BAYESIAN IMAGE ANALYSIS: AN APPLICATION TO SINGLE PHOTON EMISSION TOMOGRAPHY

Stuart Geman and Donald E. McClure, Brown University

1. INTRODUCTION

Image analysis embraces a wide range of image processing tasks, including blur and noise removal, segmentation and boundary finding, and object location and identification. Traditional approaches have addressed these problems on a largely individual basis, defining new theoretical frameworks and computational algorithms for each task. We hope to achieve a unification through a *Bayesian framework*. The idea is to articulate, in the form of a *prior probability distribution*, knowledge and expectations about the spatial smoothness of pixel intensities, the straightness or constant curvature of edges, and other such attributes of real-world images. In this set-up image analysis is by maximization, over all possible assignments of pixel intensities, locations of edges, and so on, of a *posterior distribution*, given an observed (possibly degraded) image. This program is layed out in detail in Geman & Geman [4] and Grenander [6], together with the descriptions of numerous experiments.

This paper is about an application of the approach to single photon emission tomography (SPET). The problem is to reconstruct a two or three dimensional profile of isotope intensity from a series of (respectively) one or two dimensional observations of photon emissions. The prior distribution will be on the set of possible isotope intensities. Through this distribution we will attempt to harness some very simple prior knowledge: nearby locations tend to have similar intensity levels. We shall present the results of simulation experiments in which the Bayesian restoration is compared to the maximum likelihood restoration, the latter determined by an implementation of the EM algorithm.

The next section (S.2) will provide a brief introduction to SPET, together with a mathematical formulation of the reconstruction problem. In S.3 we will examine the maximum likelihood solution and its computation by the EM algorithm. In S.4 we will turn to the Bayesian approach, and construct a simple prior distribution for this problem. The posterior distribution will be derived in S.5, and in S.6 we will examine reconstructions obtained by approximately maximizing the posterior distribution in some simulation experiments. We will conclude (in S.7) with a discussion of parameter estimation for the prior distribution.

This paper is intended as an introduction. A more complete discussion of our approach to the SPET problem will be provided in a following paper.

2. SINGLE PHOTON EMISSION TOMOGRAPHY

Single photon emission tomography is used to determine the distribution of a pharmaceutical (injected or inhaled compound) in a part of the body such as brain, liver, or heart. Depending upon the pharmaceutical used, this concentration can be taken as a measure of local blood flow ("perfusion") and/or local metabolic activity. Glucose, for example, is taken up by neuronal cells in proportion to metabolic activity, and the latter generally mirrors recent electrical activity. Thus areas of brain most used in performing a cognitive or motor task will demonstrate a relatively increased uptake of glucose immediately following the task. For the heart, pharmaceuticals can be chosen whose uptake reflects local perfusion. The concentration of these pharmaceuticals can thereby be used to assess the adequacy of blood flow to the different parts of the heart.

In SPET, pharmaceutical concentration is estimated by detecting photon emissions from an injected or inhaled quantity of the pharmaceutical that has been chemically combined with a radioactive isotope. This combined molecule is called a radiopharmaceutical. The goal of SPET is to determine radiopharmaceutical concentration (equivalently, isotope concentration) as a function of position in a target tissue, such as brain, heart, or liver. Detectors with collimators are strategically placed around the appropriate region of the body, and these are able to count photons emitted by the isotope. A detector will capture those photons which avoid attenuation, and whose trajectories carry them down the barrel of the collimator.

The determination of isotope concentration from photon counts is called *reconstruction*. The reconstructions that we develop are based on (i) a well-defined mathematical model relating isotope concentration to the observable photon counts, and (ii) classical principles of statistical inference which view the concentration as a "parameter" to be estimated.

Let $X(s)$ denote the concentration of the radiopharmaceutical at the point $s=(x,y)$ in the domain Ω of interest. For the present discussion, we shall take Ω to be a compact two-dimensional region, though for the models and methods we will describe there are no fundamental changes when Ω is three-dimensional.

