ECE598PM Computational Inference and Learning
Fall 2016

Pierre Moulin

Numerical evaluation of maximum-likelihood (ML) estimates is often difficult. The likelihood function may have multiple extrema and the parameter θ may be multidimensional, all of which are problematic for any numerical algorithm. In these notes we introduce the expectation-maximization (EM) algorithm, a powerful optimization method that has been used with great success in many applications. The idea can be traced back to Hartley [1] and was fully developed in a landmark paper by Dempster, Laird and Rubin [2]. Their EM algorithm relies on the concept of a *complete data space*, to be selected in some convenient way. The EM algorithm is iterative and alternates between conditional expectation and maximization steps. If the complete data space is properly chosen, this method can be very effective as it exploits the inherent statistical structure of the parameter estimation problem. The EM algorithm is particularly adept at dealing with problems involving *missing data* which arise in many applications in the areas of data communications, signal processing, imaging, and biology. The EM algorithm is also widely used in problems involving mixtures of distributions in exponential families. Comprehensive surveys include the tutorial paper by Meng [3] and the book by McLachlan and Krishnan [4].

Note that better alternatives to EM exist in some cases. For instance, in some problems, the loglikelihood function is concave and the feasible set for the parameter θ is convex. Then efficient convex optimization algorithms can be employed and usually converge faster than the EM algorithm. Detailed convergence analyses may be found in several papers, e.g., [6].

First recall that, given an observation y drawn from a parametric family $\{p_\theta,\, \theta \in \mathcal{S}\}$ of distributions over $\mathcal{Y}$, the ML estimator is any maximizer of the loglikelihood function

$$\hat{\theta}_{\rm ML}(y) = \arg\max_{\theta\in\mathcal{S}} \ln p_\theta(y) \tag{1}$$

(assuming a maximizer exists). If the solution is an interior point of $\mathcal{S}$, it must satisfy the likelihood equation

$$\nabla \ln p_\theta(y) = 0 \tag{2}$$

at $\theta = \hat{\theta}_{\rm ML}$.

1 General Structure of the EM Algorithm

The EM algorithm is an iterative algorithm that starts from an initial estimate of the unknown parameter θ and produces a sequence of estimates $\hat{\theta}^{(k)}$, $k = 1, 2, \cdots$. Referring to Fig. 1, the EM algorithm uses the following ingredients:

- an incomplete-data space $\mathcal{Y}$ (measurement space);
- a complete-data space $\mathcal{Z}$ (many choices are possible);
- a many-to-one mapping $h \ : \ \mathcal{Z} \to \mathcal{Y}$.

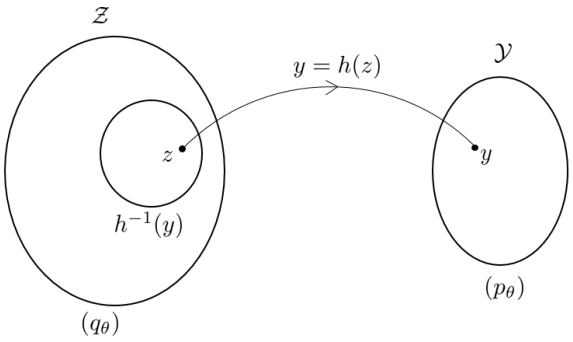

Figure 1: Complete and incomplete data spaces $\mathcal{Z}$ and $\mathcal{Y}$.

The pdf for the complete data is denoted by $q_\theta(z)$. The pdf for the incomplete data is the marginal

$$p_\theta(y) = \int_{h^{-1}(y)} q_\theta(z)\, dz,$$

where $h^{-1}(y) \triangleq \{z \in \mathcal{Z} : \ h(z) = y\}$. The joint distribution on $\mathcal{Y} \times \mathcal{Z}$ is denoted by $r_\theta(y, z) = q_\theta(z)\, \mathbb{1}_{\{y=h(z)\}}$. Throughout this chapter, the letters p, q, and r will be reserved for distributions on $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{Y} \times \mathcal{Z}$, respectively.

The incomplete-data and complete-data loglikelihood functions are respectively denoted by

$$l_{id}(\theta) \triangleq \ln p_\theta(y), \quad l_{cd}(\theta) \triangleq \ln q_\theta(z).$$

The complete data space should be chosen in such a way that maximization of $l_{cd}(\theta)$ is easy. However the complete data are not available, so one uses a *surrogate cost function*: the expectation of $l_{cd}(\theta)$ conditioned on $Y = y$ and the current estimate of θ. Maximizing this surrogate function results in an updated estimate of θ, and these steps are repeated till convergence.

Upon selection of an initial estimate $\hat{\theta}^{(0)}$, the EM algorithm alternates between Expectation (E) and Maximization (M) steps for updating the estimate $\hat{\theta}^{(k)}$ of the unknown parameter θ at iteration $k > 0$.

E-Step: Compute

$$Q(\theta|\hat{\theta}^{(k)}) \triangleq \mathbb{E}_{\hat{\theta}^{(k)}}[l_{cd}(\theta)|Y = y] \tag{3}$$

where the conditional expectation is evaluated assuming the parameter value is $\hat{\theta}^{(k)}$.

M-step:

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta \in \mathcal{S}} Q(\theta|\hat{\theta}^{(k)}). \tag{4}$$

If the EM sequence $\hat{\theta}^{(k)}$, $k \geq 0$ converges, the limit θ^* is called a stable point of the algorithm. In practice, the iterations are terminated when the estimates are stable enough (using some stopping criterion).

2 Example 1: Variance Estimation

To illustrate the EM algorithm, we consider the following variance estimation problem. We observe $Y = S + N$ where S and N are independent Gaussian random variables with common mean zero and respective variances θ and 1. We need to estimate θ. In this case, ML estimation yields a closed-form solution. Indeed $Y \sim \mathcal{N}(0, \theta + 1)$ and thus

$$\ln p_\theta(y) = -\frac{1}{2}\ln(2\pi(\theta+1)) - \frac{y^2}{2(\theta+1)}, \quad \theta \geq 0.$$

The derivative of the loglikelihood function is

$$\frac{d \ln p_\theta(y)}{d\theta} = \frac{1}{2}\left[\frac{1}{\theta+1} - \frac{y^2}{(\theta+1)^2}\right]$$

which has a single root at $\theta = y^2 - 1$ and negative slope. Hence the likelihood function is unimodal, and

$$\hat{\theta}_{\mathrm{ML}}(y) = \max\{0, y^2 - 1\}. \tag{5}$$

If $y^2 > 1$, the likelihood equation (2) is satisfied; otherwise not.

