Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

Lecture 15 : The EM algorithm

An expectation-maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. So called 'missing data' problems are widespread in many application areas of statistics.

We will consider the general set up where

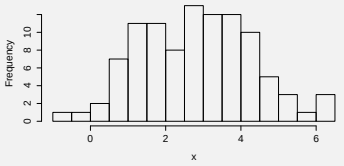
$$\begin{aligned} X &= \text{Observed data} \\ Z &= \text{Missing or unobserved data} \\ \theta &= \text{Model parameters} \\ \ell(\theta) &= \log L(\theta; X) \end{aligned}$$

Example

Suppose we observed data $X_1, \ldots, X_M$ from the following model

$$p(Z_i = 1) = p(Z_i = 2) = \frac{1}{2}, \quad X_i | Z_i \sim N(\theta Z_i, 1)$$

So we observe the X_i's but not the Z_i's or θ and want to estimate θ.



Note If we knew the Z_i's this would be an easy problem, but we don't...hence the need for the EM algorithm.

Overview of EM algorithm

The EM algorithm is an iterative algorithm for maximising the log-likelihood $\ell(\theta)$ that proceeds as follows

The EM algorithm

1. **E-Step** Calculate the expected log-likelihood of the 'full data' likelihood
$$Q(\theta|\theta_t) = \mathbb{E}_{p(Z|X,\theta_t)}\Big[\log P(X,Z|\theta)\Big]$$
where the expectation is wrt $p(Z|X,\theta_t)$ i.e. the conditional distribution of the missing data Z given the observed data X and $\theta_t =$ the current estimate.
2. **M-Step** Set
$\theta_{t+1} = \arg\max_\theta Q(\theta|\theta_t) = \arg\max_\theta \mathbb{E}_{p(Z|X,\theta_t)}\Big[\log P(X,Z|\theta)\Big]$
3. Iterate steps 2 and 3 until convergence.

Question Why does this work?

$$\ell(\theta) = \log P(X|\theta) = \log \frac{P(X,Z|\theta)}{P(Z|X,\theta)} = \log P(X,Z|\theta) - \log P(Z|X,\theta)$$

Take expectations wrt to a distribution $q(Z)$ ($\ell(\theta) = \int \ell(\theta) q(z) dz$)

$$\begin{aligned}
\ell(\theta) &= \int \log P(X,Z|\theta) q(Z) dZ - \int \log P(Z|X,\theta) q(Z) dZ \\
&= \int \log P(X,Z|\theta) q(Z) dZ - \int \log q(Z) q(Z) dz \\
&\quad + \int \log q(Z) q(Z) dz - \int \log P(Z|X,\theta) q(Z) dZ \\
&= \mathbb{E}_{q(Z)}\Big[\log P(X,Z|\theta)\Big] + H(q(Z)) + KL(q(Z)|P(Z|X,\theta)) \\
\ell(\theta) &\geq \mathbb{E}_{q(Z)}\Big[\log P(X,Z|\theta)\Big] + H(q(Z))
\end{aligned}$$

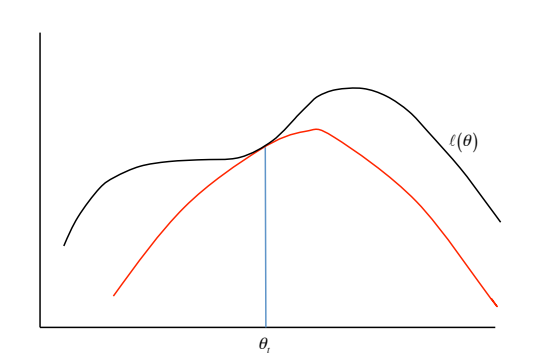
where $H(q(Z)) = -\int \log q(Z) q(Z) dz$ is known as the Entropy of $q(Z)$.

