin i that are expected to have emanated from pixel m, given that current image estimate $\hat{\mathbf{f}}^{(n)}$ is the source of these counts. Substituting Eq. (56) into Eq. (55), we obtain the final form of the E-step:

$$Q(\mathbf{f};\hat{\mathbf{f}}^{(n)}) = \sum_i \sum_m \left\{ p_{im} \ln(h_{im} f_m) - h_{im} f_m - E[\ln(s_{im}!)] \right\} \quad (57)$$

Next, we compute the M-step.

M-step

In the M-step, we find the next image estimate $\hat{\mathbf{f}}^{(n+1)}$ by maximizing $Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})$ with respect to $\mathbf{f}$. We can accomplish by setting the derivative of $Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})$ to zero and solving:

$$\frac{\partial Q(\mathbf{f};\hat{\mathbf{f}}^{(n)})}{\partial f_j} = 0 = \sum_i \left(\frac{p_{ij}}{\hat{f}_j^{(n+1)}} - h_{ij} \right) \quad (58)$$

Solving for $\hat{f}_j^{(n+1)}$ and using Eq. (56), we obtain the well-known ML-EM iteration for ET image reconstruction in Eq. (30):

$$\hat{f}_j^{(n+1)} = \frac{\hat{f}_j^{(n)}}{\sum_{i'} h_{i'j}} \sum_i h_{ij} \frac{g_i}{\sum_k h_{ik}\hat{f}_k^{(n)}}$$

Acknowledgments

Preparation of this chapter was supported in part by NIH/NHLBI grant HL65425.

References

Barrett, H. H. (1990). Objective assessment of image quality: effects of quantum noise and object variability. *J. Opt. Soc. Am. A* **7:** 1266–1278.

Barrett, H. H., Denny, J. L., Wagner, R. F., and Myers, K. J. (1995). Objective assessment of image quality, II. Fisher information, Fourier crosstalk, and figures of merit for task performance. *J. Opt. Soc. Am. A* **12:** 834–852.

Barrett, H. H., Wilson, D. W., and Tsui, B. M. W. (1994). Noise properties of the EM algorithm, I. Theory. *Phys. Med. Biol.* **39:** 833–846.

Barrett, H. H., Yao, J., Rolland, J. P., and Myers, K. J. (1993). Model observers for assessment of image quality. *Proe. Natl Acad. Sci. U.S.A.* **90:** 9758–9765.

Beekman, F., Kamphuis, C., and Viergever, M. (1996). Improved SPECT quantitation using fully three-dimensional iterative spatially variant scatter response compensation. *IEEE Trans. Med. Imaging* **15:** 491–499.

Brankov, J., Yang, Y., and Wernick, M. N. (2004). Tomographic image reconstruction based on a content-adaptive mesh model. *IEEE Trans. Med. Imaging*. **23:** 202–212.

Brankov, J. G., Yang, Y., and Wernick, M. N. (to appear). Tomographic image reconstruction based on a content-adaptive mesh model. *IEEE Trans. Med. Imaging*.

Brankov, J. G., Yang, Y., Wernick, M. N., and Narayanan, M. V. (2002). Motion compensated reconstruction of gated SPECT data. *2002 Conf. Rec. IEEE Nucl. Sci. Symp. Med. Imaging Conf.* **3:** 10–16.

Browne, J., and De Pierro, A. R. (1996). A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography. *IEEE Trans. Med. Imaging* **15:** 687–699.

Burgess, A. E., Li, X., and Abbey, C. K. (1997). Visual signal detectability with two noise components: Anomalous masking effects. *J. Opt. Soc. Am. A* **14:** 2420–2442.

Byrne, C. L. (1996). Block-iterative methods for image reconstruction from projections. *IEEE Trans. Imaging Processing* **5:** 792–794.

Censor, Y., Eggermont, P. P. B., and Gordon, D. (1983). Strong underrelaxation in Kaczmarz's method for inconsistent systems. *Numer. Math.* **41:** 83–92.

Chang, L. (1979). Attenuation and incomplete projection in SPECT. *IEEE Trans. Nucl. Sci.* **26:** 2780–2789.

Chiao, P. C., Rogers, W. L., Fessler, J. A., Clinthorne, N. H., and Hero, A. O. (1994). Model-based estimation with boundary side information or boundary regularization. *IEEE Trans. Med. Imaging* **13:** 227–234.