Let us derive an EM algorithm and see whether it converges to the above solution. Define the complete data $Z = (S, N)$. The mapping h is given by

$$Y = h(Z) = S + N.$$

We have

$$l_{cd}(\theta) = \ln q_\theta(s,n) = \ln p(n) + \ln p(s;\theta) = C - \frac{1}{2}\ln\theta - \frac{s^2}{2\theta}$$

where C denotes a term that is independent of θ.

E-Step. Taking conditional expectation, we obtain

$$Q(\theta|\hat{\theta}^{(k)}) = \mathbb{E}_{\hat{\theta}^{(k)}}[l_{cd}(\theta) \mid Y = y] = C - \frac{1}{2}\ln\theta - \frac{\mathbb{E}_{\hat{\theta}^{(k)}}[S^2 \,|\, Y = y]}{2\theta}.$$

M-Step. The maximum of Q with respect to θ is obtained by setting the derivative to zero:

$$0 = \frac{dQ(\theta|\hat{\theta}^{(k)})}{d\theta} = -\frac{1}{2\theta} + \frac{\mathbb{E}_{\hat{\theta}^{(k)}}[S^2 \,|\, Y = y]}{2\theta^2}$$

and is achieved by

$$\begin{aligned} \hat{\theta}^{(k+1)} = \mathbb{E}_{\hat{\theta}^{(k)}}[S^2 \,|\, Y = y] &= \left(\mathbb{E}_{\hat{\theta}^{(k)}}[S \,|\, Y = y]\right)^2 + \mathrm{Var}_{\hat{\theta}^{(k)}}[S \,|\, Y = y] \\ &= \left(\frac{\hat{\theta}^{(k)}}{\hat{\theta}^{(k)}+1} Y\right)^2 + \frac{\hat{\theta}^{(k)}}{\hat{\theta}^{(k)}+1} \end{aligned} \tag{6}$$

where the last line follows from the classical formulas for conditional mean and variance.

Now let us examine the convergence of the algorithm. Any limit point $\hat{\theta}^* = \lim_{k\to\infty} \hat{\theta}^{(k)}$ of the EM sequence (6) must satisfy

$$\hat{\theta}^* = \frac{\hat{\theta}^*}{\hat{\theta}^*+1}\left(\frac{\hat{\theta}^*}{\hat{\theta}^*+1}\, Y^2 + 1\right). \tag{7}$$

The only two possible solutions to (7) are $\hat{\theta}^* = 0$ and $\hat{\theta}^* = Y^2 - 1$, both of which are consistent with the ML solution (5).

If the EM algorithm is initialized with $\hat{\theta}^{(0)} = 0$, it will remain stuck there. If instead $\hat{\theta}^{(0)} > 0$, then from (6) we have $\hat{\theta}^{(k)} > 0$ for all k. We distinguish two cases:

Case I: $Y^2 \leq 1$. We have $\hat{\theta}^{(k)} > 0 \geq Y^2 - 1$, and thus (6) is a contraction mapping, which implies $\hat{\theta}^{(k+1)} < \hat{\theta}^{(k)}$. The EM sequence $\hat{\theta}^{(k)}$ is decreasing and converges to $\hat{\theta}^* = 0$.

Case II: $Y^2 > 1$. If $\hat{\theta}^{(k)} < Y^2 - 1$ for any k, then $\hat{\theta}^{(k+1)} > \hat{\theta}^{(k)}$. Thus $\hat{\theta}^* = 0$ cannot be a limit point of the EM sequence, leaving $\hat{\theta}^* = Y^2 - 1$ as the only possible limit.

In both cases, the EM algorithm converges to the ML solution.

Remark. There are many other possible choices for the complete data, for instance $Z = (S, Y)$, where S is viewed as a *missing* variable, or *latent variable*.

3 Example 2: Estimation of Gaussian Mixtures

This is one of the most famous problems to which the EM algorithm has been successfully applied [7]. Assume the data $\boldsymbol{Y} = \{Y_i,\, 1 \leq i \leq n\} \in \mathbb{R}^n$, are drawn iid from a pdf $p_\theta(y)$ which is the mixture of m univariate Gaussians with respective probabilities $\pi(j)$, means μ_j, and variances σ_j^2, for $1 \leq j \leq m$:

$$p_\theta(y) = \sum_{j=1}^{m} \pi(j)\, \phi(y; \mu_j, \sigma_j^2), \quad y \in \mathbb{R}$$

where

$$\phi(y; \mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

denotes the Gaussian pdf with mean μ and variance σ^2.

3.1 Unknown Means

To simplify notation, we initially assume that $\{\pi(j)\}$ and $\{\sigma_j^2\}$ are given and estimate the means $\{\mu_j\}$. Thus $\theta = \mu \in \mathbb{R}^m$. Unfortunately the ML estimator cannot be derived in closed form. Indeed the loglikelihood function for θ is

$$l_{id}(\theta) = \sum_{i=1}^{n} \ln p_\theta(y_i) = \sum_{i=1}^{n} \ln \sum_{j=1}^{m} \pi(j)\, \phi(y_i; \mu_j, \sigma_j^2)$$

and maximizing it is a m-dimensional, nonconcave maximization problem.

