Lecture Stat 461-561
Expectation-Maximization Algorithm

Arnaud Doucet

January 2008

MLE and EM

- The EM consists of considering maximizing with respect to θ the likelihood function

$$L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta) \text{ where } g(\mathbf{y}|\theta) = \int f(\mathbf{x},\mathbf{y}|\theta)\, d\mathbf{x}.$$

- We call
  - $\mathbf{Y}$ the incomplete data (i.e. the observed data) and $(\mathbf{X},\mathbf{Y})$ are the complete data.
- The EM proceeds as

$$\widehat{\theta}^{j+1} = \underset{\theta\in\Theta}{\arg\max} \int \log f(\mathbf{x},\mathbf{y}|\theta) \;.\; f\left(\mathbf{x}|\,\mathbf{y},\widehat{\theta}^{j}\right) d\mathbf{x}$$

Proof of Theorem for Expectation-Maximization Algorithm

- We want to show that $L\left(\widehat{\theta}^{j+1}\middle|\,\mathbf{y}\right) \geq L\left(\widehat{\theta}^{j}\middle|\,\mathbf{y}\right)$ for $\widehat{\theta}^{j+1} = \underset{\theta\in\Theta}{\arg\max}\, Q\left(\theta, \widehat{\theta}^{j}\right)$.
- *Proof*: We have

$$f\left(\mathbf{x}|\,\theta, \mathbf{y}\right) = \frac{f\left(\mathbf{x}, \mathbf{y}|\,\theta\right)}{g\left(\mathbf{y}|\,\theta\right)} \Leftrightarrow g\left(\mathbf{y}|\,\theta\right) = L\left(\theta|\,\mathbf{y}\right) = \frac{f\left(\mathbf{x}, \mathbf{y}|\,\theta\right)}{f\left(\mathbf{x}|\,\theta, \mathbf{y}\right)}$$

thus

$$\log L\left(\theta|\,\mathbf{y}\right) = \log f\left(\mathbf{x}, \mathbf{y}|\,\theta\right) - \log f\left(\mathbf{x}|\,\theta, \mathbf{y}\right)$$

and for any value $\widehat{\theta}^{j}$

$$\log L\left(\theta|\,\mathbf{y}\right) = \underbrace{\int \log f\left(\mathbf{x}, \mathbf{y}|\,\theta\right).f\left(\mathbf{x}|\,\widehat{\theta}^{j}, \mathbf{y}\right)d\mathbf{x}}_{=Q\left(\theta,\widehat{\theta}^{j}\right)} - \int \log f\left(\mathbf{x}|\,\theta, \mathbf{y}\right).f\left(\mathbf{x}|\,\widehat{\theta}^{j}, \mathbf{y}\right)d\mathbf{x}.$$

- Now the EM ensures by construction that $Q\left(\widehat{\theta}^{j+1},\widehat{\theta}^{j}\right) \geq Q\left(\widehat{\theta}^{j},\widehat{\theta}^{j}\right)$. So if we can prove that

$$\int \log f\left(\mathbf{x}|\,\widehat{\theta}^{j+1},\mathbf{y}\right).f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right)d\mathbf{x} \leq \int \log f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right).f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right)d\mathbf{x}$$

  then the Theorem is proved.
- We have

$$\begin{aligned} & \int \log \frac{f\left(\mathbf{x}|\,\widehat{\theta}^{j+1},\mathbf{y}\right)}{f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right)}.f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right)d\mathbf{x} \\ \leq \quad & \log \int \frac{f\left(\mathbf{x}|\,\widehat{\theta}^{j+1},\mathbf{y}\right)}{f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right)}.f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right)d\mathbf{x} = 0 \end{aligned}$$

  where we have use Jensen's inequality as log is concave.
- There are numerous variations of the EM in the literature. We will discuss some later.

Calculation of the Information Matrix via the EM

- The EM does not provide an obvious way to compute the observed information matrix given by

$$-\frac{\partial^2 \log L(\theta|\, \mathbf{y})}{\partial\theta^2}$$

which is an estimate of the inverse covariance matrix of the MLE.

- This quantity is very important in practice and allows us to get some approximate confidence intervals.
- Is it possible to obtain this quantity using EM-type quantities?

