Lecture Outline

EM Algorithm for MLE (maximum likelihood estimation)

- **A review of some theory**

- **An illustration involving *missing data***

- **A result showing that EM's convergence is *monotone*, and giving the *rate of convergence* for the EM algorithm in terms of the amount of *missing* information added.**

EM for *MLE* – making a one-step likelihood maximization easier through a (convergent) sequence of simpler maximizations.

Let $X_1, X_2, \ldots X_n$ be *iid* with common density function $\boldsymbol{p}(\boldsymbol{X} \mid \theta)$.

We are looking to maximize the likelihood function:

$$\hat{\theta} \;=\; \text{argmax}_{\Theta}\, \mathbf{L}(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \boldsymbol{p}(\boldsymbol{x} \mid \theta).$$

This may be hard to do as the likelihood function $\mathbf{L}(\theta \mid \boldsymbol{x})$ may be complicated.

Instead, it may be easier to work with a likelihood function augmented by data $\boldsymbol{Z}$

$$\mathbf{L}(\theta \mid \boldsymbol{x}, \boldsymbol{z})$$

to be integrated out at a later stage of computation.

This is feasible when we can write

$$\boldsymbol{p}(\boldsymbol{x} \mid \theta) = \int_{\boldsymbol{Z}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z} \mid \theta)\, d\boldsymbol{z}$$

for some convenient joint density function $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z} \mid \theta)$.

Now by the multiplication theorem for densities:

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z} \mid \theta) = \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta)\, \boldsymbol{p}(\boldsymbol{x} \mid \theta)$$

where $\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta)$ is a conditional density function for $\boldsymbol{Z}$ given $\boldsymbol{X}$ and θ.

It is the convenience of working with the joint density $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z} \mid \theta)$ and the conditional density $\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta)$ that drives *EM* calculations, as

$$\boldsymbol{p}(\boldsymbol{x} \mid \theta) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z} \mid \theta) \;/\; \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta)$$

Thus, quite generally:

(*) $$\log \mathbf{L}(\theta \mid \boldsymbol{x}) = \log \mathbf{L}(\theta \mid \boldsymbol{x}, \boldsymbol{z}) - \log \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta).$$

Following (Dempster, Laird and Rubin,1977), with θ_0 arbitrary, define the two functions:

(**) *E-step* $$\boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0) = \int_{\boldsymbol{Z}} [\log \mathbf{L}(\theta \mid \boldsymbol{x}, \boldsymbol{z})]\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta_0)\, d\boldsymbol{z}$$

and

(***) $$\boldsymbol{H}(\theta \mid \boldsymbol{x}, \theta_0) = \int_{\boldsymbol{Z}} [\log \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta_0)]\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \theta_0)\, d\boldsymbol{z}.$$

Then $$\log \mathbf{L}(\theta \mid \boldsymbol{x}) = \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0) - \boldsymbol{H}(\theta \mid \boldsymbol{x}, \theta_0).$$

Begin the iterative process by letting

M-step $\hat{\theta}_1 = \text{argmax}_{\Theta}\, \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0)$

and then replacing θ_0 with $\hat{\theta}_1$ in (**), which leads to a revised (***) in the light of (*).

Thus, $\hat{\theta}_{j+1} = \text{argmax}_{\Theta}\, \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_j)$.

(DLR) *EM -jargon*: $\mathbf{log\ L}(\theta \mid \boldsymbol{x})$ is the *incomplete* log-likelihood function.

$\mathbf{log\ L}(\theta \mid \boldsymbol{x}, \boldsymbol{z})$ is the complete log-likelihood function.

and $\boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0)$ is the *expected* log-likelihood function.

Theorem: For the sequence $\hat{\theta}_{j+1} = \mathrm{argmax}_{\Theta}\, \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \hat{\theta}_j)$, $j = 1, \ldots$.

$$\mathbf{L}(\hat{\theta}_{j+1} \mid \boldsymbol{x}) \geq \mathbf{L}(\hat{\theta}_j \mid \boldsymbol{x})$$

with equality if and only if $\boldsymbol{Q}(\hat{\theta}_{j+1} \mid \boldsymbol{x}, \hat{\theta}_j) = \boldsymbol{Q}(\hat{\theta}_j \mid \boldsymbol{x}, \hat{\theta}_j)$.

Proof: Recall that $\mathbf{log\ L}(\theta \mid \boldsymbol{x}) = \boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0) - \boldsymbol{H}(\theta \mid \boldsymbol{x}, \theta_0)$.

Then on successive iterations

$\mathbf{log\ L}(\hat{\theta}_{j+1} \mid \boldsymbol{x}) - \mathbf{log\ L}(\hat{\theta}_j \mid \boldsymbol{x}) =$

$$[\boldsymbol{Q}(\hat{\theta}_{j+1} \mid \boldsymbol{x}, \hat{\theta}_j) - \boldsymbol{Q}(\hat{\theta}_j \mid \boldsymbol{x}, \hat{\theta}_j)] - [\boldsymbol{H}(\hat{\theta}_{j+1} \mid \boldsymbol{x}, \hat{\theta}_j) - \boldsymbol{H}(\hat{\theta}_j \mid \boldsymbol{x}, \hat{\theta}_j)].$$

Evidently $[\boldsymbol{Q}(\hat{\theta}_{j+1} \mid \boldsymbol{x}, \hat{\theta}_j) - \boldsymbol{Q}(\hat{\theta}_j \mid \boldsymbol{x}, \hat{\theta}_j)] \geq 0$, by the iteration

