A Statistical Model for Positron Emission Tomography

Y. VARDI, L. A. SHEPP, and L. KAUFMAN*

Positron emission tomography (PET)—still in its research stages—is a technique that promises to open new medical frontiers by enabling physicians to study the metabolic activity of the body in a pictorial manner. Much as in X-ray transmission tomography and other modes of computerized tomography, the quality of the reconstructed image in PET is very sensitive to the mathematical algorithm to be used for reconstruction. In this article, we tailor a mathematical model to the physics of positron emissions, and we use the model to describe the basic image reconstruction problem of PET as a standard problem in statistical estimation from incomplete data. We describe various estimation procedures, such as the maximum likelihood (ML) method (using the EM algorithm), the method of moments, and the least squares method. A computer simulation of a PET experiment is then used to demonstrate the ML and the least squares reconstructions. The main purposes of this article are to report on what we believe is an important contribution of statistics to PET and to familiarize statisticians with this exciting field that can benefit from further statistical methodologies to be developed with PET problems in mind. Thus no background in physics or previous knowledge of computerized tomography is assumed. The emphasis is on the basic PET model and the statistical methodology needed for it.

KEY WORDS: Poisson point process; Estimation; Least squares; Maximum likelihood; Stein-type estimators; EM algorithm; Image reconstruction; Incomplete data; Nuclear medicine.

1. INTRODUCTION

1.1 Background

Positron emission tomography (PET)—still in its research stage—is a medical diagnostic technique that enables a physician to study blood flow in and metabolic activity of an organ in a visual way. To do so, a biochemical metabolite labeled with a *positron emitting* radioactive material is introduced into the organ under study and the radioactive emissions are then counted, using a PET scanner—a machine consisting of detector elements mounted on one or more rings, positioned so that it surrounds the patient's body. The choice of the biochemical and the radioactive tracer depends, of course, on the organ to be studied and the questions of interest. For example, since the brain uses glucose as a primary energy source, labeled glucose is often used in PET studies of the brain; metabolism of the heart, on the other hand, has been measured with labeled deoxyglucose and labeled palmitic acid (e.g., see Brownell et al. 1982). PET scans of brains of people suffering from schizophrenia and certain other diseases show distinctive metabolic patterns connected with these diseases. Thus it is hoped that these metabolic portraits obtained from PET scans will play an important role in diagnosing such diseases and in assessing the effectiveness of various treatments. An overview of the subject, including a discussion on clinical, biochemical, and other interesting aspects of PET can be found in Brownell et al. (1982) and in Ter-Pogossian et al. (1980). Hereafter, whenever we refer to a particular PET experiment, we assume that the organ of interest is the brain and that the tagged biochemical substance is glucose. This is done merely for convenience of presentation and is not intended to imply that this is the only use of PET or that the described statistical methods are limited to this application.

The Physics of PET. The positron emitting substance is deposited in the various regions of the brain in quantities proportional to the glucose uptake mechanism, and hence if it were possible to record the location (within the brain) of each positron emission, we could produce a portrait of the brain's glucose consumption. Though it is impossible to identify the exact location of a positron emission, it has been known (for some 30 years) that by positioning scintillation detectors around the patient's head, it is possible to determine a cylindrical volume in which the emission occurred. The physics behind this phenomenon is as follows: When a positron is emitted, it "finds" a nearby electron and annihilates with it. The annihilation creates two X-ray photons that fly off the point of annihilation, at the speed of light, in (nearly) opposite directions along a line with a completely random (i.e., uniformly distributed in space) orientation. There is an array of discrete detector elements surrounding the head, and the two photons are detected in coincidence by a pair of detector elements that define a cylindrical volume (to be referred to as a *detector tube* or, simply, a *tube*). Thus the only information acquired when a pair of detectors count a coincidence is that the annihilation occurred somewhere inside the tube defined by the two "firing" detectors; see Figure 1. The set of data collected, then, in a PET scan is the *tubes count* $[n^*(1), \ldots, n^*(D)]$, where $n^*(d)$ is the total number of coincidences counted by the dth detector tube and D is the total number of tubes. We note, however, that the total tubes count is typically much smaller than the total number of emissions, because all of those photons traveling along lines that do not cross the detector ring(s) or are attenuated by the body's tissues would pass undetected. These concepts are schematically described in Figure 1.

1.2 A Mathematical Model

The preceding description of the physics of PET is a slight oversimplification because of the angulation and range prob-

* Y. Vardi, L. A. Shepp, and L. Kaufman are Members of Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ 07974. The authors thank S. E. Levinson and M. M. Sondhi for a suggestion that simplified the derivations in Section 2; R. McGill, S. Morgenthaler, and A. R. Wilks for their help and interest in implementing the software used to display the reconstructions; J. Reeds for technical discussions and interest; and the editor and a referee for constructive suggestions and pointers that improved the presentation.

Journal of the American Statistical Association
March 1985, Vol. 80, No. 389, Applications

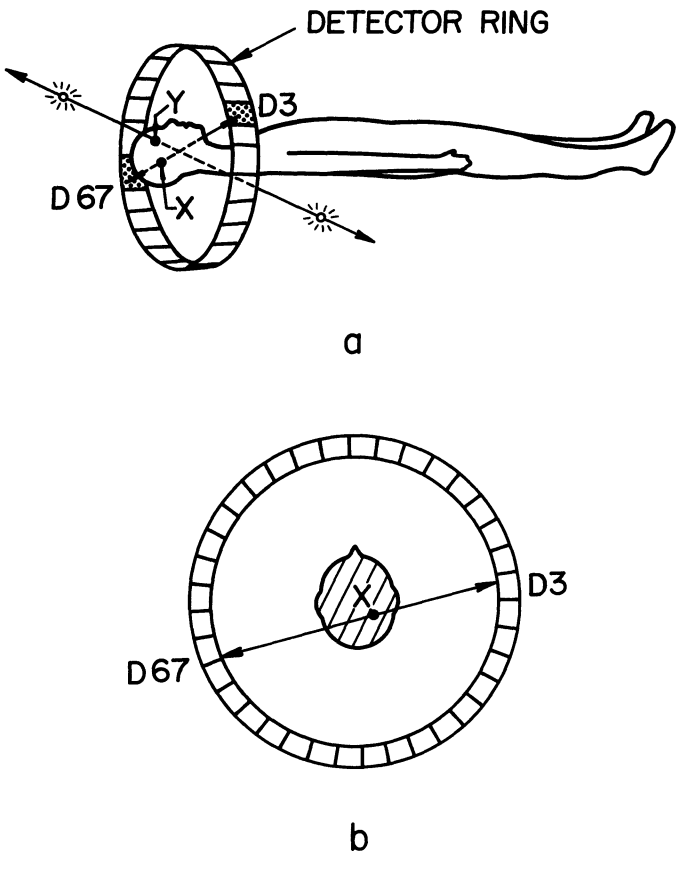

Figure 1. (a) Two annihilations: one, at x, that is detected in tube (3, 67) and the other, at y, that passes undetected because the photon path does not intersect the detector ring. (b) Top view of the detector ring "plane" of a.

lems discussed in Section 1.3. Nevertheless it is detailed enough to capture all of the important physical aspects of PET and simple enough to be modeled mathematically.

We start by assuming that emissions occur according to a spatial Poisson point process in a certain region H (the patient's head) of R^3, with an unknown intensity function $\{\lambda(x);\ x \in H\}$, which is usually referred to as the *emission density*. Each positron emission then results in an annihilation, and once a positron annihilation occurs at a point x, say, the probability that this annihilation is detected in tube d is

$$c(x, d) = \text{probability that a line through } x, \text{ with uniformly random orientation in space, intersects the two detector elements defining the tube } d. \quad (1.1)$$

Up to a normalizing constant, $c(x, d)$ is the angle of view from the point x into the tube d, and since the detector ring cannot possibly surround the head completely, we have

$$c(x, \cdot) = \sum_d c(x, d) \leq 1. \quad (1.2)$$

This, of course, means that some of the photons go undetected. The measured data set is the total number of coincidences counted in tube d, $n^*(d)$, $d = 1, \ldots, D$; and since classifying the annihilations according to the tubes that detected them amounts to a thinning of the Poisson point process, it can be shown that $n^*(d)$, $d = 1, \ldots, D$, constitute D independent Poisson random variables with means

$$\lambda^*(d) = \int_{x \in H} \lambda(x) c(x, d)\, dx, \qquad d = 1, \ldots, D. \quad (1.3)$$

Our problem is to estimate $\lambda(x)$ based on these data. [Note that the integrand $\lambda(x)c(x, d)$ is zero for x's outside the head-section defined by the detector ring(s).] If $\lambda(x)$ is allowed to be arbitrary, our problem amounts to estimating an infinite number (actually a continuum) of parameters on the basis of finite data. To avoid this situation and to simplify the mathematics, we proceed with the assumption that there exists a fine grid of B boxes and that in the bth box the value of $\lambda(x)$ is a constant. Slightly abusing the notation, we denote this constant by $\lambda(b)/V(b)$, where $V = V(b)$ is the volume of the bth box. Note that the $V(b)$ in the denominator is needed because $\lambda(b)$ represents the expected number of total counts in box b whereas $\lambda(x)$ is the density of counts at x. Thus the emission density $\{\lambda(x);\ x \in H\}$ is a step function that can be written in terms of the constants $\lambda(b)$, $b = 1, \ldots, B$, as follows:

$$\lambda(x) = \sum_{b=1}^{B} \frac{\lambda(b)}{V(b)} I[x \in b\text{th box}], \quad (1.4)$$

where, as usual, $I[\ ]$ denotes the indicator function. [We explain later in this section why choosing $\lambda(x)$ to be a step function of the form (1.4) is both natural and desirable.] Substituting (1.4) into (1.3) results in the mean $\lambda^*(d)$ of the data, $n^*(d)$, being

$$\lambda^*(d) = \sum_{b=1}^{B} \lambda(b) p(b, d), \qquad d = 1, \ldots, D, \quad (1.5)$$

where

$$p(b, d) = \frac{1}{V(b)} \int_{x \in b\text{th box}} c(x, d)\, dx, \qquad b = 1, \ldots, B, \quad d = 1, \ldots, D. \quad (1.6)$$

Here we note that the constants $p(b, d)$ can be computed from the detector ring(s) geometry, as the average angle of view of box b into tube d.