- **Missing Information Principle**

$$\underbrace{-\frac{\partial^2 \log L(\theta|\mathbf{y})}{\partial\theta^2}}_{\text{observed info.}} = \underbrace{-\int \frac{\partial^2 \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta^2} f(\mathbf{x}|\theta,\mathbf{y})\,d\mathbf{x}}_{\text{complete info.}}$$
$$-\underbrace{\left(-\int \frac{\partial^2 \log f(\mathbf{x}|\theta,\mathbf{y})}{\partial\theta^2} f(\mathbf{x}|\theta,\mathbf{y})\,d\mathbf{x}\right)}_{\text{missing info}}$$

- *Proof*: It follows straightforwardly from the following identity valid for any $\mathbf{x}$

$$\log L(\theta|\mathbf{y}) = \log f(\mathbf{x},\mathbf{y}|\theta) - \log f(\mathbf{x}|\theta,\mathbf{y}).$$

- The rate of CV of the EM is highly dependent on the 'ratio' observed info/missing info. The more informative the missing variables are, the slowler the convergence of the algorithm.

- **Proposition** : We have

$$
\begin{aligned}
&-\frac{\partial^2 \log L(\theta|\mathbf{y})}{\partial\theta^2} \\
&= -\int \frac{\partial^2 \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta^2} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x} - cov\left(\left.\frac{\partial \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta}\right|\theta,\mathbf{y}\right) \\
&= -\int \frac{\partial^2 \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta^2} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x} - \int \frac{\partial \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta}\frac{\partial \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta}^{\mathsf{T}} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x} \\
&+ \frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta}\frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta}^{\mathsf{T}} \quad \text{(Louis' identity)}
\end{aligned}
$$

as

$$
\frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta} = \int \frac{\partial \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x} \text{ (Fisher's identity)}
$$

Proof. Fisher's identity is trivial. We have

$$-\int \frac{\partial^2 \log f(\mathbf{x}|\theta,\mathbf{y})}{\partial\theta^2} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x} = \int \frac{\partial \log f(\mathbf{x}|\theta,\mathbf{y})}{\partial\theta} \frac{\partial \log f(\mathbf{x}|\theta,\mathbf{y})}{\partial\theta}^{\mathsf{T}} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x}$$

and

$$\frac{\partial \log f(\mathbf{x}|\theta,\mathbf{y})}{\partial\theta} = \frac{\partial \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta} - \frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta}.$$

So the result follows directly by noting that

$$\int \frac{\partial \log f(\mathbf{x},\mathbf{y}|\theta)}{\partial\theta} \frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta}^{\mathsf{T}} f(\mathbf{x}|\theta,\mathbf{y})\, d\mathbf{x} = \frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta} \frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta}^{\mathsf{T}}.$$

- Note that Louis' identity is not very friendly.... and is not a direct byproduct of the EM.

- **Proposition**. We have

$$\frac{\partial^2 \log L(\theta|\,\mathbf{y})}{\partial \theta^2} = \left\{ \frac{\partial^2 Q(\theta', \theta)}{\partial \theta'^2} + \frac{\partial^2 Q(\theta', \theta)}{\partial \theta' \partial \theta} \right\} \Bigg|_{\theta'=\theta}$$

- Proof. We show how it is possible to obtain such a result starting from the key identity, for any θ

$$\log L(\theta'|\,\mathbf{y}) = Q(\theta', \theta) - \int \log f(\mathbf{x}|\,\theta', \mathbf{y}) \,.f(\mathbf{x}|\,\theta, \mathbf{y})\, d\mathbf{x}.$$

Moreover We have

$$\int \frac{\partial \log f(\mathbf{x}|\theta, \mathbf{y})}{\partial \theta} .f(\mathbf{x}|\,\theta, \mathbf{y})\, d\mathbf{x} = 0,$$
$$\int \frac{\partial^2 \log f(\mathbf{x}|\theta, \mathbf{y})}{\partial \theta^2} .f(\mathbf{x}|\,\theta, \mathbf{y})\, d\mathbf{x} = -\int \frac{\partial \log f(\mathbf{x}|\theta, \mathbf{y})}{\partial \theta} \frac{\partial \log f(\mathbf{x}|\theta, \mathbf{y})}{\partial \theta}^{\mathsf{T}} .f(\mathbf{x}|\,\theta, \mathbf{y})\, d\mathbf{x} \quad (1)$$

- It follows that

$$\frac{\partial \log L(\theta'|\,\mathbf{y})}{\partial \theta'} = \frac{\partial Q(\theta', \theta)}{\partial \theta'} - \frac{\partial}{\partial \theta'} \int \log f(\mathbf{x}|\,\theta', \mathbf{y}) \,.f(\mathbf{x}|\,\theta, \mathbf{y})\, d\mathbf{x} \quad (2)$$

thus for $\theta' = \theta$ we have $\frac{\partial \log L(\theta|\mathbf{y})}{\partial \theta} = \left\{ \frac{\partial Q(\theta', \theta)}{\partial \theta'} \right\} \Big|_{\theta'=\theta}$.

