EM Algorithms from a Non-Stochastic Perspective

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

December 15, 2014

Abstract

The EM algorithm is not a single algorithm, but a template for the construction of iterative algorithms. While it is always presented in stochastic language, relying on conditional expectations to obtain a method for estimating parameters in statistics, the essence of the EM algorithm is not stochastic. The conventional formulation of the EM algorithm given in many texts and papers on the subject is inadequate. A new formulation is given here based on the notion of acceptable data.

1 Introduction

The "expectation maximization" (EM) algorithm is a general framework for maximizing the likelihood function in statistical parameter estimation [1, 2, 3]. It 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, or, as the authors of [4] put it, a "prescription for constructing an algorithm"; nevertheless, we shall continue to refer to *the* EM algorithm. As we shall demonstrate in Section 2, the essence of the EM algorithm is not stochastic. Our non-stochastic EM (NSEM) is a general approach for function maximization that has the stochastic EM methods as particular cases.

Maximizing the likelihood function is a well studied procedure for estimating parameters from observed data. When a maximizer cannot be obtained in closed form, iterative maximization algorithms, such as the expectation maximization (EM) maximum likelihood algorithms, are needed. The standard formulation of the EM algorithms postulates that finding a maximizer of the likelihood is complicated because the observed data is somehow incomplete or deficient, and the maximization

would have been simpler had we observed the complete data. The EM algorithm involves repeated calculations involving complete data that has been estimated using the current parameter value and conditional expectation.

The standard formulation is adequate for the most common discrete case, in which the random variables involved are governed by finite or infinite probability functions, but unsatisfactory in general, particularly in the continuous case, in which probability density functions and integrals are needed.

We adopt the view that the observed data is not necessarily incomplete, but just difficult to work with, while different data, which we call the preferred data, leads to simpler calculations. To relate the preferred data to the observed data, we assume that the preferred data is *acceptable*, which means that the conditional distribution of the preferred data, given the observed data, is independent of the parameter. This extension of the EM algorithms contains the usual formulation for the discrete case, while removing the difficulties associated with the continuous case. Examples are given to illustrate this new approach.

2 A Non-Stochastic Formulation of EM

The essence of the EM algorithm is not stochastic, and leads to a general approach for function maximization, which we call the "non-stochastic" EM algorithm (NSEM)[6]. In addition to being more general, this new approach also simplifies much of the development of the EM algorithm itself.

2.1 The Non-Stochastic EM Algorithm

We present now 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. For a broader treatment of the EM algorithm in the context of iterative optimization, see [5].

2.1.1 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}. Adopting such an iterative approach presupposes that maximizing $H(z^k, z)$ is simpler than maximizing $f(z)$ itself. This is the case with the EM algorithm.

The cross-entropy or Kullback-Leibler distance [7] 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 and easily proved 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*

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

This lemma can be extended to obtain the following useful identity.

Lemma 2.2 *For $f(z)$ and $b(x, z)$ as above, and z and w in Z, with $f(w) > 0$, we have*

$$KL(b(z), b(w)) = KL(f(z), f(w)) + KL(b(z), (f(z)/f(w))b(w)). \tag{2.5}$$

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

$$G(z^k, z) = KL(b(z^k), b(z)) - f(z), \tag{2.6}$$

where

$$KL(b(z^k), b(z)) = \int KL(b(x, z^k), b(x, z)) dx. \tag{2.7}$$

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

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

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

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 NSEM algorithm into the alternating minimization (AM) framework of Csiszár and Tusnády [12], to be discussed further Section 11.

2.1.2 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

$$H(z^k, z) = \sum_{x \in B} b(x, z^k) \log b(x, z) \tag{2.8}$$

to get z^{k+1}.

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

$$G(z^k, z) = KL(b(z^k), b(z)) - f(z), \tag{2.9}$$

where

$$KL(b(z^k), b(z)) = \sum_{x \in B} KL(b(x, z^k), b(x, z)). \tag{2.10}$$

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 Stochastic EM Algorithm

3.1 The E-step and M-step

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 proceeds in two steps. Having selected the preferred data X, and having found θ^k, we form the function of θ given by

$$Q(\theta|\theta^k) = E(\log f_X(x|\theta)|y, \theta^k); \tag{3.1}$$

this is the E-step of the EM algorithm. Then we maximize $Q(\theta|\theta^k)$ over all θ to get θ^{k+1}; this is the M-step of the EM algorithm. In this way, the EM algorithm based on X generates a sequence $\{\theta^k\}$ of parameter vectors.

For the discrete case of probability functions, we have

$$Q(\theta|\theta^k) = \sum_x f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta), \tag{3.2}$$

and for the continuous case of probability density functions we have

$$Q(\theta|\theta^k) = \int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx. \tag{3.3}$$

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 $\{L_y(\theta^k)\}$ non-decreasing.

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 Θ.

3.2 Difficulties with the Conventional Formulation

In [1] we are told that

$$f_{X|Y}(x|y,\theta) = f_X(x|\theta)/f_Y(y|\theta). \tag{3.4}$$

This is false; integrating with respect to x gives one on the left side and $1/f_Y(y|\theta)$ on the right side. Perhaps the equation is not meant to hold for all x, but just for some x. In fact, if there is a function h such that $Y = h(X)$, then Equation (3.4) might hold for those x such that $h(x) = y$. In fact, this is what happens in the discrete case of probabilities; in that case we do have

$$f_Y(y|\theta) = \sum_{x \in h^{-1}\{y\}} f_X(x|\theta), \tag{3.5}$$

where

$$h^{-1}\{y\} = \{x|h(x) = y\}.$$

Consequently,

$$f_{X|Y}(x|y,\theta) = \begin{cases} f_X(x|\theta)/f_Y(y|\theta), & \text{if } x \in h^{-1}\{y\}; \\ 0, & \text{if } x \notin h^{-1}\{y\}. \end{cases} \tag{3.6}$$

However, this modification of Equation (3.4) fails in the continuous case of probability density functions, since $h^{-1}\{y\}$ is often a subset of zero measure. Even if the set $h^{-1}\{y\}$ has positive measure, integrating both sides of Equation (3.4) over $x \in h^{-1}\{y\}$ tells us that $f_Y(y|\theta) \leq 1$, which need not hold for probability density functions.

3.3 An Incorrect Proof

Everyone who works with the EM algorithm will say that the likelihood is non-decreasing for the EM algorithm. The proof of this fact usually proceeds as follows; we use the notation for the continuous case, but the proof for the discrete case is essentially the same. Use Equation (3.4) to get

$$\log f_X(x|\theta) = \log f_{X|Y}(x|y,\theta) - \log f_Y(y|\theta). \tag{3.7}$$

Then replace the term $\log f_X(x|\theta)$ in Equation (3.3) with the right side of Equation (3.7), obtaining

$$\log f_Y(y|\theta) - Q(\theta|\theta^k) = -\int f_{X|Y}(x|y,\theta^k) \log f_{X|Y}(x|y,\theta) dx. \tag{3.8}$$

Jensen's Inequality tells us that

$$\int u(x) \log u(x) dx \geq \int u(x) \log v(x) dx, \tag{3.9}$$

for any probability density functions $u(x)$ and $v(x)$. Since $f_{X|Y}(x|y,\theta)$ is a probability density function, we have

$$\int f_{X|Y}(x|y,\theta^k) \log f_{X|Y}(x|y,\theta) dx \leq \int f_{X|Y}(x|y,\theta^k) \log f_{X|Y}(x|y,\theta^k) dx. \tag{3.10}$$

We conclude, therefore, that $\log f_Y(y|\theta) - Q(\theta|\theta^k)$ attains its minimum value at $\theta = \theta^k$. Then we have

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) \geq Q(\theta^{k+1}|\theta^k) - Q(\theta^k|\theta^k) \geq 0. \tag{3.11}$$

This proof is incorrect; clearly it rests on the validity of Equation (3.4), which is generally false. For the discrete case, with $Y = h(X)$, this proof is valid, when we use Equation (3.6), instead of Equation (3.4). In all other cases, however, the proof is incorrect.