It is instructive, even at the cost of duplicating some of our notation, to rederive the problem in a slightly different way: Let

$$n(b) = \text{number of emissions in box } b, \qquad b = 1, \ldots, B, \quad (1.7)$$

and assume that $n(1), \ldots, n(B)$ are independent Poisson random variables with (unknown) means $\lambda(1), \ldots, \lambda(B)$. Assume further that once an emission occurs in box b, the conditional probability that it is detected in tube d is independent of other emissions and is given by

$$p(b, d) = P(\text{detected in } d \mid \text{occurring in } b), \qquad d = 1, \ldots, D, \quad b = 1, \ldots, B, \quad (1.8)$$

where the $p(b, d)$'s are known nonnegative constants (discussed in more detail in the next section), not all of which are zero; that is,

$$0 < p(b, \cdot) \equiv \sum_{d=1}^{D} p(b, d) \leq 1, \qquad b = 1, \ldots, B. \quad (1.9)$$

Then

$$n(b, d) \equiv \text{number of emissions occurring in box } b \text{ and detected in tube } d \quad (1.10)$$

($b = 1, \ldots, B$ and $d = 1, \ldots, D$) are obtained by thinning each $n(b)$ according to the probabilities $p(b, d)$, $d = 1, \ldots, D$, and hence the $n(b, d)$'s are Poisson random variables, independent of each other, with means

$$\lambda(b, d) \equiv \lambda(b)p(b, d), \qquad b = 1, \ldots, B, \quad d = 1, \ldots, D. \quad (1.11)$$

With this description in hand, the measured data $n^*(1), \ldots, n^*(D)$ can be reexpressed as

$$n^*(d) \equiv n(\cdot, d) \equiv \sum_{b=1}^{B} n(b, d) = \text{total number of emissions detected in tube } d, \qquad d = 1, \ldots, D. \quad (1.12)$$

If we further let

$$n(b, \cdot) \equiv \sum_{d=1}^{D} n(b, d) = \text{total number of emissions that have occurred in box } b \text{ and been detected}, \qquad b = 1, \ldots, B, \quad (1.13)$$

then

$$En(b, \cdot) = \lambda(b)p(b, \cdot); \quad (1.14)$$

and since $p(b, \cdot)$ is known, we see that the PET reconstruction problem is equivalent to the problem of estimating the mean of $n(b, \cdot)$, $b = 1, \ldots, B$, on the basis of $n(\cdot, d)$, $d = 1, \ldots, D$.

1.3 More on the Model

Before using the model, it makes sense to review carefully the adequacy of our assumptions. The assumption that radioactive emissions follow Poisson statistics seems beyond challenge and requires no justification. The assumption that $\lambda(x)$ is a step function is convenient and, to a certain extent, even advantageous. Because of machine computations and display limits, discretization of the estimate at some stage of the reconstruction is necessary; and at least for the maximum likelihood method of Section 2.1, there seems to be no essential difference in the final result if we discretize either at the outset or after deriving a functional equation (using a limiting argument) that defines the estimate in the continuous case. Discretizing at the outset, however, simplifies the problem mathematically, making it into an estimation problem from incomplete data, for which a considerable amount of statistical methodology has been developed. One can also argue that for fixed values $n^*(1), \ldots, n^*(D)$, there is a certain grid fineness, B, beyond which any refinement of the grid would not improve the image resolution of the estimate of $\lambda(x)$. The coarsest such grid can define [vaguely here and somewhat more precisely when we talk about maximum likelihood estimate of $\lambda(x)$] a concept of inherent resolution of the data $n^*(1), \ldots, n^*(D)$. Finally, we note that any true emission density can be approximated arbitrarily closely by a step function of the form (1.4), and so this assumption would not detract from the relevance of the discussed methods to the medical problem of reconstructing $\lambda(x)$.

For the sake of simplicity, several assumptions concerning the physics of PET have been left out of the mathematical model of Section 1.2. The following is a short description of these assumptions and an outline of how they can be incorporated into the model. One neglected assumption is that positrons have a nonzero (2–4 mm) range before annihilation. Furthermore, the angle between the paths of the two photons is typically a few degrees less than 180°. This situation can be accounted for in our model by incorporating the positrons' behavior into the $p(b, d)$'s; in this case the $p(b, d)$ could also be positive for a box b that is close to the tube d but does not intersect it. Another aspect of the physics that our model ignored is that the body tissues can attenuate the annihilation photons, which will cause measurement errors. This can be corrected by modifying the $p(b, d)$'s, by using physical measurement procedures that involve the patient. For instance, for a pair of photons originating in box b and traveling along a line δ in tube d, the probability that neither will be absorbed or deflected (so that they generate a coincidence count in tube d) is very nearly

$$\exp\left(-\int_{\delta} \mu(y)\, dy\right) \quad (1.15)$$

(cf. Shepp and Kruskal 1978), where $\mu(y)$ is the linear attenuation function, which characterizes the patient's tissue attenuation (and will not be discussed in this article). Since the tubes are relatively narrow, we can assume, as a mathematical approximation, that the line integral of μ along δ does not vary as δ varies within the tube, and so (1.15) is a characteristic of the tube d and the patient, say $q(d, \mu)$. Thus we can modify the $p(b, d)$ of (1.6) to be

$$p(b, d) = \left(\frac{1}{V(b)} \int_{x \in b\text{th box}} c(x, d)\, dx\right) q(d, \mu), \qquad b = 1, \ldots, B, \quad d = 1, \ldots, D. \quad (1.16)$$

Indeed, in practice each PET scan is typically preceded by two separate measurements, the main purpose of which is to estimate $q(d, \mu)$. In the first measurement, there is no patient and the positron-emitting substance is contained in a hoop cocentered and surrounded by the detector ring. In the second measurement, the hoop with the positron-emitting substance and the detector ring are positioned as before, but now the patient is positioned so that he or she is surrounded by the hoop. For each tube d, the ratio of the count in the second measurement to that in the first measurement provides an estimate of $q(d, \mu)$ (see Derenzo et al. 1977, sec. 2.5), as seen from the fact that the right side of (1.16) divided by the right side of (1.6) is (1.15). This estimate can then be used to correct the $p(b, d)$'s for tissue attenuation or, alternatively (as done in practice), to inflate the $n^*(d)$'s by multiplying them by $q^{-1}(d, \mu)$'s to compensate for the attenuated photons.

There are additional situations that may call for a modifi-

cation of the $p(b, d)$'s. For instance, the differential time of flight for the two detected photons might be measured (see Snyder et al. 1981 for a mathematical treatment of such a model) so that one can pin down more accurately the region, within the tube, in which the annihilation took place. This information can be incorporated into our model by subdividing each tube d into subtubes $d_1, \ldots, d_m$ and computing $p(b, d_j)$, the probability that an emission in box b is detected in subtube d_j.

Although modifications as described will undoubtedly be incorporated into the computation of the $p(b, d)$'s in practice, for the purpose of this article, it is sufficient to work with the $p(b, d)$'s as angles of views, as defined in (1.6).

Another aspect of the physics that we have left out is the radioactive decay of the isotope. This is usually incorporated into the model by assuming that the emission density at the point x at time t is $\lambda(x)e^{-\beta t}$, where β is a known constant that depends on the isotope in use. The total emission rate during the entire scan, say from time 0 to time T, at the point x is then $\lambda(x)(1 - e^{-\beta T})\beta^{-1}$, and since β is known, we can correct our estimate of $\boldsymbol{\lambda}$. The assumption that the emission density can be factored into a time component and a metabolic activity component makes sense under certain conditions, for example, when the metabolized tracer is retained within the organ of interest during the entire measurement period. (See Ter-Pogossian et al. 1980, p. 177, for a discussion of some of the elaborate requirements involved in choosing the metabolite and the radionuclide so that the emission data supplied by a PET scan can indeed be interpreted in the manner we have described.)

2. ESTIMATING THE POISSON INTENSITY

In this section we describe various estimates of $\boldsymbol{\lambda}$, with the emphasis on the maximum likelihood (ML) estimate originally suggested in Shepp and Vardi (1982). Throughout Section 2, whenever the arguments b and d vary without a specified range, it is understood that $b = 1, \ldots, B$ and $d = 1, \ldots, D$.

2.1 Maximum Likelihood Estimate (MLE)

We recall from Section 1.2 that $n^*(d)$ are independent Poisson variables with mean $\lambda^*(d)$, $d = 1, \ldots, D$; so the likelihood of the observed data is

$$L(\boldsymbol{\lambda}) = P(\mathbf{n}^*|\boldsymbol{\lambda}) = \prod_{d=1}^{D} e^{-\lambda^*(d)} \frac{\lambda^*(d)^{n^*(d)}}{n^*(d)!}. \tag{2.1}$$

Using (1.5) we get the following expressions for the first and second derivatives of the log-likelihood function, $l(\boldsymbol{\lambda}) \equiv \log L(\boldsymbol{\lambda})$:

$$\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b_0)} = -p(b_0, \cdot) + \sum_{d=1}^{D} \frac{n^*(d)p(b_0, d)}{\sum_{b'=1}^{B} \lambda(b')p(b', d)}, \tag{2.2}$$

$$\frac{\partial^2 l(\boldsymbol{\lambda})}{\partial \lambda(b_0)\partial \lambda(b_1)} = -\sum_{d=1}^{D} \frac{n^*(d)p(b_0, d)p(b_1, d)}{\left[\sum_{b'=1}^{B} \lambda(b')p(b', d)\right]^2}. \tag{2.3}$$

Now the matrix of second derivatives is negative semidefinite, because for an arbitrary (nonzero) vector $\mathbf{z} = (z(1), \ldots, z(B))$, we have

$$\sum_{b_0=1}^{B} \sum_{b_1=1}^{B} z(b_0)z(b_1) \frac{\partial^2 l(\boldsymbol{\lambda})}{\partial \lambda(b_0)\partial \lambda(b_1)} = -\sum_{d=1}^{D} c_d^2 \leq 0, \tag{2.4}$$

where

$$c_d = \frac{\sqrt{n^*(d)}}{\sum_{b'=1}^{B} \lambda(b')p(b', d)} \sum_{b=1}^{B} z(b)p(b, d) = \frac{\sqrt{n^*(d)}}{\lambda^*(d)} \sum_{b=1}^{B} z(b)p(b, d). \tag{2.5}$$

This shows that $l(\boldsymbol{\lambda})$ is concave. Hence it follows [Zangwill 1969, Theorem 2.19 (e)] that sufficient conditions for $\hat{\boldsymbol{\lambda}}$ to be a maximizer of l are the Kuhn Tucker (KT) conditions, which for our case turn out to be, for each $b = 1, \ldots, B$,

$$0 = \lambda(b) \frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\bigg|_{\hat{\boldsymbol{\lambda}}} = -\hat{\lambda}(b)p(b, \cdot) + \sum_{d=1}^{D} \frac{n^*(d)\hat{\lambda}(b)p(b, d)}{\sum_{b'=1}^{B} \hat{\lambda}(b')p(b', d)} \tag{2.6}$$

and

$$\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\bigg|_{\hat{\boldsymbol{\lambda}}} \leq 0 \qquad \text{if } \hat{\lambda}(b) = 0. \tag{2.7}$$

Before we proceed to solve (2.6) and (2.7), we note that it is simpler—and involves no loss of generality—to assume that

$$p(b, \cdot) = \sum_{d} p(b, d) = 1, \qquad b = 1, \ldots, B, \tag{2.8}$$

because if this is not the case [and indeed, in practice, (2.8) is never the case], then by defining

$$\theta(b) = \lambda(b)p(b, \cdot), \qquad q(b, d) = p(b, d)/p(b, \cdot), \tag{2.9}$$

we can reexpress (2.6) and (2.7) as

$$-\theta(b) + \sum_{d=1}^{D} \frac{n^*(d)\theta(b)q(b, d)}{\sum_{b'=1}^{B} \theta(b')q(b', d)} = 0, \tag{2.10}$$

$$-1 + \sum_{d=1}^{D} \frac{n^*(d)q(b, d)}{\sum_{b'=1}^{B} \theta(b')q(b', d)} \leq 0 \qquad \text{if } \theta(b) = 0. \tag{2.11}$$

Equations (2.10) and (2.11) have the same form as (2.6) and (2.7) but now have the desired property

$$q(b, \cdot) = \sum_{d} q(b, d) = 1, \qquad b = 1, \ldots, B. \tag{2.12}$$

The $\boldsymbol{\lambda}$ that solves (2.6) and (2.7) is obtained from the $\boldsymbol{\theta}$ that solves (2.10) and (2.11) by the substitution (2.9). Thus for the remainder of Section 2.1, we assume, without loss of generality, that (2.8) holds. With such a relatively simple expression for the right side of (2.6), one can think of many iterative

schemes that would converge to a maximum of l. Of particular appeal is the following scheme.