The likelihood equation is

$$\begin{aligned} 0 &= \frac{\partial l_{id}(\theta)}{\partial \mu_j} \\ &= \sum_{i=1}^{n} \frac{\pi(j)\,\frac{\partial}{\partial \mu_j}\phi(y_i;\mu_j,\sigma_j^2)}{\sum_{j=1}^{m}\pi(j)\,\phi(y_i;\mu_j,\sigma_j^2)} \\ &= \frac{1}{\sigma_j^2}\sum_{i=1}^{n} \frac{\pi(j)\,(y_i-\mu_j)\,\phi(y_i;\mu_j,\sigma_j^2)}{\sum_{j=1}^{m}\pi(j)\,\phi(y_i;\mu_j,\sigma_j^2)} \\ &= \frac{1}{\sigma_j^2}\sum_{i=1}^{n}(y_i-\mu_j)\,\pi_\theta(j|y_i), \quad 1\le j\le m. \end{aligned} \tag{8}$$

In the last line we have introduced the conditional probability distribution

$$\pi_\theta(j|y) = \frac{\pi(j)\,\phi(y;\mu_j,\sigma_j^2)}{\sum_{j=1}^{m}\pi(j)\,\phi_j(y;\mu_j,\sigma_j^2)}, \quad 1\le j\le m,\ y\in\mathbb{R}$$

which admits the following interpretation. To generate Y with the prescribed mixture distribution, we may draw a random label that takes value j with probability $\pi(j)$, and then generate $Y\sim\mathcal{N}(\mu_j,\sigma_j^2)$. Then $\pi_\theta(j|y)$ represents the posterior probability of label j given observation y and θ.

We also conclude from (8) that $\hat\theta_{\rm ML}=\hat\mu_{\rm ML}$ satisfies the nonlinear system of m equations

$$\hat\mu_{{\rm ML},j} = \frac{\sum_{i=1}^{n} y_i\,\pi_{\hat\theta_{\rm ML}}(j|y_i)}{\sum_{i=1}^{n}\pi_{\hat\theta_{\rm ML}}(j|y_i)}, \quad 1\le j\le m. \tag{9}$$

which is an average of the observations y_i, weighted by their conditional probability given the random label j. The system (9) may have multiple solutions corresponding to local maxima or even local minima or saddlepoints of the likelihood function.

EM Algorithm. Let the complete data be $Z_i=(Y_i,J_i)$, $1\le i\le n$, where J_i is the random label that was drawn to produce Y_i. We have

$$\begin{aligned} l_{cd}(\theta)=\sum_{i=1}^{n}\ln r_\theta(y_i,j_i) &= \sum_{i=1}^{n}[\ln\pi(j_i)+\ln p_\theta(y_i|j_i)] \\ &= \sum_{i=1}^{n}[\ln\pi(j_i)+\ln\phi(y_i;\mu_{j_i},\sigma_{j_i}^2)] \\ &= -\frac{n}{2}\ln(2\pi)+\sum_{i=1}^{n}\left[\ln\pi(j_i)-\ln\sigma_{j_i}-\frac{(y_i-\mu_{j_i})^2}{2\sigma_{j_i}^2}\right] \end{aligned}$$

E-step:

$$\begin{aligned} Q(\theta|\hat{\theta}^{(k)}) &= \mathbb{E}_{\hat{\theta}^{(k)}}[l_{cd}(\theta)|\boldsymbol{Y}=\boldsymbol{y}] \\ &= -\frac{n}{2}\ln(2\pi) + \sum_{i=1}^{n}\mathbb{E}_{\hat{\theta}^{(k)}}\left\{\left[\ln\pi(J_i) - \ln\sigma_{J_i} - \frac{(y_i-\mu_{J_i})^2}{2\sigma_{J_i}^2}\right]\middle|\boldsymbol{Y}=\boldsymbol{y}\right] \\ &\overset{(a)}{=} -\frac{n}{2}\ln(2\pi) + \sum_{i=1}^{n}\mathbb{E}_{\hat{\theta}^{(k)}}\left\{\left[\ln\pi(J_i) - \ln\sigma_{J_i} - \frac{(y_i-\mu_{J_i})^2}{2\sigma_{J_i}^2}\right]\middle|Y_i=y_i\right] \\ &= -\frac{n}{2}\ln(2\pi) + \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\ln\pi(j) - \ln\sigma_j - \frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right]\pi_{\hat{\theta}^{(k)}}(j|y_i) \qquad (10) \end{aligned}$$

where (a) holds because J_i is conditionally independent of $\{Y_k\}_{k\neq i}$ given Y_i.

M-Step: The function $Q(\theta|\hat{\theta}^{(k)})$ is concave quadratic in $\{\mu_j\}_{j=1}^m$. Setting the partial derivatives of $Q(\theta|\hat{\theta}^{(k)})$ with respect to each μ_j to zero, we obtain

$$0 = \frac{Q(\theta|\hat{\theta}^{(k)})}{\partial\mu_j} = \frac{1}{\sigma_j^2}\sum_{i=1}^{n}(y_i-\mu_j)\,\pi_{\hat{\theta}^{(k)}}(j|y_i), \quad 1\le j\le m.$$

Hence the update

$$\hat{\mu}_j^{(k+1)} = \frac{\sum_{i=1}^n y_i\,\pi_{\hat{\theta}^{(k)}}(j|y_i)}{\sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i)}, \quad 1\le j\le m$$

which is a weighted average of the observations $\{y_i\}$.

Let's see what the sequence $\hat{\mu}^{(k)}$ might converge to. If $\hat{\mu}^{(k)}\to\hat{\mu}^*$ as $k\to\infty$, then

$$\hat{\mu}_j^* = \frac{\sum_{i=1}^n y_i\,\pi_{\hat{\theta}^*}(j|y_i)}{\sum_{i=1}^n \pi_{\hat{\theta}^*}(j|y_i)}, \quad 1\le j\le m$$

which is a solution of the nonlinear system of (9).

3.2 Unknown Mixture Probabilities, Means and Variances

If $\theta \triangleq \{\pi(j), \mu_j, \sigma_j^2,\ 1\le j\le m\}$ is unknown, the ML estimator $\hat{\theta}_{\mathrm{ML}}$ satisfies the following nonlinear system of equations:

$$\begin{aligned} \hat{\mu}_{\mathrm{ML},j} &= \frac{\sum_{i=1}^n y_i\,\pi_{\hat{\theta}^{(k)}}(j|y_i)}{\sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i)}, \\ \hat{\sigma}^2_{\mathrm{ML},j} &= \frac{\sum_{i=1}^n (y_i-\hat{\mu}_{\mathrm{ML},j})^2\,\pi_{\hat{\theta}^{(k)}}(j|y_i)}{\sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i)}, \\ \hat{\pi}_{\mathrm{ML}}(j) &= \frac{1}{n}\sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i) \quad 1\le j\le m, \end{aligned}$$

where

$$\pi_\theta(j|y_i) = \frac{\pi(j)\,\phi(y_i;\mu_j,\sigma_j^2)}{\sum_{j=1}^m \pi(j)\,\phi(y_i;\mu_j,\sigma_j^2)}, \quad 1\le j\le m.$$