E-Step

Now set $q(Z) = P(Z|X,\theta_t)$. Then we have

$$\ell(\theta) \geq \underbrace{\mathbb{E}_{P(Z|X,\theta_t)}\Big[\log P(X,Z|\theta)\Big]}_{=Q(\theta|\theta_t)} + H(q(Z))$$

with equality at $\theta = \theta_t$ i.e. we have a lower bound for $\ell(\theta)$.

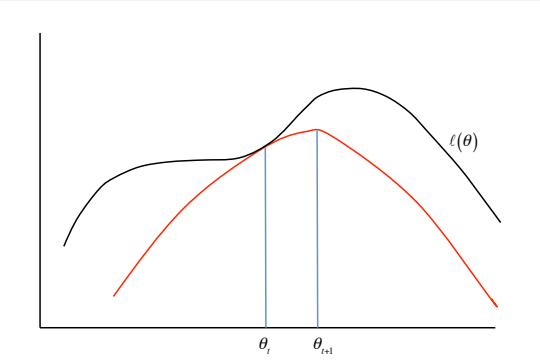


M-Step

We then obtain an updated estimate θ_{t+1} by maximizing this lower bound

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

since $H(q(Z))$ is not a function of θ.



EM algorithm

Lemma

The sequence $\ell(\theta_t)$ is non-decreasing.

Recall that

$$Q(\theta|\theta_t) = \mathbb{E}_{P(Z|X,\theta_t)}\Big[\log P(X,Z|\theta)\Big]$$

and define

$$H(\theta|\theta_t) := \mathbb{E}_{p(Z|X,\theta_t)}\Big[\log p(Z|X,\theta)\Big]$$

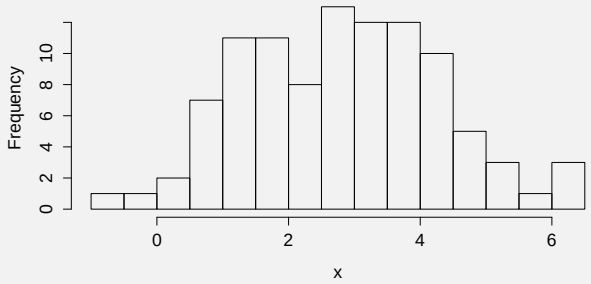
$$\begin{aligned}
\ell(\theta_{t+1}) - \ell(\theta_t) &= Q((\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) - [H(\theta_{t+1}|\theta_t) - H(\theta_t|\theta_t)] \\
&= Q((\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) \\
&\quad + \int \ln\left[\frac{p(Z|X,\theta_t)}{p(Z|X,\theta_{t+1})}\right] p(Z|\theta_t, X)dZ \\
&= Q((\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) + KL[p(Z|X,\theta_{t+1})|p(Z|X,\theta_t)]
\end{aligned}$$

Example

Suppose we observed data $X_1, \ldots, X_M$ from the following model

$$p(Z_i = 1) = p(Z_i = 2) = \frac{1}{2}, \quad X_i|Z_i \sim N(\theta Z_i, 1)$$

So we observe the X_i's but not the Z_i's or θ and want to estimate θ.



The log-likelihood is

$$\ell(\theta) = \log P(X|\theta) = \log \prod_{i=1}^{M} \sum_{k=1}^{2} P(x_i|z_i = k, \theta)p(z_i = k)$$

The full data likelihood and it's logarithm are given by

$$\begin{aligned}
P(X,Z|\theta) &= \prod_{i=1}^{M} P(x_i|z_i,\theta)p(z_i) = \prod_{i=1}^{M} N(x_i|\theta z_i, 1) \cdot \frac{1}{2} \\
\log P(X,Z|\theta) &= -\frac{1}{2}\sum_{i=1}^{M}(x_i - \theta z_i)^2 + K
\end{aligned}$$

E-Step For θ_t we then calculate the conditional distribution of the missing data $P(Z|X,\theta_t)$. This can be done one Z_i at a time.

$$p_{ikt} = P(Z_i = k|X_i,\theta_t) \quad \propto \quad \exp\Big(-\frac{1}{2}(x_i - k\theta_t)^2\Big)$$

