EM Algorithms

Charles Byrne (Charles_Byrne@uml.edu)
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854, USA

March 6, 2011

Abstract

The EM algorithm is not a single algorithm, but a framework for the design of iterative likelihood maximization methods for parameter estimation. Any algorithm based on the EM framework we refer to as an "EM algorithm". Because there is no inclusive theory that applies to all EM algorithms, the subject is a work in progress, and we find it appropriate to approach the subject through examples, each chosen to illustrate an important aspect of the subject.

We begin on a positive note with the EM algorithm for finite mixtures of Poisson random variables, which arises in single-photon emission tomography (SPECT). In this case, for which we have a nearly complete theory of convergence, there emerges quite naturally a topology on the parameter space based on the cross-entropy, or Kullback-Leibler, distance between non-negative vectors. Because the distributions involve discrete probabilities, there is no difficulty in describing the EM framework.

Our next example involves censored exponentially distributed data. Now the distributions involve probability density functions, which complicates the definition of an EM algorithm. This gives us an opportunity to discuss this issue and to illustrate how it might be resolved.

Our third example involves the finite sum of independent, uniformly distributed random variables. Once again, the use of probability density functions complicates the statement of the EM algorithm, but provides an opportunity to illustrate yet another method for overcoming this problem. As an added bonus, it provides an example of a sequence generated by an EM algorithm that does not converge to the maximum likelihood (ML) solution.

Our final example involves finite mixtures of either probability functions or probability density functions. The theory for the discrete case follows closely that for the mixtures of Poisson random variables. The theory for the continuous case is novel in ways reminiscent of that for the list-mode EM algorithm in SPECT.

Figuring out how to formulate the EM algorithm and proving convergence are not our only concerns. Convergent EM algorithms are well known to converge slowly, and there is considerable attention paid to accelerated methods. It

also is sometimes the case, particularly in image reconstruction from noisy data, that the ML solution is not particularly helpful. In such cases, it is common to regularize by adding a penalty function to the likelihood. Finding efficient and effective methods for generating regularized ML solutions is another important research area.

As such, this subject remains a work in progress. The topology on the space of parameters varies with the particular application and there is no general theory of convergence; even the basic formulation of the EM algorithm is inadequate when probability density functions are involved. Consequently, each application requires its own theory. In some cases, such as image reconstruction, the maximizer of likelihood may not provide a useful solution and regularization is needed. It is well known that EM algorithms can be slow to converge, or fail to converge at all. Research is ongoing to find appropriate assumptions to guarantee convergence and effective methods to accelerate convergence.

1 Introduction

The EM algorithm is not a single algorithm, but a framework for the design of iterative likelihood maximization methods for parameter estimation. We begin with the usual formulation of the EM algorithm, and then outline several of the issues we shall treat in more detail subsequently.

1.1 What is the EM Algorithm?

We have one realization $y \in R^N$ of the random vector Y governed by the probability function (pf) or probability density function (pdf) $g(y|x)$, where x is a vector of parameters to be determined. The maximum-likelihood estimate of x is a vector x_{ML} that maximizes the likelihood function

$$L(x) = g(y|x), \tag{1.1}$$

over $x \in X$, the set of all admissible values of the parameter vector. In those situations in which an x_{ML} maximizing $L(x)$ cannot be obtained in closed form, one can employ an iterative method, such as the Newton-Raphson algorithm, to obtain a sequence $\{x^k\}$ that, in the best cases, converges to some x_{ML}. The EM algorithm, or, more precisely, the EM "algorithm", since it is really more a template for the design of algorithms, is a method for generating such iterative procedures.

To employ the EM algorithm, we imagine that the given data vector y is somehow *incomplete*, that there is another random vector Z related to Y, the *complete data*, taking values in R^M and governed by the pf or pdf $f(z|x)$, such that, had we been able to obtain one realization z of Z, maximizing $f(z|x)$ would have been simpler than maximizing $g(y|x)$. The basic idea of the EM algorithm is to exploit this greater calculational simplicity.

When the deterministic variable z is replaced in $f(z|x)$ by the random variable Z, we obtain random variables $f(Z|x)$ and $\log f(Z|x)$, for each fixed x. Having found x^k, we calculate, for each fixed x, the expected value of the random variable $\log f(Z|x)$, conditioned on y and x^k. What we get is then a deterministic function of x, which we then maximize to get x^{k+1}.

In some applications, the complete data Z arises naturally from the problem, while in other cases the user must imagine complete data, with respect to which the obtained data is incomplete. This choice in selecting the complete data can be helpful in speeding up the algorithm.

The EM algorithm proceeds in two steps: given the data y and the current estimate x^k, the (E) step of the EM algorithm is to calculate $E(\log f(Z|x)|y, x^k)$, the conditional expected value of $\log f(Z|x)$, given y and x^k, and the (M) step is to maximize $E(\log f(Z|x)|y, x^k)$ with respect to x to obtain x^{k+1}. In order to implement this method, we need to postulate the precise manner by which Y depends on Z, and $g(y|x)$ on $f(z|x)$, so that we can calculate $E(\log f(Z|x)|y, x^k)$.

1.2 What is $E(\log f(Z|x)|y, x^k)$?

Most papers and books on the EM algorithm assume that there is a function $h : R^M \to R^N$, such that $Y = h(Z)$ and that

$$g(y|x) = \int_{\mathcal{Z}(y)} f(z|x)dz, \tag{1.2}$$

where

$$\mathcal{Z}(y) = h^{-1}(\{y\}) = \{z|h(z) = y\}. \tag{1.3}$$

This is fine in the discrete case, when both $g(y|x)$ and $f(z|x)$ are probability functions and the integral is replaced by a sum, but it can be problematic in the continuous case, when they are probability density functions. For example, in the latter case, it is quite possible that $N < M$ and for the set $\mathcal{Z}(y)$ to have measure zero in R^M, in which case the integral in Equation (1.2) has the value zero. Reformulating the EM algorithm to avoid this difficulty is one of the topics we shall take up in what follows.

1.3 The Key Issue

In order to develop a reasonable theory of the EM algorithm, it seems that we must somehow give meaning to the integral in Equation (1.2). As a reminder that the integral is not necessarily the ordinary integral on R^N, and will vary from one application to another, we shall rewrite Equation (1.2) omitting the "dz" , so that

$$g(y|x) = \int_{\mathcal{Z}(y)} f(z|x). \tag{1.4}$$

As we shall see, different applications will require different interpretations of this integral. In each case, however, if we can find a meaning for the integral, with respect to which Equation (1.4) is valid, then we can define the conditional pdf of the random vector Z, conditioned on y, as the function $b(z|y,x)$, defined for $z \in \mathcal{Z}(y)$ by

$$b(z|y,x) = f(z|x)/g(y|x). \tag{1.5}$$

It will then follow from Equation (1.4) that $b(z|y,x)$ is a pdf on the set $\mathcal{Z}(y)$, that is,

$$\int_{\mathcal{Z}(y)} b(z|y,x) = 1, \tag{1.6}$$

where the integral is interpreted as required by the problem. Then we have

$$E(\log f(Z|x)|y,x^k) = \int_{\mathcal{Z}(y)} b(z|y,x^k) \log f(z|y,x). \tag{1.7}$$

This is the (E)-step of the EM algorithm.

In the (M)-step of the EM algorithm, we maximize

$$E(\log f(Z|x)|y,x^k) = \int_{\mathcal{Z}(y)} b(z|y,x^k) \log f(z|y,x)$$

as a function of x, to get the next iterate, $x = x^{k+1}$. Before we develop the theory further, let us consider some examples of incomplete- and complete-data models.

1.4 Other Issues

It is well known that convergent EM algorithms can be slow to converge. Considerable effort has been spent on accelerating EM algorithms. One approach is to select the complete data in a way that leads to fast convergence. A good source is the 1997 paper by Meng and van Dyk [38], which also includes commentary by other experts in the field. We shall discuss briefly the issue of acceleration.

Another issue we shall touch on is the need for regularization. It can happen that the likelihood maximizer is not a useful solution; this occurs, in particular, in image processing with noisy data. In such cases, a penalty function can be added to the likelihood function to obtain a regularized solution. We shall mention some ways in which this can be done.

Convergence is the final issue we shall consider. Before we can speak of convergence of the sequence $\{x^k\}$ we need a topology on the parameter set X. It is common to assume that X is a subset of a Euclidean space, with the metric induced by the 2-norm of vectors. However, as we shall see, it is sometimes more useful to permit other distances, such as the cross-entropy, or Kullback-Leibler distance [31] between non-negative vectors. Even without a topology on X, we can ask if the sequence $\{L(x^k)\}$ converges, or if the sequence $\{f(z|x^k)\}$ converges.

To begin our discussion on a positive note, we start with the discrete case, in which the random vectors have discrete values and only probability functions are involved.

2 The Discrete Case

When Z is a discrete random variable, we do have

$$g(y|x) = \sum_{z\in\mathcal{Z}(y)} f(z|x). \tag{2.1}$$

The conditional probability function for Z, given $Y = y$, is defined by

$$b(z|y,x) = f(z|x)/g(y|x), \tag{2.2}$$

for z in the set $\mathcal{Z}(y)$, and $b(z|y,x) = 0$ otherwise. Then we have

$$f(z|x) = b(z|y,x)g(y|x), \tag{2.3}$$

for all $z \in \mathcal{Z}(y)$.

2.1 The (E) Step

Given the current estimate x^k of x, we compute the conditional expected value of the random variable $\log f(Z|x)$, conditioned on $Y = y$ and using x^k as the true value of the parameter; this is the (E) step. This gives

$$E(\log f(Z|x)|y, x^k) = \sum_z (\log f(z|x))b(z|y,x^k). \tag{2.4}$$

2.2 The (M) Step

The (M) step is to maximize $E(\log f(Z|x)|y, x^k)$ with respect to the variable x, to get the next estimate x^{k+1}. This means that we maximize

$$\sum_{z\in\mathcal{Z}(y)} (\log f(z|x))b(z|y, x^k),$$

over all admissible values of the parameter vector x. We can simplify the analysis using the Kullback-Leibler cross-entropy distance.

2.3 Cross-Entropy or the Kullback-Leibler Distance

For positive numbers u and v, the Kullback-Leibler distance from u to v is

$$KL(u, v) = u \log \frac{u}{v} + v - u. \tag{2.5}$$

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

$$KL(a, b) = \sum_{j=1}^{J} KL(a_j, b_j). \tag{2.6}$$

One of the most useful facts about the KL distance is contained in the following lemma.