The EM Algorithm. (a) Start with an initial estimate $\boldsymbol{\lambda}^{\text{old}}$, say, satisfying $\lambda^{\text{old}}(b) > 0$, $b = 1, \ldots, B$. (b) If $\boldsymbol{\lambda}^{\text{old}}$ denotes the current estimate of $\boldsymbol{\lambda}$, define a new estimate, $\boldsymbol{\lambda}^{\text{new}}$, by

$$\lambda^{\text{new}}(b) = \lambda^{\text{old}}(b) \sum_{d=1}^{D} \frac{n^*(d)p(b, d)}{\sum_{b'=1}^{B} \lambda^{\text{old}}(b')p(b', d)}, \quad b = 1, \ldots, B. \tag{2.13}$$

(The algorithm could never lead to a quotient of a positive numerator divided by a zero denominator; zero divided by zero is defined as zero.) (c) If the required accuracy for numerical convergence has been achieved, then stop. Otherwise, return to step (b), with $\boldsymbol{\lambda}^{\text{new}}$ replacing $\boldsymbol{\lambda}^{\text{old}}$. Note that numerical convergence can be determined either by testing whether $\boldsymbol{\lambda}^{\text{new}}$ is sufficiently close to $\boldsymbol{\lambda}^{\text{old}}$ or by testing whether the increment in l, which is

$$l(\boldsymbol{\lambda}^{\text{new}}) - l(\boldsymbol{\lambda}^{\text{old}}) = \sum_{d=1}^{D} n^*(d)(\log \lambda^{*\text{new}}(d) - \log \lambda^{*\text{old}}(d)),$$

is sufficiently small. (As explained later, this increment is always nonnegative.)

The reason for the appeal of (2.13) is that it is an instance of the EM algorithm, and hence it follows from Theorem 1 of Dempster et al. (1977) that

$$l(\boldsymbol{\lambda}^{\text{old}}) < l(\boldsymbol{\lambda}^{\text{new}}), \tag{2.14}$$

unless $\boldsymbol{\lambda}^{\text{old}} = \boldsymbol{\lambda}^{\text{new}}$, in which case the algorithm (2.13) has converged. To see that (2.13) is indeed an EM algorithm, the reader is invited to apply the recipe given in Dempster et al. (1977) with $\{n(b, d); b = 1, \ldots, B, d = 1, \ldots, D\}$ and $\{n(\cdot, d); d = 1, \ldots, D\}$ being the complete and incomplete data (in the terminology of the EM paper), respectively. Here we take a shortcut in explaining the rationale behind (2.13) in the context of our example: If we believed that $\boldsymbol{\lambda}^{\text{old}}$ is the true $\boldsymbol{\lambda}$, we would estimate the number of annihilations in box b to be

$$\hat{n}(b) = E[n(b, \cdot) \mid \boldsymbol{\lambda}^{\text{old}}, \mathbf{n}^*];$$

this is the E step. But if $\hat{n}(b)$ is our estimate for the emission count in box b, it should also be our estimate for the emission density $\lambda(b)$, because this is the maximum likelihood estimate if $\hat{n}(b)$ was indeed the emission count in box b; this is the M step. Combining the two steps we get

$$\begin{aligned}
\lambda^{\text{new}}(b) &= \hat{n}(b) = E[n(b, \cdot) \mid \boldsymbol{\lambda}^{\text{old}}, \mathbf{n}^*] \\
&= \sum_d E[n(b, d) \mid \boldsymbol{\lambda}^{\text{old}}, \mathbf{n}^*] \\
&\overset{(1)}{=} \sum_d E[n[b, d) \mid \boldsymbol{\lambda}^{\text{old}}, \mathbf{n}^*(d)] \\
&\overset{(2)}{=} \sum_d \frac{n^*(d)\lambda^{\text{old}}(b, d)}{\sum_{b'=1}^{B} \lambda^{\text{old}}(b', d)} \\
&= \lambda^{\text{old}}(b) \sum_d \frac{n^*(d)p(b, d)}{\sum_{b'=1}^{B} \lambda^{\text{old}}(b')p(b', d)},
\end{aligned}$$

which is (2.13). We note that in (1) we used the mutual independence of the $n^*(d)$'s, and in (2) we used the fact that if X_i are independent Poisson variables with mean a_i ($i = 1, \ldots, m$), then the conditional distribution of X_j, given $\sum X_i = x^*$, is binomial $(x^*, a_j/\sum a_i)$, so $E[X_j \mid \sum X_i = x^*] = x^* a_j/\sum a_i$.

In the following theorem, we summarize the discussion thus far and state the convergence of (2.13) to a point of maximum.

Theorem. (a) $l(\boldsymbol{\lambda})$ is concave and hence all its maxima are global maxima. (b) The EM algorithm converges [monotonically in the sense of (2.14)] to a global maximum of $l(\boldsymbol{\lambda})$. (c) The maximum of $l(\boldsymbol{\lambda})$ is unique if and only if the grid is such that the D vectors

$$(\sqrt{n^*(d)}/\lambda^*(d))(p(1, d), \ldots, p(B, d)), \quad d = 1, \ldots, D, \tag{2.15}$$

span E_B, the B-dimensional Euclidean space.

Proof. Part (a) follows from (2.4), and (c) follows from the fact that nonuniqueness could occur if and only if $c_d^2 = 0$ ($d = 1, \ldots, D$), which, from (2.5), is equivalent to having a nonzero $\mathbf{z}$ that is orthogonal to each of the vectors in (2.15). Such a $\mathbf{z}$ exists if and only if the vectors in (2.15) do not span E_B. Thus it remains to prove (b)—the convergence of (2.13) to a point of maximum. This, however, follows from Csiszàr and Tusnàdy (1982, Theorem 5). (See the following remark and the Appendix for further details.) This completes the proof.

Remark. Shepp and Vardi (1982) asserted that the sequence of estimates produced by (2.13), say $\boldsymbol{\lambda}^{(k)}$ ($k = 0, 1, 2, \ldots$), converges as $k \to \infty$ to a point of maximum. The proof given by Shepp and Vardi has a gap, as was pointed out in Lange and Carson (1984). The assertion, however, is correct; and it follows from Csiszàr and Tusnàdy's (1982) brilliant paper, which gives powerful convergence results for a class of algorithms of which (2.13) is only a special case. Another important paper on the subject is Cover's (1984), in which a similar algorithm to (2.13) is derived in connection with optimizing expected log investment. Cover's Theorem 3 gives the convergence of $l(\boldsymbol{\lambda}^{(k)})$ to the maximum of $l(\boldsymbol{\lambda})$, which implies convergence of $\boldsymbol{\lambda}^{(k)}$ to a point of maximum, say $\hat{\boldsymbol{\lambda}}$, whenever $l(\boldsymbol{\lambda})$ has a unique maximum. For the case of nonunique maximum, however, which is more relevant to the PET problem, one needs the stronger convergence result of Csiszàr and Tusnàdy. We outline in the Appendix a convergence proof based on their deep (though hard to follow) geometric argument.

A Sequence of Upper Bounds on $l(\hat{\boldsymbol{\lambda}})$ (Cover 1984, Theorem 4, and Csiszàr and Tusnàdy 1982, Theorem 5). Following Cover's (1984) Theorem 4, we get, using the concavity of $\log(x)$ and $\sum \hat{\lambda}(b) = \sum n^*(d) = N^*$,

$$\begin{aligned}
\sum_d \frac{n^*(d)}{N^*} \log \frac{\hat{\lambda}^*(d)}{\lambda^{(n)*}(d)} &\leq \log \sum_d \frac{n^*(d)}{N^*} \sum_b \frac{\hat{\lambda}(b)p(b, d)}{\lambda^{(n)*}(d)} \\
&= \log \sum_b \frac{\hat{\lambda}(b)}{N^*} \sum_d \frac{n^*(d)p(b, d)}{\lambda^{(n)*}(d)} \\
&\leq \max_b \log \sum_d \frac{n^*(d)p(b, d)}{\lambda^{(n)*}(d)}.
\end{aligned}$$

Thus

$$l(\boldsymbol{\lambda}^{(n)}) \leq l(\hat{\boldsymbol{\lambda}}) \leq l(\boldsymbol{\lambda}^{(n)}) + N^* \max_b \log \sum_d \frac{n^*(d)p(b, d)}{\lambda^{(n)*}(d)} .$$

Here the left inequality follows from the optimality of $\hat{\boldsymbol{\lambda}}$. The KT conditions (2.6) and (2.7) require that the rightmost term in the preceding inequality be ≤ 0 in the limit, and so the upper bound converges to $l(\hat{\boldsymbol{\lambda}})$ as $n \to \infty$. Note that from (2.13)

$$\frac{\lambda^{(n+1)}(b)}{\lambda^{(n)}(b)} = \sum_d \frac{n^*(d)p(b, d)}{\lambda^{(n)*}(b)} ,$$

so the above bounds can also be written as

$$l(\lambda^{(n)}) \leq l(\hat{\boldsymbol{\lambda}}) \leq l(\boldsymbol{\lambda}^{(n)}) + N^* \max_b \log \frac{\lambda^{(n+1)}(b)}{\lambda^{(n)}(b)} . \quad (2.16)$$

2.2 Further Comments on the MLE and the EM Algorithm

1. The Monotonicity of the Algorithm. The practical implication of this property is that we can initialize the EM algorithm with *any* (positive) estimate $\boldsymbol{\lambda}^0$ and then improve on it. For instance, to save on numerical computation we can choose $\boldsymbol{\lambda}^0$ to be the convolution backprojection reconstruction that is described in Section 2.3 [slightly corrected, if necessary, to satisfy the positivity required in step (a) of the algorithm] and continue with the EM iterations from there on.