The same complete data space of Sec. 3.1 can be used, again resulting in the same function $Q(\theta|\hat{\theta}^{(k)})$ as in (10) and in the following EM algorithm:

$$\begin{aligned}
\hat{\mu}_j^{(k+1)} &= \frac{\sum_{i=1}^n y_i \, \pi_{\hat{\theta}^{(k)}}(j|y_i)}{\sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i)}, \\
(\hat{\sigma}_j^2)^{(k+1)} &= \frac{\sum_{i=1}^n (y_i - \hat{\mu}_j^{(k+1)})^2 \, \pi_{\hat{\theta}^{(k)}}(j|y_i)}{\sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i)}, \\
\hat{\pi}^{(k+1)}(j) &= \frac{1}{n} \sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i), \quad 1 \le j \le m.
\end{aligned}$$

To derive the update of the mixture probabilities above, we used the extremal property (16) of cross-entropy with $\mathcal{S} = \{1, \cdots, m\}$ and

$$p_j \triangleq \frac{1}{n} \sum_{i=1}^n \pi_{\hat{\theta}^{(k)}}(j|y_i), \quad q_j = \pi(j)$$

to find the global optimum of (10) with respect to $\{\pi(j)\}_{j=1}^m$.

Note that this procedure for finding means is not the same as the well-known k-means algorithm for classification, which alternates between estimation of means and "hard" assignment of each sample y_i to a class. The EM algorithm associates posterior probabilities $\pi_{\hat{\theta}^{(k)}}(j|y_i)$ to each sample y_i and may be thought of as a "soft" version of k-means.

4 Example 3: Estimation of Exponential Mixtures

This is a generalization of the mixture-of-Gaussians example of Sec. 3. Assume the data Y_i, $1 \le i \le n$, are drawn iid from a pdf $p_\theta(y)$ which is the mixture of m regular d-dimensional exponential distributions in canonical form [5]:

$$p_\theta(y) = \sum_{j=1}^m \pi(j) \, \psi(y; \eta_j)$$

where η_j is a d-dimensional parameter vector, $A(\eta_j)$ is the normalization function, *aka* log-partition function,

$$\psi(y; \eta_j) = h(y) \exp\{\eta_j^\top T(y) - A(\eta_j)\}$$

and $\theta = \{\pi(j), \eta_j, \, 1 \le j \le m\}$. By the regularity assumption, the Jacobian $\nabla A(\cdot)$ is continuously invertible. Denote by f its inverse. The posterior probability of label j given observation y is given by

$$\pi_\theta(j|y) = \frac{\pi(j) \, \psi(y; \eta_j)}{\sum_{j=1}^m \pi(j) \, \psi(y; \eta_j)}, \quad 1 \le j \le m.$$

Any root $\hat{\theta}$ of the likelihood equation satisfies (proof as an exercise)

$$\begin{aligned}\hat{\eta}_j &= f\left[\left(\sum_{i=1}^{n}\pi_{\hat{\theta}}(j|y_i)\right)^{-1}\left(\sum_{i=1}^{n}T(y_i)\,\pi_{\hat{\theta}}(j|y_i)\right)\right],\\ \hat{\pi}(j) &= \frac{1}{n}\sum_{i=1}^{n}\pi_{\hat{\theta}}(j|y_i),\quad 1\le j\le m.\end{aligned}\tag{11}$$

EM Algorithm. As in Sec. 3.1, let the complete data be $Z_i = (Y_i, J_i)$, $1 \le i \le n$, where J_i is the random label that was drawn to produce Y_i from the corresponding mixture component $q(\cdot;\eta_{J_i})$. We have

$$\begin{aligned}l_{cd}(\theta) = \sum_{i=1}^{n}\ln r_\theta(y_i, j_i) &= \sum_{i=1}^{n}[\ln\pi(j_i) + \ln\psi(y_i;\eta_{j_i})]\\ &= C + \sum_{i=1}^{n}\left[\ln\pi(j_i) + \eta_{j_i}^{\top}T(y_i) - A(\eta_{j_i})\right]\end{aligned}$$

where $C = \sum_{i=1}^{n}\ln h(y_i)$ is independent of θ.

E-step:

$$\begin{aligned}Q(\theta|\hat{\theta}^{(k)}) &= \mathbb{E}_{\hat{\theta}^{(k)}}[l_{cd}(\theta, \boldsymbol{Z})|\boldsymbol{Y} = \boldsymbol{y}]\\ &= C + \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\ln\pi(j) + \eta_j^{\top}T(y_i) - A(\eta_j)\right]\,\pi_{\hat{\theta}(k)}(j|y_i).\end{aligned}$$

M-Step: Setting the partial derivatives of $Q(\theta|\hat{\theta}^{(k)})$ with respect to each η_j to zero, we obtain m vector equations

$$0 = \nabla_{\eta_j}Q(\theta|\hat{\theta}^{(k)}) = \sum_{i=1}^{n}[T(y_i) - \nabla A(\eta_j)]\,\pi_{\hat{\theta}(k)}(j|y_i),\quad 1\le j\le m$$

$$\Rightarrow\quad \nabla A(\eta_j)\sum_{i=1}^{n}\pi_{\hat{\theta}(k)}(j|y_i) = \sum_{i=1}^{n}T(y_i)\,\pi_{\hat{\theta}(k)}(j|y_i),\quad 1\le j\le m.$$

Hence the update

$$\begin{aligned}\hat{\eta}_j^{(k+1)} &= f\left[\left(\sum_{i=1}^{n}\pi_{\hat{\theta}(k)}(j|y_i)\right)^{-1}\left(\sum_{i=1}^{n}T(y_i)\,\pi_{\hat{\theta}(k)}(j|y_i)\right)\right],\\ \hat{\pi}^{(k+1)}(j) &= \frac{1}{n}\sum_{i=1}^{n}\pi_{\hat{\theta}(k)}(j|y_i)\quad 1\le j\le m.\end{aligned}\tag{12}$$

The stable points of the algorithm satisfy the nonlinear system (11).

