557: Mathematical Statistics II
The EM Algorithm

The EM Algorithm is a method for producing the maximum likelihood estimates in **incomplete data** problems, that is, models formulated for data that are only partially observed.

Suppose that random variables to be modelled can be partitioned $(\underset{\sim}{X}, \underset{\sim}{Y})$ where

- $\underset{\sim}{X} = (X_1, \ldots, X_m)^\mathsf{T}$ are **unobserved**, termed the **augmented data**
- $\underset{\sim}{X} = (Y_1, \ldots, Y_n)^\mathsf{T}$ are **observed**, termed the **incomplete data**
- $(\underset{\sim}{X}, \underset{\sim}{Y})$ are termed the **complete data**

where

$$f_{\underset{\sim}{Y}|\underset{\sim}{\theta}}(\underset{\sim}{y}|\underset{\sim}{\theta}) = \int f_{\underset{\sim}{X},\underset{\sim}{Y}|\underset{\sim}{\theta}}(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta})\, d\underset{\sim}{x}$$

In this formulation,

$$L(\underset{\sim}{\theta}|\underset{\sim}{y}) = f_{\underset{\sim}{Y}|\underset{\sim}{\theta}}(\underset{\sim}{y}|\underset{\sim}{\theta}).$$

is the **incomplete data** likelihood, and

$$L(\underset{\sim}{\theta}|\underset{\sim}{x}, \underset{\sim}{y}) = f_{\underset{\sim}{X},\underset{\sim}{Y}|\underset{\sim}{\theta}}(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta})$$

is the **complete data** likelihood.

Algorithm

The EM Algorithm facilitates maximization of the incomplete data likelihood $L(\underset{\sim}{\theta}|\underset{\sim}{y})$ by working with the complete data likelihood $L(\underset{\sim}{\theta}|\underset{\sim}{x}, \underset{\sim}{y})$ and the conditional distribution

$$f_{\underset{\sim}{X}|\underset{\sim}{Y},\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}) = \frac{f_{\underset{\sim}{X},\underset{\sim}{Y}|\underset{\sim}{\theta}}(\underset{\sim}{x}, \underset{\sim}{y}|\underset{\sim}{\theta})}{f_{\underset{\sim}{Y}|\underset{\sim}{\theta}}(\underset{\sim}{y}|\underset{\sim}{\theta})} = \frac{L(\underset{\sim}{\theta}|\underset{\sim}{x}, \underset{\sim}{y})}{L(\underset{\sim}{\theta}|\underset{\sim}{x})} = K(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}) \tag{1}$$

say. It follows from equation (1) that

$$\log L(\underset{\sim}{\theta}|\underset{\sim}{y}) = \log L(\underset{\sim}{\theta}|\underset{\sim}{x}, \underset{\sim}{y}) - \log K(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}) \tag{2}$$

However, the data $\underset{\sim}{x}$ are not observed, so consider replacing the right-hand side of equation (2) by the expectations with respect to the conditional density $f_{\underset{\sim}{X}|\underset{\sim}{Y},\underset{\sim}{\theta}}(\underset{\sim}{x}|\underset{\sim}{y}, \underset{\sim}{\theta}')$, for some $\underset{\sim}{\theta}' \in \Theta$. This yields

$$\log L(\underset{\sim}{\theta}|\underset{\sim}{y}) = \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y},\underset{\sim}{\theta}}}[\log L(\underset{\sim}{\theta}|\underset{\sim}{X}, \underset{\sim}{Y})|\underset{\sim}{y}, \underset{\sim}{\theta}'] - \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y},\underset{\sim}{\theta}}}[\log K(\underset{\sim}{X}|\underset{\sim}{Y}, \underset{\sim}{\theta})|\underset{\sim}{y}, \underset{\sim}{\theta}']. \tag{3}$$

Note that the notation indicates that we condition on a specific (but as yet unspecified) value of $\underset{\sim}{\theta}'$ when computing the expectations of $\log L(\underset{\sim}{\theta}|\underset{\sim}{X}, \underset{\sim}{y})$ and $\log K(\underset{\sim}{X}|\underset{\sim}{y}, \underset{\sim}{\theta})$ at the $\underset{\sim}{\theta}$ at which the likelihood on the left-hand side of equation (3) is being computed.