2. The Implication of Condition (c) of the Theorem. Since $\lambda^*(d) > 0$ and typically (in practice) $n^*(d) > 0$, we can think of (c) as saying that the maximum of $l(\boldsymbol{\lambda})$ is unique if and only if $(p(1, d), \ldots, p(B, d))$ $d = 1, \ldots, D$ span E_B. Now for any detector ring(s) design, there is a maximal grid B such that $(p(1, d), \ldots, p(B, d))$ $d = 1, \ldots, D$ span E_B; so for any refinement of the grid, the MLE of $\boldsymbol{\lambda}$ would not be unique. Although such a B is too cumbersome to compute, an upper bound is of course D, because if $B > D$, E_B cannot be spanned by D vectors. Thus if $B > D$, then the MLE is nonunique. For example, in Section 3 we shall consider a planar reconstruction based on a single ring with 128 equally spaced detectors mounted on it. The region within which we look to reconstruct $\boldsymbol{\lambda}$ is a disk of radius 1 cocentered with the detector ring, and we refer to it as the *patient circle*. The number of tubes that cross the patient circle is then $65 \times \frac{128}{2} = 4{,}160$ because there are 65 detectors opposite each one. For any grid with more than 4,160 boxes in the patient circle, the MLE of $\boldsymbol{\lambda}$ would be nonunique. Clearly in the nonuniqueness case, the point of convergence of the EM algorithm depends on the initial estimate $\boldsymbol{\lambda}^0$, and in this respect the choice of $\boldsymbol{\lambda}^0$ is somewhat akin to a choice of a Bayes prior. The question of how close the various maxima are needs to be studied using simulation. We note, however, from part (a) of the Theorem that convex combinations of maxima are also maxima, which suggests that it is possible (though we have not studied it) that by averaging these maxima, one gets (approximately) the same MLE, $\hat{\boldsymbol{\lambda}}$, that would have been obtained had we used the maximal grid, which gives a unique MLE. If this is the case, then one can legitimately call such a B the "inherent resolution of the detector ring(s) design."

3. The Connection With Gradient Methods. Comparing (2.2) and (2.13) we see that the EM is a gradient-type algorithm, and the iteration in (2.13) can be written, in matrix notation, as

$$\boldsymbol{\lambda}^{\text{new}} = \boldsymbol{\lambda}^{\text{old}} + \boldsymbol{\lambda}^{\text{old}} \operatorname{diag}\left(\frac{\partial l(\boldsymbol{\lambda}^{\text{old}})}{\partial \lambda(1)}, \ldots, \frac{\partial l(\boldsymbol{\lambda}^{\text{old}})}{\partial \lambda(B)}\right) .$$

In view of this and the algorithm's monotonicity, one would expect it to have a convergence rate similar to other gradient-type algorithms. (We plan to study this issue in the future.)

4. The Variance. Although a typical PET scan may record several million emissions, the number of tubes in which these emissions are recorded is only several thousand. This is comparable with the number of parameters we wish to estimate, so the standard asymptotic theory for maximum likelihood estimates may not always apply. Nevertheless, for those situations in which the asymptotics are relevant (e.g., in scanners capable of increasing the number of tubes by jiggling a section of the detector ring), and to satisfy our curiosity, we derive the formulas that asymptotic theory would lead us to. From (2.3) and (1.5) we get that the asymptotic covariance matrix of $\hat{\boldsymbol{\lambda}}$ is proportional to the inverse of the matrix whose typical element is

$$\sum_d \frac{p(b_0, d)p(b_1, d)}{\sum_b \lambda(b)p(b, d)} .$$

Because of the size of the matrix $p(b, d)$, this expression may not be useful in determining confidence bounds for the ML reconstruction. This and, more so, the nonapplicability of the asymptotic theory when the number of parameters is large suggest that the variability of $\hat{\boldsymbol{\lambda}}$ about the true $\boldsymbol{\lambda}$ might have to be studied by using a Monte Carlo simulation.

2.3 Other Estimates

2.3.1 Moments Estimates and Convolution Backprojection

For the moment estimator one equates the observed data $n^*(d)$ with their expectation $\lambda^*(d)$, for $d = 1, \ldots, D$, and tries to solve for $\boldsymbol{\lambda}$. This gives, using (1.5),

$$n^*(d) = \sum_{b=1}^{B} \lambda(b)p(b, d), \qquad d = 1, \ldots, D, \quad (2.17)$$

or, in matrix notation,

$$\mathbf{n}^* = \mathbf{P}'\boldsymbol{\lambda}, \quad (2.18)$$

where $\mathbf{n}^*$ and $\boldsymbol{\lambda}$ are column vectors and $\mathbf{P}'$ is the transpose of the $p(b, d)$ matrix. If (2.18) has a solution, say $\boldsymbol{\lambda}^0$, then $\boldsymbol{\lambda}^0$ is also a stationary point of (2.13), and hence the moment estimator, if it exists, is also a maximum likelihood estimate of $\boldsymbol{\lambda}$. Nevertheless, in practice, because of low count rate variations, $\mathbf{n}^*$ is sufficiently far from $\boldsymbol{\lambda}^*$ that (2.18) would typically have no solution and the moment estimate would not exist.

The convolution backprojection (CBP) is a technique that was developed to solve the reconstruction problems posed by x-ray transmission tomography. Important contributions to this method have been made by Ramachandran and Lakshminar-

ayanan (1971), Lakshminarayanan (1975), Shepp and Logan (1974), and others; a discussion of the method can be found in Shepp and Kruskal (1978). Despite the fact that the physics of transmission tomography is entirely different from that of PET, the technique has been adapted to PET and is now in use in all PET scanners. In the context of PET, the CBP technique is similar to the moment method described earlier. To estimate $\boldsymbol{\lambda}$ using the CBP method, one approximates the angles of views in (1.3) as a positive constant independent of d, for all x in d, so that up to a constant multiplier we have

$$En^*(d) \equiv \lambda^*(d) = \int_{x \in d\text{th tube}} \lambda(x)\, dx, \qquad d = 1, \ldots, D; \quad (2.19)$$

then with the same rationale as in the moment method, one tries to find a function $\lambda(x)$ whose integral along the dth tube is equal to $n^*(d)$; that is,

$$n^*(d) = \int_{x \in d\text{th tube}} \lambda(x)\, dx, \qquad d = 1, \ldots, D. \quad (2.20)$$

Equations (2.20), with $n^*(d)$ replaced by the logarithm of $I_{\text{in}}(d)/I_{\text{out}}(d)$, are those used to reconstruct the image in x-ray transmission tomography; here $I_{\text{in}}(d)$ and $I_{\text{out}}(d)$ are the input and output intensities of the x-ray beam along the tube d. Thus since (2.20) is a standard transmission tomography reconstruction problem, the CBP algorithm, which is fast and gives high-resolution reconstruction in x-ray transmission tomography, seems appropriate. There are two main drawbacks of this approach: (a) The angle of view function $c(x, d)$ [defined in (1.1)] is not constant along tubes, so the rightmost side of (2.19) is not a good approximation to $En^*(d)$; and (b) as in the method of moments, because of low count rates intrinsic to PET but not to transmission tomography, the variation of $n^*(d)$ around its mean is typically big, so $n^*(d)$ is not an accurate estimate of $En^*(d)$. As indicated in Shepp and Vardi (1982), the effect of these two errors, combined, results in CBP reconstructions that are typically noisier than the maximum likelihood reconstruction.

2.3.2 Least Squares Estimator

Since (2.17) typically has no solution, an alternative approach to solving it is to minimize $\|\mathbf{n}^* - \mathbf{P}'\boldsymbol{\lambda}\|^2$. This gives the least squares (LS) estimate $\hat{\boldsymbol{\lambda}}$ as the solution in $\boldsymbol{\lambda}$ of the normal equations:

$$\mathbf{PP}'\boldsymbol{\lambda} = \mathbf{P}\mathbf{n}^*. \quad (2.21)$$

(A solution always exists, but it may not be unique and may not satisfy $\lambda \geq 0$.) Because of the size of the matrix $\mathbf{P}$, trying to solve these equations for typical values of B and D is not practical. Instead, a direct constraint minimization of $\|\mathbf{n}^* - \mathbf{P}'\boldsymbol{\lambda}\|^2$, using an algorithm that avoids these equations, such as the conjugate-gradient method (e.g., see Ortega and Rheinboldt 1970, p. 262), could be used to find the LS estimate. This is discussed further in Section 3.

2.3.3 Stein-type Estimators

In recent years, several Stein-type estimates have been suggested for the Poisson case. What follows is a brief review of some of them and a discussion of their potential role in the PET reconstruction problem.

Consider a hypothetical situation in which the actual emission process $\mathbf{n} = \{n(b); b = 1, \ldots, B\}$ is observed and the problem is to estimate $\boldsymbol{\lambda} = \{\lambda(b); b = 1, \ldots, B\}$. Then the estimate

$$\tilde{\lambda}_P(b) = n(b) - ((B - N_0 - 2)_+/S) \sum_{k=1}^{n(b)} k^{-1}, \quad (2.22)$$

with $S \equiv \sum_{b=1}^{B} (\sum_{k=1}^{n(b)} k^{-1})^2$ and $N_0 \equiv \#\{b; n(b) = 0\}$ (Peng 1975; see also Hudson 1978), dominates (in the sense of having a uniformly lower expected loss) the MLE, $\mathbf{n}$, of $\boldsymbol{\lambda}$ with respect to the squared error loss function. The estimate

$$\tilde{\lambda}_{CZ}(b) = \left(1 - \frac{B + \beta}{\sum_{b'=1}^{B} n(b') + B + \beta}\right) n(b) \quad (2.23)$$

for an arbitrary $\beta \geq 0$ (Clevenson and Zidek 1975) dominates the MLE with respect to the loss function

$$L_1(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = \sum_{b=1}^{B} \lambda(b)^{-1}(\tilde{\lambda}(b) - \lambda(b))^2, \quad (2.24)$$

and the estimate

$$\tilde{\lambda}_{TP}(b) = n(b) - \Delta_\phi(b),$$

$$\Delta_\phi(b) \equiv \frac{\phi\left(\sum_{b'=1}^{B} n(b')\right) m(n(b) - 2)}{\sum_{b'=1}^{B} m(n(b')) - m(n(b)) + m(n(b) - 2)}, \quad (2.25)$$

where $m(n) \equiv (n + 2)(n + 1)$ and $\phi(z)$ is an arbitrary real-valued nondecreasing function of z satisfying $0 \leq \phi(z) \leq 4(B - 1)$ and $\phi \equiv 0$, dominates the MLE for the loss function

$$L_2(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = \sum_{b=1}^{B} \lambda(b)^{-2}(\tilde{\lambda}(b) - \lambda(b))^2. \quad (2.26)$$

Note that the estimate of (2.25), which was suggested by Tsui and Press (1982), may assume negative values, and so in practice it is truncated at zero.