5 Preamble: Information-Theoretic Functionals

The analysis of the EM algorithm makes uses of the information-theoretic notions of entropy, Kullback-Leibler divergence, and cross-entropy:

Definition. The entropy of a pmf $p(x)$, $x \in \mathcal{X}$ is defined as

$$H(p) \triangleq -\sum_x p(x) \ln p(x). \tag{13}$$

Definition. The Kullback-Leibler divergence (or relative entropy) of two pmf's $p(x)$, $x \in \mathcal{X}$ and $q(x)$, $x \in \mathcal{X}$ is defined as

$$D(p\|q) \triangleq \sum_x p(x) \ln \frac{p(x)}{q(x)} \geq 0, \tag{14}$$

with equality if and only if $p = q$. (The natural logarithm is used for convenience.) We also use the notational convention $0 \ln 0 = 0$.

Definition. The cross-entropy of a pmf $p(x)$, $x \in \mathcal{X}$ relative to another pmf $q(x)$, $x \in \mathcal{X}$ is defined as

$$H(p, q) \triangleq -\sum_x p(x) \ln q(x). \tag{15}$$

Note that

$$H(p, q) = H(p) + D(p\|q).$$

Since KL divergence is nonnegative, the cross-entropy satisfies the extremal property

$$H(p, q) \geq H(p) \tag{16}$$

where the lower bound is achieved by $q = p$.

6 Convergence of EM Algorithm

The convergence properties of the EM algorithm were first studied by Dempster, Laird and Rubin [2] and by Wu [8]. The main result is as follows.

Theorem. The likelihood sequence $p_{\hat{\theta}^{(k)}}(y)$, $k = 0, 1, 2, \cdots$ is nondecreasing.

Proof. Assume for notational simplicity that the random variables Y and Z are discrete. Hence their joint distribution is given by

$$r_\theta(y, z) = q_\theta(z) p(y|z) = q_\theta(z) \, \mathbb{1}_{\{y = h(z)\}} \tag{17}$$

and the following identity holds for all $z \in h^{-1}(y)$:

$$p_\theta(y) = \frac{r_\theta(z, y)}{q_\theta(z|y)} = \frac{q_\theta(z)}{q_\theta(z|y)}. \tag{18}$$

We will sometimes use the shorthand $q_{\theta,y}(z) = q_\theta(z|y)$. Taking logarithms in (18), we obtain

$$\underbrace{\ln p_\theta(y)}_{l_{id}(\theta)} = \underbrace{\ln q_\theta(z)}_{l_{cd}(\theta)} - \ln q_\theta(z|y), \quad \forall z \in h^{-1}(y). \tag{19}$$

At iteration $k+1$, the current parameter estimate $\hat{\theta}^{(k)}$ is updated as follows:

E-Step: evaluate $Q(\theta|\hat{\theta}^{(k)}) = \sum_z q_{\hat{\theta}^{(k)}}(z|y) \ln q_\theta(z)$;

M-Step: let $\hat{\theta}^{(k+1)} = \arg\max_{\theta \in \mathcal{S}} Q(\theta|\hat{\theta}^{(k)})$.

Define

$$H(\theta|\hat{\theta}^{(k)}) = -\sum_{z \in h^{-1}(y)} q_{\hat{\theta}^k}(z|y) \ln q_\theta(z|y) = H(q_{\hat{\theta}^k,y}, q_{\theta,y}) \tag{20}$$

which is viewed as a cross-entropy function (cross-entropy over parameterized distributions). Taking the conditional expectation of (19) with respect to $q_{\hat{\theta}^{(k)}}(z|y)$, we obtain

$$l_{id}(\theta) = Q(\theta|\hat{\theta}^{(k)}) + H(\theta|\hat{\theta}^{(k)}). \tag{21}$$

Hence the increment of l_{id} from step k to $k+1$ can be decomposed as

$$l_{id}(\hat{\theta}^{(k+1)}) - l_{id}(\hat{\theta}^{(k)}) = [Q(\hat{\theta}^{(k+1)}|\hat{\theta}^{(k)}) - Q(\hat{\theta}^{(k)}|\hat{\theta}^{(k)})] + [H(\hat{\theta}^{(k+1)}|\hat{\theta}^{(k)}) - H(\hat{\theta}^{(k)}|\hat{\theta}^{(k)})].$$

The first term in brackets is nonnegative because $\hat{\theta}^{(k+1)}$ is a maximizer of $Q(\cdot|\hat{\theta}^{(k)})$. The second term takes the form

$$\begin{aligned} H(\hat{\theta}^{(k+1)}|\hat{\theta}^{(k)}) - H(\hat{\theta}^{(k)}|\hat{\theta}^{(k)}) &= \sum_{z \in h^{-1}(y)} q_{\hat{\theta}^{(k)}}(z|y) \ln \frac{q_{\hat{\theta}^{(k)}}(z|y)}{q_{\hat{\theta}^{(k+1)}}(z|y)} \\ &= D(q_{\hat{\theta}^{(k)},y} \| q_{\hat{\theta}^{(k+1)},y}) \end{aligned}$$

and is nonnegative too.

Therefore the sequence of loglikelihoods $l_{id}(\hat{\theta}^{(k)})$ is nondecreasing. □

While this theorem guarantees monotonicity of the loglikelihood sequence $l_{id}(\hat{\theta}^{(k)})$, convergence to an extremum of $l_{id}(\theta)$ requires differentiability assumptions as stated in the corollary below.

Corollary. Assume that $\mathcal{S}$ is a closed, bounded subset of Euclidean space, the functions $Q(\theta|\theta')$ and $H(\theta|\theta')$ are continuously differentiable, and the loglikelihood function $l_{id}(\theta)$ is differentiable and bounded. Then the sequence $l_{id}(\hat{\theta}^{(k)})$ converges, and any limit point $\theta^* \in \text{interior}(\mathcal{S})$ of the EM sequence is a solution of the likelihood equation $\nabla l_{id}(\theta) = 0$.

Note the corollary provides no guarantee here that the Hessian $\nabla^2 l_{id}(\theta^*)$ is nonpositive definite, which is a necessary condition for a maximizer.

