EM algorithm

Guillem Riambau-Armet

Boston University. Fall 2011

October 26, 2011

Some Background

- "Broadly applicable algorithm for computing maximum likelihood estimates from incomplete data" Dempster et al. (1977) in the abstract.
- Used since the 1950s
- First formalized and denoted EM algorithm by Dempster et al. (1977)
- Other interesting articles: Hamilton (1990), Borman (2004), Bilmes (1998).
- Manuals: McLachlan and Krishnan (2008), Frühwirth-Schnatter (2006).

A particular case: finite mixtures

- Widely used for cases of missing data
- In particular, useful for estimation of mixing proportions in cases finite mixture densities.
- In that case, we do not observe from which distribution each observation comes from. The indicator function denoting what distribution it comes from is treated as the missing variable.

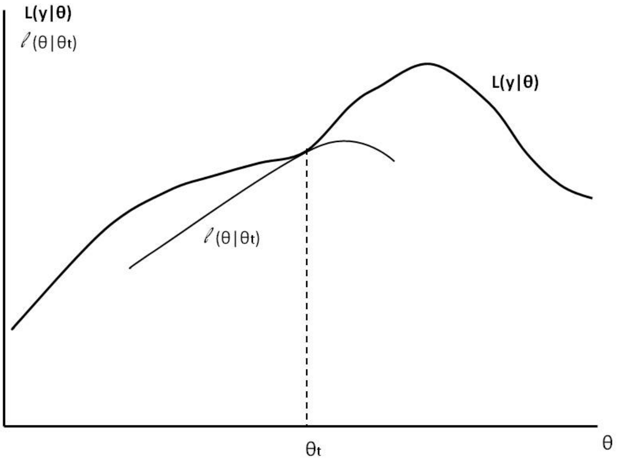
The problem

- y: observed data
- z_d=Unobserved indicator variable. Takes value 1 if the observation comes from distribution d, 0 otherwise.
- For simplicity, suppose the DGP is a mixture of two densities, denoted $f^s(\mu_s)$ and $f^c(\mu_c)$.
- Denote by π the probability that an observation is taken from distribution $f^c(\mu_c)$.
- θ: set of parameters of interest to estimate, $\theta \equiv (\mu_c, \mu_s, \pi)$
- Goal: find $\theta = \text{argmax } p(y|\theta)$.

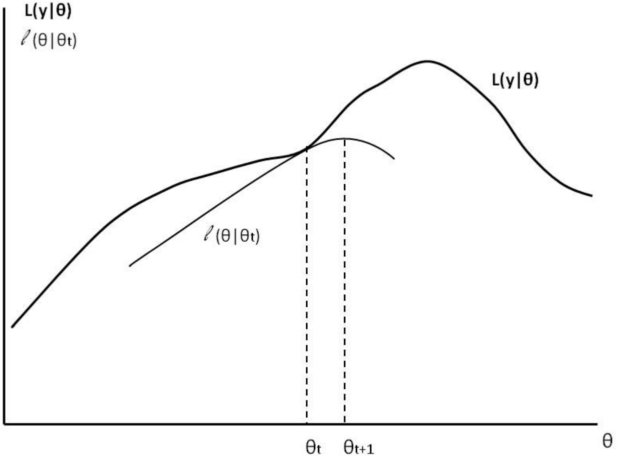
How the EM algorithm works

- $p(y|\theta)$ -alternatively called Likelihood function, $\mathscr{L}(y;\theta)$- very hard to maximise
- Solution: EM algorithm. Iterative procedure.
- E-step: At each iteration, given the current value of the parameters $\hat{\theta}_t$, define a function $\ell(\theta|\hat{\theta}_t)$ which has two properties
    - $\ell(\hat{\theta}_t|\hat{\theta}_t) = \mathscr{L}(y;\hat{\theta}_t)$
    - $\ell(\theta|\hat{\theta}_t)$ is bounded above by $\mathscr{L}(y;\theta)$
- M-step: find the value of θ that maximises $\ell(\theta|\hat{\theta}_t)$. Call it $\hat{\theta}_{t+1}$. By construction of $\ell(\theta|\hat{\theta}_t)$, we have that $\mathscr{L}(y;\hat{\theta}_{t+1}) \geq \mathscr{L}(y;\hat{\theta}_t) \quad \forall t$
- E-step again: construct a new function $\ell(\theta|\hat{\theta}_{t+1})$ and keep iterating until convergence
- Under mild conditions, $Lim_{to\rightarrow\infty}\{\theta_t\} = \theta_{mle}$ (make sure you are finding a global maximum!)