3.4 Acceptable Data

We turn now to the question of how to repair the incorrect proof. Equation (3.4) should read

$$f_{X|Y}(x|y,\theta) = f_{X,Y}(x,y|\theta)/f_Y(y|\theta), \tag{3.12}$$

for all x. In order to replace $\log f_X(x|\theta)$ in Equation (3.3) we write

$$f_{X,Y}(x,y|\theta) = f_{X|Y}(x|y,\theta) f_Y(y|\theta), \tag{3.13}$$

and

$$f_{X,Y}(x,y|\theta) = f_{Y|X}(y|x,\theta) f_X(x|\theta), \tag{3.14}$$

so that

$$\log f_X(x|\theta) = \log f_{X|Y}(x|y,\theta) + \log f_Y(y|\theta) - \log f_{Y|X}(y|x,\theta). \tag{3.15}$$

We say that the preferred data is *acceptable* if

$$f_{Y|X}(y|x,\theta) = f_{Y|X}(y|x); \tag{3.16}$$

that is, the dependence of Y on X is unrelated to the value of the parameter θ. This definition provides our generalization of the relationship $Y = h(X)$.

When X is acceptable, we have that $\log f_Y(y|\theta) - Q(\theta|\theta^k)$ again attains its minimum value at $\theta = \theta^k$. The assertion that the likelihood is non-decreasing then follows, using the same argument as in the previous incorrect proof.

4 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

$$h^{-1}\{y\} = \{x \in B | h(x) = y\}. \tag{4.1}$$

Then we have

$$f_Y(y|\theta) = \sum_{x \in h^{-1}\{y\}} f_X(x|\theta). \tag{4.2}$$

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

$$f_{X|Y}(x|y,\theta) = \frac{f_X(x|\theta)}{f_Y(y|\theta)}, \tag{4.3}$$

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

$$Q(\theta|\theta^k) = E((\log f_X(X|\theta)|y,\theta^k) = \sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k) \log f_X(x|\theta), \tag{4.4}$$

and the M-step is to maximize $Q(\theta|\theta^k)$ as a function of θ to obtain θ^{k+1}.

Using Equation (4.3), we can write

$$Q(\theta|\theta^k) = \sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k) \log f_{X|Y}(x|y,\theta) + \log f_Y(y|\theta). \tag{4.5}$$

Therefore,

$$\log f_Y(y|\theta) - Q(\theta|\theta^k) = -\sum_{x\in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta).$$

Since

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

it follows from Jensen's Inequality that

$$-\sum_{x\in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta) \geq -\sum_{x\in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta^k).$$

Therefore, $\log f_Y(y|\theta) - Q(\theta|\theta^k)$ attains its minimum at $\theta = \theta^k$. We have the following result.

Proposition 4.1 *The sequence* $\{f_Y(y|\theta^k)\}$ *is non-decreasing.*

Proof: We have

$$\log f_Y(y|\theta^{k+1}) - Q(\theta^{k+1}|\theta^k) \geq \log f_Y(y|\theta^k) - Q(\theta^k|\theta^k),$$

or

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) \geq Q(\theta^{k+1}|\theta^k) - Q(\theta^k|\theta^k) \geq 0.$$

∎

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

$$\chi_{h^{-1}\{y\}}(x) = \begin{cases} 1, & \text{if } x \in h^{-1}\{y\}; \\ 0, & \text{if } x \notin h^{-1}\{y\}. \end{cases} \tag{4.6}$$

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 see once again that the sequence $\{f_Y(y|\theta^k)\}$ is non-decreasing, and also that 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.

5 Missing Data

We say that there is *missing data* if the preferred data X has the form $X = (Y, W)$, so that $Y = h(X) = h(Y, W)$, where h is the orthogonal projection onto the first component. The case of missing data for the discrete case is covered by the discussion in Section 4, so we consider here the continuous case in which probability density functions are involved.

Once again, the E-step is to calculate $Q(\theta|\theta^k)$ given by

$$Q(\theta|\theta^k) = E(\log f_X(X|\theta)|y, \theta^k). \tag{5.1}$$

Since $X = (Y, W)$, we have

$$f_X(x|\theta) = f_{Y,W}(y, w|\theta). \tag{5.2}$$

Since the set $h^{-1}\{y\}$ has measure zero, we cannot write

$$Q(\theta|\theta^k) = \int_{h^{-1}\{y\}} f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx.$$

Instead, following [8], we write

$$Q(\theta|\theta^k) = \int f_{Y,W}(y, w|\theta^k) \log f_{Y,W}(y, w|\theta) dw / f_Y(y|\theta^k). \tag{5.3}$$

Consequently, maximizing $Q(\theta|\theta^k)$ is equivalent to maximizing

$$\int f_{Y,W}(y, w|\theta^k) \log f_{Y,W}(y, w|\theta) dw.$$

With $b(\theta) = b(\theta, w) = f_{Y,W}(y, w|\theta)$ and

$$f_Y(y|\theta) = f(\theta) = \int f_{Y,W}(y, w|\theta) dw = \int b(\theta) dw,$$

we find that maximizing $Q(\theta|\theta^k)$ is equivalent to minimizing $KL(b(\theta^k), b(\theta)) - f(\theta)$. Therefore, the EM algorithm for the case of missing data falls into the framework of the non-stochastic EM algorithm. We conclude that the sequence $\{f(\theta^k)\}$ is non-decreasing, and that the sequence $\{KL(b(\theta^k), b(\theta^{k+1}))\}$ converges to zero.

Most other instances of the continuous case in which we have $Y = h(X)$ can be handled using the missing-data model. For example, suppose that Z_1 and Z_2 are uniformly distributed on the interval $[0, \theta]$, for some positive θ, and that $Y = Z_1 + Z_2$. We may, for example, then take W to be $W = Z_1 - Z_2$ and $X = (Y, W)$ as the preferred data. We shall discuss these instances further in Section 7.

6 The Continuous Case

We turn now to the general continuous case. We 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.

6.1 Acceptable Preferred Data

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

$$Q(\theta|\theta^k) = E(\log f_X(X|\theta)|y, \theta^k) = \int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx. \tag{6.1}$$

Assuming the acceptability condition and 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

$$H(\theta^k, \theta) = \int f_{X,Y}(x, y|\theta^k) \log f_{X,Y}(x, y|\theta) dx. \tag{6.2}$$

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

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

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

In the discrete case in which $Y = h(X)$ the conditional probability $f_{Y|X}(y|x, \theta)$ is $\delta(y - h(x))$, as a function of y, for given x, and is the characteristic function of the set $h^{-1}(y)$, as a function of x, for given y. Therefore, we can write $f_{X|Y}(x|y, \theta)$ using Equation (3.6). For the continuous case in which $Y = h(X)$, the pdf $f_{Y|X}(y|x, \theta)$ is again a delta function of y, for given x; the difficulty arises when we need to view this as a function of x, for given y. The acceptability property helps us avoid this difficulty.

When X is acceptable, we have

$$f_{X|Y}(x|y, \theta) = f_{Y|X}(y|x) f_X(x|\theta) / f_Y(y|\theta), \tag{6.3}$$

whenever $f_Y(y|\theta) \neq 0$, and is zero otherwise. Consequently, when X is acceptable, we have a kernel model for $f_Y(y|\theta)$ in terms of the $f_X(x|\theta)$:

$$f_Y(y|\theta) = \int f_{Y|X}(y|x) f_X(x|\theta) dx; \tag{6.4}$$

for the continuous case we view this as a corrected version of Equation (3.5). In the discrete case the integral is replaced by a summation, of course, but when we are speaking generally about either case, we shall use the integral sign.

The acceptability of the missing data W is used in [9], but more for computational convenience and to involve the Kullback-Leibler distance in the formulation of the EM algorithm. It is not necessary that W be acceptable in order for likelihood to be non-decreasing, as we have seen.

6.2 Selecting Preferred Data

The popular example of multinomial data given below illustrates well the point that one can often choose to view the observed data as "incomplete" simply in order to introduce "complete" data that makes the calculations simpler, even when there is no suggestion, in the original problem, that the observed data is in any way inadequate or "incomplete" . It is in order to emphasize this desire for simplification that we refer to X as the preferred data, not the complete data.