We will assume that the detectors are arranged in a linear array, at equally spaced lateral sampling intervals, and that the detector array can be positioned at any orientation θ relative to the x-axis. (See Figure 1.) The detectors are assumed to be of so-called parallel-bore type, meaning that they detect only those photons in the small interval $[\theta - \frac{\Delta\theta}{2}, \theta + \frac{\Delta\theta}{2}]$ when the array has orientation θ.

Let L denote the total number of detectors in the array and let $\Delta\sigma$ denote the spacing between detectors.

The physical effects incorporated in the model are the spatial Poisson process that describes the sites of the radioactive decays from which photons emanate and the process of photon attenuation by which photons are annihilated and their energy is absorbed by matter through which their trajectories pass. Attenuation is accurately described by a linear attenuation function $\mu(s)$ on Ω. The function $\mu(\cdot)$ is assumed to be known; indeed, it can be measured by transmission tomographic methods. Attenuation is a memoryless process; consequently we can deduce the functional form of the probability that a photon survives to reach the detector array. When a photon trajectory has direction θ and it emanates from site $s = (x,y)$ in Ω, then

Reprinted from the 1985 Statistical Computing Section
Proceedings of the American Statistical Association

$$\text{P(photon survival)} = \exp\{-\int_{L(x,y)} \mu(\xi,\eta)d\ell\},$$

where the line integral is taken over the segment $L(x,y)$ from (x,y) to the detector and $d\ell$ is differential arc length (see figure 1).

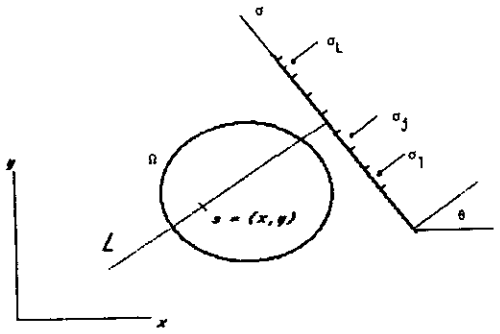


Figure 1

For our sampling design, we shall position the detector array at n equally-spaced angles θ_k for duration T time units at each angle. Then at each angle, we observe the random variables Y(t), for $t\epsilon D_k = \{(\sigma_j,\theta_k),\ j=1,\ldots,L\}$ that give the numbers of photons reaching the respective detectors during the sampling interval. Assuming that (i) photons are generated by a nonhomogeneous Poisson process with intensity X(s) per time unit, and (ii) the orientations θ of photon trajectories are uniformly distributed on $[0,2\pi)$, we can show that Y(t), for $t\epsilon D = \bigcup_{k=1}^{n} D_k$, is a Poisson process with a nonhomogeneous intensity function described in terms of the attenuated Radon transform (ART) of X(·). The ART of X(·) is defined as

$$(R_{\mu,T}X)(\sigma,\theta) = \int_L TX(x,y)\exp\{-\int_{L(x,y)} \mu(\xi,\eta)d\ell'\}d\ell$$

where L is the line with orientation θ, through point σ, of the detector array, $L(x,y)$ is the segment of L starting at point (x,y) in Ω, and $d\ell$ and $d\ell'$ are differential arc length in the two line integrals. The intensity function of Y(·) is then given by

$$EY(t) = \int_{\theta_k-\frac{\Delta\theta}{2}}^{\theta_k+\frac{\Delta\theta}{2}} \int_{\sigma_j-\frac{\Delta\sigma}{2}}^{\sigma_j+\frac{\Delta\sigma}{2}} (R_{\mu,T}X)(\sigma,\theta)d\sigma d\theta,$$

where $t = (\sigma_j,\theta_k)$.

The important feature of this representation is that the intensity function of Y is the result of a positive linear integral operator A_T applied to X:

$$EY = A_T X. \tag{2.1}$$

This model includes the predominant physical effects. Other effects, such as photon scattering and background radiation, are assumed for now to be negligible. We observe, however, that the reconstruction methods described below, since they are based on the generally applicable principles of maximum likelihood and Bayes optimality, are readily adaptable to models incorporating additional effects.

3. MAXIMUM LIKELIHOOD AND EM

A variety of reconstruction algorithms for emission tomography are described by Budinger et.al. [2]. The algorithms that are traditionally used are based on ideas of extracting a signal in the presence of noise and related methods of linear filtering.