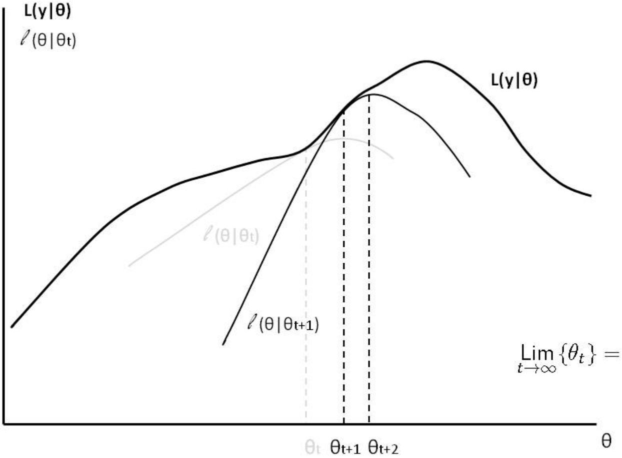
Em algorithm, (Dempster et al., 1977; Hamilton, 1990)



Em algorithm, (Dempster et al., 1977; Hamilton, 1990)



Em algorithm, (Dempster et al., 1977; Hamilton, 1990)



$$\lim_{t\to\infty}\{\theta_t\} = \theta_{mle}$$

Constructing $\ell(\theta|\theta_t)$ [E-step]

- In order to do so I construct the complete data Likelihood function
- I do not observe the z_is $\Rightarrow$ I assume $z_i \sim$ Bernoulli

$$p(z_i;\pi) = \pi^{z_i}(1-\pi)^{1-z_i}$$

- Then, the Complete data Likelihood is

$$p(y,z|\theta) = p(y|z,\theta)p(z|\theta) = \prod_{i=1}^{N} p(y_i|z_i,\theta)p(z_i|\theta)$$

$$= \prod_{i=1}^{N} \underbrace{(f_i^c(\theta))^{z_i}(f_i^s(\theta))^{(1-z_i)}}_{p(y|z,\theta)} \underbrace{\pi^{z_i}(1-\pi)^{(1-z_i)}}_{p(z|\theta)}$$

$$= \prod_{i=1}^{N} (\pi f_i^c(\theta))^{z_i}((1-\pi)f_i^s(\theta))^{(1-z_i)}$$

Constructing $\ell(\theta|\theta_t)$ [E-step]

Complete data log Likelihood:

$$log\,\mathscr{L}(y,z;\theta) = \sum_{i=1}^{N} z_i log\,(\pi) + z_i log\,\left(f_i^{col}(\theta)\right)$$
$$+ (1-z_i) log\,((1-\pi)) + (1-z_i) log\,\left(f_i^{sin}(\theta)\right)$$

[E-STEP] **E for Expectation**

$$Q(\theta|\hat{\theta}_t, y) \equiv E_{p(z)|\hat{\theta}_t, y} log\,\mathscr{L}(y,z|\theta) = \int_{\mathscr{Z}} log\,\left(p(y,z|\hat{\theta})\right) p(z|\hat{\theta}_t, y) dz$$

$$\ell(\theta|\hat{\theta}_t) = Q(\theta|\hat{\theta}_t, y) + \mathbf{C}$$

Importantly, $\theta\,\text{argmax}\,Q(\theta|\hat{\theta}_t, y) = \theta\,\text{argmax}\,\ell(\theta|\hat{\theta}_t)$

[E-step]: In practice

Compute expected value of z_i, conditional on observed data and $\hat{\theta}_t$. Using Bayes,

$$E(z_i;\hat{\theta}_t, y_i) = \frac{\pi f_i^c(\hat{\theta}_t, y_i)}{\pi f_i^c(\hat{\theta}_t, y_i) + (1-\pi) f_i^s(\hat{\theta}_t, y_i)}$$

$$\Rightarrow Q(\theta|\hat{\theta}_t, y) = \sum_{i=1}^{N} E(z_i;\hat{\theta}_t, .) log\left(\pi\right) + E(z_i;\hat{\theta}_t, .) log\left(f_i^l(\theta)\right)$$

$$+ (1 - E(z_i;\hat{\theta}_t, .)) log\left((1-\pi)\right) + (1 - E(z_i;\hat{\theta}_t, .)) log\left(f_i^s(\theta)\right)$$

[M-STEP]

[M-STEP] **M for Maximization** $\Rightarrow$ Find θ *argmax* $Q(\theta|\hat{\theta}_t, y)$.
I.e., the updated $\hat{\theta}_{t+1}$ is characterized by

$$\frac{\partial log \mathscr{L}(y, E(z; \hat{\theta}_t, .); \theta, \hat{\theta}_t)}{\partial \theta} = 0$$

- Note: $log \mathscr{L}(y; \hat{\theta}_{t+1}) \geq log \mathscr{L}(y; \hat{\theta}_t) \quad \forall t$
- Given $\hat{\theta}_{t+1}$, update the expected value of z and keep iterating until convergence.