In some applications, the preferred data X arises naturally from the problem, while in other cases the user must imagine preferred data. This choice in selecting the preferred data can be helpful in speeding up the algorithm (see [10]).

If, instead of maximizing

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

at each M-step, we simply select θ^{k+1} so that

$$\int f_{X|Y}(x|y, \theta^k) \log f_{X,Y}(x, y|\theta^{k+1}) dx - \int f_{X|Y}(x|y, \theta^k) \log f_{X,Y}(x, y|\theta^k) dx > 0,$$

we say that we are using a *generalized* EM (GEM) algorithm. It is clear from the discussion in the previous subsection that, whenever X is acceptable, a GEM also guarantees that likelihood is non-decreasing.

6.3 Preferred Data as Missing Data

As we have seen, when the EM algorithm is applied to the missing-data model, the likelihood is non-decreasing, which suggests that, for an arbitrary preferred data X,

we could imagine X as W, the missing data, and imagine applying the EM algorithm to $Z = (Y, X)$. This approach would produce an EM sequence of parameter vectors for which likelihood is non-decreasing, but it need not be the same sequence as obtained by applying the EM algorithm to X directly. It is the same sequence, provided that X is acceptable. We are not suggesting that applying the EM algorithm to $Z = (Y, X)$ would simplify calculations.

We know that, when the missing-data model is used and the M-step is defined as maximizing the function in (5.3), the likelihood is not decreasing. It would seem then that, for any choice of preferred data X, we could view this data as missing and take as our complete data the pair $Z = (Y, X)$, with X now playing the role of W. Maximizing the function in (5.3) is then equivalent to maximizing

$$\int f_{X|Y}(x|y, \theta^k) \log f_{X,Y}(x, y|\theta) dx; \tag{6.5}$$

to get θ^{k+1}. It then follows that $L_y(\theta^{k+1}) \geq L_y(\theta^k)$. The obvious question is whether or not these two functions given in (3.1) and (6.5) have the same maximizers.

For acceptable X we have

$$\log f_{X,Y}(x, y|\theta) = \log f_X(x|\theta) + \log f_{Y|X}(y|x), \tag{6.6}$$

so the two functions given in (3.1) and (6.5) do have the same maximizers. It follows once again that, whenever the preferred data is acceptable, we have $L_y(\theta^{k+1}) \geq L_y(\theta^k)$. Without additional assumptions, however, we cannot conclude that $\{\theta^k\}$ converges to θ_{ML}, nor that $\{f_Y(y|\theta^k)\}$ converges to $f_Y(y|\theta_{ML})$.

7 The Continuous Case with $Y = h(X)$

In this section we consider the continuous case in which the observed random vector Y takes values in $\mathbb{R}^N$, the preferred random vector X takes values in $\mathbb{R}^M$, the random vectors are governed by probability density functions $f_Y(y|\theta)$ and $f_X(x|\theta)$, respectively, and there is a function $h : \mathbb{R}^N \to \mathbb{R}^M$ such that $Y = h(X)$. In most cases, $M > N$ and $h^{-1}\{y\} = \{x|h(x) = y\}$ has measure zero in $\mathbb{R}^M$.

7.1 An Example

For example, suppose that Z_1 and Z_2 are independent and uniformly distributed on the interval $[0, \theta]$, for some $\theta > 0$ to be estimated. Let $Y = Z_1 + Z_2$. With

$Z = (Z_1, Z_2)$, and $h : \mathbb{R}^2 \to \mathbb{R}$ given by $h(z_1, z_2) = z_1 + z_2$, we have $Y = h(Z)$. The pdf for Z is

$$f_Z(z|\theta) = f_Z(z_1, z_2|\theta) = \frac{1}{\theta^2}\chi_{[0,\theta]}(z_1)\chi_{[0,\theta]}(z_2). \tag{7.1}$$

The pdf for Y is

$$f_Y(y|\theta) = \begin{cases} \frac{y}{\theta^2}, & \text{if } 0 \le y \le \theta; \\ \frac{2\theta - y}{\theta^2}, & \text{if } \theta \le y \le 2\theta. \end{cases} \tag{7.2}$$

It is not the case that

$$f_Y(y|\theta) = \int_{h^{-1}\{y\}} f_Z(z|\theta), \tag{7.3}$$

since $h^{-1}\{y\}$ has measure zero in $\mathbb{R}^2$.

The likelihood function is $L(\theta) = f_Y(y|\theta)$, viewed as a function of θ, and is given by

$$L(\theta) = \begin{cases} \frac{y}{\theta^2}, & \text{if } \theta \ge y; \\ \frac{2\theta - y}{\theta^2}, & \text{if } \frac{y}{2} \le \theta \le y. \end{cases} \tag{7.4}$$

Therefore, the maximum likelihood estimate of θ is $\theta_{ML} = y$.

Instead of using Z as our preferred data, suppose that we define the random variable $W = Z_2$, and let $X = (Y, W)$, a missing-data model. We then have $Y = h(X)$, where $h : \mathbb{R}^2 \to \mathbb{R}$ is given by $h(x) = h(y, w) = y$. The pdf for Y given in Equation (7.2) can be written as

$$f_Y(y|\theta) = \int \frac{1}{\theta^2}\chi_{[0,\theta]}(y - w)\chi_{[0,\theta]}(w)dw. \tag{7.5}$$

The joint pdf is

$$f_{Y,W}(y, w|\theta) = \begin{cases} 1/\theta^2, & \text{for } w \le y \le \theta + w; \\ 0, & \text{otherwise.} \end{cases} \tag{7.6}$$

7.2 Censored Exponential Data

McLachlan and Krishnan [1] give the following example of a likelihood maximization problem involving probability density functions. This example provides a good illustration of the usefulness of the missing-data model.

Suppose that Z is the time until failure of a component, which we assume is governed by the exponential distribution

$$f(z|\theta) = \frac{1}{\theta}e^{-z/\theta}, \tag{7.7}$$

where the parameter $\theta > 0$ is the expected time until failure. We observe a random sample of N components and record their failure times, z_n. On the basis of this data, we must estimate θ, 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. The *censored data* is then $y = (y_1, ..., y_N)$, where $y_n = z_n$ is the time until failure for $n = 1, ..., r$, and $y_n = T$ for $n = r + 1, ..., N$. The censored data is reasonably viewed as *incomplete*, relative to the *complete* data we would have had, had the trial lasted until all the components had failed.

Since the probability that a component will survive until time T is $e^{-T/\theta}$, the pdf for the vector y is

$$f_Y(y|\theta) = \Big(\prod_{n=1}^{r} \frac{1}{\theta} e^{-y_n/\theta} \Big) \, e^{-(N-r)T/\theta}, \tag{7.8}$$

and the log likelihood function for the censored, or incomplete, data is

$$\log f_Y(y|\theta) = -r \log \theta - \frac{1}{\theta} \sum_{n=1}^{N} y_n. \tag{7.9}$$

In this particular example we are fortunate, in that we can maximize $f_Y(y|\theta)$ easily, and find that the ML solution based on the incomplete, censored data is

$$\theta_{MLi} = \frac{1}{r} \sum_{n=1}^{N} y_n = \frac{1}{r} \sum_{n=1}^{r} y_n + \frac{N-r}{r} T. \tag{7.10}$$

In most cases in which our data is incomplete, finding the ML estimate from the incomplete data is difficult, while finding it for the complete data is relatively easy.

We say that the missing data are the times until failure of those components that did not fail during the observation time. The preferred data is the complete data $x = (z_1, ..., z_N)$ of actual times until failure. The pdf for the preferred data X is

$$f_X(x|\theta) = \prod_{n=1}^{N} \frac{1}{\theta} e^{-z_n/\theta}, \tag{7.11}$$

and the log likelihood function based on the complete data is

$$\log f_X(x|\theta) = -N \log \theta - \frac{1}{\theta} \sum_{n=1}^{N} z_n. \tag{7.12}$$

The ML estimate of θ from the complete data is easily seen to be

$$\theta_{MLc} = \frac{1}{N}\sum_{n=1}^{N} z_n. \tag{7.13}$$

In this example, both the incomplete-data vector y and the preferred-data vector x lie in $\mathbb{R}^N$. We have $y = h(x)$ where the function h operates by setting to T any component of x that exceeds T. Clearly, for a given y, the set $h^{-1}\{y\}$ consists of all vectors x with entries $x_n \geq T$ or $x_n = y_n < T$. For example, suppose that $N = 2$, and $y = (y_1, T)$, where $y_1 < T$. Then $h^{-1}\{y\}$ is the one-dimensional ray

$$h^{-1}\{y\} = \{x = (y_1, x_2) |\, x_2 \geq T\}.$$

Because this set has measure zero in $\mathbb{R}^2$, Equation (7.3) does not make sense in this case.

