Notes on the EM algorithm

Sandrine Dudoit
Division of Biostatistics, UC Berkeley

PH 296, Section 33 - October 22, 2001

The EM Algorithm

A general approach to iterative computation of maximum likelihood estimates when the observations can be viewed as incomplete data.

A. P. Dempster, N. M. Laird, and D. B. Rubin. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm *J. R. Statist. Soc. B.* **39**(1): 1-38.

The setting

- Complete data X, with density $f(x|\phi)$.
- Observed incomplete data $Y = y(X)$, with density $g(y|\phi)$

$$g(y|\phi) = \int_{\mathcal{X}(y)} f(x|\phi)dx,$$

  where $\mathcal{X}(y) = \{x : y = y(x)\}$.
- Given observed data y, the EM (Expectation Maximization) algorithm seeks to find a maximum likelihood estimate (MLE) of ϕ, i.e., a value of the parameter ϕ which maximizes $g(y|\phi)$. It makes explicit use of the associated family $f(x|\phi)$.
- There are possibly several specifications of $f(x|\phi)$, often a natural choice will be obvious.

The algorithm

Let

$$Q(\phi'|\phi) = E(\log f(X|\phi')|y, \phi).$$

Starting with an initial estimate $\phi^{(0)}$ of ϕ, the EM iteration $\phi^{(k)} \to \phi^{(k+1)}$ is as follows:

E-step. (Expectation) Compute $Q(\phi|\phi^{(k)})$.

M-step. (Maximization) Choose $\phi^{(k+1)}$ to maximize $Q(\phi|\phi^{(k)})$.

Then, the log-likelihood for the observed data $L(\phi) = \log g(y|\phi)$ is non-decreasing at each iteration of the EM algorithm.

Jensen's inequality

For a random variable Z and convex function h

$$E(h(Z)) \geq h(E(Z)),$$

with equality if and only if (iff) Z is constant almost everywhere (a.e.).

E.g. 1. $E(Z^2) \geq (E(Z))^2$.

E.g. 2. $E(\log Z) \leq \log E(Z)$ – log is a concave function.

More notation ...

- Density of X: $f(x|\phi)$;
  Density of Y: $g(y|\phi)$.
- Conditional density of X given Y:

$$k(x|y, \phi) = f(x|\phi)/g(y|\phi).$$

- Log-likelihood of Y: $L(\phi) = \log g(y|\phi)$.
- Key function in EM algorithm:

$$Q(\phi'|\phi) = E(\log f(X|\phi')|y, \phi).$$

- Function needed in proof:

$$H(\phi'|\phi) = E(\log k(X|y, \phi')|y, \phi).$$

- Thus

$$Q(\phi'|\phi) = L(\phi') + H(\phi'|\phi).$$

General properties of EM algorithm

Lemma 1. For any pair (ϕ', ϕ)

$$H(\phi'|\phi) \leq H(\phi|\phi),$$

with equality iff $k(X|y, \phi') = k(X|y, \phi)$ a.e.

General properties of EM algorithm

Proof.

$$
\begin{aligned}
LHS - RHS &= E(\log k(X|y,\phi')|y,\phi) - E(\log k(X|y,\phi)|y,\phi) \\
&= E\left(\log \frac{k(X|y,\phi')}{k(X|y,\phi)} \middle| y,\phi \right) \\
&\leq \log E\left(\frac{k(X|y,\phi')}{k(X|y,\phi)} \middle| y,\phi \right) - \text{Jensen} \\
&= \log \int_{\mathcal{X}(y)} \frac{k(x|y,\phi')}{k(x|y,\phi)} k(x|y,\phi)dx \\
&= \log \int_{\mathcal{X}(y)} k(x|y,\phi')dx \\
&= \log 1 = 0.
\end{aligned}
$$

Equality holds iff $k(X|y,\phi') = k(X|y,\phi)$ a.e. $\square$

General properties of EM algorithm

Theorem 1. The log-likelihood of the observed data is non-decreasing at every EM iteration, i.e.,

$$L(\phi^{(new)}) \geq L(\phi^{(old)}),$$

where equality holds iff both

$$\begin{aligned} Q(\phi^{(new)}|\phi^{(old)}) &= Q(\phi^{(old)}|\phi^{(old)}) \text{ and} \\ k(X|y,\phi^{(new)}) &= k(X|y,\phi^{(old)}) \text{ a.e.} \end{aligned}$$

