Convergence in Norm for Alternating Expectation-Maximization (EM) Type Algorithms[1]

Alfred O. Hero * and Jeffrey A. Fessler**

*Dept. of Electrical Engineering and Computer Science and **Division of Nuclear Medicine
The University of Michigan, Ann Arbor, MI 48109
email: hero@eecs.umich.edu, fessler@umich.edu

ABSTRACT

We provide a sufficient condition for convergence of a general class of alternating estimation-maximization (EM) type continuous-parameter estimation algorithms with respect to a given norm. This class includes EM, penalized EM, Green's OSL-EM, and other approximate EM algorithms. The convergence analysis can be extended to include alternating coordinate-maximization EM algorithms such as Meng and Rubin's ECM and Fessler and Hero's SAGE. The condition for monotone convergence can be used to establish norms under which the distance between successive iterates and the limit point of the EM-type algorithm approaches zero monotonically. For illustration, we apply our results to estimation of Poisson rate parameters in emission tomography and establish that in the final iterations the logarithm of the EM iterates converge monotonically in a weighted Euclidean norm.

[1]This research was supported in part by the National Science Foundation under grant BCS-9024370, a DOE Alexander Hollaender Postdoctoral Fellowship, DOE Grant DE-FG02-87ER60561, and NIH grant CA-60711.

I. INTRODUCTION

The maximum-likelihood (ML) expectation-maximization (EM) algorithm is a popular iterative method for finding the maximum likelihood estimate $\hat{\theta}$ of a continuous parameter θ when the likelihood function is difficult to maximize directly (e.g. Dempster, Laird, and Rubin (1977), Shepp and Vardi (1982), Lange and Carson (1984), Miller and Snyder (1987), Feder, Oppenheim, and Weinstein (1989), and Segal, Weinstein and Musicus (1991)). The penalized EM algorithm is a variant of the EM algorithm which can be used for finding *maximum a posteriori* (MAP) or *posterior mode* estimates of a random parameter (e.g. Green (1990a,b), Hebert and Leahy (1989,1992)) . To implement the EM algorithm the user first identifies a complete data space, also called an augmented data space (Wei and Tanner, (1990)), for which there exists a many-to-one mapping from the complete data to the measurement data, called the incomplete data. Then one alternates between estimating the conditional mean of the complete data log-likelihood function or log-posterior and updating the parameter estimate.

Three types of convergence results are of practical importance: conditions under which the sequence of estimates converges globally to a fixed point, norms under which the convergence is monotone; and the asymptotic convergence rate of the algorithm. A number of authors have established global convergence for the exact EM algorithm when the likelihood function satisfies conditions such as boundedness and unimodality (see Wu (1983), Boyles (1983), Lange and Carson (1984), Csiszar and Tusnady (1984)) . Sundberg (1976) and Louis (1982) have derived asymptotic convergence rates for the EM algorithm which have been used for estimating asymptotic estimator covariance (Louis (1982), Meng and Rubin (1991)) and for accelerating the basic algorithm (Meilijson (1989)). A general property of the EM algorithm is that successive iterates monotonically increase the likelihood. While increasing the likelihood is an attractive property, it does not guarantee monotone convergence of the parameter estimates: successive iterates of the EM algorithm reduce the distance to the ML estimate in some norm. In addition, for some implementations the region of convergence may only be a small subset of the entire parameter space so that global convergence may not hold. Furthermore, in some cases the EM algorithm can only be implemented by making simplifying approximations in the conditional expectation step (E) or the maximization step (M). While the resultant approximate EM algorithm has a similar alternating estimation-maximization structure, previous approaches developed to establish global convergence of the exact EM algorithm may not be effective for studying asymptotic behavior of the algorithm. In this paper we provide general conditions for monotone convergence and asymptotic convergence rates for algorithms which can be implemented via alternating estimation-maximization. The basics of this approach to EM algorithm convergence analysis were first introduced in Hero (1992).

We illustrate the application of our convergence methodology for two examples. A linear EM algorithm for a simple linear Gaussian model provides the most transparent illustration of the methodology. Then we consider the more interesting non-linear case of emission computed tomography (ECT) with Poisson statistics implemented with the EM algorithm of Shepp and Vardi (1982). For the ECT problem we show that when the EM algorithm converges to a strictly positive estimate, in the final iterations convergence is monotone in the following sense: the natural logarithm of the n-th iterate converges monotonically as $n \to \infty$ to the natural logarithm of the ML estimate in a weighted Euclidean norm.

II. AN ARCHETYPE ALGORITHM

Let $\theta = [\theta_1, \ldots, \theta_p]^T$ be a real parameter residing in an open subset Θ of the p-dimensional space $I\!R^p$. Given a general function $Q : \Theta \times \Theta \to I\!R$ and an initial point $\theta^0 \in \Theta$, consider the following recursive algorithm, called the A-algorithm:

$$\textbf{A-algorithm:} \qquad \theta^{i+1} = \mathrm{argmax}_{\theta \in \Theta} Q(\theta, \theta^i), \qquad i = 0, 1, \ldots. \tag{1}$$

If there are multiple maxima, then θ^{i+1} can be taken to be any one of them. Let $\theta^* \in \Theta$ be a fixed point of (1), i.e. θ^* satisfies: $\theta^* = \mathrm{argmax}_{\theta \in \Theta} Q(\theta, \theta^*)$

By suitable specification of the function $Q(\theta, \overline{\theta})$ the A-algorithm specializes to many popular iterative estimation algorithms. For example, for complete data $\mathbf{X}$ and incomplete data $\mathbf{Y}$ the EM algorithm is obtained by identifying $Q(\theta, \overline{\theta}) = E\{\ln f(\mathbf{X}; \theta) | \mathbf{Y}; \overline{\theta}\}$, where $f(\mathbf{X}; \theta)$ is a density function of the random variable $\mathbf{X}$ for a particular value of an unknown parameter θ. If a penalty function $P(\theta)$ is introduced then $Q(\theta, \overline{\theta}) = E\{\ln f(\mathbf{X}; \theta) | \mathbf{Y}; \overline{\theta}\} - P(\theta)$ gives the EM algorithm for penalized ML estimation, or, if $\exp(-P(\theta))$ is a prior for θ, it gives the EM algorithm for the posterior mode. Alternatively, when $Q(\theta, \overline{\theta}) = E\{\ln f(\mathbf{X}; \theta) | \mathbf{Y}; \overline{\theta}\} - (\nabla P)(\overline{\theta})[\theta - \overline{\theta}]$ we obtain the one-step-late approximation of Green (1990a,b) to the EM algorithm for the posterior mode. Likewise, the generalized EM algorithm of De Pierro (1993) and the linearized EM algorithm of Antoniadis and Hero (1994) are A-algorithms (see Hero and Fessler (1993)). Fessler and Hero (1994) extend the convergence results of this paper to the space-alternating generalized EM (SAGE) algorithm in which the functional $Q(\theta, \overline{\theta})$ changes with iteration. Similar extensions apply to the study of monotone norm convergence for the multi-cycle expectation/conditional maximization (ECM) algorithm of Meng and Rubin (1993) and the ECME algorithm of Liu and Rubin (1994).

