Non-Stochastic EM Algorithms in Optimization*

Charles Byrne† ‡

August 30, 2013

Abstract

We consider the problem of maximizing a non-negative function $f : Z \to \mathbb{R}$, where Z is an arbitrary set. We assume that there is $z^* \in Z$ with $f(z^*) \geq f(z)$, for all $z \in Z$. We assume that there is a non-negative function $b : \mathbb{R}^N \times Z \to \mathbb{R}$ such that

$$f(z) = \int b(x, z)dx.$$

Having found z^k, we maximize the function

$$H(z^k, z) = \int b(x, z^k) \log b(x, z)dx$$

to get z^{k+1}. It then follows that the sequence $\{f(z^k)\}$ is increasing and the sequence $\{b(x, z^k)\}$ is asymptotically regular, in the sense of cross-entropy; that is, the sequence $\{KL(b(x, z^k), b(x, z^{k+1}))\}$ converges to zero, where

$$KL(b(x, z^k), b(x, z^{k+1})) = \int \Big(b(x, z^k) \log \frac{b(x, z^k)}{b(x, z^{k+1})} + b(x, z^{k+1}) - b(x, z^k)\Big)dx.$$

This iterative procedure is a particular case of the alternating minimization (AM) method of Csiszár and Tusnády [12]. With additional restrictions suggested by the AM method, it follows that $\{f(z^k)\}$ converges to $f(z^*)$. We consider also a discrete version, in which the variable x takes values in a finite or countably infinite set and the integral is replaced by a sum.

For particular choices of the functions $b(x, z)$ this framework reduces to the well-known "expectation maximization" (EM) algorithm for statistical parameter estimation. It can also be used to find approximate non-negative solutions of non-negative systems of linear equations.

AMS Subject Classification: 65K10, 90C30,65F10. **Key words:** iterative optimization, likelihood maximization, cross-entropy.

*This article will appear in the Journal of Nonlinear and Convex Analysis, 2014.

†Charles_Byrne@uml.edu, Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA 01854, USA

‡I wish to thank Professor Paul Eggermont of the University of Delaware for helpful discussions.

1 Overview

The "expectation maximization" (EM) algorithm is a general framework for maximizing the likelihood in statistical parameter estimation [24], and is always presented in probabilistic terms, involving the maximization of a conditional expected value. The EM algorithm is not really a single algorithm, but a framework for the design of iterative likelihood maximization methods; nevertheless, we shall continue to refer to *the* EM algorithm. As we shall demonstrate in this paper, the essence of the EM algorithm is not stochastic, and leads to a general approach for function maximization, which we call the "generalized" EM (GEM) algorithm. In addition to being more general, this new approach also simplifies much of the development of the EM algorithm itself.

2 A Non-Stochastic EM Algorithm

In this section we present the essential aspects of the EM algorithm without relying on statistical concepts. We shall use these results later to establish important facts about the statistical EM algorithm.

2.1 The Continuous Case

The problem is to maximize a non-negative function $f : Z \to \mathbb{R}$, where Z is an arbitrary set. We assume that there is $z^* \in Z$ with $f(z^*) \geq f(z)$, for all $z \in Z$. We also assume that there is a non-negative function $b : \mathbb{R}^N \times Z \to \mathbb{R}$ such that

$$f(z) = \int b(x, z)dx.$$

Having found z^k, we maximize the function

$$H(z^k, z) = \int b(x, z^k) \log b(x, z)dx \tag{2.1}$$

to get z^{k+1}. The cross-entropy or Kullback-Leibler distance [21] is a useful tool for analyzing the EM algorithm. For positive numbers u and v, the Kullback-Leibler distance from u to v is

$$KL(u, v) = u \log \frac{u}{v} + v - u. \tag{2.2}$$

We also define $KL(0, 0) = 0$, $KL(0, v) = v$ and $KL(u, 0) = +\infty$. The KL distance is extended to nonnegative vectors component-wise, so that for nonnegative vectors

a and b we have

$$KL(a,b) = \sum_{j=1}^{J} KL(a_j, b_j). \tag{2.3}$$

One of the most useful facts about the KL distance is contained in the following lemma; we simplify the notation by setting $b(z) = b(x,z)$.

Lemma 2.1 *For non-negative vectors a and b, with $b_+ = \sum_{j=1}^{J} b_j > 0$, we have*

$$KL(a,b) = KL(a_+, b_+) + KL(a, \frac{a_+}{b_+} b). \tag{2.4}$$

Maximizing $H(z^k, z)$ is equivalent to minimizing

$$G(z^k, z) = KL(b(z^k), b(z)) - f(z), \tag{2.5}$$

where

$$KL(b(z^k), b(z)) = \int KL(b(x, z^k), b(x, z))dx. \tag{2.6}$$

Therefore,

$$-f(z^k) = KL(b(z^k), b(z^k)) - f(z^k) \geq KL(b(z^k), b(z^{k+1})) - f(z^{k+1}),$$

or

$$f(z^{k+1}) - f(z^k) \geq KL(b(z^k), b(z^{k+1})).$$

It then follows that the sequence $\{f(z^k)\}$ is increasing and bounded above, so that the sequence $\{b(x, z^k)\}$ is asymptotically regular, in the sense of cross-entropy; that is, the sequence $\{KL(b(x, z^k), b(x, z^{k+1}))\}$ converges to zero, where

$$KL(b(x, z^k), b(x, z^{k+1})) = \int \Big(b(x, z^k) \log \frac{b(x, z^k)}{b(x, z^{k+1})} + b(x, z^{k+1}) - b(x, z^k) \Big) dx.$$

Without additional restrictions, we cannot conclude that $\{f(z^k)\}$ converges to $f(z^*)$.

We get z^{k+1} by minimizing $G(z^k, z)$. When we minimize $G(z, z^{k+1})$, we get z^{k+1} again. Therefore, we can put the GEM algorithm into the alternating minimization (AM) framework of Csiszár and Tusnády [12], to be discussed further in Section 4.