Proof of Corollary. Since the sequence of loglikelihoods $l_{id}(\hat{\theta}^{(k)})$ is nondecreasing and bounded, it converges to a finite limit. By assumption, the EM sequence of estimates contains a subsequence $\hat{\theta}^{(k(j))}, j \geq 0$ that converges to a limit $\theta^* \in \text{interior}(\mathcal{S})$. Since $l_{id}(\theta)$

is differentiable and therefore continuous, we have $\lim_{j\to\infty} l_{id}(\hat{\theta}^{(k(j))}) = l_{id}(\theta^*)$. This alone does not guarantee that $\nabla l_{id}(\theta^*) = 0$. However, applying (21) in the limit as $k \to \infty$ and invoking the continuity of the functions Q and H, we have $l_{id}(\theta) = Q(\theta|\theta^*) + H(\theta|\theta^*)$. At $\theta = \theta^*$, we have $\nabla_\theta Q(\theta|\theta^*)|_{\theta=\theta^*} = 0$ (since the gradient is assumed to exist, $\theta^* = \arg\max_{\theta\in\mathcal{S}} Q(\theta|\theta^*)$, and θ^* is an interior point of $\mathcal{S}$) and $\nabla_\theta H(\theta|\theta^*)|_{\theta=\theta^*} = 0$ (by the extremal property (16) of the cross-entropy function). Hence

$$\nabla l_{id}(\theta^*) = \nabla_\theta Q(\theta|\theta^*)|_{\theta=\theta^*} + \nabla_\theta H(\theta|\theta^*)|_{\theta=\theta^*} = 0$$

which proves the claim. □

Remarks.

1. Even when the sequence $l_{id}(\hat{\theta}^{(k)})$ converges, there is in general no guarantee that $\hat{\theta}^{(k)}$ converges, e.g., the EM algorithm may jump back and forth between two attractors.
2. Even if the algorithm has a stable point, there is in general no guarantee that this stable point is a global maximum of the likelihood function, or even a local maximum. For many problems however, convergence to a local maximum has been demonstrated.
3. The solution in general depends on the initialization.
4. Several variants of the EM algorithm have been developed. One of the best ones is Fessler and Hero's SAGE algorithm (space-alternating generalized EM) whose convergence rate is often far superior to that of the EM algorithm [10].
5. There is a close relationship between convergence rate of the EM algorithm and Fisher information, as described in [2, 10].
6. Often the choice for Z is a pair (Y, U) where U are thought of as missing data, or *latent data*. Such was the case with the estimation of mixtures in Sec. 3 where the missing data were the labels associated with each observation.
7. As we saw with the mixture of exponential distributions in Sec. 4, the expectation step is relatively simple to implement when Z conditioned on Y is from an exponential family. Then the E-step is reduced to evaluating the conditional expectation of sufficient statistics, and the M-step is often tractable.

7 EM As an Alternating Maximization Algorithm

Interestingly, the EM algorithm may be formulated as an alternating maximization algorithm using an *auxiliary cost function* $L(q, \theta)$ to be maximized over both q and θ. This interpretation leads to a family of variants of the EM algorithm that in some cases are easier to implement and have good convergence properties.

We again assume for notational simplicity that the random variable Y and Z are discrete. The problem is to maximize the incomplete-data loglikelihood and evaluate (1). For fixed y, we again use the shorthand $q_{\theta,y}(z) = q_\theta(z|y)$, a pmf whose support set is $h^{-1}(y)$. We also denote by $q(z)$ a surrogate pmf over the same set.

At iteration $k+1$ of the EM algorithm, the current estimate $\hat{\theta}^{(k)}$ is updated as follows:

E-Step: evaluate $Q(\theta|\hat{\theta}^{(k)}) = \sum_z q_{\hat{\theta}^k}(z|y) \ln q_\theta(z)$;

M-Step: let $\hat{\theta}^{(k+1)} = \arg\max_{\theta \in \mathcal{S}} Q(\theta|\hat{\theta}^{(k)})$.

Let us now derive an alternative interpretation of the E- and M-steps. Denote by $\mathcal{Q}_y$ the set of pmf's over $h^{-1}(y) \subset \mathcal{Z}$. We define the function

$$L(q, \theta) \triangleq \sum_{z \in h^{-1}(y)} q(z) \ln \frac{r_\theta(y, z)}{q(z)} = \sum_{z \in h^{-1}(y)} q(z) \ln \frac{q_\theta(z)}{q(z)} \tag{22}$$

where the second equality results from (17). Note that this is not a negative Kullback-Leibler divergence between q and q_θ because q_θ does not sum to 1 over $h^{-1}(y)$.

For any pmf $q \in \mathcal{Q}_y$, we may write $l_{id}(\theta)$ as [1]

$$\begin{aligned} \ln p_\theta(y) &= \sum_{z \in h^{-1}(y)} q(z) \ln p_\theta(y) \\ &\overset{(a)}{=} \sum_z q(z) \ln \frac{r_\theta(y, z)}{q(z)} + \sum_z q(z) \ln \frac{q(z)}{q_\theta(z|y)} \\ &= L(q, \theta) + D(q \| q_{\theta, y}) \\ &\geq L(q, \theta). \end{aligned} \tag{23}$$

where (a) follows from (18). Hence $L(q, \theta)$ is the so-called *variational lower bound* on the loglikelihood, which holds for any choice of q and θ. Moreover the inequality holds with equality when $D(q \| q_{\theta,y}) = 0$, i.e., $q = q_{\theta,y} \in \mathcal{Q}_y$. Thus we have

$$\begin{aligned} \arg\max_{q \in \mathcal{Q}_y} L(q, \theta) &= q_{\theta, y} \\ \max_{q \in \mathcal{Q}_y} L(q, \theta) &= \ln p_\theta(y) \end{aligned} \tag{24}$$

for any value of θ.