Coxson, P. G., Salmeron, E. M., and Huesman, R. H. (1990). A strategy for using closed form solutions for compartmental models of dynamic PET data. *1990 Conf. Rec. IEEE. Nucl. Sci. Symp.* 1577–1578.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B* **39:** 1–38.

DiBella, E., Barclay, A., Eisner, R., and Schaefer, R. (1996). A comparison of rotation-based methods for iterative reconstruction algorithms. *IEEE Trans. Nucl. Sci.* **43:** 3370–3376.

Farncombe, T., Celler, A., Noll, J., Maeght, D., and Harrop, R. (1999). Dynamic SPECT imaging using a single camera rotation. *IEEE Trans. Nucl. Sci.*

Farncombe, T., King, M. A., Celler, A., Blinder, S. (2001). A fully 4D expectation maximization algorithm using Gaussian diffusion based detector response for slow camera rotation dynamic SPECT. Paper presented at 2001 International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine, Pacific Grove, CA.

Feng, B., Pretorius, P. H., Farncombe, T. H., Dahlberg, S. T., Narayanan, M. V., Wernick, M. N., Celler, A. M., King, M.A., and Leppo, J.A. (2002). Imaging time-varying Tc-99m teboroxime localization and cardiac runction simultaneously by five-dimensional (5D) gated-dynamic SPECT imaging and reconstruction. *Am. Soc. Nucl. Card.* **10:** S11–12.

Fessler, J. A. (1994). Penalized weighted least squares image reconstruction for positron emission tomography. *IEEE Trans. Med. Imaging* **13:** 290–300.

Fessler, J. (1996). Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Trans. Image Processing* **5:** 493–506.

Fessler, J., and Hero, A. (1994). Space-alternating generalized expectation maximization algorithm. *IEEE Trans. Sig. Processing* **42:** 2664–2677.

Fessler, J. A., and Rogers, W. L. (1996). Spatial resolution properties of penalized-likelihood image reconstruction: Space-invariant tomographs. *IEEE Trans. Imaging Processing* **5:** 1346–1358.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1:** 3–32.

Frey, E. C., Ju, Z. W., and Tsui, B. M. W. (1993). A fast projector-backprojector pair for modeling the asymmetric spatially-varying scatter response function for scatter compensation in SPECT imaging. *IEEE Trans. Nucl. Sci.* **40:** 1192–1197.

Galatsanos, N. P. Wernick, M. N., and Katsaggelos, A. K. (2000). Multichannel image recovery. *In* "Handbook of Image and Video Processing" (A. Bovik, ed.), pp. 155–168. Academic Press; San Diego, CA.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt Anal. Mach. Int.* **6:** 721–741.

Gilbert, P. (1972). Iterative methods for the reconstruction of three dimensional objects from their projections. *J. Theor. Biol.* **36:** 105–117.

Gindi, G., Lee, M., Rangarajan, A., and Zubal, I. G. (1991). Bayesian reconstruction of functional images using registered anatomical images as priors. *In* "Information Processing in Medical Imaging" (A. C. F. Colchester and D. J. Hawkes, eds.), pp. 121–131. Springer-Verlag, New York.

Golub, G. H., and Van Loan, C. F. (1989). "Matrix Computations." Johns Hopkins University Press, Baltimore.

Gordon, R., Bender, R., and Herman, G. T. (1970). Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *J. Theor. Biol.* **29:** 471–481.

Green, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imaging* **9:** 84–93.

Hebber, E., Oldenburg, D., Farncombe, T., and Celler, A. (1997). Direct estimation of dynamic parameters in SPECT tomography. *IEEE Trans. Nucl. Sci.* **44:** 2425–2430.

Hebert, T. J., and Gopal, S. S. (1992). The GEM MAP algorithm with 3D SPECT system response. *IEEE Trans. Med. Imaging* **11:** 81–90.

Hebert, T. J., and Leahy, R. (1992). Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. *IEEE Trans. Sig. Processing* **40:** 2290–2303.

Herman, G. T. (1980). "Image Reconstruction from Projections: The Fundamentals of Computerized Tomography." Academic Press, New York.

Higdon, D. M., Bowsher, J. E., Johnson, V. E., Turkington, T. G., Gilland, D. R., and Jaczszak, R. J. (1997). Fully Bayesian estimation of Gibbs hyperparameters for emission computed tomography data. *IEEE Trans. Med. Imaging* **16:** 516–526.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stati.* **2:** 360–378.