The EM algorithm is an iterative algorithm that produces a sequence of estimates that converges to the (incomplete data) maximum likelihood estimate. Generically, starting from an initial value $\underset{\sim}{\theta} = \underset{\sim}{\theta}^{(0)}$, the $(r+1)$st value in the sequence, $\underset{\sim}{\theta}^{(r+1)}$, is constructed given the rth value , $\underset{\sim}{\theta}^{(r)}$,

$$\underset{\sim}{\theta}^{(r+1)} = \underset{\underset{\sim}{\theta} \in \Theta}{\operatorname{argmax}}\, \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y},\underset{\sim}{\theta}}}[\log L(\underset{\sim}{\theta}|\underset{\sim}{X}, \underset{\sim}{Y})|\underset{\sim}{y}, \underset{\sim}{\theta}^{(r)}]$$

Two components of this calculation are

- **E-step** : compute the expected conditional log-likelihood
- **M-step** : carry out the maximization of the expectation.

In the traditional notation, we write

$$Q(\utilde{\theta}|\utilde{\theta}') = \mathrm{E}_{f_{\utilde{X}|\utilde{Y},\utilde{\theta}}}[\log L(\utilde{\theta}|\utilde{X},\utilde{Y})|\utilde{y},\utilde{\theta}']$$

We wish to show that the sequence of estimates produced by

$$\utilde{\theta}^{(r+1)} = \underset{\utilde{\theta}\in\Theta}{\text{argmax}}\, Q(\utilde{\theta}|\utilde{\theta}^{(r)}) \qquad r = 1, 2, \ldots$$

converges to the maximum likelihood estimate. First, note that for two pdfs f_1 and f_2 for random variable Z, we have by the usual argument that

$$\begin{aligned} \mathrm{E}_{f_1}[\log f_1(Z)] - \mathrm{E}_{f_1}[\log f_2(Z)] = -\mathrm{E}_{f_1}[\log\{f_2(Z)/f_1(Z)\}] &\geq -\log \mathrm{E}_{f_1}[\{f_2(Z)/f_1(Z)\}] \\ &= -\log \int_{\mathcal{Z}} \{f_2(z)/f_1(z)\} f_1(z)\, dz \\ &= -\log \int_{\mathcal{Z}} f_2(z)\, dz = 0 \end{aligned}$$

$$\therefore \qquad \mathrm{E}_{f_1}[\log f_1(Z)] \geq \mathrm{E}_{f_1}[\log f_2(Z)].$$

with equality if and only if $f_1 \equiv f_2$. Hence, for $\utilde{\theta} \in \Theta$, recalling that

$$K(\utilde{\theta}|\utilde{X},\utilde{Y}) = \frac{L(\utilde{\theta}|\utilde{X},\utilde{Y})}{L(\utilde{\theta}|\utilde{Y})} = f_{\utilde{X}|\utilde{Y},\utilde{\theta}}(\utilde{x}|\utilde{y},\utilde{\theta})$$