General properties of EM algorithm

Proof.

$$
\begin{aligned}
L(\phi^{(new)}) - L(\phi^{(old)}) &= \{Q(\phi^{(new)}|\phi^{(old)}) - Q(\phi^{(old)}|\phi^{(old)})\} \\
&- \{H(\phi^{(new)}|\phi^{(old)}) - H(\phi^{(old)}|\phi^{(old)})\}.
\end{aligned}
$$

By definition of the algorithm, the difference in Q functions is ≥ 0, and by Lemma 1, the difference in H functions is ≤ 0 with equality iff $k(X|y, \phi^{(new)}) = k(X|y, \phi^{(old)})$ a.e. $\square$

General properties of EM algorithm

Corollaries 1 and 2. An MLE is a fixed point of the EM algorithm.

Theorem 2. Conditions under which an instance of the EM algorithm converges.

N.B. Does not imply convergence to an MLE.

Theorems 3 and 4. Differentiability conditions for convergence to an MLE.

EM for Gaussian mixtures

Complete data $X = (Y, Z)$.

Observed incomplete data Y.

Assume

$pr(Z = k|\phi) = \pi_k$,

$(Y|Z = k, \phi) \sim \mathrm{N}(\mu_k, \sigma_k^2)$,

where $k = 1, \ldots, K$, and $\sum_k \pi_k = 1$.

Parameter $\phi = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$.

Problem. Given observed incomplete data $\mathbf{y} = (y_1, \ldots, y_n)$ i.i.d. with density $g(\cdot|\phi)$ find MLE of ϕ.

EM for Gaussian mixtures

Let $h(\cdot|\mu_k, \sigma_k^2)$ denote the $\mathrm{N}(\mu_k, \sigma_k^2)$ density:

$$h(y|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(y-\mu_k)^2\right).$$

Observed incomplete data log-likelihood:

$$\begin{aligned} L(\phi) &= \log g(\mathbf{y}|\phi) = \sum_{i=1}^{n} \log g(y_i|\phi) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} pr(Z_i = k|\phi)\ pr(y_i|Z_i = k, \phi) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k\ h(y_i|\mu_k, \sigma_k^2). \end{aligned}$$

Hard to maximize directly.

EM for Gaussian mixtures

Complete data log-likelihood:

$$\begin{aligned}
\log f(\mathbf{x}|\phi) &= \sum_{i=1}^{n} \log f(x_i|\phi) \qquad\qquad x_i = (y_i, z_i) \\
&= \sum_{i=1}^{n} \log \prod_{k=1}^{K} \Big(pr(Z_i = k|\phi)\ \ pr(y_i|Z_i = k, \phi) \Big)^{I(z_i = k)} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} I(z_i = k)\ \log\Big(\pi_k\ h(y_i|\mu_k, \sigma_k^2) \Big).
\end{aligned}$$

EM for Gaussian mixtures

$$
\begin{aligned}
Q(\phi|\phi^{(old)}) &= E(\log f(\mathbf{X}|\phi)|\mathbf{y}, \phi^{(old)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} pr(Z_i = k|\mathbf{y}, \phi^{(old)}) \log\Big(\pi_k \; h(y_i|\mu_k, \sigma_k^2)\Big) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} pr(Z_i = k|\mathbf{y}, \phi^{(old)}) \\
&\quad \times \Big(\log \pi_k - \log \sigma_k - \frac{1}{2\sigma_k^2}(y_i - \mu_k)^2\Big),
\end{aligned}
$$

up to an additive constant in ϕ.