Then calculate the expectation of the log of the full data likelihood

$$\begin{aligned}
\mathbb{E}\Big[\log P(X,Z|\theta)\Big] &= -\frac{1}{2}\sum_{i=1}^{M}\sum_{k=1}^{2}(x_i - \theta k)^2 p_{ikt} \\
&= -\frac{1}{2}\sum_{i=1}^{M}(x_i - \theta)^2 p_{i1t} + (x_i - 2\theta)^2 p_{i2t} \\
&= -\frac{1}{2}\theta^2 \sum_i (p_{i1t} + 4p_{i2t}) + \theta \sum_i x_i(p_{i1t} + 2p_{i2t}) + C
\end{aligned}$$

M-Step We can then update the estimate of θ by maximizing the expected log-likelihood,

$$\mathbb{E}\Big[\log P(X, Z|\theta)\Big] \quad = \quad -\frac{1}{2}\theta^2 \sum_i (p_{i1t} + 4p_{i2t}) + \theta \sum_i x_i(p_{i1t} + 2p_{i2t}) + C$$

which leads to

$$\theta_{t+1} = \frac{\sum_i x_i(p_{i1t} + 2p_{i2t})}{\sum_i (p_{i1t} + 4p_{i2t})}$$

Summary EM algorithm

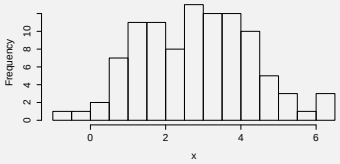
1. Set $t = 0$ and pick a initial value θ_t.
2. **E-Step** To calculate the expected log-likelihood of the 'full data' likelihood we calculate the conditional distribution of the missing data Z given and observed data X and θ_t.

$$p_{ikt} = P(Z_i = k|X_i, \theta_t) \quad \propto \quad \exp\Big(-\frac{1}{2}(x_i - k\theta_t)^2\Big)$$

3. **M-Step** We then maximise the expected log-likelihood using

$$\theta_{t+1} = \frac{\sum_i x_i(p_{i1t} + 2p_{i2t})}{\sum_i (p_{i1t} + 4p_{i2t})}$$

4. Iterate steps 2 and 3 until convergence.



On this real example the EM algorithm leads to the following sequence of estimates

i	θ_i	$\ell(\theta)$
1	1.69	-182.22
2	1.84	-178.90
3	1.90	-178.58
4	1.92	-178.59
5	1.93	-178.62
6	1.94	-178.64
7	1.94	-178.65
8	1.94	-178.66

General comments

1. The EM algorithm is only guaranteed to converge to a local maxima of the likelihood. In some situations there will be multiple-maxima. A computational strategy that is often employed to overcome this involves running the EM algorithm from many different starting points and choosing the estimate that produces the largest likelihood.
2. The convergence of the EM algorithm can sometimes be slow, and this depends upon the form of the likelihood function.

Gene counting

Suppose we take a random sample in pop. and count **phenotypes** (i.e blood types) at some autosomal locus.
Goal: estimate proportion of allelles. (If pop. size is N then $2N$ genes).
If codominant easy:

Phenotype	Genotype	count
M	M/M	119
MN	M/N	76
N	N/N	13

So for p_M estimate $\hat{p}_M = \frac{2\times 119+76}{2\times 208} = .755$
Problem if we have a **recessive** allele.
Example: ABO locus. $n_A = \#$ people of type A, n_{AB}, n_B, n_O

$$n = n_A + n_{AB} + n_B + n_O$$

but don't now how many of n_A are A/A and how many are A/O.