2.2 The Discrete Case

Again, the problem is to maximize a non-negative function $f : Z \to \mathbb{R}$, where Z is an arbitrary set. As previously, we assume that there is $z^* \in Z$ with $f(z^*) \geq f(z)$, for all $z \in Z$. We also assume that there is a finite or countably infinite set B and a non-negative function $b : B \times Z \to \mathbb{R}$ such that

$$f(z) = \sum_{x \in B} b(x, z).$$

Having found z^k, we maximize the function

$$H(z^k, z) = \sum_{x \in B} b(x, z^k) \log b(x, z) \tag{2.7}$$

to get z^{k+1}.

We set $b(z) = b(x, z)$ again. Maximizing $H(z^k, z)$ is equivalent to minimizing

$$G(z^k, z) = KL(b(z^k), b(z)) - f(z), \tag{2.8}$$

where

$$KL(b(z^k), b(z)) = \sum_{x \in B} KL(b(x, z^k), b(x, z)). \tag{2.9}$$

As previously, we find that the sequence $\{f(z^k)\}$ is increasing, and $\{KL(b(z^k), b(z^{k+1}))\}$ converges to zero. Without additional restrictions, we cannot conclude that $\{f(z^k)\}$ converges to $f(z^*)$.

3 The EM Algorithm

In statistical parameter estimation one typically has an *observable* random vector Y taking values in $\mathbb{R}^N$ that is governed by a probability density function (pdf) or probability function (pf) of the form $f_Y(y|\theta)$, for some value of the parameter vector $\theta \in \Theta$, where Θ is the set of all legitimate values of θ. Our *observed* data consists of one realization y of Y; we do not exclude the possibility that the entries of y are independently obtained samples of a common real-valued random variable. The true vector of parameters is to be estimated by maximizing the likelihood function $L_y(\theta) = f_Y(y|\theta)$ over all $\theta \in \Theta$ to obtain a maximum likelihood estimate, θ_{ML}.

To employ the EM algorithmic approach, it is assumed that there is another related random vector X, which we shall call the *preferred* data, such that, had we been able to obtain one realization x of X, maximizing the likelihood function

$L_x(\theta) = f_X(x|\theta)$ would have been simpler than maximizing the likelihood function $L_y(\theta) = f_Y(y|\theta)$. Of course, we do not have a realization x of X. The basic idea of the EM approach is to estimate x using the current estimate of θ, denoted θ^k, and to use each estimate x^k of x to get the next estimate θ^{k+1}. The EM algorithm based on X generates a sequence $\{\theta^k\}$ of parameter vectors. In decreasing order of importance and difficulty, the goals are these:

- 1. to have the sequence of parameters $\{\theta^k\}$ converging to θ_{ML};
- 2. to have the sequence of functions $\{f_X(x|\theta^k)\}$ converging to $f_X(x|\theta_{ML})$;
- 3. to have the sequence of numbers $\{L_y(\theta^k)\}$ converging to $L_y(\theta_{ML})$;
- 4. to have the sequence of numbers $\{L_y(\theta^k)\}$ increasing.

Our focus here is mainly on the fourth goal, with some discussion of the third goal. We do present some examples for which all four goals are attained. Clearly, the first goal requires a topology on the set Z.

3.1 The Discrete Case

In the discrete case, we assume that Y is a discrete random vector taking values in a finite or countably infinite set A, and governed by probability $f_Y(y|\theta)$. We assume, in addition, that there is a second discrete random vector X, taking values in a finite or countably infinite set B, and a function $h : B \to A$ such that $Y = h(X)$. We define the set

$$h^{-1}(y) = \{x \in B | h(x) = y\}. \tag{3.1}$$

Then we have

$$f_Y(y|\theta) = \sum_{x \in h^{-1}(y)} f_X(x|\theta). \tag{3.2}$$

The conditional probability function for X, given $Y = y$, is

$$f_{X|Y}(x|y,\theta) = \frac{f_X(x|\theta)}{f_Y(y|\theta)}, \tag{3.3}$$

for $x \in h^{-1}(y)$, and zero, otherwise. The so-called E-step of the EM algorithm is then to calculate

$$E((\log f_X(X|\theta)|y,\theta^k) = \sum_{x \in h^{-1}(y)} f_{X|Y}(x|y,\theta^k) \log f_X(x|\theta), \tag{3.4}$$

and the M-step is to maximize this function of θ to obtain θ^{k+1}.

Let $\chi_{h^{-1}(y)}(x)$ be the characteristic function of the set $h^{-1}(y)$, that is,

$$\chi_{h^{-1}(y)}(x) = 1,$$

for $x \in h^{-1}(y)$, and zero, otherwise. With the choices $z = \theta$, $f(z) = f_Y(y|\theta)$, and $b(z) = f_X(x|\theta)\chi_{h^{-1}(y)}(x)$, the discrete EM algorithm fits into the framework of the non-stochastic EM algorithm. Consequently, we may conclude that the sequence $\{f_Y(y|\theta^k)\}$ is increasing, and the sequence

$$\{KL(b(z^k), b(z^{k+1}))\} = \{\sum_{x \in h^{-1}(y)} KL(f_X(x|\theta^k), f_X(x|\theta^{k+1}))\}$$

converges to zero.

3.2 The Continuous Case

We now have a random vector Y taking values in $\mathbb{R}^N$ and governed by the probability density function $f_Y(y|\theta)$. The objective, once again, is to maximize the likelihood function $L_y(\theta) = f_Y(y|\theta)$ to obtain the maximum likelihood estimate of θ.

The conventional formulation of the problem, in the continuous case, presents some difficulties. One assumes that there is a random vector X, usually called the "complete data" , taking values in $\mathbb{R}^M$, where typically $M \geq N$, and often $M > N$, with probability density function $f_X(x|\theta)$, and a function $h : \mathbb{R}^M \to \mathbb{R}^N$ such that $Y = h(X)$. For example, let X_1 and X_2 be independent and uniformly distributed on $[0, \theta]$, $X = (X_1, X_2)$ and $Y = X_1 + X_2 = h(X)$. As is evident from this example, the set $h^{-1}(y)$ consists of all points (x_1, x_2) in $\mathbb{R}^2$ for which $y = x_1 + x_2$, and this set has Lebesgue measure zero in $\mathbb{R}^2$. Consequently, we cannot mimic Equation (3.2) and say that

$$f_Y(y|\theta) = \int_{h^{-1}(y)} f_X(x|\theta)dx.$$

Note that, in this example, the conditional distribution of Y, given $X = (x_1, x_2)$, is the point mass supported on the point $y = x_1 + x_2$, and is independent of the parameter θ.