We need to calculate $E(\log f_X(X|\theta)|y, \theta^k)$. Following McLachlan and Krishnan [1], we note that since $\log f_X(x|\theta)$ is linear in the unobserved data Z_n, $n = r+1, ..., N$, to calculate $E(\log f_X(X|\theta)|y, \theta^k)$ we need only replace the unobserved values with their conditional expected values, given y and θ^k. The conditional distribution of $Z_n - T$, given that $Z_n > T$, is still exponential, with mean θ. Therefore, we replace the unobserved values, that is, all the Z_n for $n = r+1, ..., N$, with $T + \theta^k$. Therefore, at the E-step we have

$$E(\log f_X(X|\theta)|y, \theta^k) = -N\log\theta - \frac{1}{\theta}\bigg(\Big(\sum_{n=1}^{N} y_n\Big) + (N-r)\theta^k\bigg). \tag{7.14}$$

The M-step is to maximize this function of θ, which leads to

$$\theta^{k+1} = \bigg(\Big(\sum_{n=1}^{N} y_n\Big) + (N-r)\theta^k\bigg)/N. \tag{7.15}$$

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

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

so that

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

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

7.3 A More General Approach

Let X take values in $\mathbb{R}^N$, and $Y = h(X)$ take values in $\mathbb{R}^M$, where $M < N$ and $h : \mathbb{R}^N \to \mathbb{R}^M$ is a (possibly) many-to-one function. Suppose that there is a second function $k : \mathbb{R}^N \to \mathbb{R}^{N-M}$ such that the function

$$G(x) = (h(x), k(x)) = (y, w) = u \tag{7.16}$$

has inverse $H(y, w) = x$. Denote by $J(y, w)$ the determinant of the Jacobian matrix associated with the transformation G. Let

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

Then

$$f_Y(y|\theta) = \int_{w \in \mathcal{W}(y)} f_X(H(y, w))J(y, w)dw. \tag{7.17}$$

Then we apply the missing-data model for the EM algorithm, with $W = k(X)$ as the missing data.

8 A Multinomial Example

In many applications, the entries of the vector y are independent realizations of a single real-valued or vector-valued random variable V, as they are, at least initially, for finite mixture problems to be considered later. This is not always the case, however, as the following example shows.

A well known example that was used in [11] and again in [1] to illustrate the EM algorithm concerns a multinomial model taken from genetics. Here there are four cells, with cell probabilities $\frac{1}{2} + \frac{1}{4}\theta_0$, $\frac{1}{4}(1 - \theta_0)$, $\frac{1}{4}(1 - \theta_0)$, and $\frac{1}{4}\theta_0$, for some $\theta_0 \in \Theta = [0, 1]$ to be estimated. The entries of y are the frequencies from a sample size of 197. We then have

$$f_Y(y|\theta) = \frac{197!}{y_1!y_2!y_3!y_4!}(\frac{1}{2} + \frac{1}{4}\theta)^{y_1}(\frac{1}{4}(1 - \theta))^{y_2}(\frac{1}{4}(1 - \theta))^{y_3}(\frac{1}{4}\theta)^{y_4}. \tag{8.1}$$

It is then supposed that the first of the original four cells can be split into two sub-cells, with probabilities $\frac{1}{2}$ and $\frac{1}{4}\theta_0$. We then write $y_1 = y_{11} + y_{12}$, and let

$$X = (Y_{11}, Y_{12}, Y_2, Y_3, Y_4), \tag{8.2}$$

where X has a multinomial distribution with five cells. Note that we do now have $Y = h(X)$.

This example is a popular one in the literature on the EM algorithm (see [11] for citations). It is never suggested that the splitting of the first group into two subgroups is motivated by the demands of the genetics theory itself. As stated in [1], the motivation for the splitting is to allow us to view the two random variables $Y_{12} + Y_4$ and $Y_2 + Y_3$ as governed by a binomial distribution; that is, we can view the value of $y_{12} + y_4$ as the number of heads, and the value $y_2 + y_3$ as the number of tails that occur in the flipping of a biased coin $y_{12} + y_4 + y_2 + y_3$ times. This simplifies the calculation of the likelihood maximizer.

9 The Example of Finite Mixtures

We say that a random vector V taking values in $\mathbb{R}^D$ is a *finite mixture* if, for $j = 1, ..., J$, f_j is a probability density function or probability function, $\theta_j \geq 0$ is a weight, the θ_j sum to one, and the probability density function or probability function for V is

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v). \tag{9.1}$$

The value of D is unimportant and for simplicity, we shall assume that $D = 1$.

We draw N independent samples of V, denoted v_n, and let y_n, the nth entry of the vector y, be the number v_n. To create the preferred data we assume that, for each n, the number v_n is a sample of the random variable V_n whose pdf or pf is f_{j_n}, where the probability that $j_n = j$ is θ_j. We then let the N entries of the preferred data X be the indices j_n. The conditional distribution of Y, given X, is clearly independent of the parameter vector θ, and is given by

$$f_{Y|X}(y|x, \theta) = \prod_{n=1}^{N} f_{j_n}(y_n);$$

therefore, X is acceptable. Note that we cannot recapture the entries of y from those of x, so the model $Y = h(X)$ does not hold here. Note also that, although the vector y is taken originally to be a vector whose entries are independently drawn samples from V, when we create the preferred data X we change our view of y. Now each entry of y is governed by a different distribution, so y is no longer viewed as a vector of independent sample values of a single random vector.

10 The EM and the Kullback-Leibler Distance

We illustrate the usefulness of acceptability and reformulate the M-step in terms of cross-entropy or Kullback-Leibler distance minimization.

10.1 Using Acceptable Data

The assumption that the data X is acceptable helps simplify the theoretical discussion of the EM algorithm.

For any preferred X the M-step of the EM algorithm, in the continuous case, is to maximize the function

$$\int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx, \tag{10.1}$$

over $\theta \in \Theta$; the integral is replaced by a sum in the discrete case. For notational convenience we let

$$b(\theta^k) = f_{X|Y}(x|y,\theta^k), \tag{10.2}$$

and

$$f(\theta) = f_X(x|\theta); \tag{10.3}$$

both functions are functions of the vector variable x. Then the M-step is equivalent to minimizing the Kullback-Leibler or cross-entropy distance

$$KL(b(\theta^k), f(\theta)) = \int f_{X|Y}(x|y,\theta^k)\log\Big(\frac{f_{X|Y}(x|y,\theta^k)}{f_X(x|\theta)}\Big)dx$$

$$= \int f_{X|Y}(x|y,\theta^k)\log\Big(\frac{f_{X|Y}(x|y,\theta^k)}{f_X(x|\theta)}\Big) + f_X(x|\theta) - f_{X|Y}(x|y,\theta^k)dx. \tag{10.4}$$

This holds since both $f_X(x|\theta)$ and $f_{X|Y}(x|y,\theta^k)$ are probability density functions or probabilities.