Let $\|\cdot\|$ denote a vector norm on $I\!R^p$. For any $p \times p$ matrix $\mathbf{A}$ the induced *matrix norm* $|\!|\!|\mathbf{A}|\!|\!|$ (see section 5.6 of Horn and Johnson (1985)) of $\mathbf{A}$ is defined as:

$$|\!|\!|\mathbf{A}|\!|\!| \stackrel{def}{=} \max_{u \in I\!R^p - \{0\}} \frac{\|\mathbf{A}u\|}{\|u\|},$$

where the maximization is over non-zero u in $I\!R^p$. A special case is the *matrix-2* norm $|\!|\!|\mathbf{A}|\!|\!|_2$ which is induced by the Euclidean vector norm $\|u\|_2^2 = u^T u$. We say that a sequence u^i, $i = 1, 2, \ldots$, converges monotonically to a point u^* in the norm $\|\cdot\|$ if:

$$\|u^{i+1} - u^*\| \le \alpha \|u^i - u^*\|, \qquad i = 1, 2, \ldots,$$

for some constant α, $\alpha \in [0,1)$. Consider the general linear iteration of the form $v^{i+1} = \mathbf{A}v^i, \quad i = 1,2,\ldots$, with $\|\mathbf{A}\| < 1$. Then, since $\|v^{i+1}\| \leq \|\mathbf{A}\| \cdot \|v^i\| < \|v^i\|$, the sequence $\{v^i\}$ converges monotonically to zero and the asymptotic rate of convergence is specified by the *root convergence factor* $\rho(\mathbf{A})$ which is defined as the largest magnitude eigenvalue of $\mathbf{A}$ (Ortega and Rheinboldt (1970, p. 301)). If $\mathbf{A}$ is real symmetric non-negative definite then $\rho(\mathbf{A}) = \|\mathbf{A}\|_2$. The above simple convergence conditions only apply to linear iterations. Theorem 1 below gives a related set of convergence conditions for the generally non-linear A-algorithm.

Assume that the function $Q(\theta,\overline{\theta})$ is twice continuously differentiable in both arguments θ and $\overline{\theta}$ over $\theta,\overline{\theta} \in \Theta$. Define the Hessian matrix of Q over $\Theta \times \Theta$ as the following block partitioned $2p \times 2p$ matrix:

$$\nabla^2 Q(\theta,\overline{\theta}) = \left[\begin{array}{cc} \nabla^{20}Q(\theta,\overline{\theta}) & \nabla^{11}Q(\theta,\overline{\theta}) \\ (\nabla^{11}Q(\theta,\overline{\theta}))^T & \nabla^{02}Q(\theta,\overline{\theta}) \end{array} \right], \tag{2}$$

where $\nabla^{20}Q(\theta,\overline{\theta}) = \nabla_\theta \nabla_\theta^T Q(\theta,\overline{\theta})$, $\nabla^{02}Q(\theta,\overline{\theta}) = \nabla_{\overline{\theta}} \nabla_{\overline{\theta}}^T Q(\theta,\overline{\theta})$, and $\nabla^{11}Q(\theta,\overline{\theta}) = \nabla_{\overline{\theta}} \nabla_\theta^T Q(\theta,\overline{\theta})$ are $p \times p$ matrices of partial derivatives $\frac{\partial^2}{\partial\theta_i \partial\theta_j} Q(\theta,\overline{\theta})$, $\frac{\partial^2}{\partial\overline{\theta}_i \partial\overline{\theta}_j} Q(\theta,\overline{\theta})$, and $\frac{\partial^2}{\partial\overline{\theta}_i \partial\theta_j} Q(\theta,\overline{\theta})$, $i,j = 1,\ldots,p$, respectively.

A *region of monotone convergence* relative to the vector norm $\|\cdot\|$ of the A-algorithm (1) is defined as any open ball $B(\theta^*,\delta) = \{\theta : \|\theta - \theta^*\| < \delta\}$ centered at $\theta = \theta^*$ with radius $\delta > 0$ such that if the initial point θ^0 is in this region then $\|\theta^i - \theta^*\|$, $i = 1,2,\ldots$, converges monotonically to zero. Note that as defined, the shape in $I\!R^p$ of the region of monotone convergence depends on the norm used. For the Euclidean norm $\|u\|^2 = u^T u$ the region of monotone convergence is a spherically shaped region in Θ. For a general positive definite matrix $\mathbf{B}$ the induced norm $\|u\|^2 = u^T\mathbf{B}u$ makes this region an ellipsoid in Θ. Since all norms are equivalent for the case of a finite dimensional parameter space, monotone convergence in a given norm implies convergence, however possibly non-monotone, in any other norm.

Define the $p\times p$ matrices obtained by averaging $\nabla^{20}Q(u,\overline{u})$ and $\nabla^{11}Q(u,\overline{u})$ over the line segments $u \in \overrightarrow{\theta\theta^*}$ and $\overline{u} \in \overrightarrow{\overline{\theta}\theta^*}$:

$$\begin{aligned} A_1(\theta,\overline{\theta}) &= -\int_0^1 \nabla^{20}Q(t\theta + (1-t)\theta^*, t\overline{\theta} + (1-t)\theta^*)dt \\ A_2(\theta,\overline{\theta}) &= \int_0^1 \nabla^{11}Q(t\theta + (1-t)\theta^*, t\overline{\theta} + (1-t)\theta^*)dt. \end{aligned} \tag{3}$$

Also, define the following set:

$$\mathcal{S}(\overline{\theta}) = \{\theta \in \Theta \;:\; Q(\theta,\overline{\theta}) \geq Q(\overline{\theta},\overline{\theta})\}.$$

By the construction of the A-algorithm (1), we have $\theta^{i+1} \in \mathcal{S}(\theta^i)$.