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

$$KL(a, b) = KL(a_+, b_+) + KL(a, \frac{a_+}{b_+}b). \tag{2.7}$$

2.4 Using the KL Distance

To simplify notation, let us write

$$b(x^k) = b(z|y, x^k),$$

and

$$f(x) = f(z|x).$$

Then the KL distance from the function of z denoted $b(x^k)$ to the function of z denoted $f(x)$ is

$$KL(b(x^k), f(x)) = \sum_z \Big(b(z|y, x^k) \log \big(\frac{b(z|y, x^k)}{f(z|x)}\big) + f(z|x) - b(z|y, x^k)\Big). \tag{2.8}$$

Minimizing $KL(b(x^k), f(x))$ is equivalent to maximizing $E(\log(f(Z|x))|y, x^k)$, since

$$\sum_z f(z|x) = 1,$$

for each x. We have the following helpful results.

Lemma 2.2 *For any x we have*

$$KL(b(x), f(x)) = -\log g(y|x). \tag{2.9}$$

Proof: Since $\sum_z f(x) = 1$ and $\sum_z b(x) = 1$, we know that

$$KL(b(x), f(x)) = \sum_z b(z|y, x) \log \frac{b(z|y, x)}{f(z|x)} = \sum_{z \in \mathcal{Z}(y)} b(z|y, x) \log \frac{b(z|y, x)}{f(z|x)}.$$

But, by Equation (2.3), we have

$$\log \frac{b(z|y, x)}{f(z|x)} = -\log g(y|x),$$

for all $z \in \mathcal{Z}(y)$. Therefore,

$$KL(b(x), f(x)) = -\log g(y|x) \sum_{z \in \mathcal{Z}(y)} b(z|y, x) = -\log g(y|x).$$

▮

Corollary 2.1 *The minimizers of $KL(b(x), f(x))$ are the maximizers of the likelihood.*

Lemma 2.3 *For any w, we have*

$$KL(b(w), f(x)) = KL(b(x), f(x)) + KL(b(w), b(x)). \tag{2.10}$$

Proof: The proof is a simple calculation. ▮

From what we have learned so far, we can say the following:

$$-\log g(y|x^k) = KL(b(x^k), f(x^k)) \geq KL(b(x^k), f(x^{k+1}))$$

$$= KL(b(x^{k+1}), f(x^{k+1})) + KL(b(x^k), b(x^{k+1})) = -\log g(y|x^{k+1}) + KL(b(x^k), b(x^{k+1})).$$

Since $g(y|x^k)$ is a probability function, we know that $0 < g(y|x^k) \leq 1$ for all y, so that $-\log g(y|x^k) \geq 0$ for all y. Consequently, the sequence $\{-\log g(y|x^k)\}$ is decreasing to a non-negative limit and the sequence $\{KL(b(x^k), b(x^{k+1}))\}$ converges to zero. So, at the very least, we can say that, using the EM algorithm in the discrete case, the likelihood increases with each step of the iteration.

Just as most books and papers that discuss the EM algorithm contain Equation (1.2), they also assert that the EM algorithm increases likelihood. As we have just seen, in the discrete case, the fact that likelihood is increasing is a consequence of Equation (2.1). In the continuous case, Equation (1.2) is always used to show that likelihood is increasing; if we cannot adopt Equation (1.2), we will need to interpret it in such a way that we can still say that likelihood is increasing.

We would like to be able to prove that the sequence $\{x^k\}$ converges to $\hat{x}$, where $\hat{x}$ is some likelihood maximizer. Failing that, we would like to prove that the sequence of discrete probabilities $\{f(x^k)\}$ converges to $f(\hat{x})$, or even that the sequence $\{g(y|x^k)\}$ converges to $g(y|x_{ML}) = g(y|\hat{x})$. However, it appears that we cannot do all of these without additional assumptions. The assumptions we shall introduce shortly will seem unmotivated and quite specialized, but are motivated by the proof of convergence given for the EM algorithm in the case of Poisson sums in Byrne (2005)[10].

2.5 The Four-Point Property

In [16] Csiszár and Tusnády consider fairly general alternating minimization methods from a geometric point of view. Their three-point and four-point properties play crucial roles in proving convergence. We consider now how their formulation translates into the problem at hand.

Our objective is to maximize the likelihood $g(y|x)$ by minimizing the function $KL(b(x), f(x))$. The EM algorithm in this case consists of minimizing $KL(b(x^{k-1}), f(z))$ to get $z = x^k$, and then minimizing $KL(b(x), f(x^k))$ to get $x = x^k$. The three-point property in this case is

$$KL(b(w), f(x)) \geq KL(b(x), f(x)) + KL(b(w), b(x)). \tag{2.11}$$

The three-point property holds here because, according to Equation (2.10), the inequality is actually an equality.

The four-point property, which may or may not hold, translates into

$$KL(b(w), b(x)) \geq KL(b(w), f(x')) - KL(b(w), f(w')), \tag{2.12}$$

where $x = w'$ minimizes $KL(b(w), f(x))$ and $w = x'$ minimizes $KL(b(x), f(w))$.

We are particularly interested in the case in which $w = \hat{x}$ is a maximizer of likelihood, so that $\hat{x}' = \hat{x}$. Then the three-point property implies that

$$KL(b(\hat{x}), f(x^k)) = KL(b(x^k), f(x^k)) + KL(b(\hat{x}), b(x^k)), \tag{2.13}$$

and the four-point property tells us that

$$KL(b(\hat{x}), b(x^{k-1})) \geq KL(b(\hat{x}), f(x^k)) - KL(b(\hat{x}), f(\hat{x})). \tag{2.14}$$

Combining Equations (2.13) and (2.14), we get

$$KL(b(\hat{x}), b(x^{k-1})) - KL(b(\hat{x}), b(x^k)) \geq KL(b(x^k), f(x^k)) - KL(b(\hat{x}), f(\hat{x})), \tag{2.15}$$

from which we can conclude that the sequence $\{KL(b(\hat{x}), f(x^k))\}$ is decreasing and the sequence $\{KL(b(x^k), f(x^k))\}$ converges to $KL(b(\hat{x}), f(\hat{x}))$. Therefore, $\{g(y|x^k)\}$ converges to $g(y|x_{ML})$.

To prove that the sequence $\{x^k\}$ converges to $\hat{x}$, where $\hat{x}$ is some likelihood maximizer, or to prove that the sequence of discrete probabilities $\{f(x^k)\}$ converges to $f(\hat{x})$, we need further assumptions.

2.6 Some Assumptions

Assuming that the four-point property holds for $\hat{x}$, we know that the sequence $\{KL(b(\hat{x}), f(x^k))\}$ is decreasing and that the sequence $\{g(y|x^k)\}$ converges to $g(y|x_{ML}) = g(y|\hat{x})$. In order to prove convergence of the EM algorithm for the discrete case, we need to make some further assumptions.

Since the variables z lie in a discrete set, the probability functions $b(x^k)$ are members of the unit ball of the Banach space l_1 of all absolutely summable sequences. We assume that there is a subsequence $\{b(x^{k_n})\}$ converging to some $b(x^*)$. Then it follows that $KL(b(x^*), f(x^*)) = KL(b(x_{ML}), f(x_{ML}))$, so that the sequence $\{KL(b(x^*), b(x^k))\}$ is decreasing. But a subsequence converges to zero, so that $\{KL(b(x^*), b(x^k))\}$ converges to zero. We assume that this implies that $\{b(x^k)\}$ converges to $b(x^*)$ in l_1. From

$$g(y|x^k)b(z|y, x^k) = f(z|x^k),$$

for all $z \in \mathcal{Z}(y)$, it follows that $\{f(z|x^k)\}$ converges to $f(z|x^*)$ for all $z \in \mathcal{Z}(y)$. We then assume that this implies that $\{x^k\}$ converges to x^* in some suitable topology for X.

We consider now several examples of the use of the EM algorithm, both to illustrate the theory for the discrete case just discussed, and to point out what we may need to do to remedy the situation in the continuous case.

3 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.

3.1 Poisson Sums

Let $Z_1, ..., Z_N$ be independent Poisson random variables with expected value $E(Z_n) = \lambda_n$. Let Z be the random vector with Z_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 Z is

$$f(z|\lambda) = \prod_{n=1}^{N} \lambda_n^{z_n} \exp(-\lambda_n)/z_n! = \exp(-\lambda_+) \prod_{n=1}^{N} \lambda_n^{z_n}/z_n! \; . \tag{3.1}$$

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

$$\begin{aligned} \text{Prob}(Y = y) &= \text{Prob}(Z_1 + ... + Z_N = y) \\ &= \sum_{z_1+...z_N=y} \exp(-\lambda_+) \prod_{n=1}^{N} \lambda_n^{z_n}/z_n! \; . \end{aligned} \tag{3.2}$$

As we shall see shortly, we have

$$\sum_{z_1+...z_N=y} \exp(-\lambda_+) \prod_{n=1}^{N} \lambda_n^{z_n}/z_n! = \exp(-\lambda_+)\lambda_+^y/y! \, . \tag{3.3}$$

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

When we observe an instance of y, we can consider the conditional distribution $f(z|\lambda, y)$ of $\{Z_1, ..., Z_N\}$, subject to $y = Z_1 + ... + Z_N$. We have

$$f(z|\lambda, y) = \frac{y!}{z_1!...z_N!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{z_1}...\Big(\frac{\lambda_N}{\lambda_+}\Big)^{z_N}. \tag{3.4}$$

This is a *multinomial distribution.*

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

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

To see why this is true, consider the marginal conditional distribution $f(z_1|\lambda, y)$ of Z_1, conditioned on y and λ, which we obtain by holding z_1 fixed and summing over the remaining variables. We have

$$f(z_1|\lambda, y) = \frac{y!}{z_1!(y-z_1)!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{z_1}\Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-z_1} \sum_{z_2+...+z_N=y-z_1} \frac{(y-z_1)!}{z_2!...z_N!} \prod_{n=2}^{N} \Big(\frac{\lambda_n}{\lambda'_+}\Big)^{z_n},$$

where

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

As we shall show shortly,

$$\sum_{z_2+...+z_N=y-z_1} \frac{(y-z_1)!}{z_2!...z_N!} \prod_{n=2}^{N} \Big(\frac{\lambda_n}{\lambda'_+}\Big)^{z_n} = 1,$$

so that

$$f(z_1|\lambda, y) = \frac{y!}{z_1!(y-z_1)!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{z_1} \Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-z_1}.$$

The random variable Z_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 Z_1 is $y\lambda_1/\lambda_+$, as claimed. In the next subsection we look more closely at the multinomial distribution.

