Optimization Methods II.
EM algorithms.

Anatoli Iambartsev
IME-USP

[RC] Missing-data models. Demarginalization.

The term *EM algorithms* has been around for a long time by [DLR].

Consider the case where the density of the observations can be expresses as

$$g_\theta(x) = \int f_\theta(x,z)dz, \quad g_\theta(x) \to f_\theta(x,z).$$

"This representation occurs in many statistical settings, including censoring models and mixtures and latent variable models (tobit, probit, arch, stochastic volatility, ect.)."

[RC] Example 5.12

The mixture model of Example 5.2 (see previous lecture),

$$0.25N(\mu_1, 1) + 0.75N(\mu_2, 1),$$

can be expressed as a missing-data model even though the (observed) likelihood can be computed in a manageable time. Indeed, if we introduce a vector $z = (z_1, \ldots, z_n) \in \{1, 2\}^n$ in addition to the sample $x = (x_1, \ldots, x_n)$ such that

$$\mathbb{P}_\theta(Z_i = 1) = 1 - \mathbb{P}_\theta(Z_i = 2) = 0.25, \quad X_i \mid Z_i = z \sim N(\mu_z, 1).$$

we recover the mixture model from the Example 5.2 as the marginal distribution of X_i. The (observed) likelihood is then obtained as $\mathbb{E}(H(x, z))$ for

$$H(x, z) \propto \prod_{i:z_i=1} \frac{1}{4} \exp\left\{-\frac{(x_i - \mu_1)^2}{2}\right\} \prod_{i:z_i=2} \frac{3}{4} \exp\left\{-\frac{(x_i - \mu_2)^2}{2}\right\}.$$

[RC] Example 5.12 The (observed) likelihood is then obtained as $\mathbb{E}(H(x,z))$ for

$$H(x,z) \propto \prod_{i:z_i=1} \frac{1}{4} \exp\left\{-\frac{(x_i-\mu_1)^2}{2}\right\} \prod_{i:z_i=2} \frac{3}{4} \exp\left\{-\frac{(x_i-\mu_2)^2}{2}\right\}:$$

Indeed,

$$\begin{aligned} g_{\mu_1,\mu_2}(x) &= \int f_{\mu_1,\mu_2}(x,z)dz = \sum_{z\in\{1,2\}^n} \frac{H(x,z)}{(\sqrt{2\pi})^n} \\ &= \sum_{z\in\{1,2\}^n} \prod_{i:z_i=1} \frac{1}{4} \frac{e^{-\frac{(x_i-\mu_1)^2}{2}}}{\sqrt{2\pi}} \prod_{i:z_i=2} \frac{3}{4} \frac{e^{-\frac{(x_i-\mu_2)^2}{2}}}{\sqrt{2\pi}} \\ &= \prod_{i=1}^n \Big(\frac{1}{4} \frac{e^{-\frac{(x_i-\mu_1)^2}{2}}}{\sqrt{2\pi}} + \frac{3}{4} \frac{e^{-\frac{(x_i-\mu_2)^2}{2}}}{\sqrt{2\pi}}\Big). \end{aligned}$$

[RC] Example 5.13

Censored data may come from experiments where some potential observations are replaced with a lower bound because they take too long to observe. Suppose that we observe $Y_1, \ldots, Y_m$, iid, from $f(y-\theta)$ and that the $(n-m)$ remaining $(Y_{m+1}, \ldots, Y_n)$ are censored at the threshold a. The corresponding likelihood function is then

$$L(\theta \mid y) = \big(1 - F(a-\theta)\big)^{n-m} \prod_{i=1}^{m} f(y_i - \theta),$$

where F is the cdf associated with f and $y = (y_1, \ldots, y_m)$.

[RC] Example 5.13

$$L(\theta \mid y) = \Big(1 - F(a-\theta)\Big)^{n-m} \prod_{i=1}^{m} f(y_i - \theta).$$

If we had observed the last $n-m$ values, say $z = (z_{m+1}, \ldots, z_n)$, with $z_i \geq a (i = m+1, \ldots, n)$, we could have constructed the (complete data) likelihood

$$L^c(\theta \mid y, z) = \prod_{i=1}^{m} f(y_i - \theta) \prod_{i=m+1}^{n} f(z_i - \theta).$$

Note that

$$L(\theta \mid y) = \mathbb{E}\Big(L^c(\theta \mid y, Z)\Big) = \int L^c(\theta \mid y, z) f(z \mid y, \theta) dz,$$

where $f(z \mid y, \theta)$ is the density of the missing data conditional on the observed data, namely the product of the $f(z_i - \theta)/(1 - F(a-\theta))$'s; i.e., $f(z-\theta)$ restricted to $(a, +\infty)$.