We need to find a condition on the preferred data X sufficient to reach the fourth goal, that the sequence $\{L_y(\theta^k)\}$ be increasing. As we shall show, in order to have $L_y(\theta^{k+1}) \geq L_y(\theta^k)$, it is sufficient that X satisfy the *acceptability condition* $f_{Y|X}(y|x, \theta) = f_{Y|X}(y|x)$. Our treatment of the EM algorithm for the continuous case differs somewhat from most conventional presentations, by which we mean that given

in [13], repeated in [24], and used in many papers on the subject subsequent to 1977, such as [34] and [25].

For the continuous case, the vector θ^{k+1} is obtained from θ^k by maximizing the conditional expected value

$$E(\log f_X(X|\theta)|y,\theta^k) = \int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx. \tag{3.5}$$

Using

$$f_{X,Y}(x,y|\theta^k) = f_{X|Y}(x|y,\theta^k)f_Y(y|\theta^k),$$

and

$$\log f_X(x|\theta) = \log f_{X,Y}(x,y|\theta) - \log f_{Y|X}(y|x),$$

we find that maximizing $E(\log f_X(x|\theta)|y,\theta^k)$ is equivalent to minimizing

$$H(\theta^k,\theta) = \int f_{X,Y}(x,y|\theta^k)\log f_{X,Y}(x,y|\theta)dx. \tag{3.6}$$

With $z = \theta$, and $b(z) = f_{X,Y}(x,y|\theta)$, this problem fits the framework of the non-stochastic EM algorithm and is equivalent to minimizing

$$G(z^k,z) = KL(b(z^k),b(z)) - f(z).$$

Once again, we may conclude that the likelihood function is increasing and that the sequence $\{KL(b(z^k),b(z^{k+1}))\}$ converges to zero.

4 Alternating Minimization

The iterative step of the GEM algorithm is to minimize $G(z^k,z)$ in Equation (2.5) to get z^{k+1}. If we then minimize $G(z,z^{k+1})$, with respect to z, we get $z = z^{k+1}$ again. Consequently, the GEM algorithm can be viewed as alternating minimization in the sense of Csiszár and Tusnády [12].

4.1 The Framework of Csiszár and Tusnády

Following [12], we take $\Psi(p,q)$ to be a real-valued function of the variables $p \in P$ and $q \in Q$, where P and Q are arbitrary sets. Minimizing $\Psi(p,q^n)$ gives p^{n+1} and minimizing $\Psi(p^{n+1},q)$ gives q^{n+1}, so that

$$\Psi(p^n,q^n) \geq \Psi(p^n,q^{n+1}) \geq \Psi(p^{n+1},q^{n+1}). \tag{4.1}$$

The objective is to find $(\hat{p}, \hat{q})$ such that

$$\Psi(p, q) \geq \Psi(\hat{p}, \hat{q}),$$

for all p and q. In order to show that $\{\Psi(p^n, q^n)\}$ converges to

$$d = \inf_{p \in P, q \in Q} \Psi(p, q)$$

the authors of [12] assume the three- and four-point properties.

If there is a non-negative function $\Delta : P \times P \to \mathbb{R}$ such that

$$\Psi(p, q^{n+1}) - \Psi(p^{n+1}, q^{n+1}) \geq \Delta(p, p^{n+1}), \tag{4.2}$$

then the *three-point property* holds. If

$$\Delta(p, p^n) + \Psi(p, q) \geq \Psi(p, q^{n+1}), \tag{4.3}$$

for all p and q, then the *four-point property* holds. Combining these two inequalities, we have

$$\Delta(p, p^n) - \Delta(p, p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p, q). \tag{4.4}$$

From the inequality in (4.4) it follows easily that the sequence $\{\Psi(p^n, q^n)\}$ converges to d. Suppose this is not the case. Then there are p', q', and $D > d$ with

$$\Psi(p^n, q^n) \geq D > \Psi(p', q') \geq d.$$

From Equation (4.4) we have

$$\Delta(p', p^n) - \Delta(p', p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p', q') \geq D - \Psi(p', q') > 0.$$

But since $\{\Delta(p', p^n)\}$ is a decreasing sequence of positive quantities, successive differences must converge to zero; that is, $\{\Psi(p^{n+1}, q^{n+1})\}$ must converge to $\Psi(p', q')$, which is a contradiction.

The *five-point property* of [12] is obtained by combining (4.2) and (4.3):

$$\Psi(p, q) + \Psi(p, q^{n-1}) \geq \Psi(p, q^n) + \Psi(p^n, q^{n-1}). \tag{4.5}$$

Note that the five-point property does not involve the second function $\Delta(p', p)$. However, assuming that the five-point property holds, it is possible to define $\Delta(p', p)$ so that both the three- and four-point properties hold. Assuming the five-point property, we have

$$\Psi(p, q^{n-1}) - \Psi(p, q^n) \geq \Psi(p^n, q^n) - \Psi(p, q), \tag{4.6}$$

from which we can show easily that $\{\Psi(p^n, q^n)\}$ converges to d.

4.2 Alternating Minimization for the GEM Algorithm

In the GEM algorithm we minimize $G(z^k, z)$ to obtain z^{k+1}. With $P = Q = Z$, $\Psi(p, q) = G(z, w)$, and

$$\Delta(p, p') = \Delta(z, z') = KL(b(z), b(z')), \tag{4.7}$$

we find that the three-point property holds. If, in addition, the four-point property holds, then we can say that the sequence $\{f(z^k)\}$ converges to $f(z^*)$. The four-point property for the GEM is

$$KL(b(z), b(z^k)) + KL(b(z), b(z')) - f(z') \geq KL(b(z), b(z^{k+1})) - f(z^{k+1}), \tag{4.8}$$

for all z and z'.

The Kullback-Leibler distance is an example of a jointly convex Bregman distance. According to a lemma of Eggermont and LaRiccia [14, 15], the four-point property holds for alternating minimization of such distances, using $\Delta(p, p') = KL(p, p')$, provided that the sets P and Q are closed convex subsets of $\mathbb{R}^N$.