- Now differentiating (2) with respect to θ' and θ, we obtain

$$\begin{aligned}
\frac{\partial \log L\left(\theta'\middle|\,\mathbf{y}\right)}{\partial \theta'^2} &= \frac{\partial Q\left(\theta', \theta\right)}{\partial \theta'^2} - \int \frac{\partial \log f\left(\mathbf{x}\middle|\,\theta', \mathbf{y}\right)}{\partial \theta'^2} . f\left(\mathbf{x}|\,\theta, \mathbf{y}\right) d\mathbf{x}, \\
\frac{\partial \log L\left(\theta'\middle|\,\mathbf{y}\right)}{\partial \theta' \partial \theta} &= 0 = \frac{\partial Q\left(\theta', \theta\right)}{\partial \theta' \partial \theta} \\
&\quad - \int \frac{\partial \log f\left(\mathbf{x}\middle|\,\theta', \mathbf{y}\right)}{\partial \theta'} . \frac{\partial \log f\left(\mathbf{x}|\,\theta, \mathbf{y}\right)^{\mathsf{T}}}{\partial \theta} f\left(\mathbf{x}|\,\theta, \mathbf{y}\right) d\mathbf{x}
\end{aligned}$$

- Substituting $\theta = \theta'$, adding the two equations and using (1), we obtain

$$\frac{\partial \log L\left(\theta|\,\mathbf{y}\right)}{\partial \theta^2} = \left\{ \frac{\partial^2 Q\left(\theta', \theta\right)}{\partial \theta'^2} + \frac{\partial^2 Q\left(\theta', \theta\right)}{\partial \theta' \partial \theta} \right\}\Bigg|_{\theta'=\theta} .$$

- This equality is valid at any point θ.

- *Example:* Remember the genetic example where

$$(Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

- The observed log-likelihood function is given by

$$\log L(\theta|\,\mathbf{y}) = cst + y_1 \log(2+\theta) + (y_2 + y_3)\log(1-\theta) + y_4 \log\theta.$$

- So we obtain via a direct calculation

$$\frac{\partial \log L(\theta|\,\mathbf{y})}{\partial\theta^2} = -\frac{y_1}{(2+\theta)^2} - \frac{y_2 + y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}$$

- Introduce the artificial missing data (X_1, X_2) such that $Y_1 = X_1 + X_2$ and define

$$\mathbf{Z} = (X_1, X_2, Y_2, Y_3, Y_4) \sim \mathcal{M}\left(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

Then

$$\log f\left(\mathbf{z}|\theta'\right) = cst + (y_2 + y_3)\log\left(1-\theta'\right) + (x_2 + y_4)\log\theta'$$

and $\mathbb{E}\left(X_2|\, y_1, \theta\right) = y_1 \frac{\theta}{2+\theta}$ so

$$Q\left(\theta', \theta\right) = cst + (y_2 + y_3)\log\left(1-\theta'\right) + \left(y_1 \frac{\theta}{2+\theta} + y_4\right)\log\theta'.$$

- The second derivatives are given by

$$\frac{\partial^2 Q\left(\theta',\theta\right)}{\partial\theta'^2} = -\frac{(y_2+y_3)}{\left(1-\theta'\right)^2} - \frac{\left(y_1\frac{\theta}{2+\theta}+y_4\right)}{\theta'^2},$$
$$\frac{\partial^2 Q\left(\theta',\theta\right)}{\partial\theta\partial\theta'} = \frac{2y_1}{(2+\theta)^2}\frac{1}{\theta'}$$

and we can indeed check that

$$\left.\left\{\frac{\partial^2 Q\left(\theta',\theta\right)}{\partial\theta'^2} + \frac{\partial^2 Q\left(\theta',\theta\right)}{\partial\theta'\partial\theta}\right\}\right|_{\theta'=\theta} = \frac{\partial \log L(\theta|\mathbf{y})}{\partial\theta^2}.$$

The EM as a simple Surrogate Optimization Approach

- The EM approach seems to be closely related to missing data problems... but it can also be seen as a simple surrogate optimization type approach.
- Assume you are interested in maximizing a general function $f(\theta)$ using an iterative algorithm generating an estimates $\widehat{\theta}^{j}$ at iteration j.
- Assume you can build a function $g\left(\theta, \widehat{\theta}^{j}\right)$ such that

$$\begin{aligned} g\left(\theta, \widehat{\theta}^{j}\right) &\leq f(\theta) \text{ for any } \theta, \\ g\left(\widehat{\theta}^{j}, \widehat{\theta}^{j}\right) &= f\left(\widehat{\theta}^{j}\right) \end{aligned}$$

then if $\widehat{\theta}^{j+1} = \text{argmax}\, g\left(\theta, \widehat{\theta}^{j}\right)$ then

$$f\left(\widehat{\theta}^{j+1}\right) \geq f\left(\widehat{\theta}^{j}\right).$$