3.2 The Multinomial Distribution

When we expand the quantity $(a_1 + ... + a_N)^y$, we obtain a sum of terms, each having the form $a_1^{z_1}...a_N^{z_N}$, with $z_1 + ... + z_N = y$. How many terms of the same form are there? There are N variables a_n. We are to use z_n of the a_n, for each $n = 1, ..., N$, to get $y = z_1 + ... + z_N$ factors. Imagine y blank spaces, each to be filled in by a variable as we do the selection. We select z_1 of these blanks and mark them a_1. We can do that in $\binom{y}{z_1}$ ways. We then select z_2 of the remaining blank spaces and enter a_2 in them; we can do this in $\binom{y-z_1}{z_2}$ ways. Continuing in this way, we find that we can select the N factor types in

$$\binom{y}{z_1}\binom{y-z_1}{z_2}...\binom{y-(z_1+...+z_{N-2})}{z_{N-1}} \tag{3.5}$$

ways, or in

$$\frac{y!}{z_1!(y-z_1)!}...\frac{(y-(z_1+...+z_{N-2}))!}{z_{N-1}!(y-(z_1+...+z_{N-1}))!} = \frac{y!}{z_1!...z_N!}. \tag{3.6}$$

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

$$(a_1 + ... + a_N)^y = \sum_{z_1+...+z_N=y} \frac{y!}{z_1!...z_N!} a_1^{z_1}...a_N^{z_N}. \tag{3.7}$$

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

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

$$= \sum_{z_1+...+z_N=y} \frac{y!}{z_1!...z_N!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{z_1}...\Big(\frac{\lambda_N}{\lambda_+}\Big)^{z_N}. \tag{3.8}$$

From this we get

$$\sum_{z_1+...z_N=y} \exp(-\lambda_+)\prod_{n=1}^{N}\lambda_n^{z_n}/z_n! = \exp(-\lambda_+)\lambda_+^y/y!\ . \tag{3.9}$$

3.3 Poisson Sums in Emission Tomography

The problem of complete versus incomplete data arises in *single-photon computed emission tomography* (SPECT) (Wernick and Aarsvold (2004) [46]). In their 1976 paper Rockmore and Makovski [42] suggested that the problem of reconstructing a tomographic image be viewed as statistical parameter estimation. Shepp and Vardi (1982) [43] expanded on this idea and suggested that the EM algorithm discussed by Dempster, Laird and Rubin (1977) [17] be used for the reconstruction. The region of interest within the body of the patient is discretized into J pixels (or voxels), with $x_j \geq 0$ the unknown amount of radionuclide within the jth pixel; we assume that x_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 $s_j = \sum_{i=1}^{I} P_{ij} > 0$.

3.3.1 The Complete Data

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

$$f(z|x) = \prod_{i=1}^{I}\prod_{j=1}^{J} \exp^{-P_{ij}x_j}(P_{ij}x_j)^{z_{ij}}/z_{ij}!\ . \tag{3.10}$$

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

$$\text{Prob}(z_{i1}, ..., z_{iJ}) = \frac{y_i!}{z_{i1}!\cdots z_{iJ}!}\prod_{j=1}^{J}\Big(\frac{P_{ij}x_j}{(Px)_i}\Big)^{z_{ij}}.$$

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

$$E(Z_{ij}|y,x^k) = x^k_j P_{ij}\Big(\frac{y_i}{(Px^k)_i}\Big),$$

and the conditional expected value of the random variable

$$\log f(Z|x) = \sum_{i=1}^{I}\sum_{j=1}^{J}(-P_{ij}x_j) + Z_{ij}\log(P_{ij}x_j) + \text{constants}$$

becomes

$$E(\log f(Z|x)|y,x^k) = \sum_{i=1}^{I}\sum_{j=1}^{J}\bigg((-P_{ij}x_j) + x^k_j P_{ij}\Big(\frac{y_i}{(Px^k)_i}\Big)\log(P_{ij}x_j)\bigg),$$

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

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

$$LL_c(x) = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(-P_{ij}x_j + z_{ij}\log(P_{ij}x_j)\Big). \tag{3.11}$$

Of course, we do not have the complete data.

3.3.2 The Incomplete Data

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

$$Y_i = \sum_{j=1}^{J} Z_{ij}. \tag{3.12}$$

This is the incomplete data. These random variables are also independent and $(Px)_i$-Poisson, where

$$(Px)_i = \sum_{j=1}^{J} P_{ij}x_j.$$

The log likelihood function for the incomplete data is

$$LL_i(x) = \sum_{i=1}^{I}\Big(-(Px)_i + y_i\log((Px)_i)\Big). \tag{3.13}$$

Maximizing $LL_c(x)$ in Equation (3.11) is easy, while maximizing $LL_i(x)$ in Equation (3.13) is harder and requires an iterative method.

The EM algorithm involves two steps: in the (E) step we compute the conditional expected value of $LL_c(x)$, conditioned on the data vector y and the current estimate

x^k of x; in the (M) step we maximize this conditional expected value to get the next x^{k+1}. Putting these two steps together, we have the following iteration:

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px^k)_i}. \tag{3.14}$$

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

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

$$g(y|x) = \sum_{z \in \mathcal{Z}(y)} f(z|x). \tag{3.15}$$

This means that for this application the integral over $\mathcal{Z}(y)$ in Equation (1.4) is just a finite sum. What required a bit of proof is to show that the $g(y|x)$ obtained by Equation (3.15) is the product of Poisson distributions.

4 Censored Exponential Data

We turn now to an example of a missing data problem that involves probability density functions. Now the integral in Equation (1.2) requires a new interpretation before we can make any progress.

Let Z be a random vector whose distribution is governed by the pdf $f(z|x)$, where $x \in X$ is a (possibly vector) parameter. Given N independent realizations of Z, denoted $z_1, z_2, ..., z_N$, a *maximum-likelihood* (ML) estimate of the parameter x can be obtained by maximizing the likelihood function

$$L(x) = \prod_{n=1}^{N} f(z_n|x), \tag{4.1}$$

over all $x \in X$, or, equivalently, by maximizing the log likelihood function

$$LL(x) = \log L(x) = \sum_{n=1}^{N} \log f(z_n|x). \tag{4.2}$$

McLachlan and Krishnan (1997) [35] give the following example. Suppose that Z is the time until failure of a component, governed by the exponential distribution

$$f(z|x) = \frac{1}{x} e^{-z/x}, \tag{4.3}$$

where $x > 0$ is the expected time until failure. We observe a sample of N components and record their failure times, z_n. On the basis of this data, we must estimate x, the mean time until failure.

It may well happen, however, that during the time allotted for observing the components, only r of the N components fail, which, for convenience, are taken to be the first r items in the record. Rather than wait longer, we record the failure times of those that failed, and record the elapsed time for the experiment, say T, for those that had not yet failed; this is *censored data.* The censored data is viewed as *incomplete*, relative to the *complete* data we would have had, had the trial lasted until all the components had failed. The log likelihood function based on the complete data is

$$LL_c(x) = -N \log x - \frac{1}{x} \sum_{n=1}^{N} z_n, \tag{4.4}$$

and the ML estimate of x is easily seen to be

$$x_{MLc} = \frac{1}{N} \sum_{n=1}^{N} z_n. \tag{4.5}$$

Since the probability that a component will survive until time T is $e^{-T/x}$, the log likelihood function for the censored, or incomplete, data is

$$LL_i(x) = -r \log x - \frac{1}{x} \sum_{n=1}^{N} y_n. \tag{4.6}$$

In this particular example we are fortunate, in that we can maximize $LL_i(x)$ easily, and find that the ML solution based on the censored data is

$$x_{MLi} = \frac{1}{r} \sum_{n=1}^{N} y_n = \frac{1}{r} \sum_{n=1}^{r} y_n + \frac{N-r}{r} T. \tag{4.7}$$

In most cases in which our data is incomplete, finding the ML estimate from the incomplete data is difficult, while find it for the complete data is relatively easy. The EM algorithm compensates for this difficulty by having us estimate the missing data using the current estimate of x, and then letting us maximize the $LL_c(x)$, given that estimated missing data, to get the next estimate of x.

In this example, both the incomplete-data vector y and the complete-data vector z lie in R^N. We have $y = h(z)$ where the function h operates by setting to T any component of z that exceeds T. Clearly, for a given y, the set $\mathcal{Z}(y)$ consists of all vectors z with $z_n \geq T$ or $z_n = y_n < T$. For example, suppose that $N = 2$, and $y = (y_1, T)$, where $y_1 < T$. Then $\mathcal{Z}(y)$ is the one-dimensional ray

$$\mathcal{Z}(y) = \{z = (y_1, z_2) | \, z_2 \geq T\}.$$

Because this set has measure zero in R^2, Equation (1.2) does not make sense in this case, but Equation (1.4) holds if we define the integral to be

$$g(y|x) = \int_{\mathcal{Z}(y)} f(z|x) = \int_T^{\infty} f(y_1, z_2|x) dz_2. \tag{4.8}$$

So, in this case, the integration is with respect to the second variable of z only.

Following McLachlan and Krishnan (1997) [35], we note that since $LL_c(x)$ is linear in the unobserved data, to calculate $E(\log f(Z|x)|y, x^k)$ we need only replace the unobserved values with their conditional expected values, given y and x^k. The conditional distribution of $z_n - T$, given that $z_n > T$, is still exponential, with mean x. Therefore, we replace the unobserved values, that is, all the $y_n = T$, with $T + x^k$. Therefore, at the (E)-step we have

$$E(\log f(Z|x)|y, x^k) = -N \log x - \frac{1}{x}\bigg(\Big(\sum_{n=1}^{N} y_n\Big) + (N-r)x^k\bigg). \tag{4.9}$$

The (M)-step is to maximize this quantity, with respect to x, which leads to

$$x^{k+1} = \bigg(\Big(\sum_{n=1}^{N} y_n\Big) + (N-r)x^k\bigg)/N. \tag{4.10}$$

Let x^* be a fixed point of this iteration. Then we have

$$x^* = \bigg(\Big(\sum_{n=1}^{N} y_n\Big) + (N-r)x^*\bigg)/N,$$

so that

$$x^* = \frac{1}{r}\sum_{n=1}^{N} y_n,$$

which, as we have seen, is the likelihood maximizer.

5 Probabilistic Mixtures

A random variable W is said to be a *mixture* (see Everett and Hand (1981) [22]; Redner and Walker (1984) [41]) if its distribution is governed by the probability density function (or probability function)

$$p(w|x) = \sum_{j=1}^{J} x_j p_j(w), \tag{5.1}$$

where the $p_j(w)$ are known pfs or pdf's and the unknown mixing proportions x_j are non-negative and sum to one. Our incomplete data are N independent realizations of W, $w_1, ..., w_N$. As we shall see, we must deal separately with the pf and pdf cases.