More recently, attention has been given to algorithms that use fuller information of the mathematical model sketched above along with the principle of maximum likelihood (ML). Shepp and Vardi [8] laid the mathematical foundations and developed effective algorithms based on EM (Dempster, Laird and Rubin [3]) for implementing ML reconstructions in positron emission tomography; in positron emission tomography, attenuation does not corrupt the model relating the radiopharmaceutical concentration to the observables. McClure and Accomando [1] have developed the foundations for applying ML to SPET reconstructions and have implemented EM algorithms on a variety of computer systems.

The likelihood function is readily expressed since the observables are mutually independent and have distributions of known form. To carry out a ML reconstruction, we first discretize the domain Ω into pixels parameterized by discrete points s in a square lattice S. Now $(X(s))_{s\epsilon S}$ represents a piecewise constant approximation of the isotope concentration on the continuuous domain. (One of the underlying theoretical questions is to understand how the quality of the reconstruction depends on the resolution of the domain Ω enforced by its discretization.) When Ω is discretized, then equation (2.1) takes the form

$$EY = \mathbf{a}_T X,$$

where $\mathbf{a}_T$ is a matrix $\mathbf{a}_T = (a(t,s))_{t\epsilon D, s\epsilon S}$; typically, the order of $\mathbf{a}_T$ is enormous.

The log-likelihood function is

$$\ln L(X) = \sum_{t\epsilon D} [-\ln(Y(t)!) + Y(t)\ln[(\mathbf{a}_T X)(t)] - (\mathbf{a}_T X)(t)].$$

The necessary conditions for maximizing ln L(X) obtained by setting derivatives to zero do not yield explicit solutions for the maximizing X. Nonetheless, -ln L(X) is globally convex, and the ML optimization problem conveniently adapts to the EM method. (In general, -ln L(X) is not strictly convex; conditions for strict convexity are analyzed by Accomando [1].)

The EM algorithm leads to an iterative reconstruction procedure. We initialize the iteration with $(X_0(s))_{s\epsilon S}$ and update X_i by the formula

$$X_{i+1} = \{[\mathbf{a}_T'(Y \oslash \mathbf{a}_T X_i)] \oslash \mathbf{a}_T'\mathbf{1}\} \otimes X_i,$$

where $\mathbf{1}$ is the vector whose components are identically one, $\oslash$ denotes component-by-component division, and $\otimes$ denotes the component-by-component product. At each step, the iteration requires two (large) matrix multiplications. The sequence of iterates is proven to converge to an X^* that maximizes ln L(X). Consistency results that depend on the sampling design and on the discretization of Ω can be proved.

4. PRIOR DISTRIBUTION

We turn now to the Bayesian reconstruction of X, based upon an observation Y and the operator A_T. We shall construct a prior distribution that captures some simple prior expectations about the form of the isotope density, X. Mainly, we wish to exploit the anticipated smoothness of X; neighboring locations will typically have similar intensity levels. But we must also accommodate sharp changes in concentration, as may be seen, for example, across an arterial wall, or across a boundary between two tissue types.

In principal, the prior could be constructed on a suitable space of functions $X:\Omega \to R$, where $\Omega \subseteq R$ is a continuous domain. It is much more convenient, however, to specialize to a discrete domain, such as the square lattice used in S.3: $S = \{(i,j): 1 \leq i,j \leq N\}$. In our experiments, N=32. The prior, therefore, is on the array $X=\{X(i,j):1 \leq i,j \leq N\}$, although we will avoid double indices by simply writing $X=\{X(s)\}_{s \in S}$. The components, X(s), will be confined to an interval, such as [0,255] which is customary in digital image processing. As a further convenience, we will restrict ourselves to *Gibbs priors*:

$$(4.1) \qquad \Pi_X(X) = \frac{1}{Z} e^{-U(X)}$$

where Z is the normalizing constant, $Z = \int e^{-U(X)} dX$,

and $U:R^{|S|} \to R$ is known as the "energy". As it stands, this is only mildly restrictive, since U is arbitrary. But we will limit U to be a "nearest neighbor energy", as we shall see shortly.