Definition 1 *For a given vector norm $\|\cdot\|$ and induced matrix norm $\|\!|\cdot|\!\|$ define $\mathcal{R}_+ \subset \Theta$ as the largest open ball $B(\theta^*, \delta) = \{\theta : \|\theta - \theta^*\| < \delta\}$ such that for each $\overline{\theta} \in B(\theta^*, \delta)$:*

$$A_1(\theta, \overline{\theta}) > 0, \qquad \textit{for all } \theta \in \mathcal{S}(\overline{\theta}) \tag{4}$$

and for some $0 \leq \alpha < 1$

$$\left|\!\left|\!\left| \left[A_1(\theta, \overline{\theta})\right]^{-1} \cdot A_2(\theta, \overline{\theta}) \right|\!\right|\!\right| \leq \alpha, \qquad \textit{for all } \theta \in \mathcal{S}(\overline{\theta}). \tag{5}$$

The following convergence theorem establishes that, if $\mathcal{R}_+$ is not empty, the region in Definition 1 is a region of monotone convergence in the norm $\|\cdot\|$ for an algorithm of the form (1). One can show that $\mathcal{R}_+$ is non-empty for sufficiently regular problems. For example, assume that: *i)* $Q(\theta, \overline{\theta})$ is continuously twice differentiable in θ and $\overline{\theta}$; *ii)* Q can be written as $Q(\theta, \overline{\theta}) = L(\theta) + H(\theta, \overline{\theta})$ where $H(\theta, \overline{\theta}) \leq H(\overline{\theta}, \overline{\theta})$ and $\nabla^{11} H(\theta, \theta) = -\nabla^{20} H(\theta, \theta) \geq 0$ (as is always the case for an EM algorithm (see Dempster, Laird, Rubin (1977))); *iii)* $L(\theta)$ has a local maximum at $\theta = \theta^*$ and *iv)* there exists a level L^* such that $L(\theta)$ is strictly concave over the set $\{\theta : L(\theta) > L^*\}$. Note that under these conditions it follows from Corollary 1 of Wu (1983), that the set $\{\theta : L(\theta) > L^*\}$ is a region of convergence to the global maximum, i.e. if the initial point θ^0 is selected from this set subsequent iterates θ^i will converge to θ^*, although it is not generally a region of monotone convergence in norm. A non-empty region of monotone convergence $\mathcal{R}_+$ is established as follows. By assumptions *i* and *iv*, for any $\epsilon > 0$ there exists a $\delta > 0$ such that if $\overline{\theta} \in B_2(\theta^*, \delta) \stackrel{def}{=} \{\theta : \|\theta - \theta^*\|_2 < \delta\}$ then $\{\theta : L(\theta) \geq L(\overline{\theta})\} \subset B_2(\overline{\theta}, \epsilon)$. Since $Q(\theta, \overline{\theta}) - Q(\overline{\theta}, \overline{\theta}) \leq L(\theta) - L(\overline{\theta})$ we have $\mathcal{S}(\overline{\theta}) \subset \{\theta : L(\theta) \geq L(\overline{\theta})\}$. Thus for $\overline{\theta} \in B_2(\theta^*, \delta)$ and $\theta \in \mathcal{S}(\overline{\theta})$ we have: $A_1(\theta, \overline{\theta}) = -\nabla^{20} Q(\theta^*, \theta^*) + O(\epsilon)$ and $\left|\!\left|\!\left|[A_1(\theta, \overline{\theta})]^{-1} A_2(\theta, \overline{\theta})\right|\!\right|\!\right| = |\!|\!|[\nabla^{20} Q(\theta^*, \theta^*)]^{-1} \nabla^{11} Q(\theta^*, \theta^*)|\!|\!| + O(\epsilon)$. By assumptions *ii)* and *iv)* the matrix $-\nabla^{20} Q(\theta^*, \theta^*)$ is symmetric positive definite and $\nabla^{11} Q(\theta^*, \theta^*)$ is symmetric non-negative definite. Hence, for sufficiently small $\epsilon > 0$, for all $\overline{\theta} \in B_2(\theta^*, \delta)$ and for all $\theta \in \mathcal{S}(\overline{\theta})$ the condition (4) is satisfied and, defining the norm $\|\cdot\|$ by $\|u\|^2 = u^T[-\nabla^{20} Q(\theta^*, \theta^*)]u$: $|\!|\!|[-\nabla^{20} Q(\theta^*, \theta^*)]^{-1} \nabla^{11} Q(\theta^*, \theta^*)|\!|\!| = \rho\left([-\nabla^{20} Q(\theta^*, \theta^*)]^{-1} \nabla^{11} Q(\theta^*, \theta^*)\right) = \rho\left([-\nabla^{20} L(\theta^*) - \nabla^{20} H(\theta^*, \theta^*)]^{-1} \, [-\nabla^{20} H(\theta^*, \theta^*)]\right) < 1$ so that the condition (5) is also satisfied. Thus $\mathcal{R}_+$ is non-empty for any EM algorithm satisfying the regularity conditions *i-iv.*

Theorem 1 *Let $\theta^* \in \Theta$ be a fixed point of the A algorithm (1), where $\theta^{i+1} = argmax_{\theta \in \Theta} Q(\theta, \theta^i)$, $i = 0, 1, \ldots$. Assume: i) for all $\overline{\theta} \in \Theta$, the maximum $\max_\theta Q(\theta, \overline{\theta})$ is achieved on the interior of the set Θ; ii) $Q(\theta, \overline{\theta})$ is twice continuously differentiable in $\theta \in \Theta$ and $\overline{\theta} \in \Theta$, and iii) the A-algorithm (1) is initialized at a point $\theta^0 \in \mathcal{R}_+$ for a norm $\|\cdot\|$.*

1. *The iterates θ^i, $i = 0, 1, \ldots$ all lie in $\mathcal{R}_+$,*

2. *the successive differences $\Delta\theta^i = \theta^i - \theta^*$ of the A algorithm obey the recursion:*

$$\Delta\theta^{i+1} = [A_1(\theta^{i+1}, \theta^i)]^{-1} A_2(\theta^{i+1}, \theta^i) \cdot \Delta\theta^i, \qquad i = 0, 1, \ldots. \tag{6}$$

3. *the norm* $\|\Delta\theta^i\|$ *converges monotonically to zero with at least linear rate, and*

4. $\Delta\theta^i$ *asymptotically converges to zero with root convergence factor*

$$\rho\left(\left[-\nabla^{20}Q(\theta^*,\theta^*)\right]^{-1}\nabla^{11}Q(\theta^*,\theta^*)\right) < 1.$$

If the iterates are initialized within a region $\mathcal{R}_+$, or for that matter if any iterate θ^i lies in $\mathcal{R}_+$, then all subsequent iterates will also lie within $\mathcal{R}_+$. Within that region, Theorem 1 provides a functional relationship (6) between successive iterates, which in turn ensures that the iterates converge monotonically in norm to θ^* with an asymptotic linear rate governed by the spectral radius of a matrix depending on the partial derivatives of Q. When specialized to the EM algorithm, the root convergence factor is equivalent to the expression obtained by Dempster, Laird and Rubin (1977) and used by Meng and Rubin (1991) to estimate the asymptotic estimator covariance matrix.