To motivate such mixture problems, we imagine that each data value is generated by first selecting one value of j, with probability x_j, and then selecting a realization of a random variable governed by $p_j(w)$. 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. The w_n are the numerical values of the colors of the selected marbles. It is also possible to formulate the previous example of Poisson sums as a mixture if we consider the data to be the list of detectors at which each photon was detected. Let the incomplete data vector be $y = (w_1, ..., w_N)$, with

$$g(y|x) = \prod_{n=1}^{N} p(w_n|x).$$

The log likelihood function is

$$LL_i(x) = \sum_{n=1}^{N} \log(\sum_{j=1}^{J} x_j p_j(w_n)). \tag{5.2}$$

To get the ML estimate of the vector x we need to maximize $LL_i(x)$, subject to the conditions that $x_j \geq 0$ and the x_j sum to one.

With $P_{nj} = p_j(w_n)$, the function $LL_i(x)$ becomes

$$LL_i(x) = \sum_{n=1}^{N} \log((Px)_n). \tag{5.3}$$

As we shall see later, the EM algorithm for this problem has the iterative step

$$x_j^{k+1} = x_j^k \sum_{n=1}^{N} P_{nj} \frac{1/N}{(Px^k)_n}. \tag{5.4}$$

We create the complete data by imagining that we could have obtained $z_n = (w_n, j_n)$, for $n = 1, ..., N$, where the selection of w_n is governed by the function $f_{j_n}(w)$. In the bowls example, j_n is the number of the bowl from which the nth marble is drawn. Since our objective is to estimate the x_j, the values w_n are irrelevant; our ML estimate of x_j is simply the proportion of times $j_n = j$.

Given $y = (w_1, ..., w_N)$, the set $\mathcal{Z}(y)$ now consists of those vectors $z = (z_1, ..., z_N)$ with $z_n = (w_n, j)$, for some $j = 1, 2, ..., J$. Therefore, $\mathcal{Z}(y)$ is a finite set, the integral in Equation (1.2) is zero, and the integral in Equation (1.4) is defined as just a finite sum.

The likelihood function for the complete data vector z is

$$f(z|x) = f(w_1, ..., w_N, j_1, ..., j_N|x) = \prod_{n=1}^{N} x_{j_n} p_{j_n}(w_n). \tag{5.5}$$

We want to show that we get $g(y|x)$ by summing $f(z|x)$ over all $z \in \mathcal{Z}(y)$, which means summing $f(z|x)$ over all possible choices for the indices $j_1, j_2, ..., j_N$. Let's fix the first $N-1$ indices and sum over $j=1,...,N$ as the values of j_N. Then we have

$$\sum_{j=1}^{J}\prod_{n=1}^{N} x_{j_n}p_{j_n}(w_n) = \Big(\prod_{n=1}^{N-1} x_{j_n}p_{j_n}(w_n)\Big)\Big(\sum_{j=1}^{J} x_j p_j(w_N)\Big)$$

$$= \Big(\prod_{n=1}^{N-1} x_{j_n}p_{j_n}(w_n)\Big)p(w_N).$$

Repeating this with each factor of the product in turn, we get

$$\sum_{z\in\mathcal{Z}(y)} f(z|x) = \prod_{n=1}^{N} p(w_n) = g(y|x),$$

which is Equation (1.4) in this case.

The precise form of the EM algorithm will depend on whether the mixture is of probability functions or probability density functions.

5.1 Finite Mixtures of Probability Vectors

We say that a discrete random variable W taking values in the set $\{i=1,...,I\}$ is a *finite mixture of probability vectors* if there are probability vectors f_j and numbers $x_j>0$, for $j=1,...,J$, such that the probability vector for W is

$$f(i) = \text{Prob}(W=i) = \sum_{j=1}^{J} x_j f_j(i). \tag{5.6}$$

We require, of course, that $\sum_{j=1}^{J} x_j = 1$.

The data are N realizations of the random variable W, denoted w_n, for $n=1,...,N$ and the incomplete data is the vector $y=(w_1,...,w_N)$. The column vector $x=(x_1,...,x_J)^T$ is the parameter vector of mixture probabilities to be estimated. The likelihood function is

$$L(x) = \prod_{n=1}^{N}\Big(x_1f_1(w_n)+...+x_Jf_J(w_n)\Big),$$

which can be written as

$$L(x) = \prod_{i=1}^{I}\Big(x_1f_1(i)+...+x_Jf_J(i)\Big)^{n_i},$$

where n_i is the cardinality of the set $\{n|\, i_n = i\}$. Then the log likelihood function is

$$LL(x) = \sum_{i=1}^{I} n_i \log\Big(x_1f_1(i)+...+x_Jf_J(i)\Big).$$

With u the column vector with entries $u_i = n_i/N$, and P the matrix with entries $P_{ij} = f_j(i)$, we see that

$$\sum_{i=1}^{I}(Px)_i = \sum_{i=1}^{I}\Big(\sum_{j=1}^{J}P_{ij}x_j\Big) = \sum_{j=1}^{J}\Big(\sum_{i=1}^{I}P_{ij}\Big) = \sum_{j=1}^{J}x_j = 1,$$

so maximizing $LL(x)$ over non-negative vectors x with $\sum_{j=1}^{J}x_j = 1$ is equivalent to minimizing the KL distance $KL(u, Px)$ over the same vectors. The restriction that the entries of x sum to one turns out to be redundant, as we show now.

From the gradient form of the Karush-Kuhn-Tucker Theorem in optimization, we know that, for any $\hat{x}$ that is a non-negative minimizer of $KL(u, Px)$, we have

$$\sum_{i=1}^{I}P_{ij}\Big(1 - \frac{u_i}{(P\hat{x})_i}\Big) \geq 0,$$

and

$$\sum_{i=1}^{I}P_{ij}\Big(1 - \frac{u_i}{(P\hat{x})_i}\Big) = 0,$$

for all j such that $\hat{x}_j > 0$. Consequently, we can say that

$$s_j\hat{x}_j = \hat{x}_j\sum_{i=1}^{I}P_{ij}\Big(\frac{u_i}{(P\hat{x})_i}\Big),$$

for all j. Since, in the mixture problem, we have $s_j = \sum_{i=1}^{I}P_{ij} = 1$ for each j, it follows that

$$\sum_{j=1}^{J}\hat{x}_j = \sum_{i=1}^{I}\Big(\sum_{j=1}^{J}\hat{x}_jP_{ij}\Big)\frac{u_i}{(P\hat{x})_i} = \sum_{i=1}^{I}u_i = 1.$$

So we know now that any non-negative minimizer of $KL(u, Px)$ will be a probability vector that maximizes $LL(x)$. Since the EM algorithm in Equation (3.14) minimizes $KL(u, Px)$, when u_i replaces y_i, it can be used to find the maximum-likelihood estimate of the mixture probabilities. It is helpful to remember that there was no mention of Poisson distributions in this example, and that the EM algorithm can be used to find likelihood maximizers in situations other than that of sums of independent Poisson random variables.

If the set of values that W can take on is infinite, say $\{i = 1, 2, ...\}$, then the f_j are infinite probability sequences. The same analysis applies to this infinite case, and again we have $s_j = 1$. The iterative scheme is given by Equation (3.14), but with an apparently infinite summation; since only finitely many of the u_i are non-zero, the summation is actually only finite.

5.2 Finite Mixtures of Probability Density Functions

For finite mixtures of probability density functions the problem is a bit more complicated. A variant of the EM algorithm still solves the problem, but this is not so obvious.

Suppose now that W is a random variable with probability density function $f(w)$ given by

$$f(w) = \sum_{j=1}^{J} x_j f_j(w), \tag{5.7}$$

where the $f_j(w)$ are known pdf's and the mixing proportions x_j are unknown. Our data is $w_1, ..., w_N$, that is, N independent realizations of the random variable W, and $y = (w_1, ..., w_N)$ is the incomplete data. With x the column vector with entries x_j, we have the likelihood function

$$L(x) = \prod_{n=1}^{N} (\sum_{j=1}^{J} x_j f_j(w_n)),$$

and the log likelihood function

$$LL(x) = \sum_{n=1}^{N} \log(\sum_{j=1}^{J} x_j f_j(w_n)).$$

We want to estimate the vector x by maximizing $LL(x)$, subject to $x_j \geq 0$ and $x_+ = \sum_{j=1}^{J} x_j = 1$.

Let $P_{nj} = f_j(z_n)$, and $s_j = \sum_{n=1}^{N} P_{nj}$. Then

$$LL(x) = \sum_{n=1}^{N} \log(Px)_n.$$

With $u_n = \frac{1}{N}$ for each n, we have that maximizing $LL(x)$, subject to $x_j \geq 0$ and $x_+ = 1$, is equivalent to minimizing

$$KL(u, Px) - \sum_{n=1}^{N} (Px)_n, \tag{5.8}$$

subject to the same constraints. Since the non-negative minimizer of the function

$$F(x) = KL(u, Px) + \sum_{j=1}^{J} (1 - s_j) x_j \tag{5.9}$$

satisfies $x_+ = 1$, it follows that minimizing $F(x)$ subject to $x_j \geq 0$ and $x_+ = 1$ is equivalent to minimizing $F(x)$, subject only to $x_j \geq 0$.

The following theorem is found in Byrne (2001) [7]:

Theorem 5.1 *Let y be any positive vector and*

$$G(x) = KL(y, Px) + \sum_{j=1}^{J} \beta_j KL(\gamma_j, x_j).$$

If $s_j + \beta_j > 0$, $\alpha_j = s_j(s_j + \beta_j)^{-1}$, and $\beta_j\gamma_j \geq 0$ for each j, then the iterative sequence generated by

$$x_j^{k+1} = \alpha_j s_j^{-1} x_j^k \Big(\sum_{n=1}^{N} P_{nj} \frac{y_n}{(Px^k)_n} \Big) + (1 - \alpha_j)\gamma_j$$

converges to a non-negative minimizer of $G(x)$.

With $y_n = u_n = \frac{1}{N}$, $\gamma_j = 0$, and $\beta_j = 1 - s_j$, it follows that the iterative sequence generated by

$$x_j^{k+1} = x_j^k \frac{1}{N} \sum_{n=1}^{N} P_{nj} \frac{1}{(Px^k)_n} \tag{5.10}$$

converges to the maximum-likelihood estimate of the mixing proportions x_j. This is the EM iteration presented in McLachlan and Krishnan, Equations (1.36) and (1.37) [35].