Main Idea of EM algorithms

Demarginalization: $g_\theta(x) \to f_\theta(x,z), g_\theta(x) = \int f_\theta(x,z)dz$.

A values from z can be generated by the conditional distribution

$$k_\theta(z \mid x) = \frac{f_\theta(x,z)}{g_\theta(x)}.$$

Take a logarithm

$$\log g_\theta(x) = \log f_\theta(x,z) - \log k_\theta(z \mid x).$$

In notations of likelihood function

$$\log L(\theta \mid x) = \log L^c(\theta \mid x, z) - \log k_\theta(z \mid x),$$

where L^c stands for complete likelihood function.

Main Idea of EM algorithms

$$\log L(\theta \mid x) = \log L^c(\theta \mid x, z) - \log k_\theta(z \mid x),$$

Let us fix a value θ_0 and calculate the expectation according to the distribution $k_{\theta_0}(z \mid x)$:

$$\begin{aligned} \log L(\theta \mid x) &= \mathbb{E}_{k,\theta_0} \log L^c(\theta \mid x, z) - \mathbb{E}_{k,\theta_0} \log k_\theta(z \mid x) \\ &=: Q(\theta \mid \theta_0, x) - H(\theta \mid \theta_0, x). \end{aligned}$$

Theorem. Let θ_1 maximizes the Q, i.e.,

$$Q(\theta_1 \mid \theta_0, x) = \max_\theta Q(\theta \mid \theta_0, x).$$

Then

$$\log L(\theta_1 \mid x) \geq \log L(\theta_0 \mid x).$$

Main Idea of EM algorithms. Proof.

$$Q(\theta_1 \mid \theta_0, x) = \max_{\theta} Q(\theta \mid \theta_0, x) \Rightarrow \log L(\theta_1 \mid x) \geq \log L(\theta_0 \mid x).$$

Proof.

$$\log L(\theta_1 \mid x) - \log L(\theta_0 \mid x)$$
$$\big(Q(\theta_1 \mid \theta_0, x) - Q(\theta_0 \mid \theta_0, x)\big) - \big(H(\theta_1 \mid \theta_0, x) - H(\theta_0 \mid \theta_0, x)\big)$$

Note that by definition of θ_1

$$Q(\theta_1 \mid \theta_0, x) - Q(\theta_0 \mid \theta_0, x) \geq 0.$$

Main Idea of EM algorithms. Proof.

$$\log L(\theta_1 \mid x) - \log L(\theta_0 \mid x)$$
$$\big(Q(\theta_1 \mid \theta_0, x) - Q(\theta_0 \mid \theta_0, x)\big) - \big(H(\theta_1 \mid \theta_0, x) - H(\theta_0 \mid \theta_0, x)\big)$$

and

$$\begin{aligned} &H(\theta_1 \mid \theta_0, x) - H(\theta_0 \mid \theta_0, x) \\ &= \mathbb{E}_{k,\theta_0} \log k_{\theta_1}(Z \mid x) - \mathbb{E}_{k,\theta_0} \log k_{\theta_0}(Z \mid x) \\ &= \mathbb{E}_{k,\theta_0} \log \frac{k_{\theta_1}(Z \mid x)}{k_{\theta_0}(Z \mid x)} \le \log \mathbb{E}_{k,\theta_0} \frac{k_{\theta_1}(Z \mid x)}{k_{\theta_0}(Z \mid x)} = \log 1 = 0, \end{aligned}$$

where Jensen inequality was used $\mathbb{E} \log \xi \le \log \mathbb{E}\xi$.

Main Idea of EM algorithms. Proof.

$$\log L(\theta_1 \mid x) - \log L(\theta_0 \mid x)$$
$$\big(Q(\theta_1 \mid \theta_0, x) - Q(\theta_0 \mid \theta_0, x)\big) - \big(H(\theta_1 \mid \theta_0, x) - H(\theta_0 \mid \theta_0, x)\big)$$

We have

$$Q(\theta_1 \mid \theta_0, x) - Q(\theta_0 \mid \theta_0, x) \geq 0,$$
$$H(\theta_1 \mid \theta_0, x) - H(\theta_0 \mid \theta_0, x) \leq 0,$$

thus

$$\log L(\theta_1 \mid x) - \log L(\theta_0 \mid x) \geq 0.$$

This completes the proof of the theorem. □

Main Idea of EM algorithms.

Each iteration EM algorithm maximizes a function Q. Let θ_t be a sequence obtained recurcively

$$Q(\theta_{t+1} \mid \theta_t, x) = \max_{\theta} Q(\theta \mid \theta_t, x).$$

This recurrent scheme consists on two steps *expectation* and *maximization*, that gives the name for the scheme: EM algorithm.