- The proof is trivial

$$\begin{aligned} & f\left(\widehat{\theta}^{j+1}\right)-f\left(\widehat{\theta}^{j}\right) \\ = \; & f\left(\widehat{\theta}^{j+1}\right)-g\left(\widehat{\theta}^{j},\widehat{\theta}^{j}\right) \\ = \; & f\left(\widehat{\theta}^{j+1}\right)-g\left(\widehat{\theta}^{j+1},\widehat{\theta}^{j}\right)+g\left(\widehat{\theta}^{j+1},\widehat{\theta}^{j}\right)-g\left(\widehat{\theta}^{j},\widehat{\theta}^{j}\right) \\ \geq \; & 0 \end{aligned}$$

  as $f\left(\widehat{\theta}^{j+1}\right) \geq g\left(\widehat{\theta}^{j+1},\widehat{\theta}^{j}\right)$ and $g\left(\widehat{\theta}^{j+1},\widehat{\theta}^{j}\right) \geq g\left(\widehat{\theta}^{j},\widehat{\theta}^{j}\right)$.
- The EM is a special case where

$$\begin{aligned} f(\theta) &= \log L(\theta|\,\mathbf{y}), \\ g\left(\theta,\widehat{\theta}^{j}\right) &= Q\left(\theta,\widehat{\theta}^{j}\right)+\int \log f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right) f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right) d\mathbf{x} \end{aligned}$$

  as

$$\begin{aligned} & \log L(\theta|\,\mathbf{y}) - \int \log \frac{f\left(\mathbf{x}|\widehat{\theta},\mathbf{y}\right)}{f\left(\mathbf{x}|\widehat{\theta}^{j},\mathbf{y}\right)} f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right) d\mathbf{x} \\ & = Q\left(\theta,\widehat{\theta}^{j}\right)+\int \log f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right) f\left(\mathbf{x}|\,\widehat{\theta}^{j},\mathbf{y}\right) d\mathbf{x} \end{aligned}$$

Application: Bradley-Terry

- You have a collection of teams $i = 1, ..., N$
- Each team i plays against the other teams (possibly several times).
- You can only win or lose: no draw.
- We are interesting in ranking the teams.

- We assign to each team i a parameter $\theta_i > 0$.
- We assume that probability that team i beats team j is

$$\frac{\theta_i}{\theta_i + \theta_j}$$

- So assuming that this happens n_{ij} times then the likelihood of $(\theta_1, ..., \theta_k)$ is

$$\prod\nolimits_{i,j;i \neq j} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{n_{ij}}$$

so for any θ_i^k, θ_j^k

$$l(\theta) = \sum\nolimits_{i,j;i \neq j} n_{ij} \left(\log \theta_i - \log (\theta_i + \theta_j) \right).$$

- We use the fact that for any $u, v > 0$

$$\log \frac{v}{u} \leq \frac{v}{u} - 1 \Rightarrow -\log v \geq -\log u - \frac{v-u}{u}$$

so for any $\theta_i^{(k)}, \theta_j^{(k)}$

$$\begin{aligned} l(\theta) &= \sum_{i,j;i\neq j} n_{ij} \left(\log \theta_i - \log\left(\theta_i + \theta_j\right)\right) \\ &\geq \sum_{i,j;i\neq j} n_{ij} \left(\log \theta_i - \log\left(\theta_i^{(k)} + \theta_j^{(k)}\right) - \frac{\left(\theta_i + \theta_j\right) - \left(\theta_i^{(k)} + \right.}{\theta_i^{(k)} + \theta_j^{(k)}} \right. \end{aligned}$$

- Maximizing the rhs, we obtain

$$\theta_i^{(k+1)} = \frac{\sum_{i\neq j} n_{ij}}{\sum_{i\neq j} \left(n_{ij} + n_{ji}\right) / \left(\theta_i^{(k)} + \theta_j^{(k)}\right)}$$

- The key to design this Majorization-Maximization algorithm consists of designing a suitable function $g\left(\theta, \theta'\right)$.
- Several 'recipes' are proposed in Hunter&Lange.
- This class of algorithms has been underused in the literature.