EM for Gaussian mixtures

E-step. Compute $z_{i,k}^{(old)} = pr(Z_i = k|\mathbf{y}, \phi^{(old)})$.

$$
\begin{aligned}
pr(Z_i = k|\mathbf{y}, \phi^{(old)}) &= \frac{pr(\mathbf{y}|Z_i = k, \phi^{(old)})\ pr(Z_i = k|\phi^{(old)})}{\sum_{k=1}^{K} pr(\mathbf{y}|Z_i = k, \phi^{(old)})\ pr(Z_i = k|\phi^{(old)})} \\
&= \frac{h(y_i|\mu_k^{(old)}, \sigma_k^{2\ (old)})\ \pi_k^{(old)}}{\sum_{k=1}^{K} h(y_i|\mu_k^{(old)}, \sigma_k^{2\ (old)})\ \pi_k^{(old)}}.
\end{aligned}
$$

EM for Gaussian mixtures

M-step. Find $\phi^{(new)}$ which maximizes

$$Q(\phi|\phi^{(old)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k}^{(old)} \Big(\log \pi_k - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_i - \mu_k)^2 \Big).$$

Set first derivatives to zero and solve.

EM for Gaussian mixtures

$$
\begin{aligned}
\pi_k^{(new)} &= \frac{\sum_{i=1}^{n} z_{i,k}^{(old)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} z_{i,k}^{(old)}} = \frac{\sum_{i=1}^{n} z_{i,k}^{(old)}}{n}; \\
\mu_k^{(new)} &= \frac{\sum_{i=1}^{n} z_{i,k}^{(old)} y_i}{\sum_{i=1}^{n} z_{i,k}^{(old)}}; \\
\sigma_k^{2\ (new)} &= \frac{\sum_{i=1}^{n} z_{i,k}^{(old)} (y_i - \mu_k^{(new)})^2}{\sum_{i=1}^{n} z_{i,k}^{(old)}}.
\end{aligned}
$$

EM for Hidden Markov Models

L. R. Rabiner. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE.* **77** (2): 257-286.

Section III C.

EM for exponential family

Suppose $f(x|\phi) = b(x)\exp(\phi t(x)^T)/a(\phi)$. Then

$Q(\phi|\phi^{(old)}) = -\log a(\phi) + \phi E(t(X)|y, \phi^{(old)})^T + E(\log b(X)|y, \phi^{(old)})$.

Maximizing $Q(\phi|\phi^{(old)})$ wrt. ϕ is equivalent to maximizing $-\log a(\phi) + \phi E(t(X)|y, \phi^{(old)})^T$ wrt. ϕ.

E-step. Compute $E(t(X)|y, \phi^{(old)})$.

M-step. Maximize $-\log a(\phi) + \phi E(t(X)|y, \phi^{(old)})^T$ wrt. ϕ, i.e., solve

$$E(t(X)|y, \phi^{(old)}) = \frac{\partial}{\partial \phi} \log a(\phi) = \phi = E(t(X)|\phi).$$

Other interpretation of E-step

The E-step can be viewed as another M-step.

$$\begin{aligned} L(\phi) &= \log g(y|\phi) = \log \int_{\mathcal{X}(y)} f(x|\phi) dx \\ &= \log \int_{\mathcal{X}(y)} h(x|y) \frac{f(x|\phi)}{h(x|y)} dx \\ &\geq \int_{\mathcal{X}(y)} h(x|y) \log \frac{f(x|\phi)}{h(x|y)} dx - \text{Jensen} \\ &= \mathcal{L}(h, \phi). \end{aligned}$$

$$LHS - RHS = L(\phi) - \mathcal{L}(h, \phi) = KL(h(\cdot|y); k(\cdot|y, \phi)).$$

Other interpretation of E-step

E-step. Find $h^{(new)} = \text{argmax}_h \, \mathcal{L}(h, \phi^{(old)})$. From result on KL divergence

$h^{(new)}(\cdot|y) = k(\cdot|y, \phi^{(old)})$.

M-step. Find $\phi^{(new)} = \text{argmax}_\phi \, \mathcal{L}(h^{(new)}, \phi)$.

Kullback-Leibler divergence

Definition. For two densities f_1 and f_2 the Kullback-Leibler divergence is defined as:

$$KL(f_1; f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx.$$

Theorem. $KL(f_1; f_2) \geq 0$ with equality iff $f_1 = f_2$.

Proof. Jensen's inequality.