Proof of Theorem 1:

Define $\Delta\theta = \theta - \theta^*$ and $\Delta\theta^i = \theta^i - \theta^*$. Convergence will be established by showing that $\|\Delta\theta^{i+1}\| \leq \alpha\|\Delta\theta^i\|$ for some $0 \leq \alpha < 1$. Define the $2p \times 1$ vectors $\xi = \left[\begin{smallmatrix}\theta \\ \theta^i\end{smallmatrix}\right]$, $\xi^* = \left[\begin{smallmatrix}\theta^* \\ \theta^*\end{smallmatrix}\right]$ and $\Delta\xi = \xi - \xi^*$. By assumption ii of the Theorem we can use the Taylor formula with remainder (Polak (1971, Eq. B.1.4))

$$h(\xi) - h(\xi^*) = \int_0^1 (\nabla h)\,(t\xi + (1-t)\xi^*)dt\;\Delta\xi$$

to expand the column vector $h(\xi) \stackrel{def}{=} [\nabla^{10}Q(\theta,\theta^i)]^T$ about the point $\xi = \xi^*$ to obtain from (3)

$$\nabla^{10}Q(\theta,\theta^i) = -A_1(\theta,\theta^i)\Delta\theta + A_2(\theta,\theta^i)\Delta\theta^i. \tag{7}$$

To obtain (7) we have used the assumption that θ^* is a fixed point of the A-algorithm: $h(\xi^*) = \nabla^{10}Q(\theta^*,\theta^*) = 0$.

Since $\theta^{i+1} = \text{argmax}_\theta Q(\theta,\theta^i)$ lies in the interior of Θ, we have $\nabla^{10}Q(\theta^{i+1},\theta^i) = 0$. Therefore from (7):

$$-A_1(\theta^{i+1},\theta^i)\Delta\theta^{i+1} + A_2(\theta^{i+1},\theta^i)\Delta\theta^i = 0. \tag{8}$$

We prove the first part of the theorem using induction. We have $\theta^0 \in \mathcal{R}_+$ by assumption. Now suppose $\theta^i \in \mathcal{R}_+$. Since $\theta^{i+1} \in \mathcal{S}(\theta^i)$, by (4) $A_1(\theta^{i+1},\theta^i)$ is invertible, so rearranging (8) shows:

$$\Delta\theta^{i+1} = [A_1(\theta^{i+1},\theta^i)]^{-1}A_2(\theta^{i+1},\theta^i)\cdot\Delta\theta^i, \tag{9}$$

and

$$\begin{aligned}
\|\Delta\theta^{i+1}\| &\leq \left\| A_1(\theta^{i+1},\theta^i)]^{-1} A_2(\theta^{i+1},\theta^i) \right\| \cdot \|\Delta\theta^i\| \\
&\leq \sup_{\theta\in\mathcal{S}(\theta^i)} \left\| [A_1(\theta,\theta^i)]^{-1} A_2(\theta,\theta^i) \right\| \cdot \|\Delta\theta^i\| \\
&\leq \alpha\|\Delta\theta^i\|, \qquad (10)
\end{aligned}$$

where the last inequality follows from (5) and the supposition that $\theta^i \in \mathcal{R}_+$. Since $\alpha < 1$ and $\mathcal{R}_+$ is an open ball centered at θ^* which contains θ^i, this implies that $\theta^{i+1} \in \mathcal{R}_+$, proving the induction step. Furthermore, from (10) we conclude that $\|\Delta\theta^i\| = \|\theta^i - \theta^*\|$ converges monotonically to zero with at least linear convergence rate.

Next we establish the asymptotic convergence rate stated in the theorem. By continuity of the derivatives of $Q(\theta,\theta^i)$ and the result (10) we obtain:

$$A_1(\theta^{i+1},\theta^i) = -\nabla^{20}Q(\theta^*,\theta^*) + O(\|\Delta\theta^i\|)$$

$$A_2(\theta^{i+1},\theta^i) = \nabla^{11}Q(\theta^*,\theta^*) + O(\|\Delta\theta^i\|).$$

Thus, by continuity of the matrix norm:

$$\alpha \geq \sup_{\theta\in\mathcal{S}(\theta^i)} \left\| [A_1(\theta,\theta^i)]^{-1} A_2(\theta,\theta^i) \right\| = \left\| \left[-\nabla^{20}Q(\theta^*,\theta^*)\right]^{-1} \nabla^{11}Q(\theta^*,\theta^*) \right\| + O(\|\Delta\theta^i\|).$$

Since $\alpha < 1$ taking the limit of the right hand side as $i \to \infty$ establishes that

$$\left\| \left[-\nabla^{20}Q(\theta^*,\theta^*)\right]^{-1} \nabla^{11}Q(\theta^*,\theta^*) \right\| < 1. \qquad (11)$$

Furthermore (9) takes the asymptotic form: $\Delta\theta^{i+1} = [-\nabla^{20}Q(\theta^*,\theta^*)]^{-1}\nabla^{11}Q(\theta^*,\theta^*)\cdot\Delta\theta^i + o(\|\Delta\theta^i\|)$. Therefore the asymptotic rate of convergence is given by the root convergence factor $\rho\left([-\nabla^{20}Q(\theta^*,\theta^*)]^{-1}\nabla^{11}Q(\theta^*,\theta^*)\right)$. For any matrix $\mathbf{A}$ we have $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$ (Horn and Johnson (1985, Thm. 5.6.9)) so that, in view of (11), the root convergence factor is less than one. 2