is itself a (conditional) pdf for all $\utilde{\theta} \in \Theta$, we have

$$\begin{aligned} Q(\utilde{\theta}|\utilde{\theta}^{(r)}) - \log L(\utilde{\theta}|\utilde{y}) &= \mathrm{E}_{f_{\utilde{X}|\utilde{Y},\utilde{\theta}}}[\log L(\utilde{\theta}|\utilde{X},\utilde{Y})|\utilde{y},\utilde{\theta}^{(r)}] - \log L(\utilde{\theta}|\utilde{y}) \\ &= \mathrm{E}_{f_{\utilde{X}|\utilde{Y},\utilde{\theta}}}\left[\log K(\utilde{\theta}|\utilde{X},\utilde{Y})|\utilde{y},\utilde{\theta}^{(r)}\right] \\ &\leq \mathrm{E}_{f_{\utilde{X}|\utilde{Y},\utilde{\theta}}}\left[\log K(\utilde{\theta}^{(r)}|\utilde{X},\utilde{Y})|\utilde{y},\utilde{\theta}^{(r)}\right] \\ &= Q(\utilde{\theta}^{(r)}|\utilde{\theta}^{(r)}) - \log L(\utilde{\theta}^{(r)}|\utilde{y}). \end{aligned}$$

Thus $\log L(\utilde{\theta}|\utilde{y}) - Q(\utilde{\theta}|\utilde{\theta}^{(r)})$ achieves its **minimum** value when $\utilde{\theta} = \utilde{\theta}^{(r)}$. Now suppose that $\utilde{\theta}^{(r+1)}$ is the value that maximizes $Q(\utilde{\theta}|\utilde{\theta}^{(r)})$ over Θ; we have that

$$\begin{aligned} \log L(\utilde{\theta}^{(r+1)}|\utilde{y}) &\equiv Q(\utilde{\theta}^{(r+1)}|\utilde{\theta}^{(r)}) + \left(\log L(\utilde{\theta}^{(r+1)}|\utilde{y}) - Q(\utilde{\theta}^{(r+1)}|\utilde{\theta}^{(r)})\right) \\ &\geq Q(\utilde{\theta}^{(r)}|\utilde{\theta}^{(r)}) + \left(\log L(\utilde{\theta}^{(r)}|\utilde{y}) - Q(\utilde{\theta}^{(r)}|\utilde{\theta}^{(r)})\right) \\ &= \log L(\utilde{\theta}^{(r)}|\utilde{y}) \end{aligned}$$

and the likelihood attained is **increasing** with the sequence $\utilde{\theta}^{(0)}, \utilde{\theta}^{(1)}, \utilde{\theta}^{(2)}, \ldots$.

EXAMPLE: Finite Mixture Model
Suppose that $Y_1 \ldots, Y_n$ are a random sample from the K component finite mixture model

$$f_{Y|\underset{\sim}{\theta}}(y|\underset{\sim}{\theta}) = \sum_{k=1}^{K} \omega_k f_k(y|\theta_k) \qquad y \in \mathbb{R}$$

where $f_1, \ldots, f_K$ are component densities, and

$$0 < \omega_k < 1 \qquad \sum_{k=1}^{K} \omega_k = 1$$

Estimation of $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_K)^\mathsf{T}$ from the likelihood $L(\underset{\sim}{\theta}|\underset{\sim}{y})$ is in general difficult. However, consider the augmented data $X_1, \ldots, X_n$, where

$$\Pr[X_i = k] = \omega_k \qquad i = 1, \ldots, K$$

are independent random variables so that

$$L(\underset{\sim}{\theta}|\underset{\sim}{X}, \underset{\sim}{Y}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\omega_k f_k(y_i|\theta_k)\}^{I_{\{k\}}(X_i)}$$

and

$$\log L(\underset{\sim}{\theta}|\underset{\sim}{X}, \underset{\sim}{Y}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I_{\{k\}}(X_i) \left(\log \omega_k + \log f_k(y_i|\theta_k)\right).$$