6 The Continuous Case

If we were to accept Equation (1.2), we could mimic the analysis for the discrete case and show that each step of the EM algorithm increases the likelihood. Because the Equation (1.2) may not make sense for our problem, however, we cannot write the conditional pdf for Z, given $Y = y$, as

$$b(z|y, x) = f(z|x)/g(y|x);$$

the problem lies with the set of z over which $b(z|y, x)$ is to be defined. If we define it only for $z \in \mathcal{Z}(y)$, then either its integral is not one, or Equation (1.2) holds. If we define $b(z|y, x)$ for all z, then $g(y|x) = 1$, which need not be the case.

In the case of censored exponential data, we found that we must reinterpret the integral in Equation (1.2) to be integration over the product of $N - r$ half-lines, where r itself depends on the observed y. In the probabilistic mixture case, the missing data is discrete, so the integral is just a sum.

It is conventional wisdom that an EM algorithm must increase likelihood at each step. However, this property depends heavily on our resolving the issue posed by Equation (1.2).

6.1 Proving Monotonicity of the EM

To examine more closely how the problem of Equation (1.2) affects the development of the theory, we look at the standard way in which it is proved that the EM increases likelihood at each step. The proof we present here is taken from McLachlan and Krishnan (1997) [35], but is essentially the proof from Dempster et al. (1977) [17]. Their notation has been changed to conform to that of this note.

Let the incomplete and complete likelihood functions now be denoted $L_i(x)$ and $L_c(x)$, respectively. Their proof begins with the definition of the conditional density of Z, given $Y = y$ and x:

$$b(z|y, x) = f(z|x)/g(y|x). \tag{6.1}$$

For $b(z|y, x)$ to be a pdf, we need its integral with respect to z, however that is defined, to be one. Since $\int f(z|x)dz = 1$ already, it must be assumed, if it is not explicitly stated, that Equation (6.1) holds only for $z \in \mathcal{Z}(y)$. Consequently, we must have

$$g(y|x) = \int_{\mathcal{Z}(y)} f(z|x), \tag{6.2}$$

which, as we have seen, brings us to the issue of how to define the integral over $\mathcal{Z}(y)$.

The proof then uses Equation (6.1) to get

$$\log L_i(x) = \log g(y|x) = \log f(z|x) - \log b(z|y, x), \tag{6.3}$$

which, of course, holds only for $z \in \mathcal{Z}(y)$. McLachlan and Krishnan rewrite Equation (6.3) as

$$\log L_i(x) = \log L_c(x) - \log b(z|y, x),$$

but this can be misleading, since

$$\log L_c(x) = \log f(z|x)$$

is used in Equation (6.3) only for those $z \in \mathcal{Z}(y)$.

They then take the expectation of both sides of Equation (6.3), conditioned on y and x^k, treating z as a random vector Z now, obtaining

$$\log L_i(x) = \log g(y|x) = Q(x, x^k) - H(x, x^k), \tag{6.4}$$

where

$$Q(x, x^k) = E_{x^k}\Big(\log f(Z|x)|y\Big), \tag{6.5}$$

and

$$H(x, x^k) = E_{x^k}\Big(\log b(Z|y, x)|y\Big). \tag{6.6}$$

Replacing x with x^{k+1} and then with x^k in Equation (6.4), we get

$$\log L_i(x^{k+1}) - \log L_i(x^k) =$$

$$\Big(Q(x^{k+1}, x^k) - Q(x^k, x^k)\Big) - \Big(H(x^{k+1}, x^k) - H(x^k, x^k)\Big). \tag{6.7}$$

Because x^{k+1} is chosen to maximize $Q(x, x^k)$, we know that

$$Q(x^{k+1}, x^k) - Q(x^k, x^k) \geq 0.$$

Therefore, the monotonicity of $L_i(x^k)$ will hold if we can show that

$$H(x^{k+1}, x^k) - H(x^k, x^k) \leq 0.$$

For any x we have

$$H(x, x^k) - H(x^k, x^k) = E_{x^k}\Big(\log(b(Z|y, x)/b(Z|y, x^k))|y\Big)$$

$$= \int_{\mathcal{Z}(y)} b(z|y, x^k) \log(b(z|y, x)/b(z|y, x^k)).$$

Since the integral over $\mathcal{Z}(y)$ has been defined to guarantee that Equation (6.2) holds and that $b(z|y, x)$ is a pdf over $\mathcal{Z}(y)$ for each x, we know from Jensen's Inequality for pdf's that

$$\int_{\mathcal{Z}(y)} b(z|y, x^k) \log(b(z|y, x)/b(z|y, x^k)) \leq 0,$$

so the proof is complete.

The important point here is that this proof that the EM algorithm increases likelihood at each step rests heavily on having resolved the issue of the integral in Equation (1.2) in such a way as to make the $b(z|y, x)$ pdf's over the set $\mathcal{Z}(y)$. Knowing that they are pdf's, we can substitute for Jensen's Inequality and write

$$\int_{\mathcal{Z}(y)} b(z|y, x^k) \log(b(z|y, x)/b(z|y, x^k)) = -KL(b(z|y, x^k), b(z|y, x)) \leq 0.$$

6.2 The Generalized Deconvolution Problem

Eggermont and LaRiccia (2009) [20] consider the *generalized deconvolution* problem

$$g(y|x) = \int k(y,z)f(z|x)dz, \tag{6.8}$$

where $k(y,z)$ is a measurable kernel defined on a subset of $R^N \times R^M$ having positive measure. We can put Equation (1.2) in this form if we take

$$k(y,z) = \chi_{\mathcal{Z}(y)}(z),$$

and require that $\mathcal{Z}(y)$ have positive measure for each y. Then the conditional density of Y, conditioned on Z, is $k(y,Z)$, and we have the conditional density of Z, conditioned on Y, is

$$b(z|y,x) = k(y,z)f(z|x)/g(y|x). \tag{6.9}$$

Now we can mimic the analysis of the discrete case.

6.3 Another Approach

We suppose that $M > N$ and that there is a second function $k : R^M \to R^{M-N}$ such that the function $G : R^M \to R^M$ given by

$$G(z) = (h(z), k(z)) = (y, w) = u$$

is invertible, with inverse H and determinant of the Jacobian matrix denoted by $J(y,w)$. For any measurable set A in R^N we have

$$\int_A g(y|x)dy = \int_{y\in A}\int_{w\in\mathcal{W}(y)} f(H(y,w)|x)J(y,w)dw,$$

where

$$\mathcal{W}(y) = \{w|w = k(z), y = h(z)\}.$$

It then follows that

$$g(y|x) = \int_{w\in\mathcal{W}(y)} f(H(y,w)|x)J(y,w)dw,$$

so that, for $z \in \mathcal{Z}(y)$,

$$b(z|y,x) = b(H(y,k(z))|y,x) = f(H(y,k(z))|x)J(y,k(z))/g(y|x)$$

defines a probability density function on $\mathcal{Z}(y)$.

For example, suppose that $Z = (Z_1, Z_2)$, where Z_1 and Z_2 are independent and uniformly distributed on the interval $[0, x]$. Suppose that $Y = Z_1 + Z_2$. The set $\mathcal{Z}(y)$ is the set of all points (z_1, z_2) for which $h(z_1, z_2) = z_1 + z_2 = y$, which is a set of planar measure zero. The function $g(y|x)$ is

$$g(y|x) = \begin{cases} y/x^2, & 0 \leq y \leq x; \\ \\ (2x - y)/x^2, & x \leq y \leq 2x. \end{cases} \tag{6.10}$$

Equation (1.2) does not hold here. How might the correct equation read?

In our example, we have $M = 2$ and $N = 1$. Let $k(z_1, z_2) = z_1 - z_2$. Then

$$G(z_1, z_2) = (z_1 + z_2, z_1 - z_2),$$

$$H(y, w) = (\frac{y + w}{2}, \frac{y - w}{2}),$$

and

$$J(y, w) = 1/2.$$

The set $\mathcal{W}(y)$ is the entire real line.

The pdf for Z is

$$f(z_1, z_2) = \frac{1}{x^2}\chi_{[0,1]}(z_1)\chi_{[0,1]}(z_2)$$

so the pdf for the random variable Y is

$$g(y) = \frac{1}{2}\int_R f(H(y, w))dw = \frac{1}{2x^2}\int_R \chi_{[0,1]}(\frac{y + w}{2})\chi_{[0,1]}(\frac{y - w}{2})dw.$$

This is easily seen to be $\frac{y}{x^2}$, for $0 \leq y \leq x$ and $\frac{2x-y}{x^2}$, for $1 \leq y \leq 2x$, which is the pdf in Equation (6.10).

6.4 Yet Another Approach

In this example, with $Z = (Z_1, Z_2)$, the pdf $f(z|x)$ is easy to calculate, but obtaining $g(y|x)$ from $f(z|x)$ using h is more complicated. We could have adopt a somewhat different approach.

We could say that the complete data random vector Z has the form $Z = (Y, W)$, where Y and W are possibly dependent random variables. The incomplete data is Y and the function h is simply the projection of Z onto its first component.

In our example, the complete data can be taken to be $Z = (Y, W) = (Z_1 + Z_2, Z_1 - Z_2)$, so that $h(Z)$ is simply the projection onto the first component. What is more complicated now is defining $f(z|x)$.

This example also provides an interesting counter-example concerning convergence.

6.5 A Counter-Example

Suppose we have an initial estimate x^0 of the parameter x. Since $y = z_1 + z_2$, it makes no sense to select a value of x^0 less than $y/2$; therefore, let us assume that $x^0 \geq y/2$. The (E) step is to calculate the conditional expected value of

$$LL_c(x) = \log \chi_{[0,x]}(Z_1) + \log \chi_{[0,x]}(Z_2) - 2 \log x, \tag{6.11}$$

conditioned on x^0 and y. For any x in the interval $[y/2, x^0)$, there will be a positive conditional probability that one or both of Z_1 or Z_2 will exceed x, so, in order for the conditional expected value to be finite, we must restrict x to the closed ray $[x^0, +\infty)$. The conditional expected value of $LL_c(x)$ is then $-2 \log x$. The maximum of $-2 \log x$ over the ray $[x^0, +\infty)$ occurs at $x = x^0$, so $x^1 = x^0$. Therefore, beginning with $x^0 \geq y/2$, we have $x^k = x^0$ for all $k = 1, 2, ...$, and so $\{x^k\}$ does not converge to x_{ML}, generally, and $\{g(y|x^k)\}$ does not converge to $g(y|x_{ML})$.

7 Historical Background

Although the Dempster, Laird and Rubin (1977) paper [17] is often cited as a fundamental work on the EM algorithm, its authors state clearly that this algorithm has a long history. We shall make no attempt here to recapitulate that history, but rather refer the reader to the book by McLachlan and Krishnan (1997) [35] and the bibliographical survey by Meng and Pedlow (1992) [37]. The article by Meng and van Dyk (1997) [38] is also a good source for the history of the EM algorithm. The editorial by Leahy and Byrne (2000) [34] provides a brief summary of the state-of-the-art in medical imaging at that time.