Example cont'd

Note that the number of A/B and O/O genotypes is exactly n_{AB}, n_O.
Let's call $x = (n_{A/A}, n_{A/B}, n_{B/B}, n_{A/O}, n_{B,O}, n_{O/O})$ the complete information (i.e. counts of **genotypes**) $y = (n_A, n_{AB}, n_B, n_O)$ the incomplete information (i.e. counts of **phenotypes**)

$$\begin{aligned}\ln f(x|p) &= n_{A/A}\ln p_A^2 + n_{A/O}\ln(2p_Ap_O) + n_{B/B}\ln p_b^2 \\ &\quad + n_{B/O}\ln(2p_bp_o) + n_{AB}\ln(2p_Ap_B) + n_O\ln p_O^2 \\ &\quad + \ln\binom{n}{n_{A/A}n_{A/O}n_{B/B}n_{B/O}n_{AB}n_O}\end{aligned}$$

E step: Let $p(t)$ be the current estimate. let us write

$$m_{A/A}(t) = \mathbb{E}(n_{A/A}|Y, p(t)) = n_A\frac{p_A(t)^2}{p_A(t)^2 + 2p_A(t)p_O(t)}$$

and in the same way we have

$$m_{A/O}(t) = \mathbb{E}(n_{A/O}|Y, p(t)) = n_A\frac{2p_A(t)p_O(t)}{p_A(t)^2 + 2p_A(t)p_O(t)}$$

E step: Let $p(t)$ be the current estimate. let us write

$$m_{A/A}(t) = \mathbb{E}(n_{A/A}|Y, p(t)) = n_A\frac{p_A(t)^2}{p_A(t)^2 + 2p_A(t)p_O(t)}$$

and in the same way we have

$$m_{A/O}(t) = \mathbb{E}(n_{A/O}|Y, p(t)) = n_A\frac{2p_A(t)p_O(t)}{p_A(t)^2 + 2p_A(t)p_O(t)}$$

and similar expressions for $m_{B/O}(t), m_{B/B}(t)$
M step:
Maximize $Q(p|p(t))$. Can be done introducing Lagrange multiplier and finding stationary point in the unconstrained

$$H(p, \lambda) = Q(p|p(t)) + \lambda(p_A + p_B + p_O - 1)$$

get

- $p_A(t+1) = \frac{2m_{A/A}(t)+m_{A/O}(t)+n_{AB}}{2n}$
- $p_B(t+1) = \frac{2m_{B/B}(t)+m_{B/O}(t)+n_{AB}}{2n}$
- $p_O(t+1) = \frac{m_{A/O}(t)+m_{B/O}(t)+2n_O}{2n}$

The missing information principle

Recall

$$\ln p(\theta|x) = \ln p(\theta, z|x) - \ln p(z|x, \theta)$$

Differentiate twice to get

$$-\frac{\partial^2 \ln p(\theta|x)}{\partial\theta\partial\theta'} = -\frac{\partial^2 \ln p(\theta, z|x)}{\partial\theta\partial\theta'} + \frac{\partial^2 \ln p(z|x, \theta)}{\partial\theta\partial\theta'}$$

LHS is not a function of z. Integrate both side wrt $p(z|x, \theta)$

$$\begin{aligned}-\frac{\partial^2 \ln p(\theta|x)}{\partial\theta\partial\theta'} &= -\int \frac{\partial^2 \ln p(\theta, z|x)}{\partial\theta\partial\theta'}p(z|x, \theta)dz \\ &\quad + \int \frac{\partial^2 \ln p(z|x, \theta)}{\partial\theta\partial\theta'}p(z|x, \theta)dz \\ &= -\frac{\partial^2 Q(\theta|\theta)}{\partial\theta\partial\theta'} - \left[-\frac{\partial^2 H(\theta|\theta)}{\partial\theta\partial\theta'}\right]\end{aligned}$$

The missing information principle

- $-\frac{\partial^2 \ln p(\theta|x)}{\partial\theta\partial\theta'} = I(\theta|x)$ Observed information
- $-\frac{\partial^2 Q(\theta|\theta)}{\partial\theta\partial\theta'} = I_(_c(\theta|x)$ Complete information
- $\left[-\frac{\partial^2 H(\theta|\theta)}{\partial\theta\partial\theta'}\right] = I_m(\theta|x)$ missing information

$$I(\theta|x) = I_c(\theta|x) - I_m(\theta|x)$$