EM algorithm, Proof based on Hamilton (1990) and Borman (2004)

- Goal: Maximise $p(y|\theta)$ (incomplete or observed data)
- We're instead maximising

$$Q(\theta|\theta_t, y)$$

$$\equiv E_{p(z)|\theta_t,y} log\ p(y,z|\theta) = \sum_{i=1}^{N} E(z_i;\hat{\theta}_t,.) log\,(\pi) + E(z_i;\hat{\theta}_t,.) log\,(f_i^c(\theta))$$

$$+ (1 - E(z_i;\hat{\theta}_t,.)) log\,((1-\pi)) + (1 - E(z_i;\hat{\theta}_t,.)) log\,(f_i^s(\theta))$$

$$= \int_{\mathscr{Z}} log\,(p(y,z|\theta))\, p(z|\theta_t, y) dz$$

- Let $\theta_{t+1} = \operatorname{argmax} Q(\theta|\theta_t, y)$. NTS
  - (i) $p(y|\theta_{t+1}) \geq p(y|\theta_t)$
  - (ii) $Lim_{t\to\infty}\theta_t \to \theta_{MLE}$, where $\theta_{MLE} = \operatorname{argmax} p(y|\theta)$

EM algorithm, Proof of (i): $p(y|\theta_{t+1}) \geq p(y|\theta_t)$

- **(a) NTS** $\boldsymbol{\theta_{t+1}}$**argmax** $\mathbf{Q}(\theta|\theta_t, \mathbf{y}) = \theta_{t+1}$**argmax** $\mathfrak{l}(\theta|\theta_t, \mathbf{y})$
- (b) $\mathfrak{l}(\theta|\theta_t, y) \leq log\ p(y|\theta)$
- (c) $\mathfrak{l}(\theta_t|\theta_t, y) = p(y|\theta_t)$

$$
\begin{aligned}
\text{(a)}\ & argmax_\theta \int_{\mathscr{Z}} log\left[p(y, z|\theta)\right] p(z|\theta_t, y) dz \\
&= argmax_\theta \int_{\mathscr{Z}} log\left[p(y|z, \theta) p(z|\theta)\right] p(z|\theta_t, y) dz \\
&= argmax_\theta \left\{ \int_{\mathscr{Z}} log\left(\frac{p(y|z, \theta) p(z|\theta)}{p(y|\theta_t) p(z|y, \theta_t)} \right) p(z|\theta_t, y) dz + log\ p(y|\theta_t) \right\} \\
&\equiv argmax_\theta \mathfrak{l}(\theta|\theta_t, y)
\end{aligned}
$$

EM algorithm, Proof of (i): $p(y|\theta_{t+1}) \geq p(y|\theta_t)$

- (a) $\theta_{t+1} argmax\ Q(\theta|\theta_t, y) = \theta_{t+1} argmax\ \mathfrak{l}(\theta|\theta_t, y)$
- **(b) NTS** $\mathfrak{l}(\boldsymbol{\theta}|\boldsymbol{\theta_t}, \mathbf{y}) \leq\ \mathbf{p}(\mathbf{y}|\boldsymbol{\theta})$
- (c) $\mathfrak{l}(\theta_t|\theta_t, y) = p(y|\theta_t)$

$$\mathfrak{l}(\theta|\theta_t, y) = log\ p(y|\theta_t) + \int_{\mathscr{Z}} log \left(\frac{p(y|z,\theta)p(z|\theta)}{p(y|\theta_t)p(z|y,\theta_t)} \right) p(z|\theta_t, y)dz$$

$$= log\ p(y|\theta_t) + \int_{\mathscr{Z}} log \left(\frac{p(y|z,\theta)p(z|\theta)}{p(z|y,\theta_t)} \right) p(z|\theta_t, y) - log\ p(y|\theta_t)dz$$

$$\leq [Jensen]\ \ log \int_{\mathscr{Z}} \left(\frac{p(y|z,\theta)p(z|\theta)}{p(z|,\theta_t, y)} \right) p(z|\theta_t, y)dz$$

$$= [rearranging]\ \ log \int_{\mathscr{Z}} \frac{p(z|\theta_t, y)}{p(z|\theta_t, y)} p(y|z,\theta)p(z|\theta)dz = log\ p(y|\theta)$$