Thus, we must show that:

$$\int_{\boldsymbol{Z}} [\mathbf{log}\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1}) - \mathbf{log}\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})]\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})\, d\boldsymbol{z}. \le 0.$$

Or, $$\int_{\boldsymbol{Z}} \mathbf{log}\, [\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1}) / \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})]\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})\, d\boldsymbol{z}. \le 0.$$

Recall, ***K-L*** information is non-negative and 0 only for identical distributions.

$$\mathrm{E}_{\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})}\, \mathbf{log}\, [\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j}) / \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1})] \ge 0.$$

Aside: This follows by Jensen’s Inequality, twice, noting that for positive rv’s
$1/\mathrm{E}[\boldsymbol{X}] < \mathrm{E}[1/X]$ and that $\mathrm{E}[\mathbf{log}\, \boldsymbol{X}] < \mathbf{log}\, \mathrm{E}[X]$.

So,
$$\begin{aligned} 0 &\ge -\mathrm{E}_{\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})}\, \mathbf{log}\, [\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j}) / \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1})] \\ &= \mathrm{E}_{\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})} -\mathbf{log}\, [\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j}) / \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1})] \\ &= \mathrm{E}_{\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})}\, \mathbf{log}\, [\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1}) / \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})] \\ &= \int_{\boldsymbol{Z}} \mathbf{log}\, [\boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j+1}) / \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})]\, \boldsymbol{h}(\boldsymbol{z} \mid \boldsymbol{x}, \hat{\theta}_{j})\, d\boldsymbol{z} \end{aligned}$$

To insure that the sequence $<\hat{\theta}_j>$ converges the following result helps:

Theorem: (Boyles, 1983; Wu, 1983)

If the expected log-likelihood function $\boldsymbol{Q}(\theta \mid \boldsymbol{x}, \theta_0)$ is continuous in both θ and θ_0, then all limit points of an *EM* sequence $<\hat{\theta}_j>$ are *stationary points* of $\mathbf{L}(\theta \mid \boldsymbol{x})$ and $\mathbf{L}(\hat{\theta}_j \mid \boldsymbol{x})$ converges monotonically to $\mathbf{L}(\hat{\theta} \mid \boldsymbol{x})$ for some *stationary point* $\hat{\theta}$.

That is, then $$\left.\frac{\partial \log p(\theta \mid x)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0.$$

EM with *missing-data*.

One-way layout with missing data:

Let X_{ij} denote the response variable of the j^{th} subject among those receiving treatment dose-i.

Statistical model: Assume $X_{ij} \sim \boldsymbol{N}(\mu_i, \sigma^2)$; $i = 1, \ldots, k$; $j = 1, \ldots, n_i$.

The μ_i are the parameters of interest: average effects of a given treatment dose.

Let $\bar{\mu}$ be an average of average dose effects so that: $\mu_i = \bar{\mu} + \alpha_i$, where $\sum_i \alpha_i = 0$.

That is $\bar{\mu} = \sum_i \mu_i / k$ and $\alpha_i = \mu_i - \bar{\mu}$.

Note well the relation to the k-MoG problem!

The least squares estimator of μ_i is (evidently) $\bar{x}_i = (1/n_i) \sum_{j=1}^{n_i} x_{ij}$.

And the minimum variance (unbiased) estimators for the other parameters are:

$$\hat{\mu} = (1/k)\sum_i \bar{x}_i \quad \text{and} \quad \hat{\alpha}_i = \bar{x}_i - \hat{\mu}$$

However, when the sample sizes (n_i) are not all equal, the vectors of the coefficients of the X_{ij} in the $\hat{\alpha}_i$ are not orthogonal to the respective vector of coefficients of $\hat{\mu}$. Thus, $\hat{\mu}$ is not independent of the $\hat{\alpha}_i$.

Suppose we have 4 treatment groups, with outcomes

TREATMENTS

T1	T2	T3	T4
x_{11}	x_{21}	x_{31}	x_{41}
x_{12}	x_{22}	x_{32}	x_{42}
z_1	x_{23}	z_3	x_{43}

Observe X_{ij} and use the ***Z***s as the *dummy* missing values to create a balanced sample.

Thus, $X_{ij} \sim \boldsymbol{N}(\bar{\mu}+\alpha_i, \sigma^2)$ and our dimensional parameter $\theta = (\bar{\mu}, \sigma^2, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$.

The *incomplete* likelihood is:

$$\mathbf{L}(\theta \mid \boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x} \mid \theta) = \sqrt{}\,(1/2\pi\sigma^2)^{10} \exp[\textstyle\sum_{i=1}^{4}\sum_{j=1}^{n_i}(x_{ij} - \bar{\mu} - \alpha_i)^2 / \sigma^2]$$

The *complete* likelihood is:

$$\mathbf{L}(\theta \mid \boldsymbol{x},\boldsymbol{z}) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{z} \mid \theta) = \sqrt{}\,(1/2\pi\sigma^2)^{12} \exp[\textstyle\sum_{i=1}^{4}\sum_{j=1}^{3}(x_{ij} - \bar{\mu} - \alpha_i)^2 / \sigma^2]$$

where, of course, $x_{13} = z_1$ and $x_{33} = z_3$.

Now, run the *EM* algorithm with the augmented data (x,z) and simplified likelihood (based on a balanced sample) in order to find the MLE for $\mathbf{L}(\theta, x)$.