Moreover, for any $q \in \mathcal{Q}_y$, it follows from (22) that

$$\begin{aligned} L(q, \theta) &= \sum_{z \in h^{-1}(y)} q(z) \ln q_\theta(z) - \sum_{z \in h^{-1}(y)} q(z) \ln q(z) \\ &= \mathbb{E}_q[l_{cd}(\theta)] + H(q) \end{aligned}$$

where the second term is independent of θ. Define the shorthand

$$q^{(k)}(z) = q_{\hat{\theta}^{(k)}, y}(z) = q_{\hat{\theta}^{(k)}}(z|y). \tag{25}$$

Observe that the E-step consists in evaluating the expected complete-data loglikelihood with respect to the distribution $q^{(k)}$, which by (24) is the distribution that maximizes $L(\cdot, \hat{\theta}^{(k)})$ over $\mathcal{Q}_y$:

$$q^{(k)} = \arg\max_{q \in \mathcal{Q}_y} L(q, \hat{\theta}^{(k)}).$$

[1] Alternatively, the variational lower bound (23) could be derived using Jensen's inequality:

$$\ln p_\theta(y) = \ln \sum_{z \in h^{-1}(y)} q_\theta(z) = \ln \sum_{z \in h^{-1}(y)} q(z) \frac{q_\theta(z)}{q(z)} \geq \sum_{z \in h^{-1}(y)} q(z) \ln \frac{q_\theta(z)}{q(z)} = L(q, \theta)$$

with equality if $q(z) = q_\theta(z)\alpha(y, \theta)\ \forall z \in h^{-1}(y)$ for some function $\alpha(y, \theta)$ that does not depend on z.

Thus the E-step may be thought of as resulting from an optimization of the auxiliary function over q, when the second argument is fixed at $\hat{\theta}^{(k)}$.

Now consider the M-step. Since

$$Q(\theta|\hat{\theta}^{(k)}) \quad = \quad \mathbb{E}_{q^{(k)}}[l_{cd}(\theta)] = L(q^{(k)}, \theta) - H(q^{(k)}), \tag{26}$$

we conclude that

$$\arg\max_{\theta\in\mathcal{S}} Q(\theta|\hat{\theta}^{(k)}) = \arg\max_{\theta\in\mathcal{S}} L(q^{(k)}, \theta).$$

Hence the EM algorithm is equivalent to the *alternating maximization algorithm* of Table 1, in which the function $L(q, \theta)$ is maximized over q with θ fixed, then over θ with q fixed, and this two-step process is repeated until convergence.

Initialize $\hat{\theta}^{(0)}$.
Then for $k = 0, 1, \cdots$, perform the following operations:
$q^{(k)} = \arg\max_{q\in\mathcal{Q}_y} L(q, \hat{\theta}^{(k)}) \quad \Longrightarrow \quad L(q^{(k)}, \hat{\theta}^{(k)}) = \ln p_{\hat{\theta}^{(k)}}(y);$
$\hat{\theta}^{(k+1)} = \arg\max_{\theta\in\mathcal{S}} L(q^{(k)}, \theta).$

Table 1: Alternating maximization of $L(q, \theta)$.

Therefore the sequence of loglikelihoods $\ln p_{\hat{\theta}^{(k)}}(y)$ is nondecreasing, and convergence of the EM algorithm to a stable value of the loglikelihood function is guaranteed.

8 Alternating Minimization of Kullback-Leibler Divergence

The alternating maximization procedure of Sec. 7 admits another interesting interpretation. The function $L(q, \theta)$ of (22) resembles a negative Kullback-Leibler divergence between two distributions on Z but is not one because $\sum_{z\in h^{-1}(y)} q_\theta(z) = p_\theta(y) \neq 1$, i.e., $q_\theta(z)$ is *not* a probability distribution over $z \in h^{-1}(y)$. However we can write $L(q, \theta)$ as a negative Kullback-Leibler divergence between two probability distributions on (Y, Z). To do so, we exploit the fact that a negative loglikelihood may be viewed as a Kullback-Leibler divergence: $-\ln p_\theta(y) = D(\mathbb{1}_y \| p_\theta)$ where $\mathbb{1}_y$ is the degenerate distribution on $\mathcal{Y}$ that puts all of its mass at the point y. Hence

$$\begin{aligned} L(q, \theta) &= \sum_{z\in h^{-1}(y)} q(z) \ln \frac{r_\theta(y, z)}{q(z)} \\ &= \sum_{y'\in\mathcal{Y}} \sum_{z\in h^{-1}(y')} \mathbb{1}_{\{y'=y\}}\, q(z)\, \ln \frac{r_\theta(y', z)}{\mathbb{1}_{\{y'=y\}}\, q(z)} \\ &= -D(r\|r_\theta) \end{aligned}$$

where r and r_θ are the following distributions over $\mathcal{Y} \times \mathcal{Z}$:

$$\begin{aligned} r(y', z) \triangleq \mathbb{1}_y(y')q(z) &= \mathbb{1}_{\{y'=y\}}\, q(z) \\ r_\theta(y', z) &= \mathbb{1}_{\{y'=h(z)\}}\, q_\theta(z). \end{aligned}$$

Denote by $\mathcal{R}$ the set of joint pmf's whose support set is $\{y\} \times h^{-1}(y) \subset \mathcal{Y} \times \mathcal{Z}$, and by $\mathcal{R}_\mathcal{S}$ the set of joint pmf's $r_\theta(y,z)$ parameterized by $\theta \in \mathcal{S}$; this is also called the *model family* . Using the shorthand

$$\hat{r}^{(k)}(y,z) = r_{\hat{\theta}^{(k)}}(y,z)$$

and observing that $\mathbb{1}_y q^{(k)} \in \mathcal{R}$, we see that the algorithm of Table 1 is equivalent to the algorithm of Table 2 for minimizing Kullback-Leibler divergence. This iterative minimization procedure is illustrated in Fig. 2.

Initialize $\hat{r}^{(0)} \in \mathcal{R}_\mathcal{S}$.
Then for $k = 0, 1, \cdots$, perform the following operations:
$r^{(k)} = \arg\min_{r \in \mathcal{R}} D(r \| \hat{r}^{(k)})$
$\hat{r}^{(k+1)} = \arg\min_{\hat{r} \in \mathcal{R}_\mathcal{S}} D(r^{(k)} \| \hat{r})$.

Table 2: Alternating minimization of Kullback-Leibler divergence.