In [2] the authors consider the problem of minimizing a function of the form

$$\Lambda(p, q) = \phi(p) + \psi(q) + D_g(p, q), \tag{4.9}$$

where ϕ and ψ are convex and differentiable on $\mathbb{R}^N$, D_g is a Bregman distance, and $P = Q$ is the interior of the domain of g. In [7] it was shown that, when D_g is jointly convex, the function $\Lambda(p, q)$ has the five-point property of [12], which is equivalent to the three- and four-point properties taken together. If Z is a closed, convex subset of $\mathbb{R}^N$, and f is concave and differentiable, then $G(z, w)$ has the form of $\Lambda(p, q)$ in Equation (4.9).

As we saw previously, to have $\Psi(p^n, q^n)$ converging to d, it is sufficient that the five-point property hold. It is conceivable, then, that the five-point property may hold for Bregman distances under somewhat more general conditions than those employed in the Eggermont-LaRiccia Lemma.

5 Sequential Unconstrained Minimization

The GEM algorithm can viewed as a particular case of sequential unconstrained minimization, in which, at the kth step, we minimize the function

$$G_k(z) = s(z) + g_k(z) \tag{5.1}$$

to get z^{k+1}. Here $s : Z \to \mathbb{R}$ is an arbitrary function, and the auxiliary function $g_k(z)$ is non-negative, with $g_k(z^k) = 0$. It then follows that the sequence $\{s(z^k)\}$ is decreasing.

Typically, sequential unconstrained minimization algorithms, such as barrier-function and penalty-function methods, are used to solve constrained minimization problems, with the $g_k(z)$ chosen to impose constraints or penalize violation of the constraints [18]. These methods can also be used to facilitate computation, with the $g_k(z)$ chosen so that z^{k+1} can be expressed in closed form at each step [6].

5.1 SUMMA

In [6] a subclass of sequential unconstrained minimization algorithms called the SUMMA class was presented. In order for a sequential unconstrained minimization method to be in the SUMMA class it is required that the $g_k(z)$ be chosen so that

$$G_k(z) - G_k(z^{k+1}) \geq g_{k+1}(z), \tag{5.2}$$

for all $z \in Z$. For iterations in the SUMMA class it can be shown that the sequence $\{s(z^k)\}$ converges to $\inf_z s(z)$.

In [7] it was shown that alternating minimization methods can be reformulated as sequential unconstrained minimization, and that those with the five-point property are in the SUMMA class. In [8] the quite general forward-backward splitting methods [11] were also shown to be members of the SUMMA class. These notions are discussed in greater detail in [9].

5.2 SUMMA for GEM

For the kth step of the GEM algorithm we minimize $G(z^k, z)$, which can be written as

$$G(z^k, z) = G_k(z) = -f(z) + KL(b(z^k), b(z)), \tag{5.3}$$

so that $s(z) = -f(z)$ and $g_k(z) = KL(b(z^k), b(z))$. Clearly, the GEM fits into the sequential minimization framework. In order for the GEM algorithm to be in the SUMMA class we need

$$KL(b(z^k), b(z)) - f(z) - KL(b(z^k), b(z^{k+1})) + f(z^{k+1}) \geq KL(b(z^{k+1}), b(z)), \tag{5.4}$$

for all $z \in Z$. When this condition holds, we can conclude that the sequence $\{f(z^k)\}$ converges to $f(z^*)$.

We turn now to several examples in which the sequence $\{z^k\}$ converges to a maximizer of $f(z)$.

6 Sums of Independent Poisson Random Variables

The EM is often used with aggregated data. The case of sums of independent Poisson random variables is particularly important.

6.1 Poisson Sums

Let $X_1, ..., X_N$ be independent Poisson random variables with expected value $E(X_n) = \lambda_n$. Let X be the random vector with X_n as its entries, λ the vector whose entries are the λ_n, and $\lambda_+ = \sum_{n=1}^N \lambda_n$. Then the probability function for X is

$$f_X(x|\lambda) = \prod_{n=1}^N \lambda_n^{x_n} \exp(-\lambda_n)/x_n! = \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \tag{6.1}$$

Now let $Y = \sum_{n=1}^N X_n$. Then, the probability function for Y is

$$\begin{aligned} \text{Prob}(Y = y) &= \text{Prob}(X_1 + ... + X_N = y) \\ &= \sum_{x_1+...x_N=y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \end{aligned} \tag{6.2}$$

As we shall see shortly, we have

$$\sum_{x_1+...x_N=y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y! \,. \tag{6.3}$$

Therefore, Y is a Poisson random variable with $E(Y) = \lambda_+$.

When we observe an instance of Y, we can consider the conditional distribution $f_{X|Y}(x|y, \lambda)$ of $\{X_1, ..., X_N\}$, subject to $y = X_1 + ... + X_N$. We have

$$f_{X|Y}(x|y, \lambda) = \frac{y!}{x_1!...x_N!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1} ... \Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \tag{6.4}$$

This is a *multinomial distribution.*

Given y and λ, the conditional expected value of X_n is then

$$E(X_n|y, \lambda) = y\lambda_n/\lambda_+.$$

To see why this is true, consider the marginal conditional distribution $f_{X_1|Y}(x_1|y, \lambda)$ of X_1, conditioned on y and λ, which we obtain by holding x_1 fixed and summing over the remaining variables. We have

$$f_{X_1|Y}(x_1|y, \lambda) = \frac{y!}{x_1!(y - x_1)!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1} \Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-x_1} \sum_{x_2+...+x_N=y-x_1} \frac{(y - x_1)!}{x_2!...x_N!} \prod_{n=2}^N \Big(\frac{\lambda_n}{\lambda'_+}\Big)^{x_n},$$

where

$$\lambda'_+ = \lambda_+ - \lambda_1.$$

As we shall show shortly,

$$\sum_{x_2+...+x_N=y-x_1} \frac{(y-x_1)!}{x_2!...x_N!} \prod_{n=2}^{N} \Big(\frac{\lambda_n}{\lambda'_+}\Big)^{x_n} = 1,$$

so that

$$f_{X_1|Y}(x_1|y,\lambda) = \frac{y!}{x_1!(y-x_1)!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}\Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-x_1}.$$

The random variable X_1 is equivalent to the random number of heads showing in y flips of a coin, with the probability of heads given by λ_1/λ_+. Consequently, the conditional expected value of X_1 is $y\lambda_1/\lambda_+$, as claimed. In the next subsection we look more closely at the multinomial distribution.