We use the Gibbs formulation because it is easier to construct an energy function, U, than it is to directly construct a distribution, Π_X. We will design U so that the expected configurations have low energy, as they do in a real physical system. The expected configurations are those for which typical neighboring sites, $s,t \in S$, have similar intensities, X(s), X(t). This is a local constraint, and it is conveniently captured by a locally-composed energy function U:

$$(4.2)\ U(X) = \sum_{<s,t>} \beta\Phi(X(s)-X(t)) + \sum_{\{s,t\}} \frac{\beta}{\sqrt{2}}\Phi(X(s)-X(t)).$$

Here, we use <s,t> to indicate that s and t are nearest horizontal or vertical neighbors in the square lattice, and {s,t} to indicate diagonal neighbors. The constant β is positive and the function $\Phi(\xi)$ is even and minimized by $\xi=0$. Thus U is minimized by configurations of constant intensity. Under the Gibbs distribution, (4.1), the more likely isotope densities are those with small site-to-site variation in intensity.

The exact form for Φ is probably not important, but its qualitative features can make a difference. We have experimented with Φ's that are increasing in ξ for $\xi \geq 0$ (recall that Φ is even): roughly speaking, the more different two neighboring intensities, the less likely. An obvious choice is $\Phi(\xi) = \xi^2$, but then under Π_X large intensity gradients, as would be associated with certain natural boundaries, are exceedingly unlikely. Instead, we will use here functions of the form

$$(4.3) \qquad \Phi(\xi) = \frac{\xi^2}{1 + (\frac{\xi}{\delta})^2}$$

where δ, like β, is a constant which will be fixed later.

Depending upon δ and β, a "typical" sample from Π_X will have large regions of near-constant intensity, with rather discrete changes of intensity defining region boundaries. A sample from Π_X with δ =.7, β=5, and X(s) confined to [0,15] $s \in S$, is shown in figure 2, panel A. The sample was generated by an iterative Monte Carlo technique called Stochastic Relaxation (SR), details of which are layed out in Geman and Geman [4], Geman and Hwang [5], and Grenander [6]. SR is a modification of the well-known Metropolis algorithm (see [7]) for sampling from Gibbs distributions.

5. THE POSTERIOR DISTRIBUTION AND THE MAP ESTIMATOR

Recall that Y(t) is Poisson, with mean the t-component of $\mathbf{a}_T X$:

$$\sum_{s \in S} \alpha(t,s)X(s)$$

where, as before, $\mathbf{a}_T = \{\alpha(t,s)\}_{t \in D, s \in S}$. Given X, the components of Y are independent, and hence their joint distribution is

$$\Pi_{Y|X}(Y) = \prod_{t \in D} \frac{[(\mathbf{a}_T X)(t)]^{Y(t)}}{Y(t)!} e^{-(\mathbf{a}_T X)(t)}$$

$$(5.1) \qquad = \exp\Big(\sum_{t \in D} [-\ln(Y(t)!) + Y(t)\ln[(\mathbf{a}_T X)(t)] - (\mathbf{a}_T X)(t)]\Big)$$

The *posterior distribution*, $\Pi_{X|Y}$, is therefore

$$(5.2) \qquad \frac{1}{Z(Y)} \exp\{-U(X) + \sum_{t \in D} [Y(t)\ln[(\mathbf{a}_T X)(t)] - (\mathbf{a}_T X)(t)]\}$$

where Z(Y) is a normalizing constant and depends on Y.