The conditional distribution $f_{X|Y,\underset{\sim}{\theta}}(x|y, \underset{\sim}{\theta})$ is a discrete distribution on the set $\{1, 2, \ldots, K\}$ where for each $i = 1, \ldots, n$

$$\Pr[X_i = k|\underset{\sim}{Y}, \underset{\sim}{\omega}, \underset{\sim}{\theta}] = \frac{\omega_k f_k(y_i|\theta_k)}{\sum\limits_{j=1}^{K} \omega_j f_j(y|\theta_j)} = \varpi_k(y_i, \underset{\sim}{\theta}) \qquad k = 1, \ldots, K$$

where $X_1, \ldots, X_n$ are conditionally independent. Thus

$$E_{f_{X_i|Y_i,\underset{\sim}{\theta},\underset{\sim}{\omega}}}[I_{\{k\}}(X_i)|y_i, \underset{\sim}{\theta}, \underset{\sim}{\omega}] = \varpi_k(y_i, \underset{\sim}{\theta})$$

and hence

$$\begin{aligned}
Q(\underset{\sim}{\theta}, \underset{\sim}{\omega}|\underset{\sim}{\theta}^{(r)}, \underset{\sim}{\omega}^{(r)}) &= E_{f_{\underset{\sim}{X}|Y,\underset{\sim}{\theta},\underset{\sim}{\omega}}}[\log L(\underset{\sim}{\theta}|\underset{\sim}{X}, \underset{\sim}{Y})|\underset{\sim}{y}, \underset{\sim}{\theta}^{(r)}, \underset{\sim}{\omega}^{(r)}] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)}) \left(\log \omega_k + \log f_k(y_i|\theta_k)\right) \\
&= \sum_{k=1}^{K} \left\{ \sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)}) \right\} \log \omega_k + \sum_{k=1}^{K} \sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)}) \log f_k(y_i|\theta_k) \qquad (4)
\end{aligned}$$

We seek to maximize over $(\underset{\sim}{\theta}, \underset{\sim}{\omega})$ to obtain $(\underset{\sim}{\theta}^{(r+1)}, \underset{\sim}{\omega}^{(r+1)})$ presuming that the values $\varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})$ are fixed. From the form of equation (4) it is evident that the function is sum of two parts, the first only depending on $\underset{\sim}{\omega}$, the second only dependent on $\underset{\sim}{\theta}$. We can therefore maximize the two parts separately to obtain $(\underset{\sim}{\theta}^{(r+1)}, \underset{\sim}{\omega}^{(r+1)})$.

The first part of equation (4) is of the form of a multinomial likelihood in $\underset{\sim}{\omega}$, therefore, by previous results, it follows that

$$\omega_k^{(r+1)} = \frac{\sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})}{\sum_{j=1}^{K}\sum_{i=1}^{n} \varpi_j^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})} \qquad k = 1, \ldots, K$$

The second part of equation (4) is the sum of K log-likelihoods for the K mixture components which can be maximized separately

$$\theta_k^{(r+1)} = \underset{\theta_k}{\text{argmax}} \sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)}) \log f_k(y_i|\theta_k) \tag{5}$$

For certain choices of the component densities, this maximization can be carried out analytically. For example, if $f_k(y|\theta_k)$ is the normal density with expectation μ_k and variance σ_k^2, it follows that the new maximizing value equals $\theta_k^{(r+1)} = (\mu_k^{(r+1)}, \sigma_k^{(r+1)})$ where

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)}) y_i}{\sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})}$$

and

$$\sigma_k^{(r+1)} = \sqrt{\frac{\sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})(y_i - \mu_k^{(r+1)})^2}{\sum_{i=1}^{n} \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})}}$$

Note that in the normal model the terms in (5) correspond to likelihood components of the form

$$\{f_k(y_i|\theta_k)\}^{\varpi_k^{(r)}} = \left(\frac{1}{2\pi\sigma_k^2}\right)^{\varpi_k^{(r)}/2} \exp\left\{-\frac{\varpi_k^{(r)}}{2\sigma_k^2}(y_i - \mu_k^{(r)})^2\right\}$$

so the terms $\varpi_k^{(r)} \equiv \varpi_k^{(r)}(y_i, \underset{\sim}{\theta}^{(r)})$ are acting as weighting factors.