As will be seen in the next section, to apply Theorem 1 it is sometimes useful to make a transformation of parameters $\theta \to \tau$. Consider a smooth invertible functional transformation g: $\tau = g(\theta)$. Then θ^i can be represented as $g^{-1}(\tau^i)$, where g^{-1} is the inverse of g, the sequence $\{\tau^i\}$ is generated by an analogous A-algorithm:

$$\tau^{i+1} = \mathrm{argmax}_{\tau\in g(\Theta)} \tilde{Q}(\tau,\tau^i), \qquad i = 0, 1, \ldots,$$

and

$$\tilde{Q}(\tau,\tau^i) \stackrel{def}{=} Q\left(g^{-1}(\tau), g^{-1}(\tau^i)\right) = Q(\theta,\theta^i)\big|_{\theta=g^{-1}(\tau),\theta^i=g^{-1}(\tau^i)}.$$

The convergence properties of the sequence $\tau^i = g(\theta^i)$ can be studied using Theorem 1 with A_1 and A_2 defined in terms of the mixed partial derivatives of $\tilde{Q}$:

$$\begin{aligned} \nabla^{11}\tilde{Q}(\tau,\tau^i) &= J^{-T}(\tau)\left[\nabla^{11}Q\left(g^{-1}(\tau), g^{-1}(\tau^i)\right)\right] J^{-1}(\tau^i) & (12)\\ \nabla^{20}\tilde{Q}(\tau,\tau^i) &= J^{-T}(\tau)\left[\nabla^{20}Q\left(g^{-1}(\tau), g^{-1}(\tau^i)\right)\right] J^{-1}(\tau), & (13) \end{aligned}$$

where $J(\tau) = \nabla g(\theta)|_{\theta=g^{-1}(\tau)}$ is the $p \times p$ Jacobian matrix of partial derivatives of g.

III. Examples

To illustrate the usefulness of Theorem 1 we consider two examples. For more details we refer the reader to Hero and Fessler (1993).

III.a. Linear Gaussian Model

Consider the following model:

$$\mathbf{Y} = \mathbf{G}\theta + \mathbf{W}_y$$

where $\mathbf{G}$ is a known $m \times p$ matrix with full column rank $p \le m$, and $\mathbf{W}_y$ is an m-dimensional zero mean Gaussian noise with known positive definite covariance matrix Λ_{yy}. The ML estimator of θ given $\mathbf{Y}$ is the weighted least squares estimator which is the solution θ^* to the normal equations:

$$[\mathbf{G}^T\Lambda_{yy}^{-1}\mathbf{G}]\theta^* = \mathbf{G}^T\Lambda_{yy}^{-1}\mathbf{Y}. \tag{14}$$

An EM algorithm for estimating θ can be derived by decomposing the matrix $\mathbf{G}$ into the matrix product: $\mathbf{G} = \mathbf{BC}$, where the $m \times n$ matrix $\mathbf{B}$ has full row rank m, the $n \times p$ matrix $\mathbf{C}$ has full column rank p, and $p \le m \le n$. With this decomposition we define the hypothetical observations $\mathbf{X} = \mathbf{C}\theta + \mathbf{W}_x$ where $\mathbf{W}_x$ is a zero mean Gaussian noise with θ-independent positive definite covariance matrix Λ_{xx}. We assume that $\mathbf{W}_x$ and $\mathbf{W}_y$ are statistically independent. Using $(\mathbf{X},\mathbf{Y})$ as a complete data set, the EM algorithm takes the form of the A-algorithm (1) with $Q(\theta,\overline{\theta}) = E\{\ln f(\mathbf{X};\theta)|\mathbf{Y};\overline{\theta}\}$ given by:

$$Q(\theta,\overline{\theta}) = \theta^T\mathbf{F}_X\overline{\theta} - \theta^T\mathbf{F}_Y\overline{\theta} + \theta^T\mathbf{G}^T\Lambda_{yy}^{-1}\mathbf{y} - \tfrac{1}{2}\theta^T\mathbf{F}_X\theta, \tag{15}$$

where $\mathbf{F}_X = E\{-\nabla_\theta^2 \ln f(\mathbf{X};\theta)\} = \mathbf{C}^T\Lambda_{xx}^{-1}\mathbf{C}$ and $\mathbf{F}_Y = E\{-\nabla_\theta^2 \ln f(\mathbf{Y};\theta)\} = \mathbf{G}^T\Lambda_{yy}^{-1}\mathbf{G}$ are respectively the Fisher information matrices for θ associated with data sets $\mathbf{X}$ and $\mathbf{Y}$. Since the Q function (15) is quadratic the M step is in closed form and we have the EM recursion:

$$\theta^{i+1} = [I - \mathbf{F}_X^{-1}\mathbf{F}_Y]\theta^i + \mathbf{F}_X^{-1}\mathbf{G}^T\Lambda_{yy}^{-1}\mathbf{Y}. \tag{16}$$

For diagonal $\mathbf{F}_X$ the EM recursion is equivalent to the well known Jacobi iterations technique (Golub and Van Loan (1989, Sec. 10.1.2)) for solving linear equations of the type (14). The advantages of Jacobi iterations relative to direct solution of (14) are: i) the computations in (16) are parallelizable; ii) if the iterations of (16) converge rapidly, a good approximation to (14) can be obtained with fewer floating point operations, particularly if $\mathbf{G}$ is large but sparse.

The convergence properties of the Jacobi iteration (16) are well known. However, due to the simplicity of this example it is instructive to illustrate how Theorem 1 directly applies. It is easy to see that $A_1(\theta,\overline{\theta}) = -\int_0^1 \nabla^{20}Q(\theta,\overline{\theta})dt = \mathbf{F}_X$ and $A_2(\theta,\overline{\theta}) = \int_0^1 \nabla^{11}Q(\theta,\overline{\theta})dt = \mathbf{F}_X - \mathbf{F}_Y$. The condition $A_1(\theta,\overline{\theta}) > 0$ (4) is satisfied since $\mathbf{C}$ is full rank. Thus we obtain directly from Theorem 1 the recursion for $\Delta\theta^i = \theta^i - \theta^*$:

$$\Delta\theta^{i+1} = (\mathbf{I} - \mathbf{F}_X^{-1}\mathbf{F}_Y)\Delta\theta^i.$$

We remark that, unless $\mathbf{I} - \mathbf{F}_X^{-1}\mathbf{F}_Y$ is symmetric, convergence of $\Delta\theta^i$ is not monotone with respect to the unweighted Euclidean norm.