For acceptable X we have

$$\log f_{X,Y}(x,y|\theta) = \log f_{X|Y}(x|y,\theta) + \log f_Y(y|\theta) = \log f_{Y|X}(y|x) + \log f_X(x|\theta). \tag{10.5}$$

Therefore,

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta) =$$

$$KL(b(\theta^k), f(\theta)) - KL(b(\theta^k), f(\theta^{k+1})) + KL(b(\theta^k), b(\theta^{k+1})) - KL(b(\theta^k), b(\theta)). \tag{10.6}$$

Since $\theta = \theta^{k+1}$ minimizes $KL(b(\theta^k), f(\theta))$, we have that

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) =$$

$$KL(b(\theta^k), f(\theta^k)) - KL(b(\theta^k), f(\theta^{k+1})) + KL(b(\theta^k), b(\theta^{k+1})) \geq 0. \tag{10.7}$$

This tells us, once again, that the sequence of likelihood values $\{\log f_Y(y|\theta^k)\}$ is increasing, and that the sequence of its negatives, $\{-\log f_Y(y|\theta^k)\}$, is decreasing. Since we assume that there is a maximizer θ_{ML} of the likelihood, the sequence $\{-\log f_Y(y|\theta^k)\}$ is also bounded below and the sequences $\{KL(b(\theta^k), b(\theta^{k+1}))\}$ and $\{KL(b(\theta^k), f(\theta^k)) - KL(b(\theta^k), f(\theta^{k+1}))\}$ converge to zero.

Without some notion of convergence in the parameter space Θ, we cannot conclude that $\{\theta^k\}$ converges to a maximum likelihood estimate θ_{ML}. Without some additional assumptions, we cannot even conclude that the functions $f(\theta^k)$ converge to $f(\theta_{ML})$.

11 The Approach of Csiszár and Tusnády

For acceptable X the M-step of the EM algorithm is to minimize the function $KL(b(\theta^k), f(\theta))$ over $\theta \in \Theta$ to get θ^{k+1}. To put the EM algorithm into the framework of the *alternating minimization* approach of Csiszár and Tusnády [12], we need to view the M-step in a slightly different way; the problem is that, for the continuous case, having found θ^{k+1}, we do not then minimize $KL(b(\theta), f(\theta^{k+1}))$ at the next step.

11.1 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 and minimizing $\Psi(p^n, q)$ gives q^{n+1}, so that

$$\Psi(p^n, q^n) \geq \Psi(p^n, q^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}). \tag{11.1}$$

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

$$\Psi(p, q^{n+1}) - \Psi(p^{n+1}, q^{n+1}) \geq \Delta(p, p^{n+1}), \tag{11.2}$$

then the *three-point property* holds. If

$$\Delta(p, p^n) + \Psi(p, q) \geq \Psi(p, q^{n+1}), \tag{11.3}$$

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

$$\Delta(p, p^n) - \Delta(p, p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p, q). \tag{11.4}$$

From the inequality in (11.4) 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 (11.4) 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 (11.2) and (11.3):

$$\Psi(p, q) + \Psi(p, q^{n-1}) \geq \Psi(p, q^n) + \Psi(p^n, q^{n-1}). \tag{11.5}$$

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

$$\Psi(p, q^{n-1}) - \Psi(p, q^n) \geq \Psi(p^n, q^n) - \Psi(p, q), \tag{11.6}$$

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

11.2 Alternating Minimization for the EM Algorithm

Assume that X is acceptable. We define the function $F(\theta)$ to be

$$F(\theta) = \int f_{X|Y}(x|y, \theta) \log f_{Y|X}(y|x) dx, \tag{11.7}$$

for the continuous case, with a sum replacing the integral for the discrete case. Using the identities

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

we then have

$$\log f_Y(y|\theta) = F(\theta') + KL(b(\theta'), b(\theta)) - KL(b(\theta'), f(\theta)), \tag{11.8}$$

for any parameter values θ and θ'. With the choice of $\theta' = \theta$ we have

$$\log f_Y(y|\theta) = F(\theta) - KL(b(\theta), f(\theta)). \tag{11.9}$$

Therefore, subtracting Equation 11.9 from Equation 11.8, we get

$$\Big(KL(b(\theta'), f(\theta)) - F(\theta')\Big) - \Big(KL(b(\theta), f(\theta)) - F(\theta)\Big) = KL(b(\theta'), b(\theta)). \tag{11.10}$$

Now we can put the EM algorithm into the alternating-minimization framework.

Define

$$\Psi(b(\theta'), f(\theta)) = KL(b(\theta'), f(\theta)) - F(\theta'). \tag{11.11}$$

We know from Equation (11.10) that

$$\Psi(b(\theta'), f(\theta)) - \Psi(b(\theta), f(\theta)) = KL(b(\theta'), b(\theta)). \tag{11.12}$$

Therefore, we can say that the M-step of the EM algorithm is to minimize $\Psi(b(\theta^k), f(\theta))$ over $\theta \in \Theta$ to get θ^{k+1} and that minimizing $\Psi(b(\theta), f(\theta^{k+1}))$ gives us $\theta = \theta^{k+1}$ again. Because the EM algorithm can be viewed as an alternating minimization method, it is also a particular case of the sequential unconstrained minimization techniques [13], and of "optimization transfer" [4].

With the choice of

$$\Delta(b(\theta'), b(\theta)) = KL(b(\theta'), b(\theta)),$$

Equation (11.12) becomes

$$\Psi(b(\theta'), f(\theta)) - \Psi(b(\theta), f(\theta)) = \Delta(b(\theta'), b(\theta)), \tag{11.13}$$

which is the three-point property.

With $P = \mathcal{B}(\Theta)$ and $Q = \mathcal{F}(\Theta)$ the collections of all functions $b(\theta)$ and $f(\theta)$, respectively, we can view the EM algorithm as alternating minimization of the function

$\Psi(p, q)$, over $p \in P$ and $q \in Q$. As we have seen, the three-point property holds. What about the four-point property?

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 and convex subsets of $\mathbb{R}^N$. In the continuous case of the EM algorithm, we are not performing alternating minimization on the function $KL(b(\theta), f(\theta'))$, but on $KL(b(\theta), f(\theta')) + F(\theta)$. In the discrete case, whenever $Y = h(X)$, the function $F(\theta)$ is always zero, so we are performing alternating minimization on the KL distance $KL(b(\theta), f(\theta'))$. In [16] the authors consider the problem of minimizing a function of the form

$$\Lambda(p, q) = \phi(p) + \psi(q) + D_g(p, q), \tag{11.14}$$

where ϕ and ψ are convex and differentiable on $\mathbb{R}^J$, D_g is a Bregman distance, and $P = Q$ is the interior of the domain of g. In [13] 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. In some particular instances, the collection of the functions $f(\theta)$ is a convex subset of $\mathbb{R}^J$, as well, so the three- and four-point properties hold.

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.

The five-point property for the EM case is the following:

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

$$\Big(KL(b(\theta^k), f(\theta^k)) - F(\theta^k)\Big) - \Big(KL(b(\theta), f(\theta)) - F(\theta)\Big). \tag{11.15}$$

12 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.

12.1 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

$$f_X(x|\lambda) = \prod_{n=1}^N \lambda_n^{x_n} \exp(-\lambda_n)/x_n! = \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \tag{12.1}$$

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

$$\begin{aligned} \text{Prob}(Y = y) &= \text{Prob}(X_1 + ... + X_N = y) \\ &= \sum_{x_1+...x_N=y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \end{aligned} \tag{12.2}$$

As we shall see shortly, we have

$$\sum_{x_1+...x_N=y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y! \,. \tag{12.3}$$

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

$$f_{X|Y}(x|y, \lambda) = \frac{y!}{x_1!...x_N!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1} ... \Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \tag{12.4}$$

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.

12.2 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

$$\binom{y}{x_1}\binom{y-x_1}{x_2}...\binom{y-(x_1+...+x_{N-2})}{x_{N-1}} \tag{12.5}$$

ways, or in

$$\frac{y!}{x_1!(y-x_1)!}...\frac{(y-(x_1+...+x_{N-2}))!}{x_{N-1}!(y-(x_1+...+x_{N-1}))!} = \frac{y!}{x_1!...x_N!}. \tag{12.6}$$

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

$$(a_1+...+a_N)^y = \sum_{x_1+...+x_N=y} \frac{y!}{x_1!...x_N!} a_1^{x_1}...a_N^{x_N}. \tag{12.7}$$

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

$$\begin{aligned} 1 = 1^y &= \Big(\frac{\lambda_1}{\lambda_+}+...+\frac{\lambda_N}{\lambda_+}\Big)^y \\ &= \sum_{x_1+...+x_N=y} \frac{y!}{x_1!...x_N!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}...\Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \end{aligned} \tag{12.8}$$

From this we get

$$\sum_{x_1+\ldots x_N=y} \exp(-\lambda_+) \prod_{n=1}^{N} \lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y! \ . \tag{12.9}$$

13 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) [17].