EM algorithm.

Choose the initial parameter θ_0 and repeat:

- **E-step**. Calculate the expectation

$$Q(\theta \mid \theta_t, x) = \mathbb{E}_{k,\theta_t} \log L^c(\theta \mid x, Z)$$

  with respect to the distribution $k_{\theta_t}(z \mid x)$.

- **M-step**. Maximize $Q(\theta \mid \theta_t, x)$ on θ and determine the next value

$$\theta_{t+1} = \arg\max_{\theta} Q(\theta \mid \theta_t, x),$$

  define $t = t + 1$, return to the E-step

[RC]. Example 5.14 (Cont. of Example 5.13)

Let again $Y_1, \ldots, Y_m$ are iid com density $f(y-\theta)$ and others $Y_{m+1}, \ldots, Y_n$ are censured at the level a. The likelihood function

$$L(\theta \mid y) = \big(1 - F(a-\theta)\big)^{n-m} \prod_{i=1}^{m} f(y_i - \theta),$$

where $F(a-\theta) = \mathbb{P}(Y_i \leq a)$. If we had observed the last $n-m$ values, say $z = (z_{m+1}, \ldots, z_n)$, with $z_i \geq a (i = m+1, \ldots, n)$, we could have constructed the (complete data) likelihood

$$L^c(\theta \mid y, z) = \prod_{i=1}^{m} f(y_i - \theta) \prod_{i=m+1}^{n} f(z_i - \theta).$$

and

$$k_\theta(z \mid y) = \prod_{i=1}^{n-m} \frac{f(z_i - \theta)}{1 - F(a-\theta)}.$$

[RC]. Example 5.14 (Cont. of Example 5.13)

Suppose that $f(y-\theta)$ corresponds to the $N(\theta, 1)$ distribution, the complete-data likelihood is

$$L^c(\theta \mid y, z) \propto \prod_{i=1}^{m} e^{-(y_i-\theta)^2/2} \prod_{i=m+1}^{n} e^{-(z_i-\theta)^2/2},$$

resulting in the expected complete-data log-likelihood

$$Q(\theta \mid \theta_0, y) = -\frac{1}{2}\sum_{i=1}^{m}(y_i-\theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n} \mathbb{E}_{k,\theta_0}\big((Z_i-\theta)^2\big),$$

where the missing observations Z_i are distributed from a normal $N(\theta, 1)$ distribution truncated in a. Doing the M-step (i.e., differentiating the function $Q(\theta \mid \theta_0, y)$ in θ) and setting it equal to 0 then leads to the EM update

$$\hat{\theta} = \frac{m\bar{y} + (n-m)\mathbb{E}_{k,\theta_0}(Z_1)}{n}.$$

[RC]. Example 5.14 (Cont. of Example 5.13)

Doing the M-step ... the EM update

$$\hat{\theta} = \frac{m\bar{y} + (n-m)\mathbb{E}_{k,\theta_0}(Z_1)}{n}.$$

Since $\mathbb{E}_{k,\theta_0}(Z_1) = \theta + \frac{\phi(a-\theta)}{1-\Phi(a-\theta)}$, where ϕ and Φ are the normal pdf and cdf, respectively, the EM sequence is

$$\theta_{t+1} = \frac{m}{n}\bar{y} + \frac{n-m}{n}\left(\theta_t + \frac{\phi(a-\theta_t)}{1-\Phi(a-\theta_t)}\right).$$

Principle of missing information (informally)

$$\log L(\theta \mid x) = \log L^c(\theta \mid x, z) - \log k_\theta(z \mid x)$$
$$\Rightarrow -\frac{\partial^2 \log L(\theta \mid x)}{\partial \theta^2} = -\frac{\partial^2 \log L^c(\theta \mid x, z)}{\partial \theta^2} + \frac{\partial^2 \log k_\theta(z \mid x)}{\partial \theta^2}$$

Observed information = Complete information –
– Missing information

Informally about a convergence rate of EM algorithms: if a proportion of missing information increases with iterations, then a rate of convergence of an algorithm decreases.

EM algorithms for exponential family models.