The EM algorithm has become a popular tool in many areas of applications. It was introduced into the medical tomography literature by Shepp and Vardi (1982) [43], with subsequent elaboration by Lange and Carson (1984) [32], Vardi, Shepp and Kaufman (1985) [45], Lange, Bahn and Little (1987) [33], and many others. These papers describe the physics of the transmission and emission tomographic problems, set up the complete data models to be used in the EM algorithm, provide proofs of convergence, sometimes with gaps, and suggest regularization methods to reduce sensitivity to noise in the data.

8 Mathematical Modeling and Applications

The problems of reconstructing an image from single-photon computed emission tomography (SPECT) data and of estimating the mixing proportions for finite mixtures provide insight into the workings of the EM algorithm in important applications.

8.1 The EM Algorithm for the Poisson Model

The underlying model for the SPECT problem is that of sums of independent Poisson random variables. For $i = 1, ..., I$ and $j = 1, ..., J$, the *complete data* random variables Z_{ij} are independent Poisson, with mean values

$$E(Z_{ij}) = P_{ij}x_j,$$

where $0 \leq P_{ij} \leq 1$, and $x_j \geq 0$. The parameter vector is $x = (x_1, ..., x_J)^T$. The probability function for the random vector

$$Z = \{Z_{ij}|i = 1, ..., I, j = 1, ..., J\},$$

given x, is

$$f(z|x) = \prod_{i=1}^{I}\prod_{j=1}^{J} \exp(-P_{ij}x_j)(P_{ij}x_j)^{z_{ij}}/z_{ij}!.$$

Because of the independence, we can write

$$f(z|x) = \prod_{1=1}^{I}\prod_{j=1}^{J} f_{ij}(z_{ij}|x),$$

where

$$f_{ij}(z_{ij}|x) = \exp(-P_{ij}x_j)(P_{ij}x_j)^{z_{ij}}/z_{ij}!.$$

The *incomplete data* random variables are

$$Y_i = \sum_{j=1}^{J} Z_{ij}.$$

The Y_i are also independent and Poisson, with mean values

$$E(Y_i) = (Px)_i = \sum_{j=1}^{J} P_{ij}x_j.$$

The objective is to estimate x, given one instance $y = (y_1, ..., y_I)^T$ of the random vector Y.

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

$$\text{Prob}(z_{i1}, ..., z_{iJ}) = \frac{y_i!}{z_{i1}! \cdots z_{iJ}!} \prod_{j=1}^J \Big(\frac{P_{ij}x_j}{(Px)_i}\Big)^{z_{ij}}.$$

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

$$E(Z_{ij}|y, x^k) = x_j^k P_{ij}\Big(\frac{y_i}{(Px^k)_i}\Big),$$

and the conditional expected value of the random variable

$$\log f(Z|x) = \sum_{i=1}^I \sum_{j=1}^J (-P_{ij}x_j) + Z_{ij}\log(P_{ij}x_j) + \text{constants}$$

becomes

$$E(\log f(Z|x)|y, x^k) = \sum_{i=1}^I \sum_{j=1}^J \bigg((-P_{ij}x_j) + x_j^k P_{ij}\Big(\frac{y_i}{(Px^k)_i}\Big)\log(P_{ij}x_j)\bigg),$$

omitting terms that do not involve the parameter vector x. In the EM algorithm, we obtain the next estimate x^{k+1} by maximizing $E(\log f(Z|x)|y, x^k)$. As is the case with the general EM algorithm, the Kullback-Leibler distance plays an important role.

Maximizing $E(\log f(Z|x)|y, x^k)$ is equivalent to minimizing $KL(r(x^k), q(x))$, where $r(x)$ and $q(x)$ are I by J arrays with entries

$$r(x)_{ij} = x_j P_{ij}\Big(\frac{y_i}{(Px)_i}\Big),$$

and

$$q(x)_{ij} = x_j P_{ij}.$$

The iterative step of the EM algorithm is given in Equation (3.14).

Any likelihood maximizer x_{ML} is also a non-negative minimizer of the KL distance $KL(y, Px)$, so the EM algorithm can be thought of as a method for finding a non-negative solution (or approximate solution) for a system $y = Px$ of linear equations in which $y_i > 0$ and $P_{ij} \geq 0$ for all indices.

9 Regularization

In most applications the measured data is noisy. Algorithms such as the EM algorithm that work well in theory, on noise-free data, may work poorly on actual noisy data, necessitating the use of *regularization*.

9.1 The Need for Regularization

Maximizing the likelihood function $g(y|x)$ sounds like a good idea, but there are circumstances in which the resulting estimator x_{ML} is not useful. One particular case is in medical image reconstruction, for example, in SPECT, where the parameter vector x represents the vectorization of the two-dimensional array of pixel intensities that comprises the desired image. The following theorem describes the problem. As previously, we assume that the vector y has positive entries and the matrix P has non-negative entries. If the matrix P and every matrix obtained from P by deleting columns have full rank, we say that P has the *full-rank property.*

Theorem 9.1 *([5]) Let P have the full-rank property. If the system $y = Px$ has no non-negative solution, then there is a subset S of $\{j = 1, ..., J\}$, with cardinality at most $I - 1$, such that any non-negative minimizer of the function $KL(y, Px)$ is supported on S, and so there is a unique non-negative minimizer of $KL(y, Px)$.*

In fact, this result is not limited to KL distances and non-negative systems, nor is it a property of the EM algorithm; it holds more generally, as the following theorem shows.

Theorem 9.2 *([12]) Let A be an arbitrary I by J matrix. If A and every matrix obtained from A by deleting columns has full rank, and if the system $Ax = b$ has no non-negative solutions, then there is a subset S of $\{j = 1, ..., J\}$, with cardinality at most $I - 1$, such that any non-negative minimizer of the function $\|Ax - b\|_2$ is supported on S, and so there is a unique non-negative minimizer of $\|Ax - b\|_2$.*

Whether we are minimizing $KL(y, Px)$ or $\|Ax-b\|$ over non-negative x, we are trying to fit the model Px or Ax to the measured data, y or b. When the data are noisy, we are often over-fitting noisy data to the model, resulting in noisy answers. The theorems above make this precise. Another way to think about the problem is this: for non-negative x the vectors of the form Px or Ax constitute a cone. When the data vector falls outside this cone, the point in the cone that is closest will lie on a face of the cone, and so will be a convex combination of a small number of the vectors that generated the cone. Therefore, most of the coefficients will be zero.

In the case of SPECT the larger the number of pixels J, the smaller the pixels, so it is natural to assume that a larger J means greater resolution. However, as Theorem 9.1 shows, when y is noisy, which is the typical case, increasing the number of pixels can lead to a maximum-likelihood image in which many of the pixels have the value zero. These tend to be scattered throughout the image, creating a picture that resembles stars in the night sky. One way to remedy this is simply to stop the iterative procedure early. There are other ways, as we shall see shortly.

9.2 Penalized EM

The iterative step of the EM algorithm requires us to minimize the function $KL(r(x^k), q(x))$ over $x \geq 0$ to get x^{k+1}. A regularized version of the EM algorithm requires us to minimize

$$KL(r(x^k), q(x)) + \sum_{j=1}^{J} w_j KL(p_j, x_j),$$

where $w_j > 0$ are weights and the vector p with entries $p_j > 0$ is a prior estimate of the desired x (Byrne 1993) [5]. The resulting iterative step is

$$x_j^{k+1} = \frac{w_j}{w_j + s_j} p_j + \frac{1}{w_j + s_j} x_j^k \sum_{i=1}^{I} P_{ij} \Big(\frac{y_i}{(Px^k)_i} \Big).$$

This regularization algorithm is equivalent to the Bayesian maximum a posteriori method with a prior gamma distribution of Lange, Bahn and Little (1987) [33]. The sequence $\{x^k\}$ converges to the non-negative minimizer of the function

$$KL(y, Px) + \sum_{j=1}^{J} w_j KL(p_j, x_j).$$

9.3 Other Regularization Methods

As we have already seen, regularization can be achieved through the use of penalty functions. If one has prior knowledge about the solution x being sought, such as that it is the vectorization of a locally smooth image, then one can use a penalty function that penalizes images that are not locally smooth. This is usually done with what is called a *Gibbs prior* in Geman and Geman (1984) [24]. There is a difficulty with the use of general penalty functions, however.

Suppose that we wish to regularize the EM algorithm to find the penalized maximum-likelihood solution that minimizes the function

$$KL(y, Px) + g(x).$$

It is natural to set up an iterative algorithm in which, at the kth step, we minimize

$$KL(r(x^k), q(x)) + g(x).$$

When we take the gradient with respect to the vector variable x, we find that we cannot obtain a closed-form expression for the next x^{k+1}; the problem is that the equation to be solved contains the gradient of $g(x)$, evaluated at x^{k+1}. The *one-step-late* approach of Green (1990) [25] partially resolves this difficulty by evaluating this gradient at x^k instead.

9.4 De Pierro's Surrogate-Function Method

De Pierro (1995) [18] presents a modified EM algorithm that includes regularization in the form of a penalty function. His objective is to embed the penalty term in the alternating minimization framework in such a way as to make it possible to obtain the next iterate in closed form. Because his *surrogate function* method has been used subsequently by Fessler et al. (Fessler et al. (1997) [23], Ahn et al. (2006) [1]) and others to obtain penalized likelihood algorithms, we consider his approach in some detail.

Let x and z be vector variables and $H(x, z) > 0$. Mimicking what occurs in penalized EM when the gamma prior distribution is used, we require that if we fix z and minimize $H(x, z)$ with respect to x, the solution should be $x = z$, the vector we fixed; that is, $H(x, z) \geq H(z, z)$ always. If we fix x and minimize $H(x, z)$ with respect to z, we should get something new; call it Tx. As with the EM, the algorithm will have the iterative step $x^{k+1} = Tx^k$.

Summarizing, we see that we need a function $H(x, z)$ with the properties (1) $H(x, z) \geq H(z, z)$ for all x and z; (2) $H(x, x)$ is the function $F(x)$ we wish to minimize; and (3) minimizing $H(x, z)$ with respect to z for fixed x is easy.

The function to be minimized is

$$F(x) = KL(y, Px) + g(x), \tag{9.12}$$

where $g(x) \geq 0$ is some penalty function. De Pierro uses penalty functions $g(x)$ of the form

$$g(x) = \sum_{l=1}^{p} f_l(\langle s_l, x\rangle\,). \tag{9.13}$$

Let us define the matrix S to have for its lth row the vector s_l^T. Then $\langle s_l, x\rangle = (Sx)_l$, the lth entry of the vector Sx. Therefore,

$$g(x) = \sum_{l=1}^{p} f_l((Sx)_l). \tag{9.14}$$

Let $\lambda_{lj} > 0$ with $\sum_{j=1}^{J} \lambda_{lj} = 1$, for each l.