In trying to assess the potential importance of such estimates to the PET reconstruction problem, several points have to be taken into account. First, we do not observe the emission process, so the estimates do not apply directly. A reasonable strategy to overcome this is first to derive the MLE, $\hat{\boldsymbol{\lambda}}$, and then to use $\hat{\lambda}(b)$ in place of $n(b)$ in the definition of the estimates. A second point for consideration is the special character of the PET reconstruction problem, which involves a large number of parameters and a substantially larger number of emission counts, with typical numbers like $B = 5{,}000$ and $\sum n^*(d) = 10{,}000{,}000$. From the discussion in Clevenson and Zidek (1975, sec. 3), it seems that there is little to be gained by using $\tilde{\boldsymbol{\lambda}}_{CZ}$ instead of $\tilde{\boldsymbol{\lambda}}$ if the ratio $a \equiv \sum \hat{\lambda}(b)/B$ ($= 10^7/5{,}000 = 2{,}000$ in our case) is big. Similar "back-of-the-envelope" calculations need to be derived for other estimates such as $\tilde{\boldsymbol{\lambda}}_P$ and $\tilde{\boldsymbol{\lambda}}_{TP}$ to assess the saving in expected loss when

they are compared with $\hat{\boldsymbol{\lambda}}$. A third point to reckon with is the displaying technique. For instance, the method used in Section 3 is such that positive multiples of $\hat{\boldsymbol{\lambda}}$ would give the same picture as $\hat{\boldsymbol{\lambda}}$, so $\tilde{\boldsymbol{\lambda}}_{CZ}$ would be indistinguishable from $\hat{\boldsymbol{\lambda}}$. In this connection it is perhaps worth pointing out that the final display is often done in colors and is preceded by processing of the estimate to be displayed, say $\hat{\boldsymbol{\lambda}}$, in which smoothing and grouping into colors are often done to enhance details and increase contrast. In a certain respect this is in the spirit of $\tilde{\boldsymbol{\lambda}}_P$ and $\tilde{\boldsymbol{\lambda}}_{TP}$, which reduce larger observations more than they reduce smaller observations, thereby having the effect of nonlinear smoothers applied to the estimate. Finally, we can ask whether these Stein-type modifications would produce better reconstructions of $\boldsymbol{\lambda}$ than the MLE, $\hat{\boldsymbol{\lambda}}$. This, of course, depends on whether the particular loss function that gave rise to a given estimate is a better analytic representation of human visual perception than is a likelihood function. Conclusive answers can be given only after some more experiments, so we leave this general area for further investigation with the hope that this short discussion will stimulate additional research.

2.3.4 A Bayesian Approach

Conceptually it is simple to give a framework for Bayesian reconstruction of the emission density $\boldsymbol{\lambda}$, though it might prove difficult to carry out the computations involved. One can assume that $\boldsymbol{\lambda} = (\lambda(1), \ldots, \lambda(B))$ is a random vector from a known distribution F, say, and proceed to compute quantities such as the posterior expectation of $\boldsymbol{\lambda}$, $E_F(\boldsymbol{\lambda} \mid \mathbf{n}^*)$, or the posterior mode, $\max_{\boldsymbol{\lambda}} P_F(\boldsymbol{\lambda} \mid \mathbf{n}^*)$ (see Fortes 1980, ch. 3). The justification for such an approach could be that it incorporates prior knowledge about the brain's metabolism into the reconstruction. This, however, suggests that the prior distribution F should indeed be chosen to reflect such knowledge and not simply to facilitate the computations, as is often done in statistical applications. Since human brains come in different sizes and shapes, the problem of finding such a prior F is perhaps not a practical one and an alternative approach could be to introduce a penalty term into the likelihood function in which deviations of $\hat{\boldsymbol{\lambda}}$ from the a priori expected pattern of $\boldsymbol{\lambda}$ are penalized. (See Good and Gaskins 1980 for a related approach applied to density estimation.) An estimate derived as a maximizer of such a penalized likelihood function would be a compromise between a pure likelihood approach and a Bayes approach because it enables us to incorporate prior knowledge about the general behavior of $\boldsymbol{\lambda}$ into the reconstruction. We leave this general area open for further investigation.

In thinking about the potential role that Bayes reconstructions can play in PET, we are led to think about Bayes large-sample theory. One can argue that if the total count, $\sum n^*(d)$, is large and the prior law, F, satisfies some (weak) smoothness conditions (e.g., see Lindley 1965 and Walker 1969), then the Bayes estimate would be very close to the ML estimate, $\hat{\boldsymbol{\lambda}}$. Specifically it would have an approximate Gaussian distribution with mean $\hat{\boldsymbol{\lambda}}$ and covariance matrix whose inverse has the typical element

$$-\frac{\partial^2 l(\boldsymbol{\lambda})}{\partial\lambda(b_0)\partial\lambda(b_1)}\bigg|_{\hat{\boldsymbol{\lambda}}} = \sum_{d=1}^{D} \frac{n^*(d)p(b_0, d)p(b_1, d)}{\left(\sum_{b'=1}^{B} \hat{\lambda}(b')p(b', d)\right)^2}.$$

If anything, however, we find this to be a strong argument for preferring the MLE rather than a particular Bayes procedure. Note that the preceding asymptotic argument is driven by sending $\sum n^*(d)$ to infinity and holding D fixed. Because the number of estimated parameters, B, is large, D is fixed (although large too), and the priors suitable for PET are probably not smooth, the relevance of the large-sample results needs further study.

3. AN EXAMPLE

In this section we describe a computer simulation of a PET scan and use the EM algorithm and the (conjugate-gradient approximation to the) LS method to reconstruct the emission density. We assume a *single* detector ring of radius $\sqrt{2}$ with 128 equally spaced detectors mounted on it, and consequently our problem is reconstructing $\boldsymbol{\lambda}$ over a head section, which is a two-dimensional reconstruction. The emission density, $\lambda(x, y)$, which is used to generate the data and which we want to reconstruct, is as described in Figure 2; the nomenclature in the tomography literature for this quantity, or in general for the mathematical model that simulates a body section, is a *mathematical phantom* or, simply, a *phantom*. The phantom of Figure 2 is made up of eight ellipses and is chosen as a simplified imitation of the brain's metabolic activity, where the skull metabolizes at a low rate of .1 and the ventricles, tumors, and so on, metabolize at rates between .3 and 2.0. The region within which we look to reconstruct the phantom is a disc of radius 1 cocentered with the detector ring; we call it the *patient circle*. Overlaid on the patient circle (and circumscribed by the detector ring) is a 128 × 128 grid of display boxes that cover the square $|x|, |y| \leq 1$. In this setup, tubes that do not cross the patient circle will show no count [$n^*(d) = 0$], so D can be taken to be the number of tubes that cross the patient circle. Since each detector faces 65 opposite detectors, we get $D = 65 \times 64$. In addition, the number of

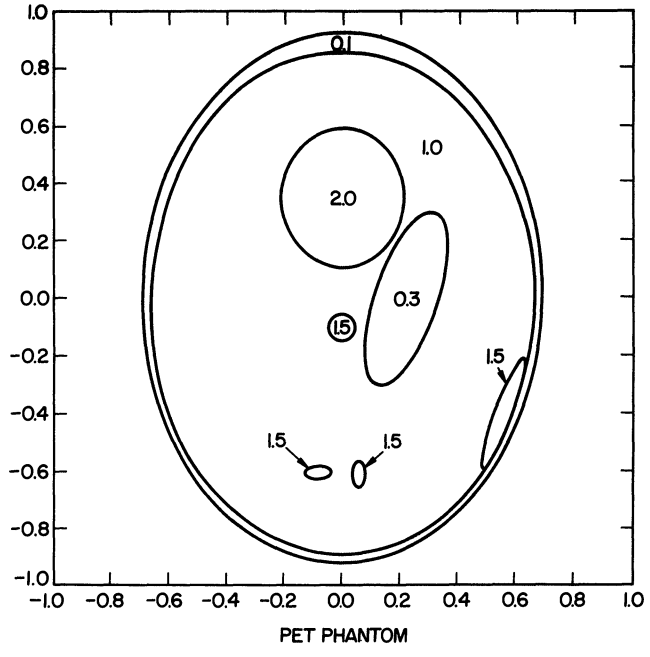

Figure 2. The phantom used in the computer simulation of the PET experiment.

display boxes B is approximately $128^2\pi/4$, because for those boxes b outside the patient circle, we know that $\lambda(b) = 0$.

We used the phantom of Figure 2 to generate a total of 10^7 detected emissions in a manner that agrees with the Poisson model. The details of the data generation method are as in Shepp and Vardi (1982). A "histogram" of the emission process, or more accurately a photodisplay of the observed emissions grouped into the display boxes of the 128 × 128 grid, is given in Figure 3a. Note that this is the *unobserved* process that takes place in the brain. For each emission, once it occurs at a point (x, y), say, we choose a line δ through (x, y) with a (uniformly) random orientation in the plane and add increments to the count for the tube corresponding to the two detector intervals through which δ passes. This gives us the tubes count $n^*(d)$, $d = 1, \ldots, D$. These data are then processed via the EM algorithm (2.13) (for 64 iterations) to find an MLE and via the conjugate-gradient algorithm (for 32 iterations) to find an LSE. The $p(b, d)$'s are computed as the angle of view from the center of box b into tube d. (In the future, we plan to study the improvement in resolution resulting from taking the $p(b, d)$'s to be the *average* angle of view from box b into tube d, where the average is taken over a grid of points within each box.) Photodisplays of the resulting EM reconstruction after 16, 32, and 64 iterations and the LS estimate after 32 iterations (of the conjugate-gradient algorithm) are given in Figure 3b–e, respectively. These displays, as well as that of the histogram in Figure 3a, were created by linearly mapping each reconstruction $\hat{\boldsymbol{\lambda}}$ onto a linear gray scale from 0 (white) to 127 (black) and then, to enhance details, remapping all the gray levels in the range 0–30 into white (0) and all the gray levels in the range 70–127 into black (127). We note that the LS optimization would typically result in some of the $\bar{\lambda}(b)$'s being negative. In the reconstruction of Figure 3e these values were set equal to zero at the end of the 32nd iteration. This problem does not occur in the ML reconstruction, because in each iteration we have $\hat{\lambda}^{\text{new}}(b) \geq 0$ for all b's. Furthermore, the LS reconstruction introduces some "noise" outside the boundary of the phantom (the "skull"). As the number of EM iterations increases, the ML reconstruction becomes more oscillatory and the checkerboard display (within homogeneous regions of the phantom), which characterizes the LS reconstruction, starts to show also in the ML reconstruction. This phenomenon becomes even stronger as the number of EM iterations goes beyond 64, and it suggests that some smoothing might be desirable for the final EM reconstruction. Alternatively, one can start the algorithm with a uniform $\boldsymbol{\lambda}^0$ and then control the degree of smoothness of the reconstruction by controlling the number of EM iterations. Nevertheless, the resolution of the ML reconstruction and its ability to pick up details such as the low density at the "skull" [$\lambda(x, y) = .1$ in this region] seem better than for the LS reconstruction, as is evident from the line plots of Figure 4. To check the accuracy of the preceding reconstructions, we drew line plots through the y axis of the histogram and each of the reconstructed images (i.e., a cut from bottom to top through the center of each of Figure 3a–e). These line plots are given in Figure 4a–e, respectively. Both the photodisplays and the line plots suggest that the EM reconstruction is preferable to the LS reconstruction. This is further emphasized by Figure 5, which is made up of a superposition of Figure 4 a and d.