It is more easy to implement an algorithm when a complete data (z, x) has the exponential family type of distribution in a canonic form:

$$p(z, x \mid \theta) = b(z, x)\frac{\exp(\theta^T s(z, x))}{a(\theta)}$$

Let $\mathbf{y} = (z, x)$ be a vector of complete data. We have

$$\begin{aligned} \log p(\mathbf{y} \mid \theta) &= \log b(\mathbf{y}) + \theta^T s(\mathbf{y}) - \log a(\theta) \\ \frac{\partial}{\partial\theta} \log p(\mathbf{y} \mid \theta) &= s(\mathbf{y}) - \frac{1}{a(\theta)} \frac{\partial a(\theta)}{\partial\theta} \end{aligned}$$

Remember that

$$a(\theta) = \int b(\mathbf{y}) \exp(\theta^T s(\mathbf{y})) d\mathbf{y}$$

EM algorithms for exponential family models.

$$\frac{\partial}{\partial\theta}\log p(\mathbf{y}\mid\theta) = s(\mathbf{y}) - \frac{1}{a(\theta)}\frac{\partial a(\theta)}{\partial\theta}, a(\theta) = \int b(\mathbf{y})\exp(\theta^T s(\mathbf{y}))d\mathbf{y}$$

we have

$$\begin{aligned}
\frac{\partial \log a(\theta)}{\partial\theta} &= \frac{1}{a(\theta)}\frac{\partial a(\theta)}{\partial\theta} = \frac{1}{a(\theta)}\int b(\mathbf{y})\frac{\partial \exp(\theta^T s(\mathbf{y}))}{\partial\theta}d\mathbf{y} \\
&= \frac{1}{a(\theta)}\int b(\mathbf{y})s(\mathbf{y})\exp(\theta^T s(\mathbf{y}))d\mathbf{y} \\
&= \int s(\mathbf{y})\frac{b(\mathbf{y})\exp(\theta^T s(\mathbf{y}))}{a(\theta)}d\mathbf{y} = \mathbb{E}(s(\mathbf{y})\mid\theta)
\end{aligned}$$

Thus,

$$\frac{\partial}{\partial\theta}\log p(\mathbf{y}\mid\theta) = 0 \Leftrightarrow s(\mathbf{y}) = \mathbb{E}(s(\mathbf{y})\mid\theta)$$

EM algorithms for exponential family models.

Implementation of EM algorithm: E-step

$$\begin{aligned} Q(\theta \mid \theta_t) &= \int \log p(z, x \mid \theta) p(z \mid \theta_t, x) dz \\ &= \int b(z, x) p(z \mid \theta_t, x) dz + \theta^T \int s(z, x) p(z \mid \theta_t, x) dz - \log a(\theta) \end{aligned}$$

Note that the first term will not participate in M-step. M-step: we obtain extreme point

$$\begin{aligned} \frac{\partial Q(\theta \mid \theta_t)}{\partial \theta} &= \int s(z, x) p(z \mid \theta_t, x) dz - \frac{1}{a(\theta)} \frac{\partial a(\theta)}{\partial \theta} \\ &= \mathbb{E}(s(z, x) \mid \theta_t, x) - \mathbb{E}(s(z, x) \mid \theta) \end{aligned}$$

Remember that $\mathbb{E}(s(z, x) \mid \theta) = \frac{1}{a(\theta)} \frac{\partial a(\theta)}{\partial \theta}$

EM algorithms for exponential family models.

Thus the maximization of $Q(\theta \mid \theta_t, x)$ in M-step is equivalent to solve the following equation

$$\mathbb{E}(s(z,x) \mid \theta_t, x) = \mathbb{E}(s(z,x) \mid \theta)$$

If the solution exists, then it is unique.

Monte Carlo for E-step.

Given θ_t we need to calculate $Q(\theta \mid \theta_t, x) = \mathbb{E}_{k,\theta_t} \log L^c(\theta \mid Z, x)$. When it is difficult to calculate explicitly we can calculate approximately using Monte Carlo:

1. generate $z_1, \ldots, z_m$ according $k_\theta(z \mid x)$;

2. calculate $Q(\theta \mid \theta_t) = \frac{1}{m} \sum_{i=1}^m \log L^c(\theta \mid z_i, x)$;

during M-step one maximizes Q by θ in order to obtain θ_{t+1}.

EM algorithm.

[H] Hunter, D.R. *On the Geometry of EM algorithms.*: this paper demonstrates how the geometry of EM algorithms can help explain how their rate of convergence is related to the proportion of missing data.

In footnote [H] wrote: "In a footnote, [DLR] refer to the comment of a referee, who noted that the use of the word "algorithm" may be criticized since EM is not, strictly speaking, an algorithm. However, EM *is* a recipe for creating algorithms, and thus we consider the set of "EM algorithms" to consist of all algorithms baked according to the EM recipe."

References.

[DLR] Dempster, A.P., Laird, N.M., and Rubin, D.B. *Maximum likelihood from incomplete data via the EM algorithm*, J.Roy. Statist. Soc. Ser. B, **39**, 1-38, 1977.

[H] David R. Hunter. *On the Geometry of EM algorithms.* Technical Report 0303, Dep. of Stat., Penn State University. February, 2003.

[RC] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Series “Use R!” . Springer