Assume that the functions f_l are convex. Therefore, for each l, we have

$$f_l((Sx)_l) = f_l(\sum_{j=1}^J S_{lj}x_j) = f_l(\sum_{j=1}^J \lambda_{lj}(S_{lj}/\lambda_{lj})x_j) \tag{9.15}$$

$$\leq \sum_{j=1}^J \lambda_{lj} f_l((S_{lj}/\lambda_{lj})x_j). \tag{9.16}$$

Therefore,

$$g(x) \leq \sum_{l=1}^p \sum_{j=1}^J \lambda_{lj} f_l((S_{lj}/\lambda_{lj})x_j). \tag{9.17}$$

So we have replaced $g(x)$ with a related function in which the x_j occur separately, rather than just in the combinations $(Sx)_l$. But we aren't quite done yet.

We would like to take for De Pierro's $H(x, z)$ the function $KL(r(x^k), q(x))$ used in the EM algorithm, plus the function

$$\sum_{l=1}^p \sum_{j=1}^J \lambda_{lj} f_l((S_{lj}/\lambda_{lj})z_j). \tag{9.18}$$

But there is one slight problem: we need $H(z, z) = F(z)$, which we don't have yet. De Pierro's clever trick is to replace $f_l((S_{lj}/\lambda_{lj})z_j)$ with

$$f_l\Big((S_{lj}/\lambda_{lj})z_j - (S_{lj}/\lambda_{lj})x_j\Big) + f_l((Sx)_l). \tag{9.19}$$

So, De Pierro's function $H(x, z)$ is the sum of the $H(x, z)$ used in the EM case and the function

$$\sum_{l=1}^p \sum_{j=1}^J \lambda_{lj} f_l\Big((S_{lj}/\lambda_{lj})z_j - (S_{lj}/\lambda_{lj})x_j\Big) + f_l((Sx)_l). \tag{9.20}$$

Now he has the three properties he needs. Once he has computed x^k, he minimizes $H(x^k, z)$ by taking the gradient and solving the equations for the correct $z = Tx^k = x^{k+1}$. For the choices of f_l he discusses, these intermediate calculations can either be done in closed form (the quadratic case) or with a simple Newton-Raphson iteration (the logcosh case).

10 Accelerating the EM Algorithm

It is well known that the EM algorithm can be slow to convergence. For that reason, there is great interest in finding methods to accelerate the EM algorithm. We focus here on block-iterative methods, in which some, but not necessarily all the data is used at each step.

The paper of Holte, Schmidlin, *et al.* (1990) [28] compares the performance of Schmidlin's method of (1972) [44] with the EM algorithm. Almost as an aside, they notice the accelerating effect of what they call *projection interleaving*, that is, the use of blocks. This paper contains no explicit formulas, however, and presents no theory, so one can only make educated guesses as to the actual iterative methods employed. Somewhat later, it was noticed that useful images could be obtained quickly if, in the implementation of the EM algorithm, the summation was performed only over those i in a subset, or block, of the detector indices; then a new block was selected and the process repeated. This *ordered-subset* (OSEM) method of Hudson and Larkin (1994) [29] quickly became the algorithm of choice, at first, for researchers, and a bit later, for the clinic.

The absence of a solid mathematical foundation for the OSEM led several groups to reexamine other block-iterative methods, particularly BI-MART, the block-iterative version of the multiplicative algebraic reconstruction technique (Censor and Segman (1987) [15]). Unlike the OSEM, the BI-MART always converges to a non-negative solution of the system $y = Px$, whenever there is a non-negative solution, regardless of how the blocks are selected. This suggested that the OSEM is not the correct block-iterative version of the EM. This problem was resolved with the appearance, in Browne and DePierro (1996) [4], of the RAMLA and, in Byrne (1996) [6], of the rescaled BI-EM (RBI-EM) method.

Block-iterative methods do not necessarily converge faster than simultaneous ones that use all the equations at each step. The block-iterative methods do provide the opportunity for a rescaling of the equations, which, as we shall see, does lead to significant acceleration of the algorithms.

Throughout this section, we assume that the set $\{i = 1, ..., I\}$ is the (not necessarily disjoint) union of subsets, or blocks, denoted B_n, for $n = 0, 1, ..., N - 1$.

We let P_n be the matrix and y^n the vector obtained from P and y, respectively, by removing all the rows except for those whose index i is in the set B_n. For each n and j, we let

$$s_{nj} = \sum_{i \in B_n} P_{ij},$$

$$m_n = \max\{s_{nj},\ j = 1, ..., J\},$$

and

$$\mu_n = \max\{s_{nj} s_j^{-1},\ j = 1, ..., J\}.$$

When $N = 1$, $s_{nj} = s_j$, so $\mu = \mu_n = 1$ and

$$m = m_n = \max\{s_j,\ j = 1, ..., J\}.$$

When $N = I$, and $n = i$, $s_{nj} = P_{ij}$, so

$$\mu_i = \mu_n = \max\{P_{ij} s_j^{-1},\ j = 1, ..., J\},$$

and

$$m_i = m_n = \max\{P_{ij},\ j = 1, ..., J\}.$$

We say that the system $Px = y$ is consistent if it has solutions x whose entries are all non-negative. The norm $||x||$ is the Euclidean norm.

10.1 The Block-Iterative EM Algorithm

For $k = 1, 2, ...$, let $n = k(\mathrm{mod}\, N)$. The block-iterative EM (BI-EM) has the iterative step

$$x_j^k = (1 - \gamma_n \delta_j s_{nj}) x_j^{k-1} + x_j^{k-1} \gamma_n \delta_j \sum_{i \in B_n} P_{ij} \frac{y_i}{(Px^{k-1})_i}, \tag{10.21}$$

with $\gamma > 0$ chosen so that

$$s_{nj} \delta_j \gamma_n \leq 1.$$

The *rescaled* BI-EM (RBI-EM) uses the largest values of γ_n consistent with these constraints.

The analogue of the MART is the EMART, with the iterative step

$$x_j^k = (1 - \gamma_i \delta_j P_{ij}) x_j^{k-1} + x_j^{k-1} \gamma_i \delta_j P_{ij} \frac{y_i}{(Px^{k-1})_i}, \tag{10.22}$$

with $P_{ij} \delta_j \gamma_i \leq 1$ and $i = k(\mathrm{mod}\, I)$. We have the following result concerning the BI-EM.

Theorem 10.1 *When the system $y = Px$ is consistent, the BI-EM sequence $\{x^k\}$ converges to a non-negative solution of $y = Px$, for any choice of blocks and any $x^0 > 0$.*

The inequality in the following lemma is the basis for the convergence proof.

Lemma 10.1 *Let $y = Px$ for some nonnegative x. Then for $\{x^k\}$ as in Equation (10.21) we have*

$$\sum_{j=1}^{J} \delta_j^{-1} KL(x_j, x_j^{k-1}) - \sum_{j=1}^{J} \delta_j^{-1} KL(x_j, x_j^k) \geq \tag{10.23}$$

$$\gamma_n \sum_{i \in B_n} KL(y_i, (Px^k)_i). \tag{10.24}$$

Proof: From the iterative step

$$x_j^k = x_j^{k-1}(1 - \delta_j \gamma_n \sigma_{nj}) + x_j^k \delta_j \gamma_n \sum_{i \in B_n} P_{ij} \frac{y_i}{(Px^k)_i} \tag{10.25}$$

we have

$$\log(x_j^k / x_j^{k-1}) = \log\Big((1 - \delta_j \gamma_n \sigma_{nj}) + \delta_j \gamma_n \sum_{i \in B_n} P_{ij} \frac{y_i}{(Px^k)_i}\Big). \tag{10.26}$$

By the concavity of the logarithm we obtain the inequality

$$\log(x_j^k / x_j^{k-1}) \geq \Big((1 - \delta_j \gamma_n \sigma_{nj}) \log 1 + \delta_j \gamma_n \sum_{i \in B_n} P_{ij} \log \frac{y_i}{(Px^k)_i}\Big), \tag{10.27}$$

or

$$\log(x_j^k / x_j^{k-1}) \geq \delta_j \gamma_n \sum_{i \in B_n} P_{ij} \log \frac{y_i}{(Px^k)_i}. \tag{10.28}$$

Therefore

$$\sum_{j=1}^J \delta_j^{-1} x_j \log(x_j^{k+1} / x_j^k) \geq \gamma_n \sum_{i \in B_n} (\sum_{j=1}^J x_j P_{ij}) \log \frac{y_i}{(Px^k)_i}. \tag{10.29}$$

Also

$$\sum_{j=1}^J \delta_j^{-1} (x_j^k - x_j^{k-1}) = \gamma_n \sum_{i \in B_n} ((Px^k)_i - y_i). \tag{10.30}$$

This concludes the proof of the lemma. ∎

From the inequality in (10.24) we can conclude several things:

- the sequence $\{\sum_{j=1}^J \delta_j^{-1} KL(x_j, x_j^k)\}$ is decreasing;
- the sequence $\{x^k\}$ is therefore bounded; and
- the sequence $\{\sum_{i \in B_n} KL(y_i, (Px^{mN+n-1})_i)\}$ is converging to zero.

Let x^* be any cluster point of the sequence $\{x^\}$. Then it is not difficult to show that $y = Px^*$. Replacing x with x^* we have that the sequence $\{\sum_{j=1}^J \delta_j^{-1} KL(x_j^*, x_j^k)\}$ is decreasing; since a subsequence converges to zero, so does the whole sequence. Therefore x^* is the limit of the sequence $\{x^k\}$. This proves that the algorithm produces a nonnegative solution of $y = Px$. We have been unable, so far, to replace the inequality in (10.24) with an equation in which the right side is independent of the particular solution x chosen. Therefore, in contrast with the BI-MART, we do not know which solution the BI-EM gives us, how the solution depends on the starting vector x^0, nor how the solution may depend on the choice of blocks.

The behavior of BI-EM illustrates once again that using block-iterative methods does not, by itself, lead to faster convergence. It seems that the main advantage of the use of these block-iterative methods is the opportunity to select the parameters. As with BI-MART, the key to accelerating the convergence of BI-EM is the proper choice of the parameters γ_n and δ_j. Recall that we must have

$$\gamma_n \delta_j s_{nj} \leq 1,$$

for all n and j. When we select $\delta_j = s_j^{-1}$, we must then have $\gamma_n \leq \mu_n^{-1}$. When we have $\delta_j = 1$, we need $\gamma_n \leq m_n^{-1}$. Generally speaking, the larger the γ_n the faster the convergence. The *rescaled* BI-EM (RBI-EM) uses the largest acceptable value of the γ_n.