Hudson, H. M., and Larkin, R. S. (1994). Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans. Med. Imaging* **13:** 601–609.

Huesman, R. H., Gullberg, G. T., Greenberg, W. L., and Budinger, T. F. (1977). "User Manual, Donner Algorithms for Reconstruction Tomography." University of California, Lawrence Berkeley Laboratory.

Hutton, D. A., and Strickland, R. N. (1997). Channelized detection filters for detecting tumors in nuclear medical images. *Proc. SPIE* **3034:** 457–466.

Jinyi, Q. (2003). A unified noise analysis for iterative image estimation. *Phys. Med. Biol*. **48:** 3505–3519.

Johnson, V. E., Wong, W. H., Hu, X., and Chen, C. T. (1991). Image restoration using Gibbs priors: Boundary modeling, treatment of blurring, and selection of hyperparameters. *IEEE Trans. Patt. Anal. Mach. Int.* **13:** 413–425.

Kaczmarz, S. (1937). Angenaherte Auflosung von Systemen linearer Gleichungen. *Bull. Acad. Polon. Sci.* **A35:** 355–357.

Kalbfleisch, J. G. (1985). "Probability and Statistical Inference, Vol. 1. Probability." 2nd ed. Springer-Verlag, New York.

Kao, C.-M., Yap, J.-T., Mukherjee, J., and Wernick, M. N. (1997). Image reconstruction for dynamic PET based on low-order approximation and restoration of the sinogram. *IEEE Trans. Med. Imaging* **16:** 738–749.

Kaufman, L. (1987). Implementing and accelerating the EM algorithm for positron emission tomography. *IEEE Trans. Med. Imaging* **6:** 37–51.

Kaufman, L. (1993). Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Trans. Med. Imaging* **12:** 200–214.

Kay, S. M. (1993). "Fundamentals of Statistical Signal Processing: Estimation Theory." Prentice Hall, New York.

King, M. A., and Miller, T. R. (1985). Use of a nonstationary temporal Wiener filter in nuclear medicine. *Eur. J. Nuce. Med.* **10:** 458–461.

Klein, G. J., Reutter, B. W., and Huesman, R. H. (1997). Non-rigid summing of gated PET via optical flow. *IEEE Trans. Nucl. Sci.* **44:** 1509–1512.

Lalush, D. S., and Tsui, B. M. W. (1992). Simulation evaluation of Gibbs prior distributions for use in maximum *a posteriori* SPECT reconstructions. *IEEE Trans. Med. Imaging* **11:** 267–275.

Lalush, D. S., and Tsui, B. M. W. (1993). A generalized Gibbs prior for maximum a posteriori reconstruction in SPECT. *Phys. Med. Biol.* **38:** 729–741.

Lalush, D. S., and Tsui, B. M. W. (1994). Improving the convergence of iterative filtered backprojection algorithms. *Med. Phys.* **21:** 1283–1286.

Lalush, D. S., and Tsui, B. M. W. (1995). A fast and stable maximum a posteriori conjugate gradient reconstruction algorithm. *Med. Phys.* **22:** 1273–1284.

Lalush, D. S., and Tsui, B. M. W. (1996). Space-time Gibbs priors applied to gated SPECT myocardial perfusion studies. *In* "Three-dimensional Image Reconstruction in Radiology and Nuclear Medicine" (P. Grangeat and J. L. Amans eds.), pp. 209–224. Kluwer Academic, Dordrecht.

Lalush, D. S., and Tsui, B. M. W. (1998a). Block-iterative techniques for fast 4D reconstruction using *a priori* motion models in gated cardiac SPECT. *Phys. Med. Biol.* **43:** 875–887.

Lalush, D. S., and Tsui, B. M. W. (1998b). Mean-variance analysis of block-iterative reconstruction algorithms modeling 3D detector response in SPECT. *IEEE Trans. Nucl. Sci.* **45:** 1280–1287.

Lalush, D. S., and Tsui, B. M. W. (2000). Performance of ordered-subset reconstruction algorithms under conditions of extreme attenuation and truncation in myocardial SPECT. *J. Nucl. Med.* **41:** 737–744.

Lange, K. (1990). Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Trans. Med. Imaging* **9:** 439–446.

Lange, K., Bahn, M., and Little, R. (1987). A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. *IEEE Trans. Med. Imaging* **6:** 106–114.