13.1 The SPECT Reconstruction Problem

In their 1976 paper Rockmore and Makovski [18] suggested that the problem of reconstructing a tomographic image be viewed as statistical parameter estimation. Shepp and Vardi [19] expanded on this idea and suggested that the EM algorithm discussed by Dempster, Laird and Rubin [11] 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 $s_j = \sum_{i=1}^{I} P_{ij} > 0$.

13.1.1 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

$$f_X(x|\lambda) = \prod_{i=1}^{I}\prod_{j=1}^{J} \exp^{-P_{ij}\lambda_j}(P_{ij}\lambda_j)^{x_{ij}}/x_{ij}! \ . \tag{13.1}$$

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

$$LL_x(\lambda) = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(-P_{ij}\lambda_j + X_{ij}\log(P_{ij}\lambda_j)\Big). \tag{13.2}$$

Of course, we do not have the complete data.

13.1.2 The Incomplete Data

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

$$Y_i = \sum_{j=1}^{J} X_{ij}; \tag{13.3}$$

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

$$LL_y(\lambda) = \sum_{i=1}^{I}\Big(-(P\lambda)_i + y_i\log((P\lambda)_i)\Big). \tag{13.4}$$

Maximizing $LL_x(\lambda)$ in Equation (13.2) is easy, while maximizing $LL_y(\lambda)$ in Equation (13.4) 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:

$$\lambda_j^{k+1} = \lambda_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\lambda^k)_i}. \tag{13.5}$$

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

$$f_Y(y|\lambda) = \sum_{x \in h^{-1}\{y\}} f_X(x|\lambda). \tag{13.6}$$

13.2 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

$$\lambda_j^{k+1} = \lambda_j^k \sum_{i=1}^{I} P_{i,j} \frac{y_i}{(P\lambda^k)_i}. \tag{13.7}$$

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 [20]:

$$KL(y, P\lambda^k) - KL(y, P\lambda^{k+1}) = KL(r(\lambda^k), r(\lambda^{k+1})) + KL(\lambda^{k+1}, \lambda^k); \tag{13.8}$$

and

$$KL(\lambda_{ML}, \lambda^k) - KL(\lambda_{ML}, \lambda^{k+1}) \geq KL(y, P\lambda^k) - KL(y, P\lambda_{ML}). \tag{13.9}$$

14 Non-Negative Solutions for Linear Equations

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, more generally, 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.

14.1 The General Case

Suppose we want a non-negative solution x for a system $Ax = b$ of real equations; unless b is positive and A has only non-negative entries, we cannot use the EMML algorithm directly. We may, however, be able to transform $Ax = b$ to $P\lambda = y$.

Suppose that, by rescaling the equations in $Ax = b$, we can make $c_j = \sum_{i=1}^I A_{ij} > 0$, for each $j = 1, ..., J$, and $b_+ = \sum_{i=1}^I b_i > 0$. Now replace A_{ij} with $G_{ij} = A_{ij}/c_j$, and x_j with $z_j = c_j x_j$; then $Gz = Ax = b$ and $\sum_{i=1}^I G_{ij} = 1$, for all j. We also know now that $b_+ = z_+ > 0$, so z_+ is now known.

Let U and u be the matrix and column vector whose entries are all one, respectively, and let $t > 0$ be large enough so that all the entries of $B = G + tU$ and $(tz_+)u$ are positive. Now

$$Bz = Gz + (tz_+)u = b + (tz_+)u.$$

We then solve $Bz = b + (tz_+)u$ for z. It follows that $Ax = Gz = b$ and $x \geq 0$. Finally, we let $P = B$, $\lambda = z$, and $y = b + (tb_+)u$.

14.2 Regularization

It is often the case, as in tomography, that the entries of the vector y are obtained by measurements, and are therefore noisy. Finding an exact solution of $y = P\lambda$ or even minimizing $KL(y, P\lambda)$ may not be advisable in such cases. To obtain an approximate solution that is relatively insensitive to the noise in y we *regularize*. One way to do that is to minimize not $KL(y, P\lambda)$, but

$$F_\alpha(\lambda) = (1 - \alpha)KL(y, P\lambda) + \alpha KL(p, \lambda), \tag{14.1}$$

where $\alpha \in (0,1)$ and $p > 0$ is a prior estimate of the desired λ. The iterative step of the regularized EMML algorithm is now

$$\lambda_j^{k+1} = (1-\alpha)\Big(\lambda_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\lambda^k)_i}\Big) + \alpha p_j. \tag{14.2}$$

As was shown in [20], the sequence $\{\lambda^k\}$ converges to a minimizer of $F_\alpha(\lambda)$.

14.3 Acceleration

When the system $y = P\lambda$ is large, the EMML algorithm can be slow to converge. One method that has been used to accelerate convergence to a solution is the use of block iteration [21, 22, 23].

We begin by writing the index set $\{i = 1, 2, ..., I\}$ as the (not necessarily disjoint) union of $B_n, n = 1, 2, ..., N$. Of particular interest is the *row-action* EMML, obtained by letting each block be a singleton. At each step of the iteration we employ only those equations whose index is a member of the current block. We then cycle through the blocks.

An obvious way to impose blocks would seem to be to modify the EMML iteration as follows:

$$\lambda_j^{k+1} = \lambda_j^k s_{n,j}^{-1} \sum_{i \in B_n} P_{ij} \frac{y_i}{(P\lambda^k)_i}, \tag{14.3}$$

where

$$s_{n,j} = \sum_{i \in B_n} P_{ij}.$$

This doesn't work, though.

Let $H_i = \{z \geq 0 | (Pz)_i = y_i\}$. Note that, for a fixed $x > 0$, we cannot calculate in closed form the vector $z \in H_i$ that minimizes $KL(z, \lambda)$. However, the vector $z = z^i$ in H_i that minimizes the weighted KL distance

$$\sum_{j=1}^{J} P_{ij} KL(z_j, \lambda_j^k)$$

is given by

$$z_j^i = \lambda_j^k \frac{y_i}{(P\lambda^k)_i}. \tag{14.4}$$

The iterative step of the EMML algorithm can then be interpreted as saying that λ^{k+1} is a weighted arithmetic mean of the z^i; that is,

$$\lambda_j^{k+1} = s_j^{-1} \sum_{i=1}^{I} P_{ij} z_j^i. \tag{14.5}$$

This suggests a different form for a block-iterative version of the EMML.

For $k = 0, 1, , , ,$ and $n = n(k) = k(\mathrm{mod}\, N) + 1$, let

$$\lambda_j^{k+1} = (1 - m_n^{-1} s_{nj})\lambda_j^k + m_n^{-1}\lambda_j^k \sum_{i \in B_n} P_{ij} \frac{y_i}{(P\lambda^k)_i}, \tag{14.6}$$

where $m_n = \max_j s_{nj}$. This is the *rescaled* block-iterative EMML (RBI-EMML) algorithm. The sequence $\{\lambda^k\}$ converges to a non-negative solution of the system $y = P\lambda$, for any choice of blocks, whenever the system has a non-negative solution [21].

When each block is a singleton, that is, $B_n = B_i = \{i\}$, for $i = 1, 2, ..., I = N$, the RBI-EMML becomes the EMART algorithm, with the iterative step

$$\lambda_j^{k+1} = (1 - m_i^{-1} P_{ij})\lambda_j^k + \lambda_j^k m_i^{-1} P_{ij} \frac{y_i}{(P\lambda^k)_i}, \tag{14.7}$$

where $m_i = \max_j P_{ij} > 0$. It is interesting to compare the EMART algorithm with the *multiplicative algebraic reconstruction technique* (MART)[24], which has the iterative step

$$\lambda_j^{k+1} = \lambda_j^k \Big(\frac{y_i}{(P\lambda^k)_i}\Big)^{P_{ij}/m_i}, \tag{14.8}$$

so that

$$\lambda_j^{k+1} = \Big(\lambda_j^k\Big)^{1 - P_{ij}/m_i} \Big(\lambda_j^k \frac{y_i}{(P\lambda^k)_i}\Big)^{p_{ij}/m_i}, \tag{14.9}$$

or

$$\log \lambda_j^{k+1} = (1 - m_i^{-1} P_{ij}) \log \lambda_j^k + m_i^{-1} P_{ij} \log \Big(\lambda_j^k \frac{y_i}{(P\lambda^k)_i}\Big). \tag{14.10}$$

The difference between the MART and the EMART is then the difference between a geometric mean and an arithmetic mean.