The $\Delta\theta^{i+1}$ recursion is equivalent to

$$\mathbf{F}_X^{\frac{1}{2}}\Delta\theta^{i+1} = \mathbf{F}_X^{-\frac{1}{2}}[\mathbf{F}_X - \mathbf{F}_Y]\mathbf{F}_X^{-\frac{1}{2}} \cdot \mathbf{F}_X^{\frac{1}{2}}\Delta\theta^i$$

Take the Euclidean norm of both sides to obtain

$$\|\Delta\theta^{i+1}\| \quad \leq \quad \left|\!\left|\!\left| \mathbf{F}_X^{-\frac{1}{2}}[\mathbf{F}_X - \mathbf{F}_Y]\mathbf{F}_X^{-\frac{1}{2}} \right|\!\right|\!\right|_2 \cdot \|\Delta\theta^i\|$$

where $|\!|\!|\cdot|\!|\!|_2$ is the matrix-2 norm and $\|\cdot\|$ is the weighted Euclidean norm defined on vectors $u \in I\!R^p$

$$\|u\|^2 \stackrel{def}{=} u^T\mathbf{F}_X u. \tag{17}$$

Since $|\!|\!|\mathbf{A}|\!|\!|_2 = \rho(\mathbf{A})$ for symmetric nonnegative definite $\mathbf{A}$

$$\left|\!\left|\!\left| \mathbf{F}_X^{-\frac{1}{2}}[\mathbf{F}_X - \mathbf{F}_Y]\mathbf{F}_X^{-\frac{1}{2}} \right|\!\right|\!\right|_2 \quad = \quad \rho\left(\mathbf{F}_X^{-\frac{1}{2}}[\mathbf{F}_X - \mathbf{F}_Y]\mathbf{F}_X^{-\frac{1}{2}}\right) = \rho\left(\mathbf{I} - \mathbf{F}_X^{-1}\mathbf{F}_Y\right) < 1,$$

where the strict inequality follows from the fact that the eigenvalues of $\mathbf{I} - \mathbf{F}_X^{-1}\mathbf{F}_Y$ all lie in the interval $[0,1)$ due to nonnegative definiteness of $\mathbf{F}_X - \mathbf{F}_Y$. Thus conditions (4) and (5) hold for all $\theta,\overline{\theta}$ and the region of monotone convergence $\mathcal{R}_+$ is the entire parameter space $\Theta = I\!R^p$. By part 2 of Theorem 1, convergence of the EM algorithm is monotone in the weighted Euclidean norm (17) and by part 4 the root convergence factor is the maximum eigenvalue of $\mathbf{I} - \mathbf{F}_X^{-1}\mathbf{F}_Y$.

III.b. ECT Image Reconstruction

In the ECT problem the objective is to estimate the intensity vector $\theta = [\theta_1, \ldots, \theta_p]^T$, $\theta_b \geq 0$, governing the number of gamma-ray emissions $\mathbf{N} = [\mathbf{N}_1, \ldots, \mathbf{N}_p]^T$ over an imaging volume of p pixels. The estimate of θ must be based on the projection data $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_m]^T$. The elements $\mathbf{N}_b$ of $\mathbf{N}$ are independent Poisson distributed with rate parameters θ_b, and the elements $\mathbf{Y}_d$ of $\mathbf{Y}$ are independent Poisson distributed with rate parameters $\mu_d(\theta) = \sum_{b=1}^p P_{d|b}\theta_b$, where $P_{d|b}$ is the transition probability corresponding to emissions from pixel b being detected at detector module d. We will only consider the unpenalized EM algorithm here. A similar treatment of penalized EM is contained in Hero and Fessler (1993). To ensure a unique ML estimate we assume that $m \geq p$, the $m \times p$ system matrix $(P_{d|b}, d = 1, \ldots, m; b = 1, \ldots, p)$ has full column rank, and $(\mu_d(\theta),\ \mathbf{Y}_d)$ are strictly positive for all $d = 1, \ldots, m$. We also assume that the ML estimate θ^* lies in the interior, $\theta_b^* > 0$, $b = 1, \ldots, p$, of the parameter space.

The standard choice of complete data $\mathbf{X}$ for estimation of θ via the EM algorithm is the set $\{\mathbf{N}_{db}\}_{d=1,b=1}^{m,p}$, where $\mathbf{N}_{db}$ denotes the number of emissions in pixel b which are detected at detector d (see Lange and Carson (1984)). These complete data are related to the incomplete data via the deterministic many-to-one mapping: $\mathbf{Y}_d = \sum_{b=1}^p \mathbf{N}_{db}$, $d = 1, \ldots, m$. It is easily established that $\{\mathbf{N}_{db}\}$ are independent Poisson random variables with intensity $E_\theta\{\mathbf{N}_{db}\} = P_{d|b}\theta_b$, $d = 1, \ldots, m$, $b = 1, \ldots, p$, and that the Q function in the A-algorithm (1) is (Green, (1990a,b))

$$Q(\theta, \theta^i) = E\{\ln f(\mathbf{X}; \theta)|\mathbf{Y}; \theta^i\} = \sum_{d=1}^m \sum_{b=1}^p \left[\frac{\mathbf{Y}_d P_{d|b}\theta_b^i}{\mu_d(\theta^i)} \ln(P_{d|b}\theta_b) - P_{d|b}\theta_b\right].$$

By solving for $\theta = \theta^{i+1}$ in the equation $\nabla_\theta Q(\theta, \theta^i) = 0$ the EM algorithm is obtained:

$$\theta_b^{i+1} = \frac{\theta_b^i}{P_b} \sum_{d=1}^m \frac{\mathbf{Y}_d P_{d|b}}{\mu_d(\theta^i)}, \qquad b = 1, \ldots, p, \tag{18}$$

where $P_b \stackrel{def}{=} \sum_{b=1}^p P_{d|b}$ is positive under the assumption that $P_{d|b}$ has full column rank.