Lange, K., and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assist. Tomogr.* **8:** 306–316.

Leahy, R., and Yan, X. (1991). Incorporation of anatomical MR data for improved functional imaging with PET. *In* "Information Processing in Medical Imaging" (A. C. F. Colchester and D. J. Hawkes, eds.), pp. 105–120. Springer-Verlag, New York.

Lee, S. J., Rangarajan, A., and Gindi, G. (1995). Bayesian image reconstruction in SPECT using higher order mechanical models as priors. *IEEE Trans. Med. Imaging* **14:** 669–680.

Levitan, E., and Herman, G. T. (1987). A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography. *IEEE Trans. Med. Imaging* **6:** 185–192.

Lewitt, R. (1992). Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys. Med. Biol.* **37:** 705–716.

Lewitt, R. M., and Muehllehner, G. (1986). Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation. *IEEE Trans. Med. Imaging* **5:** 16–22.

Liang, Z. (1993). Compensation for attenuation, scatter, and detector response in SPECT reconstruction via iterative FBP methods. *Med. Phys.* **20:** 1097–1106.

Liang, Z., Jaszczak, R., and Greer, K. (1989). On Bayesian image reconstruction from projections: uniform and nonuniform *a priori* source information. *IEEE Trans. Med. Imaging* **8:** 227–235.

Limber, M. A., Limber, M. N., Celler, A., Barney, J. S., and Borwein, J. M. (1995). Direct reconstruction of functional parameters for dynamic SPECT. *IEEE Trans. Nucl. Sci.* **42:** 1249–1256.

Liow, J., and Strother, S. C. (1993). The convergence of object dependent resolution in maximum likelihood based tomographic image reconstruction. *Phys. Med. Biol.* **38:** 55–70.

Llacer, J., and Veklerov, E. (1989). Feasible images and practical stopping rules for iterative algorithms in transmission tomography. *IEEE Trans. Med. Imaging* **8:** 186–193.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distribution. *Astrophys. J.* **79:** 745–754.

Mailloux, G., Noumeir, R., and Lemieux, R. (1993). Deriving the multiplicative algebraic reconstruction algorithm (MART) by the method of convex projections (POCS). *Proc. IEEE Int. Conf. Acoust., Speech Sig. Processing*, **5:** 457.

Mair, B. A., Gilland, D. R., and Cao, Z. (2002). Simultaneous motion estimation and image reconstruction from gated data. *Proc. IEEE Int. Symp. Biomed. Imaging* 661–664.

Malko, J. A., Van Heertum, R. L., Gullberg, G. T., and Kowalsky, W. P. (1986). SPECT liver imaging using an iterative attenuation correction algorithm and an external flood source. *J. Nucl. Med.* **27:** 701–705.

Maze, A., Le Cloirec, J., Collorec, R., Bizais, Y., Briandet, P., and Bourguet, P. (1993). Iterative reconstruction methods for nonuniform attenuation distribution in SPECT. *J. Nucl. Med.* **34:** 1204–1209.

McLachlan, G. J., and Krishnan, T. (1997). "The EM Algorithm and Extensions." John Wiley, New York.

Metz, C. E. (1978). Basic principles of ROC analysis. *Sem. Nucl. Medi.* **8:** 283–298.

Metz, C. E. (1986). ROC methodology in radiologic imaging. *Invest. Radiol.* **21:** 720–733.

Metz C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol.* **24:** 234–245.

Moore, S. C., Brunelle, J. A., and Kirsch, C. M. (1987). Quantitative multi-detector emission computerized tomography using iterative attenuation compensation. *J. Nucl. Med.* **23:** 706–714.

Mumcuoglu, E. U., Leahy, R., Cherry, S. R., and Zhou, Z. (1994). Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images. *IEEE Trans. Med. Imaging* **13:** 687–701.

Myers, K. J., and Barrett, H. H. (1987). Addition of a channel mechanism to the ideal-observer model. *J. Opt. Am. A* **4:** 2447–2457.

Myers, K. J., Rolland, J. P., and Barrett, H. H. (1990). Aperture optimization for emission imaging: Effect of a spatially varying background. *J. Opt. Soc. Am. A* **7:** 1279–1293.