6.2 The Multinomial Distribution

When we expand the quantity $(a_1 + ... + a_N)^y$, we obtain a sum of terms, each having the form $a_1^{x_1}...a_N^{x_N}$, with $x_1 + ... + x_N = y$. How many terms of the same form are there? There are N variables a_n. We are to use x_n of the a_n, for each $n = 1, ..., N$, to get $y = x_1 + ... + x_N$ factors. Imagine y blank spaces, each to be filled in by a variable as we do the selection. We select x_1 of these blanks and mark them a_1. We can do that in $\binom{y}{x_1}$ ways. We then select x_2 of the remaining blank spaces and enter a_2 in them; we can do this in $\binom{y-x_1}{x_2}$ ways. Continuing in this way, we find that we can select the N factor types in

$$\binom{y}{x_1}\binom{y-x_1}{x_2}...\binom{y-(x_1+...+x_{N-2})}{x_{N-1}} \tag{6.5}$$

ways, or in

$$\frac{y!}{x_1!(y-x_1)!}...\frac{(y-(x_1+...+x_{N-2}))!}{x_{N-1}!(y-(x_1+...+x_{N-1}))!} = \frac{y!}{x_1!...x_N!}. \tag{6.6}$$

This tells us in how many different sequences the factor variables can be selected. Applying this, we get the multinomial theorem:

$$(a_1 + ... + a_N)^y = \sum_{x_1+...+x_N=y} \frac{y!}{x_1!...x_N!} a_1^{x_1}...a_N^{x_N}. \tag{6.7}$$

Select $a_n = \lambda_n/\lambda_+$. Then,

$$1 = 1^y = \Big(\frac{\lambda_1}{\lambda_+} + ... + \frac{\lambda_N}{\lambda_+}\Big)^y$$

$$= \sum_{x_1+\ldots+x_N=y} \frac{y!}{x_1!\ldots x_N!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1} \ldots \Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \tag{6.8}$$

From this we get

$$\sum_{x_1+\ldots x_N=y} \exp(-\lambda_+) \prod_{n=1}^{N} \lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y! \ . \tag{6.9}$$

7 Poisson Sums in Emission Tomography

Sums of Poisson random variables and the problem of complete versus incomplete data arise in single-photon computed emission tomography (SPECT) (Wernick and Aarsvold (2004) [33]).

7.1 The SPECT Reconstruction Problem

In their 1976 paper Rockmore and Makovski [30] suggested that the problem of reconstructing a tomographic image be viewed as statistical parameter estimation. Shepp and Vardi (1982) [31] expanded on this idea and suggested that the EM algorithm discussed by Dempster, Laird and Rubin (1977) [13] be used for the reconstruction. The region of interest within the body of the patient is discretized into J pixels (or voxels), with $\lambda_j \geq 0$ the unknown amount of radionuclide within the jth pixel; we assume that λ_j is also the expected number of photons emitted from the jth pixel during the scanning time. Emitted photons are detected at any one of I detectors outside the body, with $y_i > 0$ the photon count at the ith detector. The probability that a photon emitted at the jth pixel will be detected at the ith detector is P_{ij}, which we assume is known; the overall probability of detecting a photon emitted from the jth pixel is $s_j = \sum_{i=1}^{I} P_{ij} > 0$.

7.1.1 The Preferred Data

For each i and j the random variable X_{ij} is the number of photons emitted from the jth pixel and detected at the ith detector; the X_{ij} are assumed to be independent and $P_{ij}\lambda_j$-Poisson. With x_{ij} a realization of X_{ij}, the vector x with components x_{ij} is our preferred data. The pdf for this preferred data is a probability vector, with

$$f_X(x|\lambda) = \prod_{i=1}^{I} \prod_{j=1}^{J} \exp^{-P_{ij}\lambda_j} (P_{ij}\lambda_j)^{x_{ij}}/x_{ij}! \ . \tag{7.1}$$

Given an estimate λ^k of the vector λ and the restriction that $Y_i = \sum_{j=1}^{J} X_{ij}$, the random variables $X_{i1}, ..., X_{iJ}$ have the multinomial distribution

$$\text{Prob}(x_{i1}, ..., x_{iJ}) = \frac{y_i!}{x_{i1}! \cdots x_{iJ}!} \prod_{j=1}^{J} \Big(\frac{P_{ij}\lambda_j}{(P\lambda)_i}\Big)^{x_{ij}}.$$

Therefore, the conditional expected value of X_{ij}, given y and λ^k, is

$$E(X_{ij}|y, \lambda^k) = \lambda_j^k P_{ij}\Big(\frac{y_i}{(P\lambda^k)_i}\Big),$$

and the conditional expected value of the random variable

$$\log f_X(X|\lambda) = \sum_{i=1}^{I}\sum_{j=1}^{J}(-P_{ij}\lambda_j) + X_{ij}\log(P_{ij}\lambda_j) + \text{constants}$$

becomes

$$E(\log f_X(X|\lambda)|y, \lambda^k) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left((-P_{ij}\lambda_j) + \lambda_j^k P_{ij}\Big(\frac{y_i}{(P\lambda^k)_i}\Big)\log(P_{ij}\lambda_j)\right),$$

omitting terms that do not involve the parameter vector λ. In the EM algorithm, we obtain the next estimate λ^{k+1} by maximizing $E(\log f_X(X|\lambda)|y, \lambda^k)$.

The log likelihood function for the preferred data X (omitting constants) is

$$LL_x(\lambda) = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(-P_{ij}\lambda_j + X_{ij}\log(P_{ij}\lambda_j)\Big). \tag{7.2}$$

Of course, we do not have the complete data.

7.1.2 The Incomplete Data

What we do have are the y_i, values of the random variables

$$Y_i = \sum_{j=1}^{J} X_{ij}; \tag{7.3}$$

this is the given data. These random variables are also independent and $(P\lambda)_i$-Poisson, where

$$(P\lambda)_i = \sum_{j=1}^{J} P_{ij}\lambda_j.$$

The log likelihood function for the given data is

$$LL_y(\lambda) = \sum_{i=1}^{I}\Big(-(P\lambda)_i + y_i\log((P\lambda)_i)\Big). \tag{7.4}$$

Maximizing $LL_x(\lambda)$ in Equation (7.2) is easy, while maximizing $LL_y(\lambda)$ in Equation (7.4) is harder and requires an iterative method.