We will experiment with the *MAP* estimator of X given Y. This is the maximum aposteriori X given Y, which is to say the most likely X under the posterior distribution $\Pi_{X|Y}$. Notice that *MAP* estimation amounts to minimizing

$$(5.3) \qquad U(X) - \sum_{t \in D} [Y(t)\ln[(\mathbf{a}_T X)(t)] - (\mathbf{a}_T X)(t)],$$

which quantity might be called the "posterior energy". Notice also that we have the usual equivalence between Bayesian *MAP* estimation and so-called penalized maximum likelihood. Maximum likelihood maximizes

$$\sum_{t \in D} [Y(t)\ln[(\mathbf{a}_T X)(t)] - (\mathbf{a}_T X)(t)],$$

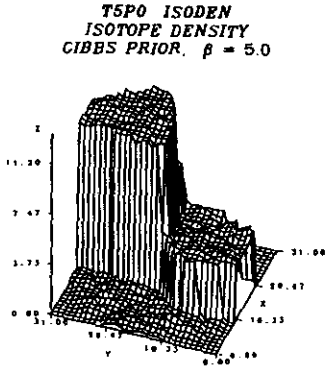
whereas *MAP* estimation includes the "penalty term" -U(X), which penalizes lack of smoothness. One advantage, we believe, of the Bayesian viewpoint is that it suggests mechanisms for *estimating* the required degree of smoothness, which amounts to estimating the pivotal parameter β appearing in the prior (see 4.2). We will return to the estimation problem in S.7.

6. EXPERIMENTS

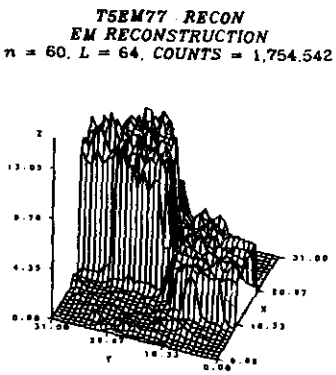
We will present the results of two simulation experiments. In the first, a sample drawn from the prior distribution (4.1), with U as in (4.2) and with β=5 and δ=.7, was used as the isotope intensity function, X. This is shown in figure 2, panel A. Y was generated, Monte Carlo, using the conditional Y-distribution (5.1), given the isotope density X

displayed in 2A, and matrix $\mathbf{a}_T$ resulting from the attenuation function μ displayed in 2D (see S.2). Panel B is the maximum likelihood reconstruction, obtained by an implementation of EM as described in S.3. Maximum likelihood was achieved after 77 iterations of the algorithm. Panel C is an approximate *MAP* reconstruction. This was obtained by a gradient descent of the posterior energy (5.3), starting with the maximum likelihood reconstruction. Of course the experiment is highly artificial, since the prior was known exactly.

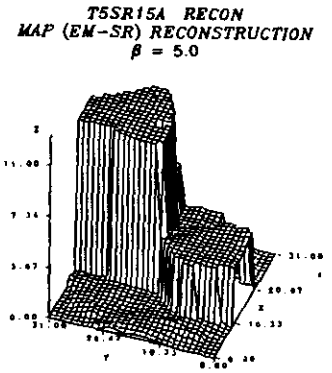
The second experiment was based on the isotope density shown in figure 3, Panel A. This also is artificial, but was not drawn from a prior distribution. The Poisson process Y was again Monte Carlo generated using the conditional Y-distribution (5.1), based upon the attenuation function shown in 2D, but this time conditioned on the density shown in 3A. Panel B is the maximum likelihood reconstruction, achieved after 54 iterations of EM. Panels A, B, and C of figure 4 are approximate *MAP* reconstructions, under the prior distribution (4.1) with δ=.7. Each of these was obtained by gradient descent of the posterior energy, starting at maximum likelihood. In panel A, β=.25; in B, β=2; and in C, β=6. Obviously, β is an important parameter. We will discuss a possible estimation technique in the next section.