We have:

$$\begin{aligned} -\nabla^{20} Q(\theta, \theta^i) &= \operatorname{diag}_b\left(\frac{\theta_b^i}{\theta_b}\right) \cdot [\mathbf{B}(\theta^i) + \mathbf{C}(\theta^i)] \cdot \operatorname{diag}_b\left(\frac{\theta_b^i}{\theta_b}\right) && (19) \\ \nabla^{11} Q(\theta, \theta^i) &= \operatorname{diag}_b\left(\frac{\theta_b^i}{\theta_b}\right) \cdot \mathbf{C}(\theta^i) && (20) \end{aligned}$$

where, similar to the definition in Green (1990a), $\mathbf{B}(\theta^i)$ is the positive definite $p \times p$ matrix:

$$\mathbf{B}(\theta^i) \stackrel{def}{=} \sum_{d=1}^m \frac{\mathbf{Y}_d}{[\mu_d(\theta^i)]^2} P_{d|*} P_{d|*}^T,$$

$P_{d|*} = [P_{d|1}, \ldots, P_{d|p}]^T$, and $\mathbf{B}(\theta^i) + \mathbf{C}(\theta^i)$ is the $p \times p$ positive definite matrix

$$\mathbf{B}(\theta^i) + \mathbf{C}(\theta^i) \stackrel{def}{=} \operatorname{diag}_b\left(\frac{1}{\theta_b^i} \sum_{d=1}^m \frac{\mathbf{Y}_d P_{d|b}}{\mu_d(\theta^i)}\right).$$

From (19) and (20) it can be shown that for any θ^i, the norm $\sup_{\theta\in\mathcal{S}(\theta^i)} \|A_1(\theta,\theta^i)]^{-1}A_2(\theta,\theta^i)\|_2$ is greater than or equal to $2\rho\left([\mathbf{B}(\theta^*)+\mathbf{C}(\theta^*)]^{-1}\mathbf{C}(\theta^*)\right)$. Now $\rho\left([\mathbf{B}(\theta^*)+\mathbf{C}(\theta^*)]^{-1}\mathbf{C}(\theta^*)\right)<1$ but it is typically greater than 0.5 and Theorem 1 cannot be applied to establish monotone convergence of θ^i in Euclidean norm. The principal difficulty lies in the unboundedness of (19) and (20) as a function of θ.

Consider the alternative parameterization defined by the logarithmic transformation g:

$$\tau \;=\; \ln\theta = [\ln\theta_1,\ldots,\ln\theta_p]^T.$$

Using the relations (12)-(13), and the identities (19)-(20):

$$-\nabla^{20}Q(\tau,\tau^i) \;=\; \mathrm{diag}_b\left(e^{\tau_b^i}\right)\cdot\left[\mathbf{B}\left(e^{\tau^i}\right)+\mathbf{C}\left(e^{\tau^i}\right)\right]\cdot\mathrm{diag}_b\left(e^{\tau_b^i}\right) \tag{21}$$

$$\nabla^{11}Q(\tau,\tau^i) \;=\; \mathrm{diag}_b\left(e^{\tau_b^i}\right)\cdot\mathbf{C}\left(e^{\tau^i}\right)\cdot\mathrm{diag}_b\left(e^{\tau_b^i}\right). \tag{22}$$

Note that unlike (19) and (20), which are in the original parameter coordinates, the matrices (21) and (22) are constant and bounded in the transformed parameter $\tau=\ln\theta$.

Let $A_1(\tau,\overline{\tau})$ and $A_2(\tau,\overline{\tau})$ be defined as in (3) with the integrands (21) and (22), respectively. If θ^i lies in the interior of Θ, $-\nabla^{20}Q(\tau,\tau^i)$ (21) is positive definite. In this case the recursion (6) of Theorem 1 applies to $\Delta\tau^i=\Delta\ln\theta^i=\ln(\theta^i/\theta^*)$. After some algebraic manipulations we obtain:

$$\Delta\ln\theta^{i+1} = [\tilde{\mathbf{B}}(\theta^i)+\tilde{\mathbf{C}}(\theta^i)]^{-1}\tilde{\mathbf{C}}(\theta^i)\cdot\Delta\ln\theta^i, \tag{23}$$

where

$$\begin{aligned}\tilde{\mathbf{B}}(\theta^i)+\tilde{\mathbf{C}}(\theta^i) &= \mathrm{diag}_b\left(\sum_{d=1}^m \mathbf{Y}_d\int_0^1 \frac{P_{d|b}(\theta_b^i/\theta_b^*)^t\theta_b^*}{\sum_{b=1}^p P_{d|b}(\theta_b^i/\theta_b^*)^t\theta_b^*}\,dt\right) \\ \tilde{\mathbf{C}}(\theta^i) &= \sum_{d=1}^m \mathbf{Y}_d\left(\left(\int_0^1 \frac{P_{d|j}(\theta_j^i/\theta_j^*)^t\theta_j^*}{\sum_{b=1}^p P_{d|b}(\theta_b^i/\theta_b^*)^t\theta_b^*}\cdot\frac{P_{d|k}(\theta_k^i/\theta_k^*)^t\theta_k^*}{\sum_{b=1}^p P_{d|b}(\theta_b^i/\theta_b^*)^t\theta_b^*}dt\right)\right)_{j,k=1,\ldots,p}\end{aligned} \tag{24}$$

For simplicity, in the sequel we suppress the functional dependence on θ^i in the notation for $\tilde{\mathbf{B}}(\theta^i)$ and $\tilde{\mathbf{C}}(\theta^i)$. The recursion (23) is equivalent to:

$$[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{\frac{1}{2}}\Delta\ln\theta^{i+1} = [\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-\frac{1}{2}}\tilde{\mathbf{C}}[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-\frac{1}{2}}\cdot[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{\frac{1}{2}}\Delta\ln\theta^i.$$

Taking the Euclidean norm of both sides we obtain:

$$\begin{aligned}\left[\Delta\ln\theta^{i+1}\right]^T\;[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]\;[\Delta\ln\theta^{i+1}] &\le \left\|[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-\frac{1}{2}}\tilde{\mathbf{C}}[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-\frac{1}{2}}\right\|_2\cdot\left[\Delta\ln\theta^i\right]^T\;[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]\;[\Delta\ln\theta^i] \\ &= \rho\left([\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-1}\tilde{\mathbf{C}}\right)\cdot\left[\Delta\ln\theta^i\right]^T\;[\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]\;[\Delta\ln\theta^i].\end{aligned} \tag{25}$$

It can easily be shown that if θ^i is in the interior of Θ then $\tilde{\mathbf{B}}$ is positive definite, $\tilde{\mathbf{C}}$ is non-negative definite and therefore $\rho\left([\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-1}\tilde{\mathbf{C}}\right) < 1$.