The EM algorithm involves two steps: in the E-step we compute the conditional expected value of $LL_x(\lambda)$, conditioned on the data vector y and the current estimate λ^k of λ; in the M-step we maximize this conditional expected value to get the next λ^{k+1}. Putting these two steps together, we have the following EMML iteration:

$$\lambda_j^{k+1} = \lambda_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\lambda^k)_i}. \tag{7.5}$$

For any positive starting vector λ^0, the sequence $\{\lambda^k\}$ converges to a maximizer of $LL_y(\lambda)$, over all non-negative vectors λ.

Note that, because we are dealing with finite probability vectors in this example, it is a simple matter to conclude that

$$f_Y(y|\lambda) = \sum_{x \in h^{-1}(y)} f_X(x|\lambda). \tag{7.6}$$

7.2 Using the KL Distance

In this subsection we assume, for notational convenience, that the system $y = P\lambda$ has been normalized so that $s_j = 1$ for each j. Maximizing $E(\log f_X(X|\lambda)|y, \lambda^k)$ is equivalent to minimizing $KL(r(\lambda^k), q(\lambda))$, where $r(\lambda)$ and $q(\lambda)$ are I by J arrays with entries

$$r(\lambda)_{ij} = \lambda_j P_{ij} \Big(\frac{y_i}{(P\lambda)_i} \Big),$$

and

$$q(\lambda)_{ij} = \lambda_j P_{ij}.$$

In terms of our previous notation we identify $r(\lambda)$ with $b(\theta)$, and $q(\lambda)$ with $f(\theta)$. The set $\mathcal{F}(\Theta)$ of all $f(\theta)$ is now a convex set and the four-point property of [12] holds. The iterative step of the EMML algorithm is then

$$\lambda_j^{k+1} = \lambda_j^k \sum_{i=1}^{I} P_{i,j} \frac{y_i}{(P\lambda^k)_i}. \tag{7.7}$$

The sequence $\{\lambda^k\}$ converges to a maximizer λ_{ML} of the likelihood for any positive starting vector.

As we noted previously, before we can discuss the possible convergence of the sequence $\{\lambda^k\}$ of parameter vectors to a maximizer of the likelihood, it is necessary to have a notion of convergence in the parameter space. For the problem in this section,

the parameter vectors λ are non-negative. Proof of convergence of the sequence $\{\lambda^k\}$ depends heavily on the following identities [4]:

$$KL(y, P\lambda^k) - KL(y, P\lambda^{k+1}) = KL(r(\lambda^k), r(\lambda^{k+1})) + KL(\lambda^{k+1}, \lambda^k); \tag{7.8}$$

and

$$KL(\lambda_{ML}, \lambda^k) - KL(\lambda_{ML}, \lambda^{k+1}) \geq KL(y, P\lambda^k) - KL(y, P\lambda_{ML}). \tag{7.9}$$

Any likelihood maximizer λ_{ML} is also a non-negative minimizer of the KL distance $KL(y, P\lambda)$, so the EMML algorithm can be thought of as a method for finding a non-negative solution (or approximate solution) for a system $y = P\lambda$ of linear equations in which $y_i > 0$ and $P_{ij} \geq 0$ for all indices. This will be helpful when we consider mixture problems.

8 Finite Mixture Problems

Estimating the combining proportions in probabilistic mixture problems shows that there are meaningful examples of our acceptable-data model, and provides important applications of likelihood maximization.

8.1 Mixtures

We say that a random vector V taking values in $\mathbb{R}^D$ is a *finite mixture* (see [16, 29]) if there are probability density functions or probabilities f_j and numbers $\theta_j \geq 0$, for $j = 1, ..., J$, such that the probability density function or probability function for V has the form

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v), \tag{8.1}$$

for some choice of the $\theta_j \geq 0$ with $\sum_{j=1}^{J} \theta_j = 1$. We shall assume, without loss of generality, that $D = 1$.

8.2 The Likelihood Function

The data are N realizations of the random variable V, denoted v_n, for $n = 1, ..., N$, and the given data is the vector $y = (v_1, ..., v_N)$. The column vector $\theta = (\theta_1, ..., \theta_J)^T$

is the generic parameter vector of mixture combining proportions. The likelihood function is

$$L_y(\theta) = \prod_{n=1}^{N} \Big(\theta_1 f_1(v_n) + ... + \theta_J f_J(v_n)\Big). \tag{8.2}$$

Then the log likelihood function is

$$LL_y(\theta) = \sum_{n=1}^{N} \log \Big(\theta_1 f_1(v_n) + ... + \theta_J f_J(v_n)\Big).$$

With u the column vector with entries $u_n = 1/N$, and P the matrix with entries $P_{nj} = f_j(v_n)$, we define

$$s_j = \sum_{n=1}^{N} P_{nj} = \sum_{n=1}^{N} f_j(v_n).$$

Maximizing $LL_y(\theta)$ is equivalent to minimizing

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J} (1 - s_j)\theta_j. \tag{8.3}$$

8.3 A Motivating Illustration

To motivate such mixture problems, we imagine that each data value is generated by first selecting one value of j, with probability θ_j, and then selecting a realization of a random variable governed by $f_j(v)$. For example, there could be J bowls of colored marbles, and we randomly select a bowl, and then randomly select a marble within the selected bowl. For each n the number v_n is the numerical code for the color of the nth marble drawn. In this illustration we are using a mixture of probability functions, but we could have used probability density functions.

8.4 The Acceptable Data

We approach the mixture problem by creating acceptable data. We imagine that we could have obtained $x_n = j_n$, for $n = 1, ..., N$, where the selection of v_n is governed by the function $f_{j_n}(v)$. In the bowls example, j_n is the number of the bowl from which the nth marble is drawn. The acceptable-data random vector is $X = (X_1, ..., X_N)$, where the X_n are independent random variables taking values in the set $\{j = 1, ..., J\}$. The value j_n is one realization of X_n. Since our objective is to estimate the true θ_j, the values v_n are now irrelevant. Our ML estimate of the true θ_j is simply the proportion of times $j = j_n$. Given a realization x of X, the conditional pdf or pf of Y does not involve the mixing proportions, so X is acceptable. Notice also that it is not possible to calculate the entries of y from those of x; the model $Y = h(X)$ does not hold.