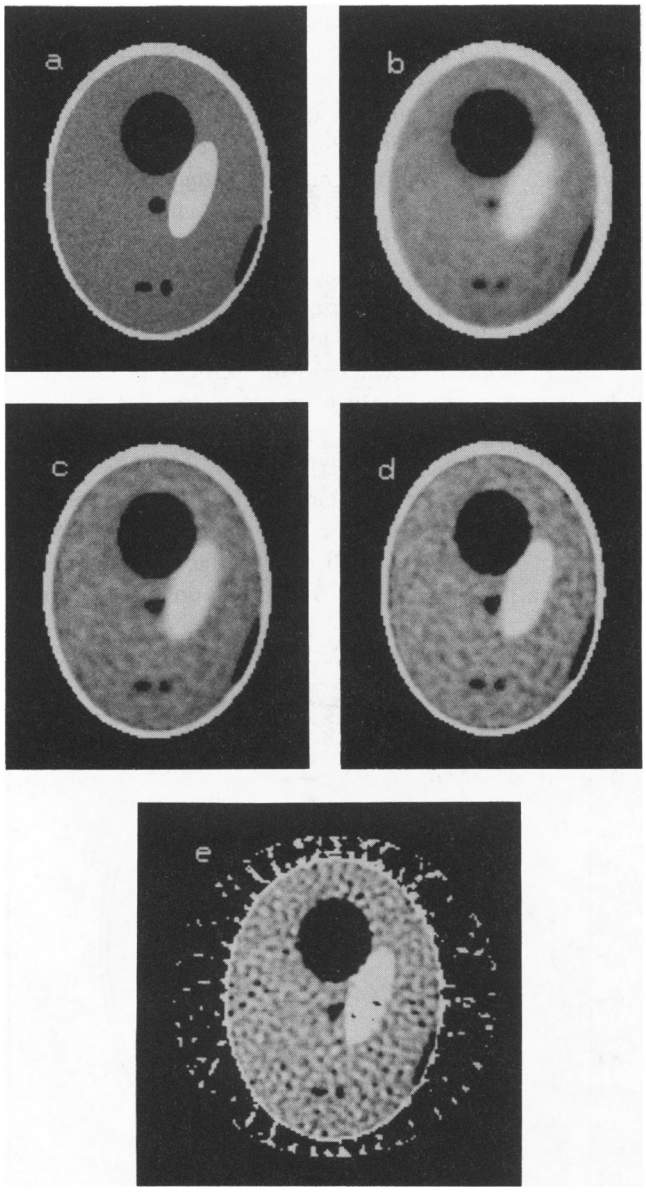

Figure 3. (a) The histogram n(b), b = 1, . . . , B (= 128²), of the 10⁷ counts drawn from the phantom of Figure 2 at a rate proportional to λ(x, y) at each point. (b), (c), and (d) The reconstruction using the EM algorithm after 16, 32, and 64 iterations, respectively. (e) Reconstruction using conjugate gradient approximation to least squares, after 32 iterations.

4. SOME CONCLUDING REMARKS

In Shepp and Vardi (1982), the CBP method was chosen for comparison with the MLE because the CBP is the technique presently in use in PET reconstructions. In Shepp et al. (1984), a further comparison between these two methods is carried out with real PET data, rather than computer-generated data. In this article, we chose to compare the MLE with the LSE because of the wide use of the latter method in statistical practice. All comparisons indicate that the MLE gives a better reconstruction. Based on some further experiments with nonlinear smoothers applied to the ML reconstruction, we feel that either a slightly smoothed version of the MLE or, alternatively, an

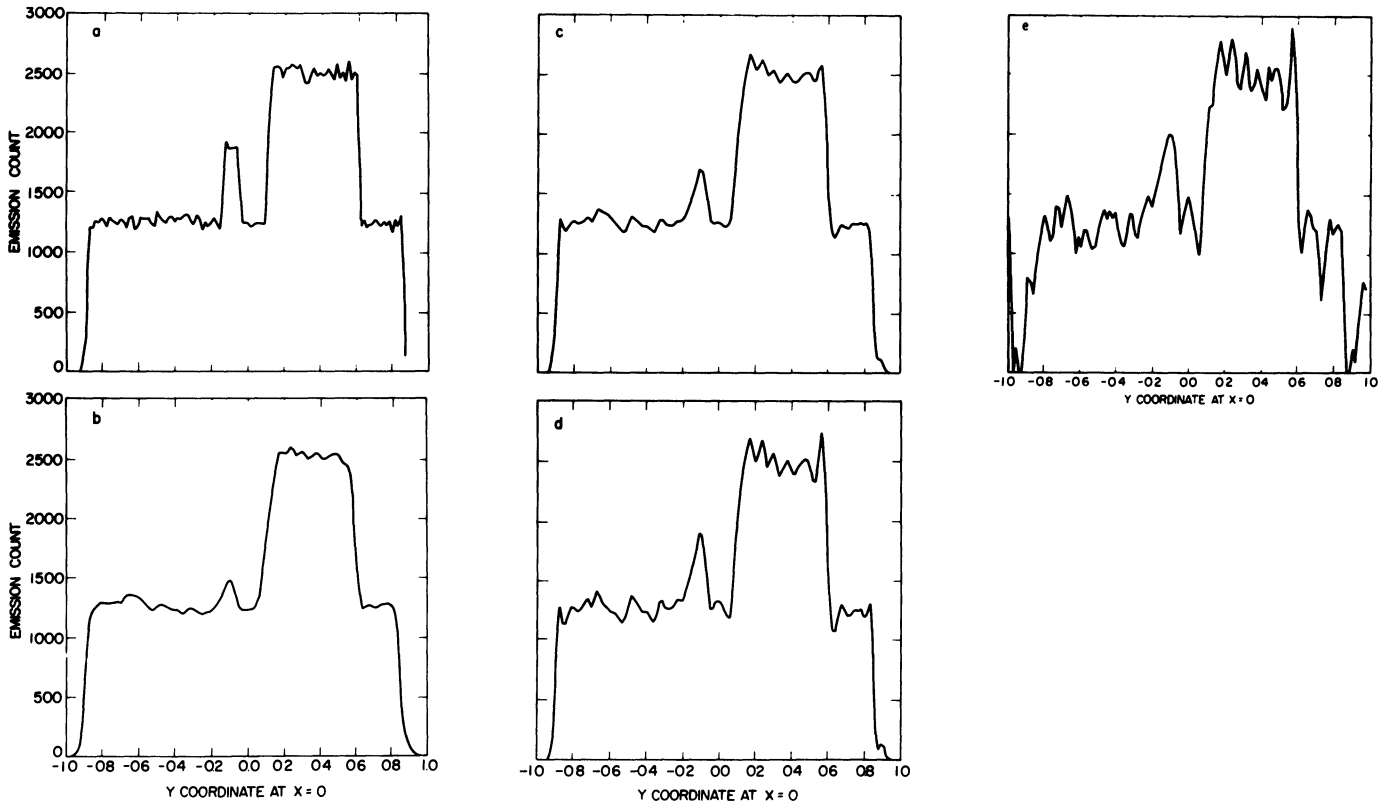

Figure 4. (a–e) Line plots through y axis of the photodisplays in Figure 3 a–e, respectively.

EM reconstruction that starts with a uniform λ^{old} and is run for a limited number of iterations (our experience suggests about 50 iterations) gives very good reconstructions. As is apparent from our discussion of Stein-type, Bayes, and other estimates, there are other techniques worth trying, and we hope that our article will stimulate further research in this direction. In pursuing such techniques, however, we hope that statisticians will not stop after proving that "method A" has a uniformly lower expected loss when compared with "method B" but will, rather, continue to compare *pictures*. It is conceivable that the loss function used for comparing method A with method B is not a good mathematical formulation for our visual perception. Indeed, the question of what *is* a good mathematical formulation for visual perception (either monochromatic or polychromatic) is an interesting line of research to pursue.

We have said very little, so far, about the computational aspects of the EM reconstruction that we described. Although some PET centers are now experimenting with implementing it, the computational complexity and running time are still bigger, by an order of magnitude, than those of the CBP. We feel that further research should go into making the EM algorithm and subsequent variants more computationally efficient. Presently, for the example of Section 3, it takes our CRAY 1 computer about 9 seconds to execute the first EM iteration and about .8 second for each subsequent iteration. Note that in the first iteration, we compute all of the $p(b, d)$'s and store all of the nonzero ones—approximately 240,000—whereas in subsequent iterations the $p(b, d)$'s are called from core memory. To be able to implement the algorithm on a VAX 780, we made an attempt to reduce the memory requirements and running time. By using a circular display grid that takes advantage of the circular symmetry of the detector ring, we cut the memory requirement by a factor of 8. The running time on our CRAY 1 for this setup is about 1 second for the first iteration [because the number of distinct nonzero $p(b, d)$'s that have to be computed is now approximately 30,000] and about .95 second for subsequent iterations. For the VAX 780 these running times should be multiplied by a factor of more than 50, so the program is still slow but perhaps within acceptable

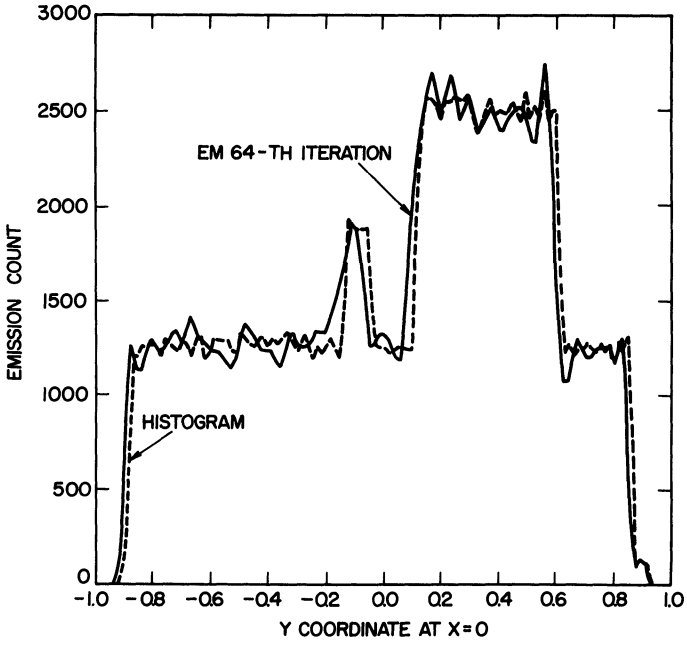

Figure 5. Superposition of Figure 4 a and d, which compares the EM reconstruction with the original histogram.

limits. Further advances in computer technology and further software improvements should make the EM reconstruction more practically appealing.