Narayanan, M. V., Gifford, H. C., King, M. A., Pretorius, P. H., Farncombe, T. H., Bruyant, P., and Wernick, M. N. (2002). Optimization of iterative reconstructions of Tc-99m cardiac SPECT studies using numerical observers. *IEEE Trans. Nucl. Sci.* **49:** 2355–2360.

Narayanan, V. M., King, M. A., Soares, E., Byrne, C., Pretorius, H., and Wernick, M. N. (1999). Application of the Karhunen-Loeve transform to 4D reconstruction of gated cardiac SPECT images. *IEEE Trans. Nucl. Sci.* **46:** 1001–1008.

Nichols, T. E., Qi, J., and Leahy. R. M. (1999). Continuous time dynamic PET imaging using list mode data. *In* "Information Processing in Medical Imaging" (A. C. F. Colchester and D. J. Hawkes, eds.), pp. 98–111. Springer-Verlag, New York.

Ouyang, X., Wong, W. H., Johnson, V. E., Hu, X. P., and Chen, C. T. (1994). Incorporation of Correlated Structural Images in PET Image Reconstruction. *IEEE Trans. Med. Imaging* **13:** 627–640.

Pan, T.-S., and Yagle, A. E. (1991). Numerical study of multigrid implementations of some iterative image reconstruction algorithms. *IEEE Trans. Med. Imaging* **10:** 572–588.

Pan, X., Wong, W. H., Chen, C.-T., and Jun, L. (1993). Correction for photon attenuation in SPECT: Analytical framework, average attenuation factors, and a new hybrid approach. *Phys. Med. Biol.* **38:** 1219–1234.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). "Numerical Recipes in C." Cambridge University Press, Cambridge, UK.

Qi, J. and Huesman, R. H. (2001). Theoretical study of lesion detectability of MAP reconstruction using computer observers. *IEEE Trans. Med. Imaging* **20:** 815–22.

Qi, J., Leahy, R. M., Hsu, C., Farquhar, T. H., and Cherry, S. R. (1996). Fully 3D Bayesian image reconstruction for the ECAT EXACT HR+. *IEEE Trans. Nucl. Sci.* **45:** 1096–1103.

Qi, J. and Leahy, R. M. (1999). Fast computation of the covariance of MAP reconstructions of PET images. *Proc. SPIE* **3661:** 344–355.

Rajeevan, N., Rajgopal, K., and Krishna, G. (1992). Vector-extrapolated fast maximum likelihood estimation algorithms for emission tomography. *IEEE Trans. Med. Imaging* **11:** 9–20.

Ranganath, M. V., Dhawan, A. P., and Mullani, N. (1988). A multigrid expectation maximization reconstruction algorithm for positron emission tomography. *IEEE Trans. Med. Imaging* **7:** 273–278.

Reutter, B. W., Gullberg, G. T., and Huesman, R. H. (1998). Kinetic parameter estimation from dynamic cardiac patient SPECT projection measurements. *1998 Conf. Rec. IEEE Nucl. Sci. Symp.* **3:** 1953–1958.

Reutter, B. W., Gullberg, G. T., Huesman, R. H. (2000). Direct least-squares estimation of spatiotemporal distributions from dynamic SPECT projections using a spatial segmentation and temporal B-splines. *IEEE Trans. Med. Imaging* **19:** 434–450.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am. A*, **62:** 55–59.

Rosenfeld, A., and Kak, A. C. (1982). "Digital Picture Processing." Academic Press, Orlando, FL.

Shepp, L. A., and Vardi, Y. (1982). Maximum likelihood estimation for emission tomography. *IEEE Trans. Med. Imaging* **1:** 113–121.

Smith, M. F., Floyd Jr., C. E., Jaszczak, R. J., and Coleman, R. E. (1992). Reconstruction of SPECT images using generalized matrix inverses. *IEEE Trans. Med. Imaging* **11:** 165–175.

Snyder, D. L., and Miller, M. I. (1985). The use of sieves to stabilize images produced with the EM algorithm for emission tomography. *IEEE Trans. Nucl. Sci.* **32:** 3864–3872.

Stamos, J. A., Rogers, W. L., Clinthorne, N. H., and Koral, K. F. (1988). Object-dependent performance comparison of two iterative reconstruction algorithms. *IEEE Trans. Nucl. Sci*. **35:** 611–614.

Stark, H., and Yang, Y. (1998). "Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics." John Wiley, New York.