8.5 The Mix-EM Algorithm

Using this acceptable data, we derive the EM algorithm, which we call the Mix-EM algorithm.

With N_j denoting the number of times the value j occurs as an entry of x, the likelihood function for X is

$$L_x(\theta) = f_X(x|\theta) = \prod_{j=1}^{J} \theta_j^{N_j}, \tag{8.4}$$

and the log likelihood is

$$LL_x(\theta) = \log L_x(\theta) = \sum_{j=1}^{J} N_j \log \theta_j. \tag{8.5}$$

Then

$$E(\log L_x(\theta)|y, \theta^k) = \sum_{j=1}^{J} E(N_j|y, \theta^k) \log \theta_j. \tag{8.6}$$

To simplify the calculations in the E-step we rewrite $LL_x(\theta)$ as

$$LL_x(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{J} X_{nj} \log \theta_j, \tag{8.7}$$

where $X_{nj} = 1$ if $j = j_n$ and zero otherwise. Then we have

$$E(X_{nj}|y, \theta^k) = \text{prob}\,(X_{nj} = 1|y, \theta^k) = \frac{\theta_j^k f_j(v_n)}{f(v_n|\theta^k)}. \tag{8.8}$$

The function $E(LL_x(\theta)|y, \theta^k)$ becomes

$$E(LL_x(\theta)|y, \theta^k) = \sum_{n=1}^{N} \sum_{j=1}^{J} \frac{\theta_j^k f_j(v_n)}{f(v_n|\theta^k)} \log \theta_j. \tag{8.9}$$

Maximizing with respect to θ, we get the iterative step of the Mix-EM algorithm:

$$\theta_j^{k+1} = \frac{1}{N} \theta_j^k \sum_{n=1}^{N} \frac{f_j(v_n)}{f(v_n|\theta^k)}. \tag{8.10}$$

We know from our previous discussions that, since the preferred data X is acceptable, likelihood is increasing for this algorithm. We shall go further now, and show that the sequence of probability vectors $\{\theta^k\}$ converges to a maximizer of the likelihood.

8.6 Convergence of the Mix-EM Algorithm

As we noted earlier, maximizing the likelihood in the mixture case is equivalent to minimizing

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}(1 - s_j)\theta_j,$$

over probability vectors θ. It is easily shown that, if $\hat{\theta}$ minimizes $F(\theta)$ over all non-negative vectors θ, then $\hat{\theta}$ is a probability vector. Therefore, we can obtain the maximum likelihood estimate of θ by minimizing $F(\theta)$ over non-negative vectors θ.

The following theorem is found in [5].

Theorem 8.1 *Let u be any positive vector, P any non-negative matrix with $s_j > 0$ for each j, and*

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}\beta_j KL(\gamma_j, \theta_j).$$

If $s_j + \beta_j > 0$, $\alpha_j = s_j/(s_j + \beta_j)$, and $\beta_j\gamma_j \geq 0$, for all j, then the iterative sequence given by

$$\theta_j^{k+1} = \alpha_j s_j^{-1}\theta_j^k\Big(\sum_{n=1}^{N} P_{n,j}\frac{u_n}{(P\theta^k)_n}\Big) + (1 - \alpha_j)\gamma_j \tag{8.11}$$

converges to a non-negative minimizer of $F(\theta)$.

With the choices $u_n = 1/N$, $\gamma_j = 0$, and $\beta_j = 1 - s_j$, the iteration in Equation (8.11) becomes that of the Mix-EM algorithm. Therefore, the sequence $\{\theta^k\}$ converges to the maximum likelihood estimate of the mixing proportions.

9 More on Convergence

There is a mistake in the proof of convergence given in Dempster, Laird, and Rubin (1977) [13]. Wu (1983) [34] and Boyles (1983) [3] attempted to repair the error, but also gave examples in which the EM algorithm failed to converge to a global maximizer of likelihood. In Chapter 3 of McLachlan and Krishnan (1997) [24] we find the basic theory of the EM algorithm, including available results on convergence and the rate of convergence. Because many authors rely on the continuous version of Equation (3.2), it is not clear that these results are valid in the generality in which they are presented. There appears to be no single convergence theorem that is relied on universally; each application seems to require its own proof of convergence. When the use of the EM

algorithm was suggested for SPECT and PET, it was necessary to prove convergence of the resulting iterative algorithm in Equation (7.5), as was eventually achieved in a sequence of papers (Shepp and Vardi (1982) [31], Lange and Carson (1984) [22], Vardi, Shepp and Kaufman (1985) [32], Lange, Bahn and Little (1987) [23], and [4]). When the EM algorithm was applied to list-mode data in SPECT and PET (Barrett, White, and Parra (1997) [1, 28], and Huesman et al. (2000) [20]), the resulting algorithm differed slightly from that in Equation (7.5) and a proof of convergence was provided in [5]. The convergence theorem in [5] also establishes the convergence of the iteration in Equation (8.10) to the maximum-likelihood estimate of the mixing proportions.

10 Open Questions

As we have seen, the conventional formulation of the EM algorithm presents difficulties when probability density functions are involved. We have shown here that the use of acceptable preferred data can be helpful in resolving this issue, but other ways may also be useful.

Proving convergence of the sequence $\{\theta^k\}$ appears to involve the selection of an appropriate topology for the parameter space Θ. While it is common to assume that Θ is a subset of Euclidean space and that the usual norm should be used to define distance, it may be helpful to tailor the metric to the nature of the parameters. In the case of Poisson sums, for example, the parameters are non-negative vectors and we found that the cross-entropy distance is more appropriate. Even so, additional assumptions appear necessary before convergence of the $\{\theta^k\}$ can be established. To simplify the analysis, it is often assumed that cluster points of the sequence lie in the interior of the set Θ, which is not a realistic assumption in some applications.

It may be wise to consider, instead, convergence of the functions $f_X(x|\theta^k)$, or maybe even to identify the parameters θ with the functions $f_X(x|\theta)$. Proving convergence to $L_y(\theta_{ML})$ of the likelihood values $L_y(\theta^k)$ is also an option.