The simultaneous MART (SMART) is analogous to the EMML and uses all the equations at each step [25, 26, 20]. The iterative step for the SMART is

$$\lambda_j^{k+1} = \lambda_j^k \exp\Big(s_j^{-1} \sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(P\lambda^k)_i}\Big). \tag{14.11}$$

Block-iterative versions of the MART (RBI-SMART) have been considered by [27] and [21]. When $y = P\lambda$ has non-negative solutions, the RBI-SMART sequence converges to the non-negative solution of $y = P\lambda$ for which the cross-entropy $KL(\lambda, \lambda^0)$ is minimized. When there are no non-negative solutions of $y = P\lambda$, the SMART converges to the non-negative minimizer of $KL(P\lambda, y)$ for which $KL(\lambda, \lambda^0)$ is minimized [28].

14.4 Using Prior Bounds on λ

The EMML algorithm finds an approximate non-negative solution of $y = P\lambda$. In some applications it is helpful to be able to incorporate upper and lower bounds on the λ [29].

The SMART, EMML, MART and EMART methods are based on the Kullback-Leibler distance between nonnegative vectors. To impose more general constraints on the entries of λ we derive algorithms based on shifted KL distances, also called *Fermi-Dirac generalized entropies*.

For a fixed real vector u, the shifted KL distance $KL(x-u, z-u)$ is defined for vectors x and z having $x_j \geq u_j$ and $z_j \geq u_j$. Similarly, the shifted distance $KL(v-x, v-z)$ applies only to those vectors x and z for which $x_j \leq v_j$ and $z_j \leq v_j$. For $u_j \leq v_j$, the combined distance

$$KL(x-u, z-u) + KL(v-x, v-z)$$

is restricted to those x and z whose entries x_j and z_j lie in the interval $[u_j, v_j]$. Our objective is to mimic the derivation of the SMART and EMML methods, replacing KL distances with shifted KL distances, to obtain algorithms that enforce the constraints $u_j \leq \lambda_j \leq v_j$, for each j. The algorithms that result are the ABMART and ABEMML block-iterative methods. These algorithms were originally presented in [30], in which the vectors u and v were called a and b, hence the names of the algorithms. We shall assume that the entries of the matrix P are nonnegative. We shall denote by B_n, $n = 1, ..., N$ a partition of the index set $\{i = 1, ..., I\}$ into blocks. For $k = 0, 1, ...$ let $n = n(k) = k(\mathrm{mod}\, N) + 1$.

14.4.1 The ABMART Algorithm

We assume that $(Pu)_i \leq y_i \leq (Pv)_i$ and seek a solution of $P\lambda = y$ with $u_j \leq \lambda_j \leq v_j$, for each j. The algorithm begins with an initial vector λ^0 satisfying $u_j \leq \lambda_j^0 \leq v_j$, for each j. Having calculated λ^k, we take

$$\lambda_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{14.12}$$

with $n = n(k)$,

$$\alpha_j^k = \frac{c_j^k \prod^n (d_i^k)^{P_{ij}}}{1 + c_j^k \prod^n (d_i^k)^{A_{ij}}}, \tag{14.13}$$

$$c_j^k = \frac{(\lambda_j^k - u_j)}{(v_j - \lambda_j^k)}, \tag{14.14}$$

and

$$d_j^k = \frac{(y_i - (Pu)_i)((Pv)_i - (P\lambda^k)_i)}{((Pv)_i - y_i)((P\lambda^k)_i - (Pu)_i)}, \tag{14.15}$$

where $\prod^n$ denotes the product over those indices i in $B_{n(k)}$. Notice that, at each step of the iteration, λ_j^k is a convex combination of the endpoints u_j and v_j, so that λ_j^k always lies in the interval $[u_j, v_j]$.

We have the following theorem concerning the convergence of the ABMART algorithm:

Theorem 14.1 *If there is a solution of the system $P\lambda = y$ that satisfies the constraints $u_j \leq \lambda_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABMART sequence converges to that constrained solution of $P\lambda = y$ for which the Fermi-Dirac generalized entropic distance from λ to λ^0, given by*

$$KL(\lambda - u, \lambda^0 - u) + KL(v - \lambda, v - \lambda^0),$$

is minimized. If there is no constrained solution of $P\lambda = y$, then, for $N = 1$, the ABMART sequence converges to the minimizer of

$$KL(P\lambda - Pu, y - Pu) + KL(Pv - P\lambda, Pv - y)$$

for which

$$KL(\lambda - u, \lambda^0 - u) + KL(v - \lambda, v - \lambda^0)$$

is minimized.

The proof is in [30].

14.4.2 The ABEMML Algorithm

We make the same assumptions as previously. The iterative step of the ABEMML algorithm is

$$\lambda_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{14.16}$$

where

$$\alpha_j^k = \gamma_j^k / d_j^k, \tag{14.17}$$

$$\gamma_j^k = (\lambda_j^k - u_j) e_j^k, \tag{14.18}$$

$$\beta_j^k = (v_j - \lambda_j^k) f_j^k, \tag{14.19}$$

$$d_j^k = \gamma_j^k + \beta_j^k, \tag{14.20}$$

$$e_j^k = \left(1 - \sum_{i \in B_n} P_{ij}\right) + \sum_{i \in B_n} P_{ij} \left(\frac{y_i - (Pu)_i}{(P\lambda^k)_i - (Pu)_i}\right), \tag{14.21}$$

and

$$f_j^k = \left(1 - \sum_{i \in B_n} P_{ij}\right) + \sum_{i \in B_n} P_{ij} \left(\frac{(Pv)_i - y_i}{(Pv)_i - (P\lambda^k)_i}\right). \tag{14.22}$$

The following theorem concerns the convergence of the ABEMML algorithm:

Theorem 14.2 *If there is a solution of the system $P\lambda = y$ that satisfies the constraints $u_j \leq \lambda_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABEMML sequence converges to such a constrained solution of $P\lambda = y$. If there is no constrained solution of $P\lambda = y$, then, for $N = 1$, the ABEMML sequence converges to a constrained minimizer of*

$$KL(y - Pu, P\lambda - Pu) + KL(Pv - y, Pv - P\lambda).$$

The proof is found in [30]. In contrast to the ABMART theorem, this is all we can say about the limits of the ABEMML sequences.

15 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.

15.1 Mixtures

We say that a random vector V taking values in $\mathbb{R}^D$ is a *finite mixture* [31, 32] 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

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v), \tag{15.1}$$

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

15.2 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

$$L_y(\theta) = \prod_{n=1}^{N} \Big(\theta_1 f_1(v_n) + ... + \theta_J f_J(v_n) \Big). \tag{15.2}$$

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

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J} (1 - s_j)\theta_j. \tag{15.3}$$

15.3 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.

15.4 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.

15.5 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

$$L_x(\theta) = f_X(x|\theta) = \prod_{j=1}^{J} \theta_j^{N_j}, \tag{15.4}$$

and the log likelihood is

$$LL_x(\theta) = \log L_x(\theta) = \sum_{j=1}^{J} N_j \log \theta_j. \tag{15.5}$$

Then

$$E(\log L_x(\theta)|y, \theta^k) = \sum_{j=1}^{J} E(N_j|y, \theta^k) \log \theta_j. \tag{15.6}$$

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

$$LL_x(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{J} X_{nj} \log \theta_j, \tag{15.7}$$

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

$$E(X_{nj}|y, \theta^k) = \text{prob}\,(X_{nj} = 1|y, \theta^k) = \frac{\theta_j^k f_j(v_n)}{f(v_n|\theta^k)}. \tag{15.8}$$

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

$$E(LL_x(\theta)|y, \theta^k) = \sum_{n=1}^{N} \sum_{j=1}^{J} \frac{\theta_j^k f_j(v_n)}{f(v_n|\theta^k)} \log \theta_j. \tag{15.9}$$

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

$$\theta_j^{k+1} = \frac{1}{N}\theta_j^k \sum_{n=1}^{N} \frac{f_j(v_n)}{f(v_n|\theta^k)}. \tag{15.10}$$

We know from our previous discussions that, since the preferred data X is acceptable, likelihood is non-decreasing 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.