Tanaka, E. (1987). A fast reconstruction algorithm for stationary positron emission tomography based on a modified EM algorithm. *IEEE Trans. Med. Imaging* **6:** 98–105.

Tekalp, M. A. (1995). "Digital Video Processing." Prentice-Hall, New York.

Tsui, B. M. W., Gullberg, G. T., Edgerton, E. R., Ballard, J. G., Perry, J. R., McCartney, W. H., and Berg, J. (1989). Correction of nonuniform attenuation in cardiac SPECT imaging. *J. Nucl. Med.* **30:** 497–507.

Tsui, B. M. W., Zhao, X.-D., Frey, E. C., and Gullberg, G. T. (1991). Comparison between EM and CG algorithms for SPECT image reconstruction. *IEEE Trans. Nucl. Sci.* **38:** 1766–1772.

Van Trees, H. L. (1968). "Detection, Estimation, and Modulation Theory, Part I." John Wiley & Sons, New York.

Wallis, J., and Miller, T. R. (1997). An optimal rotator for iterative reconstruction. *IEEE Trans. Med. Imaging* **16:** 118–123.

Wallis, J. W., and Miller, T. R. (1993). Rapidly converging iterative reconstruction algorithms in single-photon emission computed tomography. *J. Nucl. Med.* **34:** 1793–1800.

Wallis, J. W., Miller, T. R., and Dai, G. M. (1998). Comparison of the convergence properties of the It-W and OS-EM algorithms in SPECT. *IEEE Trans. Nucl. Sci.* **45:** 1317–1323.

Walters, E., Simon, W., Chesler, D. A., and Correia, J. A. (1981). Attenuation correction in gamma emission computed tomography. *J. Comput. Assist. Tomogr.* **5:** 89–94.

Wernick, M. N., Infusino, E. J., and Milosevic, M. (1999). Fast spatio-temporal image reconstruction for dynamic PET. *IEEE Trans. Med. Imaging* **18:** 185–195.

Wilson, D. W., Tsui, B. M. W., and Barrett, H. H. (1994). Noise properties of the EM algorithm, II. Monte Carlo simulations. *Phys. Med. Biol.* **39:** 847–871.

Wollenweber, S. D., Tsui, B. M. W., Lalush, D. S., Frey, E. C., LaCroix, K. J., and Gullberg, G. T. (1999). Comparison of Hotelling observer models and human observers in defect detection from myocardial SPECT imaging. *IEEE Trans. Nucl. Sci.* **46:** 2098–2103.

Xing, Y., and Gindi, G. (2002). Rapid calculation of detectability in Bayesian SPECT *Proc. IEEE Int. Symp. on Biomed. Imaging*. 78–81.

Xu, X.-L., Liow, J.-S., and Strother, S. C. (1993). Iterative algebraic reconstruction algorithms for emission computed tomography. *Med. Phys.* **20:** 1675–1684.

Yavus, M., and Fessler, J. A. (1996). Objective functions for tomographic reconstruction from randoms-precorrected PET scans. *1996 Conf. Rec. Nucl. Sci. Symp. Med. Imaging Conf.* **2:** 1067–1071.

Zeng, G. L., Gullberg, G. T., and Huesman, R. H. (1995). Using linear time-invariant system theory to estimate kinetic parameters directly from projection measurements. *IEEE Trans. Nucl. Sci.* **42:** 2339–2346.

Zeng, G. L., Gullberg, G. T., Tsui, B. M. W., and Terry, J. A. (1991). Three-dimensional iterative reconstruction algorithms with attenuation and geometric point response correction. *IEEE Trans. Nucl. Sci.* **38:** 693–702.

Zeng, G. L., Hsieh, Y., and Gullberg, G. (1994). A rotating and warping projector backprojector for fan-beam and cone-beam iterative algorithm. *IEEE Trans. Nucl. Sci.* **41:** 2807–2811.

Zhou, Z., Leahy, R. M., and Mumcuoglu, E. U. (1995). Maximum likelihood hyperparameter estimation for Gibbs priors with applications to PET. *In* "Information Processing in Medical Imaging" (Y. Bizais, C. Barillot, and R. DiPaola, eds.), pp. 39–52. Kluwer Academic, Dordrecht.

Zubal, I. G., Harrell, C. R., Smith, E. O., Rattner, Z., Gindi, G. R., and Hoffer, P. B. (1994). Computerized three-dimensional segmented human anatomy. *Med. Phys.* **21:** 299–302.