Because it is a growing discipline within the field of nuclear medicine, PET has largely been developed in or in affiliation with large medical centers; it attracts financial resources and scientists from various fields. Many of these medical centers have biostatistics departments that can open the door for more statistical involvement in PET projects, a situation that will benefit both disciplines. The reconstruction problem is only one of the areas of PET in which statistical methodology can make an important contribution; there are many others. For instance, PET scanning is a valuable research tool in diseases that are difficult to study, such as aging, schizophrenia, manic–depressive illness, and various heart diseases. Finding good statistical models that correlate the images obtained from PET scanning with symptoms of such diseases is a major challenge that would increase the value of PET scanning as a diagnostic tool. Such models, if developed, will be based on medical profiles of hundreds of patients, and statistical expertise will be needed in their analysis. Another example that is statistical in nature involves PET scanners that measure time-of-flight information (see Sec. 1.3) and their instrumentation. Here the problem (very simplified and in a nutshell) is that the scintillation detectors used in connection with time-of-flight tomography are inherently less efficient than those used in scanners that do not measure time-of-flight information, so the merit of time-of-flight tomography depends on the trade-off between time-of-flight information and coincidence-count data, a trade-off that is hard to measure. A figure of merit for this trade-off is an interesting statistical problem that will help in evaluating the usefulness of time-of-flight tomography.

Since the whole procedure of PET scanning—starting with the metabolism itself, continuing with the emission and annihilation processes, and moving to the traveling and detection of photons—involves a stochastic behavior at each stage, we can point to numerous additional examples of statistical problems arising in PET. In most of these problems, however, the statistical aspect is only one of many different aspects, and separating it from its full context involves the risk of posing the wrong questions and misrepresenting the subject matter. We therefore refer the reader to the review articles by Brownell et al. (1982) and Ter-Pogossian et al. (1980) for further discussion, including past development, current status, and future directions of PET.

APPENDIX: AN OUTLINE OF THE PROOF OF CONVERGENCE OF (2.13)

Let $p_{b,d} \geq 0$ for $b = 1, \ldots, B$, $d = 1, \ldots, D$, $p_{b+} \equiv \sum_d p_{b,d} = 1$ for all b, and $n_d^* \geq 0$ for all d, and without loss of generality, normalize to have $\sum_d n_d^* = 1$. Suppose $\lambda_b^{(0)} > 0$ for all b and define

$$\lambda_b^{(k+1)} \equiv \sum_d \lambda_b^{(k)} p_{b,d}\, n_d^* / \lambda_d^{(k)*}, \tag{A.1}$$

where the superscript * is defined by

$$\lambda_d^{(k)*} \equiv \sum_b \lambda_b^{(k)} p_{b,d}, \qquad d = 1, \ldots, D.$$

Note that (A.1) is the same as (2.13), so it never leads to a quotient of a positive numerator divided by a zero denominator; zero divided by zero is defined as zero.

Theorem A.1 (Csiszàr and Tusnàdy 1982). The sequence $\boldsymbol{\lambda}^{(k)}$, $k = 0, 1, \ldots$, converges to a limit point $\hat{\boldsymbol{\lambda}}$, which maximizes $l(\boldsymbol{\lambda})$.

Although Csiszàr and Tusnàdy's proof is self-contained, the proof we outline below refers to some arguments from Cover (1984) and Shepp and Vardi (1982). This is done so that we could shorten the presentation and highlight the key ideas of Csiszàr and Tusnàdy's clever and deep geometric argument (Lemma A.1).

An Outline of Proof for Theorem A.1. The proof breaks into two parts:

1. Showing that if $\boldsymbol{\lambda}^{(k)}$, $k = 0, 1, \ldots$, converges, then the limit point, $\hat{\boldsymbol{\lambda}}$ say, maximizes $l(\boldsymbol{\lambda})$.
2. Showing that the sequence $\boldsymbol{\lambda}^{(k)}$, $k = 0, 1, \ldots$, converges to a unique limit point.

Part 1 is proved by showing that such a limit point $\hat{\boldsymbol{\lambda}}$ must satisfy the KT conditions (2.6) and (2.7), and hence it is a point of maximum. The details of the argument are given in Cover (1984, Theorem 3) and in Shepp and Vardi (1982). To continue with the proof let $d(\ ,\ )$ be the Kullback Leibler information divergence. That is, for any two probability measures $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_B)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_B)$, we define

$$d(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_b \mu_b \log \frac{\mu_b}{\nu_b}.$$

N.B.: $d(\ ,\ )$ is continuous nonnegative, and

$$d(\boldsymbol{\mu}, \boldsymbol{\nu}) = 0 \text{ iff } \boldsymbol{\mu} \equiv \boldsymbol{\nu}. \tag{A.2}$$

Csiszàr and Tusnàdy's key to proving part 2 is the following lemma.

Lemma A.1. If $\boldsymbol{\lambda}$ is a limit point of the sequence $\boldsymbol{\lambda}^{(k)}$, $k = 0, 1, \ldots$, then

$$d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k+1)}) \leq d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k)}), \qquad k = 0, 1, \ldots. \tag{A.3}$$

Using the lemma, part 2 is proved as follows: Because of compactness, there exists a convergent subsequence $\boldsymbol{\lambda}^{(k')}$ with a limit point $\bar{\boldsymbol{\lambda}}$, and because of (A.2), we get $d(\bar{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{(k')}) \to 0$. The lemma now implies that $d(\bar{\boldsymbol{\lambda}}, \boldsymbol{\lambda}^{(k)}) \to 0$ for the full sequence $\boldsymbol{\lambda}^{(k)}$, and again, because of (A.2) $\boldsymbol{\lambda}^{(k)} \to \bar{\boldsymbol{\lambda}}$. This proves part 2.

It remains to prove the lemma. Consider the set $\mathcal{C}$ of all nonnegative matrices $c_{b,d} \geq 0$ such that $\sum_{bd} c_{b,d} = 1$, and let

$$\mathcal{Q} = \{c \in \mathcal{C};\ c_{b,d} = \lambda_b p_{b,d}, \text{ where } \lambda_b \geq 0 \text{ and } \sum \lambda_b = 1\}$$

$$\mathcal{P} = \{c \in \mathcal{C};\ c_{+d} \equiv \sum_b c_{b,d} = n_d^*\}.$$

For any probability measure $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_b)$, define

$$q(\boldsymbol{\lambda}) \in \mathcal{Q} \quad \text{by} \quad q(\boldsymbol{\lambda})_{b,d} \equiv \lambda_b p_{b,d} \tag{A.4}$$

and

$$\pi(\boldsymbol{\lambda}) \in \mathcal{P} \quad \text{by} \quad \pi(\boldsymbol{\lambda})_{b,d} \equiv \lambda_b p_{b,d} n_d^* / \lambda_d^*. \tag{A.5}$$

[Note that if $\lambda_d^* = 0$, then the numerator is also zero, in which case we define $\pi(\boldsymbol{\lambda})_{b,d} = 0$.] We further define

$$D(\pi, q) = \sum_{b,d} \pi_{b,d} \log \frac{\pi_{b,d}}{q_{b,d}} \tag{A.6}$$

for any $\pi, q \in \mathcal{C}$.

Proof of Lemma A.1. Denote for each k,

$$q^k \equiv q(\boldsymbol{\lambda}^{(k)}), \qquad \pi^{k+1} \equiv \pi(\boldsymbol{\lambda}^{(k)}).$$

Proposition. If $\boldsymbol{\lambda}$ is a limit point of $\boldsymbol{\lambda}^{(k)}$, $k = 0, 1, \ldots,$ then the following two inequalities hold:

$$d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k+1)}) \leq D(\pi(\boldsymbol{\lambda}), \pi^{k+1}), \tag{A.7}$$

$$D(\pi(\boldsymbol{\lambda}), \pi^{k+1}) + D(\pi^{k+1}, q^k) \leq D(\pi(\boldsymbol{\lambda}), q^k). \tag{A.8}$$

Note that (A.8) is Csiszàr and Tusnàdy's (1982) "3 points property."

Using these two inequalities, we have for any $\boldsymbol{\lambda}$ that is a limit point of the sequence $\boldsymbol{\lambda}^{(k)}$, $k = 0, 1, \ldots,$

$$d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k+1)}) \leq D(\pi(\boldsymbol{\lambda}), \pi^{k+1}) \leq D(\pi(\boldsymbol{\lambda}), q^k) - D(\pi^{k+1}, q^k)$$

$$= \sum_{b,d} \pi(\boldsymbol{\lambda})_{b,d} \log \frac{\lambda_b n_d^*}{\lambda_b^{(k)} \lambda_d^*} - \sum_{b,d} \pi(\boldsymbol{\lambda}^{(k)})_{b,d} \log \frac{n_d^*}{\boldsymbol{\lambda}_d^{(k)*}}. \tag{A.9}$$

Now from Cover (1984, Theorem 3), if $\boldsymbol{\lambda}$ is a limit point of $\{\boldsymbol{\lambda}^{(k)}, k = 0, 1, \ldots\}$, then it is also a fixed point of the algorithm, and therefore

$$\sum_d \pi(\boldsymbol{\lambda})_{b,d} = \lambda_b. \tag{A.10}$$

Applying this to (A.9) we get:

rightmost side of (A.9)

$$= \sum_b \lambda_b \log \frac{\lambda_b}{\lambda_b^{(k)}} + \sum_d n_d^* \log \frac{n_d^*}{\lambda_d^*} - \sum_d n_d^* \log \frac{n_d^*}{\lambda_d^{(k)*}}$$

$$= d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k)}) + \sum_d n_d^* \log \frac{\lambda_d^{(k)*}}{\lambda_d^*} \leq d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k)}). \tag{A.11}$$

Note that the last inequality follows from the monotonicity of the likelihood and the fact that $\boldsymbol{\lambda}$ is a limit point of $\boldsymbol{\lambda}^{(k)}$, so

$$\sum_d n_d^* \log \lambda_d^{(k)*} \leq \sum_d n_d^* \log \lambda_d^*.$$

Combining (A.9) and (A.11), the lemma is proved.