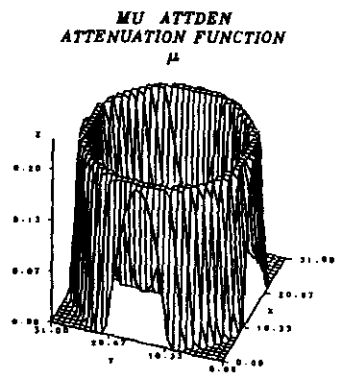
2A



2B



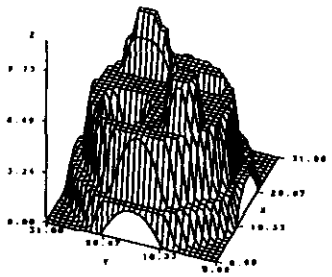
2C



2D

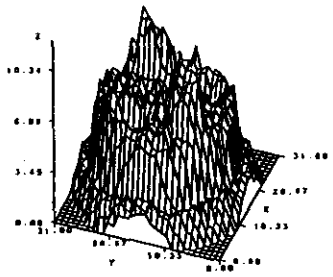
Figure 2

2D ISODEN
ISOTOPE DENSITY



3A

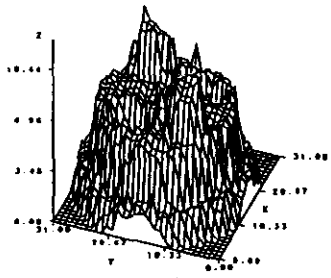
2DEM54 RECON
EM RECONSTRUCTION
n = 60, L = 64, *COUNTS* = 1,998,189



3B

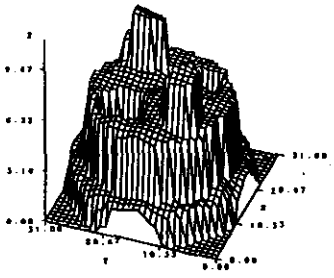
Figure 3

2D10SRC RECON
MAP (EM-SR) RECONSTRUCTION
β = 0.25



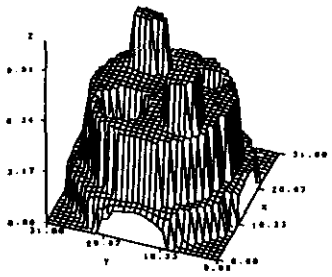
4A

2D10SRD RECON
MAP (EM-SR) RECONSTRUCTION
β = 2.0



4B

2D12SRF RECON
MAP (EM-SR) RECONSTRUCTION
β = 6.0



4C

Figure 4

7. PARAMETER ESTIMATION

The choice of β (see 4.2) is critical. With β=0 the estimator is undersmoothed, and in fact *MAP* estimation is just MLE, since the prior is uniform. If β is too large, the observation, Y, has little influence; the estimator is too faithful to the prior, and too smooth. The parameter δ (see 4.3) is also important, although it appears to be less critical than β. In any case, the brief discussion here will be about β, which is easier from the point of view of parameter estimation.

Suppose that we are given samples $\tilde{Y}_1,\ldots,\tilde{Y}_n$, which are the observables in a series of tomography experiments. The isotope densities, $\tilde{X}_1,\ldots,\tilde{X}_n$ are assumed to have been drawn, independently, from a common (prior) distribution (4.1), which is known except for β. Our main concern is with reconstruction: we want to estimate $\tilde{X}_1,\ldots,\tilde{X}_n$. However, our Bayesian program will have to begin with a suitable choice for β, and this we intend to estimate. We are often interested in the case n=1, which is not hopeless since an observation, Y, contains a large amount of data. In what follows we will specialize to this case and denote the observation by Y and the isotope density by $\tilde{X}$; the extension to n > 1 is trivial.

To be more explicit about the dependency on β of the prior and posterior distributions, we introduce the function

$$V(X) = \sum_{<s,t>} \phi(X(s)-X(t)) + \frac{1}{\sqrt{2}} \sum_{\{s,t\}} \phi(X(s)-X(t)).$$

V is just U/β (see 4.2) and ϕ is as it was in (4.3). The prior is now written

$$\Pi_X(X) = \frac{1}{Z_\beta} e^{-\beta V(X)} \tag{7.1}$$

and the posterior, given $\tilde{Y}$, is

$$\Pi_{X|Y}(X) = \frac{1}{Z_\beta(\tilde{Y})} \exp\{-\beta V(X) + \sum_{t\in D} [\tilde{Y}(t)\ln[(\mathbf{a}_T X)(t)] - (\mathbf{a}_T X)(t)]\} \tag{7.2}$$

We use the subscript β, in Z_β and $Z_\beta(\tilde{Y})$, to remind the reader of the dependency on β of these normalizing constants ("partition functions" in the language of statistical mechanics). This latter dependency is quite complicated, and is the chief reason that the problem of estimating β is hard.