15.6 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 [33].

Theorem 15.1 *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

$$\theta_j^{k+1} = \alpha_j s_j^{-1}\theta_j^k \Big(\sum_{n=1}^{N} P_{n,j} \frac{u_n}{(P\theta^k)_n} \Big) + (1 - \alpha_j)\gamma_j \tag{15.11}$$

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 (15.11) becomes that of the Mix-EM algorithm. Therefore, the sequence $\{\theta^k\}$ converges to the maximum likelihood estimate of the mixing proportions.

16 More on Convergence

There is a mistake in the proof of convergence given in [11]. Wu [34] and Boyles [35] 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 the book by McLachlan and Krishnan [1] 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 (3.4), 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 (13.5), as was eventually achieved in a sequence of papers [19], [36],[37],[38], and [20]). When the EM algorithm was applied to list-mode data in SPECT and PET [39, 40, 41], the resulting algorithm differed slightly from that in Equation (13.5) and a proof of convergence was provided in [33]. The convergence theorem in [33] also establishes the convergence of the iteration in Equation (15.10) to the maximum-likelihood estimate of the mixing proportions.

17 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 conver-

gence to $L_y(\theta_{ML})$ of the likelihood values $L_y(\theta^k)$ is also an option.

18 Conclusion

Difficulties with the conventional formulation of the EM algorithm in the continuous case of probability density functions (pdf) has prompted us to adopt a new definition, that of acceptable data. As we have shown, this model can be helpful in generating EM algorithms in a variety of situations. For the discrete case of probability functions (pf), the conventional approach remains satisfactory. In both cases, the two steps of the EM algorithm can be viewed as alternating minimization of the Kullback-Leibler distance between two sets of parameterized pf or pdf, along the lines investigated by Csiszár and Tusnády [12]. In order to use the full power of their theory, however, we need the sets to be closed and convex. This does occur in the important special case of sums of independent Poisson random variables, but is not generally the case.

19 Acknowledgment

I wish to thank Professor Paul Eggermont of the University of Delaware for helpful discussions on these matters.

AMS Subject Classification: 49M37, 62L12, 90C25.

Cross-References

EM Algorithms
Iterative Solution Methods
Large-Scale Inverse Problems in Imaging
Linear Inverse Problems
Mathematical methods in PET and Spect Imaging

References

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

[2] Meng, X., and Pedlow, S. : EM: a bibliographic review with missing articles. Proc. Stat. Comput. Sect., American Statistical Association, American Statistical Association, Alexandria, VA. (1992)

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

[4] Becker, M., Yang, I., and Lange, K. : EM algorithms without missing data. Stat. Methods Med. Res. 6, 38–54 (1997)

[5] Byrne, C. : Iterative Optimization in Inverse Problems. Taylor and Francis, Publ. (2014)

[6] Byrne, C. : Non-stochastic EM algorithms in optimization. Journal of Nonlinear and Convex Analysis (to appear, 2015).

[7] Kullback, S. and Leibler, R. : On information and sufficiency. Annals of Math. Stat. 22, 79–86 (1951)

[8] Hogg, R., McKean, J., and Craig, A. : Introduction to Mathematical Statistics, 6th edition. Prentice Hall (2004)

[9] Byrne, C. and Eggermont, P. : EM algorithms. In: Otmar Scherzer (ed.) Handbook of Mathematical Methods in Imaging. Springer-Science (2010)

[10] Fessler, J., Ficaro, E., Clinthorne, N., and Lange, K. : Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. IEEE Trans. Med. Imag. 16 (2), 166–175 (1997)

[11] Dempster, A., Laird, N., and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B 37, 1–38 (1977)

[12] Csiszár, I. and Tusnády, G. : Information geometry and alternating minimization procedures. Statistics and Decisions Supp. 1, R. Oldenbourg Verlag, München, 205–237 (1984)

[13] Byrne, C. : Alternating and sequential unconstrained minimization algorithms. J. Opt. Th. and Appl., electronic 154(3), DOI 10.1007/s1090134-2; hardcopy 156(2) (2012)

[14] Eggermont, P. and LaRiccia, V. : Smoothed maximum likelihood density estimation for inverse problems. Annals of Stat. 23, 199–220 (1995)

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

[16] Bauschke, H., Combettes, P., and Nol,l D. : Joint minimization with alternating Bregman proximity operators. Pacific J. of Opt. 2, 401–424 (2006)

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

[18] Rockmore, A. and Macovski, A. : A maximum likelihood approach to emission image reconstruction from projections. IEEE Trans. Nucl. Sc. NS-23, 1428–1432 (1976)

[19] Shepp, L. and Vardi, Y. (1982) Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imag. MI-1, 113–122 (1982)

[20] Byrne, C. : Iterative image reconstruction algorithms based on cross-entropy minimization. IEEE Trans. Image Proc. IP-2, 96–103 (1993)

[21] Byrne, C. : Block-iterative methods for image reconstruction from projections. IEEE Trans. Image Proc. IP-5, 792–794 (1996)

[22] Byrne, C. : Convergent block-iterative algorithms for image reconstruction from inconsistent data. IEEE Trans. Image Proc. IP-6, 1296–1304 (1997)

[23] Byrne, C. : Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods. IEEE Trans. Image Proc. IP-7, 100–109 (1998)

[24] Gordon, R., Bender, R., and Herman, G.T. : Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography. J. Theoret. Biol. 29, 471–481 (1970)

[25] Darroch, J. and Ratcliff, D. : Generalized iterative scaling for log-linear models. Annals of Math. Stat. 43, 1470–1480 (1972)

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

[27] Censor, Y. and Segman, J. : On block-iterative maximization. J. of Inf. and Opt. Sciences 8, 275–291 (1987)

[28] Byrne, C. : Iterative reconstruction algorithms based on cross-entropy minimization. In: Shepp, L., and Levinson, S.E. (eds.) Image Models (and their

Speech Model Cousins). IMA Volumes in Mathematics and its Applications 80, Springer-Verlag, New York, 1–11 (1996)

[29] Narayanan, M., Byrne, C., and King, M. : An interior point iterative maximum-likelihood reconstruction algorithm incorporating upper and lower bounds with application to SPECT transmission imaging. IEEE Trans. Med. Imag. TMI-20 (4), 342–353 (2001)

[30] Byrne, C. : Iterative algorithms for deblurring and deconvolution with constraints. Inverse Problems 14, 1455–1467 (1998)

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

[32] Redner, R. and Walker, H. : Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26(2), 195–239 (1984)

[33] Byrne, C. : Likelihood maximization for list-mode emission tomographic image reconstruction. IEEE Trans. Med. Imag. 20(10), 1084–1092 (2001)

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

[35] Boyles, R. : On the convergence of the EM algorithm. J. Roy. Statist. Soc. B 45, 47–50 (1983)

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

[37] Vardi, Y., Shep,p .L, and Kaufman, L. : (1985) A statistical model for positron emission tomography. J. American Statistical Association 80, 8–20 (1985)

[38] 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), 106–114 (1987)

[39] Barrett, H., White, T., and Parra, L. : List-mode likelihood. J. Opt. Soc. Am. A 14, 2914–2923 (1997)

[40] Parra, L. and Barrett, H. : List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET. IEEE Trans. Med. Imag. 17, 228–235 (1998)

[41] Huesman, R., Klein, G., Moses, W., Qi, J., Ruetter, B., and Virador, P. : List-mode maximum likelihood reconstruction applied to positron emission mammography (PEM) with irregular sampling. IEEE Trans. Med. Imag. 19 (5), 532–537 (2000)