It remains to prove (A.7) and (A.8).

Proof of (A.7). Being a limit point of $\boldsymbol{\lambda}^{(k)}$, $\boldsymbol{\lambda}$ is also a fixed point of the algorithm (Cover 1984, Theorem 3), so

$$\pi(\boldsymbol{\lambda})_{b+} \equiv \sum_d \pi(\boldsymbol{\lambda})_{b,d} = \lambda_b.$$

Since

$$\pi_{b+}^{k+1} \equiv \sum_d \pi_{b,d}^{k+1} = \lambda_b^{(k+1)},$$

we have

$$D(\pi(\boldsymbol{\lambda}), \pi^{k+1}) = \sum_{b,d} \pi(\boldsymbol{\lambda})_{b+} \pi(\boldsymbol{\lambda})_{d|b} \log \frac{\pi(\boldsymbol{\lambda})_{b+} \pi(\boldsymbol{\lambda})_{d|b}}{\pi_{b+}^{k+1} \pi_{d|b}^{k+1}}$$

$$\overset{\text{(from A.2)}}{\geq} \sum_b \pi(\boldsymbol{\lambda})_{b+} \log \frac{\pi(\boldsymbol{\lambda})_{b+}}{\pi_{b+}^{k+1}} = \sum_b \lambda_b \log \frac{\lambda_b}{\lambda_b^{(k+1)}}$$

$$= d(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k+1)}),$$

as desired. Here and in what follows, we use the subscript notation $d|b$ for conditional probability, so $\pi_{d|b} \equiv \pi_{b,d}/\pi_{b+}$.

Proof of (A.8). First observe that π^{k+1} is the minimizer of $D(c, q^k)$ over all $c \in \mathcal{P}$. This is seen as follows:

$$D(c, q^k) = \sum_d c_{+d} \log \frac{c_{+d}}{q_{+d}^k} + \sum_d c_{+d} \sum_d c_{b|d} \log \frac{c_{b|d}}{q_{b|d}^k}$$

$$\geq \sum_d c_{+d} \log \frac{c_{+d}}{q_{+d}^k}$$

with equality iff $c_{b|d} = q_{b|d}^k$. Now $q_{b|d}^k = \lambda_b^{(k)} p_{b,d}/\lambda_d^{(k)*}$, and since $c \in \mathcal{P}$, $c_{+d} = n_d^*$; so the minimum is achieved for

$$c_{b,d} \equiv c_{b|d} c_{+d} = q_{b|d}^k n^* = \pi(\boldsymbol{\lambda}^{(k)})_{b,d} = \pi_{b,d}^{k+1}.$$

Next, let $f(t) = D(t\pi^{k+1} + (1 - t)\pi(\boldsymbol{\lambda}), q^k)$. Then because π^{k+1} minimizes $D(c, q^k)$ for $c \in \mathcal{P}$, we have $f(1) = \min_{0 \leq t \leq 1} f(t)$; so

$$O \geq f'(1) = \sum_{b,d} (\pi_{b,d}^{k+1} - \pi(\boldsymbol{\lambda})_{b,d}) \log \frac{\pi_{b,d}^{k+1}}{q_{b,d}^k} + \sum_{b,d} (\pi_{b,d}^{k+1} - \pi(\boldsymbol{\lambda})_{b,d})$$

$$= D(\pi^{k+1}, q^k) + D(\pi(\boldsymbol{\lambda}), \pi^{k+1}) - D(\pi(\boldsymbol{\lambda}), q^k) + \sum_{b,d} (\pi_{b,d}^{k+1} - \pi(\boldsymbol{\lambda})_{b,d}).$$

The last sum, however, is zero because $\pi^{k+1}, \pi(\boldsymbol{\lambda}) \in \mathcal{C}$; so (A.8) is proved. This completes the proof of the theorem.

[*Received September 1982. Revised February 1984.*]

REFERENCES

Brownell, G. L., Budinger, T. F., Lauterbur, P. C., and McGeer, P. L. (1982), "Positron Tomography and Nuclear Magnetic Resonance Imaging," *Science*, 215, 619–626.

Clevenson, L., and Zidek, J. V. (1975), "Simultaneous Estimation of the Means of Independent Poisson Laws," *Journal of the American Statistical Association*, 70, 698–705.

Cover, T. M. (1984), "An Algorithm for Maximizing Expected Log Investment Return," *IEEE Transactions on Information Theory*, IT-30, 369–373.

Csiszàr, I., and Tusnàdy, G. (1982), "Information Geometry and Alternating Minimization Procedures," Technical Report, Mathematical Institute of the Hungarian Academy of Sciences (also to appear in *Statistics and Decisions*).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Derenzo, S. E., Banchers, P. G., Cahoon, J. L., Huesman, R. H., Vuletich, T., and Budinger, T. F. (1977), "Design and Construction of the Donner 280-Crystal Positron Ring for Dynamic Transverse Section Emission Imaging," *Proceedings of the IEEE Conference on Decision and Control*, Publication 77CH1269-0CS, Silver Spring, MD: IEEE Computer Society Press.

Fortes, J. M. P. (1980), "An Estimation Approach to 3-D Reconstruction Problems Involving Counting Statistics With Applications to Medical Imaging," unpublished Ph.D. dissertation, Stanford University, Dept. of Electrical Engineering.

Good, I. J., and Gaskins, R. A. (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and

Meteorite Data" (with discussion), *Journal of the American Statistical Association*, 75, 42–70.
Hudson, H. M. (1978), "A Natural Identity for Exponential Families With Applications in Multiparameter Estimation," *Annals of Statistics*, 6, 473–484.
Lakshminarayanan, A. V. (1975), "Reconstruction From Divergent Ray Data," Technical Report No. 92, State University of New York at Buffalo, Computer Science Dept.
Lange, K., and Carson, R. (1984), "EM Reconstruction Algorithms for Emission and Transmission Tomography," *Journal of Computer Assisted Tomography*, 8, 302–316.
Lindley, D. V. (1965), *Introduction to Probability and Statistics From a Bayesian Viewpoint. Part 2: Inference*, Cambridge, U.K.: Cambridge University Press.
Ortega, J. M., and Rheinboldt, W. C. (1970), *Iterative Solutions of Nonlinear Equations in Several Variables*, New York: Academic Press.
Peng, J. C. M. (1975), "Simultaneous Estimation of the Parameters of Independent Poisson Distributions," Technical Report No. 78, Stanford University, Dept. of Statistics.
Ramachandran, G. N., and Lakshminarayanan, A. V. (1971), "Three Dimensional Reconstruction From Radiographs and Electron Micrographs: Application of Convolutions Instead of Fourier Transform," *Proceedings of the National Academy of Sciences, U.S.A.*, 68, 2236–2240.
Shepp, L. A., and Kruskal, J. B. (1978), "Computerized Tomography: The New Medical X-ray Technology," *American Mathematical Monthly*, 85, 420–439.
Shepp, L. A., and Logan, B. F. (1974), "The Fourier Reconstruction of a Head Section," *IEEE Transactions on Nuclear Science*, NS-21, 21–43.
Shepp, L. A., and Vardi, Y. (1982), "Maximum Likelihood Reconstruction in Positron Emission Tomography," *IEEE Transactions on Medical Imaging*, 1, 113–122.
Shepp, L. A., Vardi, Y., Ra, J. B., Hilal, S. K., and Cho, Z. H. (1984), "Maximum Likelihood PET With Real Data," *IEEE Transactions on Nuclear Science*, NS-31, 910–913.
Snyder, D. L., Thomas, L. J., Jr., and Ter-Pogossian, M. M. (1981), "A Mathematical Model for Positron-Emission Tomography Systems Having Time-of-Flight Measurements," *IEEE Transactions on Nuclear Science*, NS-28, 3575–3583.
Ter-Pogossian, M. M., Raichle, M. E., and Sobel, B. E. (1980), "Positron Emission Tomography," *Scientific American*, 243 (4), 170–181.
Tusi, K. W., and Press, S. J. (1982), "Simultaneous Estimation of Several Poisson Parameters Under k-Normalized Squared Error Loss," *Annals of Statistics*, 10, 93–100.
Walker, A. M. (1969), "On the Asymptotic Behaviour of Posterior Distributions," *Journal of the Royal Statistical Society*, Ser. B, 31, 80–88.
Zangwill, W. I. (1969), *Nonlinear Programming: A Unified Approach*, Englewood Cliffs, NJ: Prentice-Hall.

Comment

The EM Parametric Image Reconstruction Algorithm

RICHARD E. CARSON and KENNETH LANGE*

Positron emission tomography provides the capability to quantitatively measure the local concentration of a variety of radionuclides in humans and animals. This high-quality data can be used to elucidate the subtleties of physiology and biochemistry in normal and diseased states. Accurate interpretation of the data depends upon quantitative image reconstruction methods combined with the techniques of tracer kinetic modeling.

The EM algorithm for image reconstruction in positron emission tomography was first presented by Shepp and Vardi (1982) and further developed in the preceding article here. Independently, Lange and Carson (1984) developed EM algorithms for image reconstruction for emission and transmission tomography. The definitions from the latter paper are as follows:

i—projection line index
j—pixel index
Y_i—counts collected along projection line i (random variable)
λ_j—emission rate of pixel j (unknown parameters)
c_{ij}—probability of an emission from pixel j being detected along projection line i (times counting time)
I_i—the set of pixels that contribute to projection line i
J_j—the set of projection lines that contribute to pixel j

The physical model for the tomographic projection measurements is

$$Y_i \sim \text{Poisson}\left(\sum_{j \in I_i} c_{ij}\lambda_j\right). \tag{1}$$

The c_{ij} are assumed to be known exactly and depend upon the physical factors affecting tomographic projection measurements: radioactive decay, projection sampling scheme, detector efficiency and location, spatial resolution, attenuation, scatter, accidental coincidences, positron range, and angulation. The use of a physical model for the observations provides the EM algorithm with a different philosophy than Fourier reconstruction techniques (see Shepp and Logan 1974)—that is, to fit the algorithm to the data, rather than "correcting" the data to suit the assumptions of the algorithm.

To use the EM algorithm (Dempster et al. 1977) to develop a maximum likelihood estimate of λ, the observations Y with likelihood function $g(y \mid \lambda)$ must be viewed as incomplete data embedded in a complete data space X with likelihood function $f(x \mid \lambda)$. For emission tomography, define the complete data X_{ij} to be the random number of emissions from pixel j that are detected along projection line i. The log likelihood of the com-

* Richard E. Carson is in the Department of Nuclear Medicine, National Institutes of Health, Bethesda, MD 20205. Kenneth Lange is in the Department of Biomathematics, UCLA, Los Angeles, CA 90024.

Journal of the American Statistical Association
March 1985, Vol. 80, No. 389, Applications