11 Conclusion

The essential aspects of the EM algorithm are non-stochastic and are more simply described in terms of a more general optimization procedure that we call the generalized EM (GEM) method. The EM algorithm for the discrete case of probabilities fits into the GEM framework, which allows us to conclude that likelihood is increasing.

Difficulties with the conventional formulation of the EM algorithm in the con-

tinuous case of probability density functions (pdf) has prompted us to adopt a new definition, that of acceptable data. This new formulation of the EM algorithm for the continuous case then fits into the GEM framework as well, from which we conclude, once again, that likelihood is increasing.

In both the discrete and continuous cases, the two steps of the EM algorithm can be viewed as alternating minimization, along the lines investigated by Csiszár and Tusnády [12]. The GEM can also be viewed as sequential unconstrained minimization. If the five-point property holds in the AM formulation, or the SUMMA condition holds in the sequential unconstrained minimization formulation, then the sequence $\{f(z^k)\}$ converges to $f(z^*)$.

References

1. Barrett, H., White, T., and Parra, L. (1997) "List-mode likelihood." *J. Opt. Soc. Am. A* **14**, pp. 2914–2923.

2. Bauschke, H., Combettes, P., and Noll, D. (2006) "Joint minimization with alternating Bregman proximity operators." *Pacific Journal of Optimization*, **2**, pp. 401–424.

3. Boyles, R. (1983) "On the convergence of the EM algorithm." *Journal of the Royal Statistical Society B*, **45**, pp. 47–50.

4. Byrne, C. (1993) "Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Transactions on Image Processing* **IP-2**, pp. 96–103.

5. Byrne, C. (2001) "Likelihood maximization for list-mode emission tomographic image reconstruction." *IEEE Transactions on Medical Imaging* **20(10)**, pp. 1084–1092.

6. Byrne, C. (2008) "Sequential unconstrained minimization algorithms for constrained optimization." *Inverse Problems*, **24(1)**, article no. 015013.

7. Byrne, C. (2012) "Alternating and sequential unconstrained minimization algorithms." *Journal of Optimization Theory and Applications*, **156(2)**, and DOI 10.1007/s1090134-2.

8. Byrne, C. (2013) "An elementary proof of convergence of the forward-backward splitting algorithm." to appear in the *Journal of Nonlinear and Convex Analysis*.

9. Byrne, C. (2014) *Iterative Optimization in Inverse Problems*, Taylor and Francis.

10. Byrne, C., and Eggermont, P. (2011) "EM Algorithms." in *Handbook of Mathematical Methods in Imaging*, Otmar Scherzer, ed., Springer-Science.

11. Combettes, P., and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

12. Csiszár, I. and Tusnády, G. (1984) "Information geometry and alternating minimization procedures." *Statistics and Decisions* **Supp. 1**, pp. 205–237.

13. 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, Series B* **37**, pp. 1–38.

14. Eggermont, P.P.B., LaRiccia, V.N. (1995) "Smoothed maximum likelihood density estimation for inverse problems." Annals of Statistics **23**, pp. 199–220.

15. Eggermont, P., and LaRiccia, V. (2001) *Maximum Penalized Likelihood Estimation, Volume I: Density Estimation*, New York: Springer.

16. Everitt, B. and Hand, D. (1981) *Finite Mixture Distributions*, London: Chapman and Hall.

17. Fessler, J., Ficaro, E., Clinthorne, N., and Lange, K. (1997) "Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction." *IEEE Transactions on Medical Imaging*, **16 (2)**, pp. 166–175.

18. Fiacco, A., and McCormick, G. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Philadelphia, PA: SIAM Classics in Mathematics (reissue).

19. Hogg, R., McKean, J., and Craig, A. (2004) *Introduction to Mathematical Statistics*, 6th edition, Prentice Hall.

20. Huesman, R., Klein, G., Moses, W., Qi, J., Ruetter, B., and Virador, P. (2000) "List-mode maximum likelihood reconstruction applied to positron emission mammography (PEM) with irregular sampling." *IEEE Transactions on Medical Imaging* **19 (5)**, pp. 532–537.

21. Kullback, S. and Leibler, R. (1951) "On information and sufficiency." *Annals of Mathematical Statistics* **22**, pp. 79–86.

22. Lange, K. and Carson, R. (1984) "EM reconstruction algorithms for emission and transmission tomography." *Journal of Computer Assisted Tomography* **8**, pp. 306–316.

23. Lange, K., Bahn, M. and Little, R. (1987) "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography." *IEEE Trans. Med. Imag.* **MI-6(2)**, pp. 106–114.

24. McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*, New York: John Wiley and Sons, Inc.

25. Meng, X., and Pedlow, S. (1992) "EM: a bibliographic review with missing articles." *Proceedings of the Statistical Computing Section, American Statistical Association*, American Statistical Association, Alexandria, VA.

26. Meng, X., and van Dyk, D. (1997) "The EM algorithm- An old folk-song sung to a fast new tune." *J. R. Statist. Soc. B*, **59(3)**, pp. 511–567.

27. Narayanan, M., Byrne, C. and King, M. (2001) "An interior point iterative maximum-likelihood reconstruction algorithm incorporating upper and lower bounds with application to SPECT transmission imaging." *IEEE Transactions on Medical Imaging* **TMI-20 (4)**, pp. 342–353.

28. Parra, L. and Barrett, H. (1998) "List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET." *IEEE Transactions on Medical Imaging* **17**, pp. 228–235.

29. Redner, R., and Walker, H. (1984) "Mixture Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review*, **26(2)**, pp. 195–239.

30. Rockmore, A., and Macovski, A. (1976) "A maximum likelihood approach to emission image reconstruction from projections." *IEEE Transactions on Nuclear Science*, **NS-23**, pp. 1428–1432.

31. Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Transactions on Medical Imaging*, **MI-1**, pp. 113–122.

32. Vardi, Y., Shepp, L.A. and Kaufman, L. (1985) "A statistical model for positron emission tomography." *Journal of the American Statistical Association* **80**, pp. 8–20.

33. Wernick, M. and Aarsvold, J., (eds.) (2004) *Emission Tomography: The Fundamentals of PET and SPECT*, San Diego: Elsevier Academic Press.

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