From (24) we obtain the small $\Delta\theta^i$ asymptotic forms:

$$\begin{aligned}\tilde{\mathbf{B}}+\tilde{\mathbf{C}} &= \mathrm{diag}_b\left(\theta_b^*\sum_{d=1}^m \mathbf{Y}_d \frac{P_{d|b}}{\mu_d(\theta^*)}\right) + \mathbf{I}\,O(\|\Delta\theta^i\|_2) \\ \rho\left([\tilde{\mathbf{B}}+\tilde{\mathbf{C}}]^{-1}\tilde{\mathbf{C}}\right) &= \rho\left([\mathbf{B}+\mathbf{C}]^{-1}\mathbf{C}\right) + O(\|\Delta\theta^i\|_2).\end{aligned}$$

where, as long as θ^* is in the interior of Θ, $\rho\left([\mathbf{B}+\mathbf{C}]^{-1}\mathbf{C}\right) = \alpha < 1$. Furthermore, since θ^* is a stationary point of (18): $\sum_{d=1}^p \mathbf{Y}_d \frac{P_{d|b}}{\mu_d(\theta^*)} = P_b$. Thus to order $O(\|\Delta\theta^i\|_2)$ (25) is equivalent to:

$$\sum_{b=1}^p P_b\theta_b^*\left(\ln\theta_b^{i+1} - \ln\theta_b^*\right)^2 \quad\leq\quad \alpha\sum_{b=1}^p P_b\theta_b^*\left(\ln\theta_b^i - \ln\theta_b^*\right)^2$$

We thus obtain the following theorem.

Theorem 2 *Assume that the unpenalized ECT EM algorithm specified by (18) converges to the strictly positive limit θ^*. Then, for some sufficiently large positive integer M:*

$$\|\ln\theta^{i+1} - \ln\theta^*\| \leq \alpha\|\ln\theta^i - \ln\theta^*\|, \qquad i \geq M,$$

where $\alpha = \rho([\boldsymbol{B}+\boldsymbol{C}]^{-1}\boldsymbol{C})$, $\boldsymbol{B} = \boldsymbol{B}(\theta^*)$, $\boldsymbol{C} = \boldsymbol{C}(\theta^*)$, the norm $\|\bullet\|$ is defined as:

$$\|u\|^2 \stackrel{def}{=} \sum_{b=1}^p P_b\theta_b^*\; u_b^2, \tag{26}$$

and $P_b \stackrel{def}{=} \sum_{d=1}^m P_{d|b}$.

Lange and Carson (1984) showed that the ECT EM algorithm converges to the maximum likelihood estimate. As long as θ^* is strictly positive, the theorem asserts that in the final iterations of the algorithm the logarithmic differences $\ln\theta^i - \ln\theta^*$ converge monotonically to zero relative to the norm (26).

IV. CONCLUDING COMMENTS

We have presented a general methodology for studying the norm convergence properties of EM-type algorithms. Since Theorem 1 can specify a norm relative to which convergence of a properly implemented EM algorithm must be monotone our results may provide a practical verification tool, similar to checking the increasing-likelihood property, for testing for errors in algorithm implementation. To perform such a test the algorithm should be run to its convergence limit whereby the final iterations can be checked for the norm reducing property.

A weakness of the method given here is that it does not apply to cases where the maximization in the M step is achieved on a boundary of the parameter space. While there are a certain number of such problems where this method will not apply, we believe that the method will nonetheless be useful for a number of applications areas.

Acknowledgement

The authors would like to thank the chair editor, the associate editor, and the anonymous reviewers for their helpful comments and suggestions on this paper.

References

[] Antoniadis, N. and Hero, A. O. (1994). Time delay estimation for filtered Poisson processes using an EM-type algorithm. *Signal Processing*, to appear.

[] Boyles, R. A. (1983). On the convergence properties of the EM algorithm. *J. Royal Statistical Society, Ser. B* **1**, 47–50.

[] Csiszar, I. and Tusnady, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplement Issue No. **1**, 205–237.

[] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Ser. B* **39**, 1–38.

[] De Pierro, A. R. (1993). A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans. on Medical Imaging*, in review.

[] Feder, M., Oppenheim, A., and Weinstein, E. (1989). Maximum likelihood noise cancellation using the EM algorithm. *IEEE Trans. Acoust., Speech, and Sig. Proc.* **37**, 204–216.

[] Fessler, J. A. and Hero, A. O. (1994). Space-alternating generalized EM algorithm. *IEEE Trans. on Signal Processing*, to appear.

[] Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations (2nd Edition).* The Johns Hopkins University Press, Baltimore.

[] Green, P. J. (1990a). On the use of the EM algorithm for penalized likelihood estimation. *J. Royal Statistical Society, Ser. B* **52**, 443–452.

[] Green, P. J. (1990b). Bayesian reconstructions from emission tomography using a modified EM algorithm. *IEEE Trans. on Medical Imaging* **11**, 81–90.

[] Hebert, T. and Leahy, R. (1989). A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. on Medical Imaging* **8**, 194–203.

[] Hebert, T. and Leahy, R. (1992). Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. *IEEE Trans. on Signal Processing* **40**, 2290–2302.

[] ero, A. O. (1992). The influence of the choice of complete data on convergence of E-M type algorithms. Proceedings of 1992 IEEE Workshop on Statistical Signal and Array Processing, Victoria B.C.

[] Hero, A. O. and Fessler, J. A. (1993). Asymptotic convergence properties of EM-type algorithms. Technical Report 282, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor.

[] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis.* Cambridge.

[] Lange, K. and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomography* **8**, 306–316.

[] Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, to appear.

[] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Royal Statistical Society, Ser. B* **44**, 226–233.

[] Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. Royal Statistical Society, Ser. B* **51**, 127–138.

[] Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Assoc.*, **86**, 899–909.

[] Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–728.

[] Miller, M. I. and Snyder, D. L. (1987). The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances. *IEEE Proceedings* **75**, 892–907.

[] Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York.

[] Polak, E. (1971). *Computational methods in optimization: a unified approach.* Academic Press, Orlando.

[] Segal, M., Weinstein, E., and Musicus, B. (1991). Estimate-maximize algorithms for multichannel time delay and signal estimation. *IEEE Trans. Acoust., Speech, and Sig. Proc.* **39**, 1–16.

[] Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. on Medical Imaging* **1**, 113–122.

[] Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Commun. Statist.-Simula. Computa.* **B5**, 55–64.

[] Wei, G. C. and Tanner, M. A. (1990). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Assoc.* **85**, 699–704.

[] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.