EM algorithm, Proof of (i): $p(y|\theta_{t+1}) \geq p(y|\theta_t)$

- (a) $\theta_{t+1} argmax\ Q(\theta|\theta_t, y) = \theta_{t+1} argmax\ \mathfrak{l}(\theta|\theta_t, y)$
- (b)$\mathfrak{l}(\theta|\theta_t, \mathbf{y}) \leq \mathbf{p}(\mathbf{y}|\theta)$
- **(c) NTS** $\mathfrak{l}(\theta_\mathbf{t}|\theta_\mathbf{t}, \mathbf{y}) = \mathbf{p}(\mathbf{y}|\theta_\mathbf{t})$

$$\mathfrak{l}(\theta|\theta_t, y) = log\ p(y|\theta_t) + \int_{\mathscr{Z}} log\left(\frac{p(y|z,\theta)p(z|\theta)}{p(y|\theta_t)p(z|y,\theta_t)}\right) p(z|\theta_t, y)dz$$

$$\Rightarrow \mathfrak{l}(\theta_t|\theta_t, y) = log\ p(y|\theta_t) + \int_{\mathscr{Z}} log\left(\frac{p(y|z,\theta_t)p(z|\theta_t)}{p(y|\theta_t)p(z|y,\theta_t)}\right) p(z|\theta_t, y)dz$$

$$= log\ p(y|\theta_t) + \int_{\mathscr{Z}} log\left(\frac{p(y,z|\theta_t)}{p(y,z|\theta_t)}\right) p(z|\theta_t, y)dz$$

$$= log\ p(y|\theta_t)$$

EM algorithm, Proof of (i)

Recall θ_{t+1} argmax $\mathfrak{l}(\theta|\theta_t, y)$. Hence,

$$\begin{aligned} &\text{if} \qquad \mathfrak{l}(\theta_{t+1}|\theta_t, y) > \mathfrak{l}(\theta_t|\theta_t, y) = log\ p(y|\theta_t) \\ &\text{and} \qquad \mathfrak{l}(\theta|\theta_t, y) \leq log\ p(y|\theta) \quad \forall\theta \\ &\text{it must be that} \\ &log\ p(y|\theta_{t+1}) > log\ p(y|\theta_t) \quad \square \end{aligned}$$

EM algorithm, Proof of (ii) Based on Hamilton, 1990

- NTS $Lim_{t\to\infty}\theta_t \to \theta_{MLE}$ (where $\theta_{MLE} = \text{argmax}\, p(y|\theta)$)
- Note: $\theta\ argmax\ Q(\cdot) = \theta\ argmax\ Q(\cdot)p(y|\theta_t) \equiv Q^*(\cdot)$
- Let θ_{t+1} be the parameter value at which convergence has been achieved. I.e. $\frac{\partial Q^*(y,\theta|\theta_t)}{\partial\theta}\Big|_{\theta=\theta_t} = 0$ (that is, the 'old' value θ_t which maximized $Q^*(y,\theta|\theta_{t-1})$, still maximizes $Q^*(y,\theta|\theta_{t-1})$. Before convergence, $\neq 0$)

$$\frac{\partial Q^*(y,\theta|\theta_t)}{\partial\theta}\Big|_{\theta=\theta_t} = \int_{\mathscr{Z}} \frac{\partial log\,(p(y,z|\theta))}{\partial\theta}\Big|_{\theta=\theta_t} p(y,z|\theta_t)dz$$

$$= \int_{\mathscr{Z}} \frac{\partial p(y,z|\theta)}{\partial\theta}\frac{1}{p(y,z|\theta_t)}\Big|_{\theta=\theta_t} p(y,z|\theta_t)dz = \int_{\mathscr{Z}} \frac{\partial p(y,z|\theta)}{\partial\theta}\Big|_{\theta=\theta_t} dz$$

$$= \frac{\partial p(y|\theta)}{\partial\theta}\Big|_{\theta=\theta_t}$$

Since the LHS is 0, so must be the RHS □

References

- Dempster et al. (1977)
- Hamilton (1990)
- Borman (2004)
- Bilmes (1998)

Frühwirth-Schnatter, Sylvia: *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, New York (2006).

McLachlan, Geoffrey, and Krishnan, Thriyambakam: *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, New York (2008).