In Meidunas (2001) [36] the RBI-EM was used to obtain sub-pixel resolution for satellite imaging.

10.2 The RAMLA

We must mention a method that closely resembles the EMART, the *row-action maximum likelihood algorithm* (RAMLA) [4], which was discovered independently by Browne and De Pierro (1996). The RAMLA avoids the limit cycle in the inconsistent case by using strong underrelaxation involving a decreasing sequence of relaxation parameters λ_k. The RAMLA is the following:

Algorithm 10.1 (RAMLA) *Let x^0 be an arbitrary positive vector, and $n = k(\mathrm{mod}\, N)$. Let the positive relaxation parameters λ_k converge to zero, with $\sum_{k=0}^{+\infty} \lambda_k = +\infty$. Then,*

$$x_j^k = (1 - \lambda_k s_{nj}) x_j^{k-1} + \lambda_k x_j^{k-1} \sum_{i \in B_n} P_{ij} \Big(\frac{y_i}{(Px^{k-1})_i} \Big). \tag{10.31}$$

11 The Paradigm of Alternating Minimization

When we formulate the problem in terms of minimizing the KL distance $KL(b(x^k), f(x))$ at each step, we are employing what might be called the *alternating minimization paradigm* of Csiszár and Tusnády (1984) [16]. Since $KL(b(x), f(x)) = -\log g(y|x)$, we have

$$-\log g(y|x^k) = KL(b(x^k), f(x^k)) \geq KL(b(x^k), f(x^{k+1}))$$
$$\geq KL(b(x^{k+1}), f(x^{k+1})) = -\log g(y|x^{k+1}),$$

so the likelihood function $g(y|x^k)$ is increasing.

The basic idea is to consider the two sets $B(X) = \{b(x)|x \in X\}$ and $F(X) = \{f(x)|x \in X\}$ and to find a pair $\{b, f\}$ where b is the member of $B(X)$ closest to $F(X)$, and f is the member of $F(X)$ closest to $B(X)$, where "closest" is yet to be defined. The sets and distance measure are defined in such a way that the optimal pair provides a solution to the original problem. Once a "distance" $d(b, f)$ is defined for $b \in B(X)$ and $f \in F(X)$, the kth step of the alternating minimization algorithm first has us minimize $d(b, f^k)$ to get b^k and then minimize $d(b^k, f)$ to get f^{k+1}. In order to mimic the case of the KL distance, we want $d(b(x), f(x))$ to be the objective function to be minimized, and

$$d(b(x), f(x^k)) \geq d(b(x^k), f(x^k)),$$

for all x and all k.

12 More on Convergence

There is a mistake in the proof of convergence given in Dempster, Laird, and Rubin (1977) [17]. Wu (1983) [47] 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) [35] we find the basic theory of the EM algorithm, including available results on convergence and the rate of convergence. Because many authors rely on Equation (1.2), 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 (3.14), as was eventually achieved in a sequence of papers (Shepp and Vardi (1982) [43], Lange and Carson (1984) [32], Vardi, Shepp and Kaufman (1985) [45], Lange, Bahn and Little (1987) [33], and Byrne (1993) [5]). When the EM algorithm was applied to list-mode data in SPECT and PET (Barrett, White, and Parra (1997) [2], and Huesman et al. (2000) [30], the resulting algorithm differed slightly from that in Equation (3.14) and a proof of convergence was provided in Byrne (2001) [7]. The convergence theorem in Byrne (2001) also establishes the convergence of the iteration in Equation (5.10) to the maximum-likelihood estimate of the mixing proportions, for the case of finite mixtures of probability density functions.

To illustrate a possible problem, we return to the example of $Y = Z_1 + Z_2$, the sum of two independent random variables uniformly distributed on $[0, x]$. Maximizing

$g(y|x)$ given by Equation (6.10), we find that the maximum-likelihood estimate of x, given the incomplete data, is $x_{ML} = y$. Now let us consider the EM algorithm for this case.

Suppose we have a current estimate x^k of the parameter x. The (E) step is to calculate the conditional expected value of

$$LL_c(x) = \log \chi_{[0,x]}(Z_1) + \log \chi_{[0,x]}(Z_2) - 2 \log x. \tag{12.32}$$

If $x^k > x$, then with positive probability Z_1 and Z_2 may exceed x, and the value of $LL_c(x)$ would be $-\infty$. Therefore, the conditional expected value of $LL_c(x)$ is finite, and equals $-2 \log x$, if and only if $x \geq x^k$. From our knowledge of y, we infer that $x \geq y/2$ also. The maximum of $-2 \log x$ then occurs when $x = x^{k+1}$ is the maximum of x^k and $y/2$. The sequence $\{x^k\}$ does not converge to $x_{ML} = y$.

13 Open Questions

As we have seen, even the basic formulation of the EM algorithm presents difficulties when probability density functions are involved. We have suggested several ways to avoid the difficulties, but other ways may also be useful.

Proving convergence of the sequence $\{x^k\}$ appears to involve the selection of an appropriate topology for the parameter space X. While it is common to assume that X 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, a number of additional assumptions appear necessary before convergence of the $\{x^k\}$ can be established. To simplify the analysis, it is often assumed that cluster points of the sequence lie in the interior of the set X, which is not a realistic assumption in some applications.

It may be wise to consider, instead, convergence of the functions $f(z|x^k)$, or maybe even to identify the parameters x with the functions $f(z|x)$. Proving convergence of the likelihood values $L(x^k)$ is also an option.

Accelerating convergence is an important area of research. The use of block-iterative methods has shown some promise, but the issue of subsequential convergence to a limit cycle, rather than to a single likelihood maximizer, is still a concern.

Regularization appears to be important in many applications, but when penalty functions are included the M step of the algorithm can no longer be performed without iteration or some other approximate method. Efficient regularization methods are still needed.

14 Conclusion

As I hope the reader is now convinced, the EM algorithm is still a work in progress and there are many gaps in the theory to be filled in. Because it is not really an algorithm, but a template for the design of algorithms, each particular application will generate its own difficulties. Because there is no general convergence theory that applies to all cases, each application will require its own convergence theory.

References

[1] Ahn, S., Fessler, J., Blatt, D., and Hero, A. (2006) “Convergent incremental optimization transfer algorithms: application to tomography.” *IEEE Transactions on Medical Imaging*, **25(3)**, pp. 283–296.

[2] Barrett, H., White, T., and Parra, L. (1997) “List-mode likelihood.”*J. Opt. Soc. Am. A* **14**, pp. 2914–2923.

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

[4] Browne, J. and A. DePierro, A. (1996) “A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography.”*IEEE Trans. Med. Imag.* **15**, pp. 687–699.

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

[6] Byrne, C. (1996) “Block-iterative methods for image reconstruction from projections.”*IEEE Transactions on Image Processing* **IP-5**, pp. 792–794.

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

[8] Byrne, C. (2002) “Iterative oblique projection onto convex sets and the split feasibility problem.”*Inverse Problems* **18**, pp. 441–453.

[9] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction."*Inverse Problems* **20**, pp. 103–120.

[10] Byrne, C. (2005) "Choosing parameters in block-iterative or ordered-subset reconstruction algorithms." *IEEE Transactions on Image Processing*, **14 (3)**, pp. 321–327.

[11] Byrne, C. (2005) *Signal Processing: A Mathematical Approach*, AK Peters, Publ., Wellesley, MA.

[12] Byrne, C. (2007) *Applied Iterative Methods*, AK Peters, Publ., Wellesley, MA.

[13] Byrne, C. " 'Csiszár and Tusnády' Revisited." *preprint.*

[14] Censor, Y., Eggermont, P.P.B., and Gordon, D. (1983) "Strong underrelaxation in Kaczmarz's method for inconsistent systems."*Numerische Mathematik* **41**, pp. 83–92.

[15] Censor, Y. and Segman, J. (1987) "On block-iterative maximization."*J. of Information and Optimization Sciences* **8**, pp. 275–291.

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

[17] 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.

[18] De Pierro, A. (1995) "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography."*IEEE Transactions on Medical Imaging* **14**, pp. 132–137.

[19] De Pierro, A., and Yamaguchi, M. (2001) "Fast EM-like methods for maximum 'a posteriori' estimates in emission tomography" *Transactions on Medical Imaging*, **20 (4)**.

[20] Eggermont, P., and LaRiccia, V. (2009) *unpublished book.*

[21] Erdogan, H., and Fessler, J. (1999) "Fast monotonic algorithms for transmission tomography" *IEEE Transactions on Medical Imaging*, **18(9)**, pp. 801–814.

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

[23] 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.

[24] Geman, S., and Geman, D. (1984) “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images.”*IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, pp. 721–741.

[25] Green, P. (1990) “Bayesian reconstructions from emission tomography data using a modified EM algorithm.”*IEEE Transactions on Medical Imaging* **9**, pp. 84–93.

[26] Hebert, T. and Leahy, R. (1989) “A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors.”*IEEE Transactions on Medical Imaging* **8**, pp. 194–202.

[27] Herman, G.T., Censor, Y., Gordon, D., and Lewitt, R. (1985) “Comment on the paper by Vardi, Shepp and Kaufman.” *Journal of the American Statistical Association* **80**, pp. 22–25.

[28] Holte, S., Schmidlin, P., Linden, A., Rosenqvist, G. and Eriksson, L. (1990) “Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems.”*IEEE Transactions on Nuclear Science* **37**, pp. 629–635.

[29] Hudson, H.M. and Larkin, R.S. (1994) “Accelerated image reconstruction using ordered subsets of projection data.”*IEEE Transactions on Medical Imaging* **13**, pp. 601–609.

[30] 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.

[31] Kullback, S. and Leibler, R. (1951) “On information and sufficiency.”*Annals of Mathematical Statistics* **22**, pp. 79–86.

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

[33] 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.

[34] Leahy, R. and Byrne, C. (2000) "Guest editorial: Recent development in iterative image reconstruction for PET and SPECT."*IEEE Trans. Med. Imag.* **19**, pp. 257–260.

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

[36] Meidunas, E. (2001) *Re-scaled Block Iterative Expectation Maximization Maximum Likelihood (RBI-EMML) Abundance Estimation and Sub-pixel Material Identification in Hyperspectral Imagery*, MS thesis, Department of Electrical Engineering, University of Massachusetts Lowell.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

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

[44] Schmidlin, P. (1972) “Iterative separation of sections in tomographic scintigrams.”*Nucl. Med.* **15(1)**.

[45] 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.

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

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