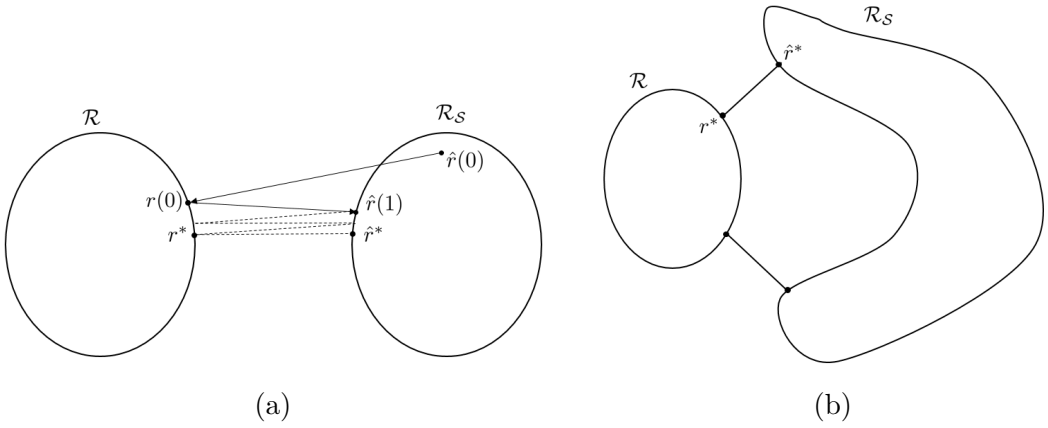

Figure 2: A pictorial view of alternating minimization of Kullback-Leibler divergence. The set $\mathcal{R}$ is convex. (a) For convex $\mathcal{R}_\mathcal{S}$, convergence to a global minimizer $(r^*, \hat{r}^*)$ is guaranteed; (b) for nonconvex $\mathcal{R}_\mathcal{S}$, different initializations will lead to different local minima.

The Kullback-Leibler divergence $D(r\|\hat{r})$ is convex in the pair $(r, \hat{r})$. As shown by Csiszár and Tusnady [11], if the sets $\mathcal{R}$ and $\mathcal{R}_\mathcal{S}$ are convex, the algorithm of Table 2 is guaranteed to converge to a global minimum (over $\mathcal{R} \times \mathcal{R}_\mathcal{S}$) of the function $D(r\|\hat{r})$. In our problem, $\mathcal{R}$ is convex, so a sufficient condition for convergence to a global minimum is that the model family $\mathcal{R}_\mathcal{S}$ be convex. In terms of our original EM problem, this would ensure convergence to a global maximum of the loglikelihood function $\ln p_\theta(y)$. However, in most problems of practical interest (e.g., mixture of Gaussians), the model class is nonconvex, and convergence to a global maximum is unlikely.

9 Extensions

In some cases, finding the maximizing θ in the M-step of the EM algorithm is difficult. In other cases, computing the expectation in the E-step is difficult because the conditional

expectation $q_{\hat{\theta}^{(k)}}(z|y)$ does not admit a simple form. Several generalizations of the EM algorithm have been proposed to deal with such difficulties. See Gunawardana and Byrne [9] for an elegant analysis of the convergence properties of such generalizations.

- *Generalized EM*: In the M-step of the EM algorithm, one does not perform a full optimization over θ but instead look for some $\hat{\theta}^{(k+1)}$ that increases the value of the cost function:
$$Q(\hat{\theta}^{(k+1)}|\hat{\theta}^{(k)}) \geq Q(\hat{\theta}^{(k)}|\hat{\theta}^{(k)})$$
or equivalently, by (26),
$$L(q^{(k)}, \hat{\theta}^{(k+1)}) \geq L(q^{(k)}, \hat{\theta}^{(k)}).$$
From the algorithm in Table 1, we see that the monotonicity property of the likelihood function is preserved.

- *Variational EM* [12]: Use a tractable surrogate distribution $q^{(k)}$ in place of the ideal $q_{\hat{\theta}^{(k)}}(z|y)$. For instance, choose $q^{(k)}$ from an exponential family, or choose $q^{(k)}$ to be a product distribution. Given a family $\mathcal{Q}$ of tractable candidate distributions, variational EM selects $q^{(k)} \in \mathcal{Q}$ that maximizes $L(\cdot, \hat{\theta}^{(k)})$. Unlike the EM algorithm, monotonicity of the likelihood function is not guaranteed (unless $\mathcal{Q}$ is the class of all pmf's, in which case variational EM trivially reduces to EM).

References

[1] H. Hartley, "Maximum likelihood estimation from incomplete data," *Biometrics*, Vol. 14, pp. 174—194, 1958.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," (with discussion) J. Roy. Stat. Soc. B., Vol. 39, No. 1, pp. 1—38, 1977.

[3] X.-L. Meng and D. van Dijk, "The EM algorithm: An Old Folk Song Sung to a Fast New Tune, (with discussion) J. Roy. Stat. Soc. B., Vol. 59, No. 3, pp. 511—567, 1997.

[4] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd Ed., Wiley, 2008.

[5] O. Barndorff-Nielsen, *Information and Exponential Families: In Statistical Theory*, Wiley, 1978.

[6] L. Xu and M. I. Jordan, "On Convergence Properties of the EM Algorithm for Gaussian mixtures," *Neural Computation*, 8, pp. 129—151, 1996.

[7] R. J. Hathaway, "Another Interpretation of the EM Algorithm for Mixture Distributions," *Stat. Prob. Letters*, Vol. 4, pp. 53—56, 1986.

[8] C. F. J. Wu, "On the Convergence Properties of the EM Algorithm," *Ann. Stat.*, Vol. 11, No. 1, pp. 95—103, 1983.

[9] A. Gunawardana and W. Byrne, "Convergence Theorems for Generalized Alternating Minimization Procedures," *J. of Machine Learning Research*, Vol. 6, pp. 2049—2073, 2005.

[10] J. A. Fessler and A. O. Hero, "Space-Alternating Generalized Expectation-Maximization Algorithm," *IEEE Trans. on Sig. Proc.*, Vol. 42, No. 10, pp. 2664—2677, Oct. 1994.

[11] I. Csiszár and G. Tusnady, "Information Geometry and Alternating Minimization Procedures," *Statistics and Decisions*, Vol. 1, supplement issue, pp. 205—237, 1984.

[12] R. Neal and G. Hinton, "A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants," in *Learning in Graphical Models*, M. Jordan, Ed., Kluwer, 1998.