We shall present a method for estimating β that was suggested to us by Donald Geman. Since this is a maximum likelihood (ML) based method, we shall begin with a discussion of the ML estimator of β given Y. Let $E_\beta[\cdot]$ and $E_\beta[\cdot|\tilde{Y}]$ denote expectations with respect to the prior (7.1) and posterior (7.2) distributions. The subscript β is again used to emphasize dependency on β. Unfortunately, the likelihood equation,

$$E_\beta[V(X)] = E_\beta[V(X)|\tilde{Y}], \tag{7.3}$$

typically has multiple solutions, reflecting non-concavity of the likelihood. The basic problem is that we can not observe $\tilde{X}$. If we could, then the likelihood would be concave, and the likelihood equation would be

$$E_\beta[V(X)] = V(\tilde{X}). \tag{7.4}$$

It is not hard to show that the left hand side is monotonic (decreasing) in β. Given $\tilde{X}$, we could in principal find the unique β satisfying (7.4). In practice, we would proceed as follows: (1) Using SR, generate multiple samples of X from (7.1) at each of a range of values of β; (2) estimate the monotonic function $E_\beta[V(X)]$ by evaluating V at the sampled values of X; (3) From the curve, β versus $E_\beta[V(X)]$, find $\hat{\beta}$ such that (7.4) is satisfied. Of course, steps (1) and (2) require a great deal of computation, but this is to be done only once, "off-line". Thereafter, given a new observation of $\tilde{X}$, ML estimation of β is trivial.

Of course we wish to solve (7.3) instead, and this is complicated by the fact that the RHS is also a function of β. Even EM (see, for example [3]), which is designed for analogous situations, is computationally prohibitive in this case. Instead, D. Geman has proposed the following variation on EM, which we intend to implement in the coming weeks and will report on in our talk.

As with the full observation case, we begin by constructing the graph of $E_\beta[V(X)]$ versus β (steps (1) and (2) of the previous paragraph). The estimate of β is then derived iteratively:

(a) Derive the ML estimate (call it X_1) of $\tilde{X}$ given $\tilde{Y}$;

(b) Treat X_1 as the actual isotope density, and determine the ML estimate of β (call it β_1), as in step (3) of the previous paragraph;

(c) Use SR, with starting configuration X_1, to sample from the posterior (7.2) with $\beta=\beta_1$ (call the sample X_2);

(d) Repeat (b), using X_2 in place of X_1, and call the result β_2 instead of β_1;

(e) Repeat (c), starting with X_2, and setting $\beta=\beta_2$, and call the result X_3;

etc.

D. Geman has performed preliminary experiments with one dimensional Ising model priors, and these suggest fast convergence and good resulting reconstructions.

Acknowledgements

This research was supported in part by the Army Research Office, Contract DAAG-29-83-K-0116, The National Science Foundation, DMS-8306507, The National Science Foundation, DMS-8352087, and the Machine Vision Products Group of Analog Devices, Inc.

References

[1] N. Accomando, "Maximum likelihood reconstruction of a two dimensional Poisson intensity function from attenuated projections," Ph.D. Thesis, Division of Applied Mathematics, Brown University, 1984.

[2] T. Budinger, G. Gullberg, and R. Huesman, "Emission computed tomography," in Image Reconstruction from Projections: Implemen- and Application, G. Herman, Ed., Springer-Verlag, 1979.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. of the Royal Stat. Society, Ser. B, 39, 1-38, 1977.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE-PAMI, vol. 6, 721-741, 1984.

[5] S. Geman and C.-R. Hwang, "Diffusions for global optimization," SIAM J. on Control and Optimization (to appear).

[6] U. Grenander, "Tutorial in Pattern Theory", Division of Applied Mathematics, Brown University, 1984.

[7] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," J. Chem. Phys., vol. 21, 1087-1091, 1953.

[8] L. Shepp and Y. Vardi, "Maximum likelihood reconstruction in positron emission tomography," IEEE Tranactions on Medical Imaging, 1, 113-122, 1982.