Lecture Notes on the Expectation-Maximization Algorithm

John C. Chao

Econ 721 Lecture Notes

November 22, 2022

EM Algorithm

- **A Motivating Example:** Suppose we want to estimate the following two-component mixture model:
  Let
  $$W_i = (B_i, Y_{1,i}, Y_{2,i})' \text{ for } i = 1, ..., n.$$
  Suppose that
  $$\{W_i\} \equiv i.i.d.$$
  and suppose that
  $$\begin{aligned} Y_{1,i} &\sim N(\mu_1, \sigma_1^2), \; Y_{2,i} \sim N(\mu_2, \sigma_2^2), \\ B_i &= \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases} \end{aligned}$$
  Moreover, let
  $$Y_i = (1 - B_i) Y_{1,i} + B_i Y_{2,i}$$
  and assume that B_i, $Y_{1,i}$, and $Y_{2,i}$ are mutually independent and that we only observe Y_i and not B_i, $Y_{1,i}$, and $Y_{2,i}$ separately.

EM Algorithm

- **Remarks:**
  1. We can think of the underlying generating mechanism is one where a Bernoulli random variable $B_i \in \{0, 1\}$ is first generated with probability p; and, then, depending on the outcome delivers either $Y_{1,i}$ or $Y_{2,i}$.
  2. Many economic/econometric models come in the form of a mixture (e.g. models of regime switching, unobserved heterogeneity, etc.)
- **Probability Density Function:**
  The pdf of Y_i is given by

$$g_Y(y) = (1-p)\,\phi_{\theta_1}(y) + p\phi_{\theta_2}(y),$$

where for $s = 1, 2$

$$\begin{aligned} \theta_s &= \left(\mu_s, \sigma_s^2\right)', \\ \phi_{\theta_s}(y) &= \frac{1}{\sigma_s\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_s^2}(y-\mu_s)^2\right\}. \end{aligned}$$

EM Algorithm

- **Additional Notations:** Further define

$$\theta = \begin{pmatrix} p \\ \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} p \\ \mu_1 \\ \sigma_1^2 \\ \mu_2 \\ \sigma_2^2 \end{pmatrix} \text{ and } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

- **Log-likelihood Function**: Given the notations developed above, the log-likelihood function for this model can be written as

$$\ell(\theta, Y) = \sum_{i=1}^{n} \ln \left\{ (1-p)\,\phi_{\theta_1}(y) + p\phi_{\theta_2}(y) \right\}.$$

Note that this log-likelihood function will be difficult to maximize (say, by Newton's method) because of the sum of terms within the log function.

EM Algorithm

- **The Case Where B_i Is Observed:** To understand the idea behind the expectation-maximization algorithm, consider the case where we could observe the values of B_i (the complete data case). Then, the problem would be a lot easier, since if $B_i = 1$; then, Y_i comes from model 2; otherwise, it comes from model 1. Hence, if the values of B_i are observable, then the probability density function of the data would take the form

$$\begin{aligned} & f\left(B_i, Y_i \middle| \theta\right) \\ = \; & f\left(Y_i \middle| B_i, \theta_1, \theta_2\right) f\left(B_i \middle| p\right) \\ = \; & \left\{\left[\phi_{\theta_1}\left(y_i\right)\right]^{(1-B_i)}\left[\phi_{\theta_2}\left(y_i\right)\right]^{B_i}\right\}\left\{\left[1-p\right]^{(1-B_i)} p^{B_i}\right\} \end{aligned}$$

EM Algorithm

- Hence, in this case, the likelihood function and the log-likelihood function can be written down, respectively, as

$$
\begin{aligned}
L_0(\theta, B, Y) &= \prod_{i=1}^{n} \left[\phi_{\theta_1}(y_i)\right]^{(1-B_i)} \left[\phi_{\theta_2}(y_i)\right]^{B_i} [1-p]^{(1-B_i)} p^{B_i}, \\
\ell_0(\theta, B, Y) &= \sum_{i=1}^{n} \left\{(1-B_i) \ln \phi_{\theta_1}(y_i) + B_i \ln \phi_{\theta_2}(y_i)\right\} \\
&\quad + \sum_{i=1}^{n} \left\{(1-B_i) \ln(1-p) + B_i \ln p\right\},
\end{aligned}
$$

where $B = (B_1, ..., B_n)'$.

EM Algorithm

- Since in reality the values of B_i are typically unknown, the EM algorithm proceeds in an iterative manner, substituting for each B_i with its expected value

$$\begin{aligned}
\gamma_i(\theta) &= E[B_i|\theta, Y] \\
&= E[B_i|\theta, Y_i] \quad \text{(by independence)} \\
&= \Pr(B_i = 1|\theta, Y_i) \\
&= \frac{f(B_i = 1 \cap Y_i|\theta)}{f(Y_i|\theta)} \\
&= \frac{f(B_i = 1|p)\, f(Y_i|B_i = 1,\ \theta)}{f(Y_i|\theta)} \\
&= \frac{p\phi_{\theta_2}(y_i)}{(1-p)\,\phi_{\theta_1}(y_i) + p\phi_{\theta_2}(y_i)}.
\end{aligned}$$

- **Remark:** The quantity $\gamma_i(\theta) = E[B_i|\theta, Y]$ is often called the **responsibility** of model 2 for observation i.

EM Algorithm

- In addition, taking conditional expectation of the log-likelihood function, we get

$$
\begin{aligned}
& E\left[\ell_0\left(\theta, B, Y\right)|\widehat{\theta}, Y\right] \\
= & \sum_{i=1}^{n}\left\{\left(1-E\left[B_i|\widehat{\theta}, Y_i\right]\right)\ln\phi_{\theta_1}\left(y_i\right)+E\left[B_i|\widehat{\theta}, Y_i\right]\ln\phi_{\theta_2}\left(y_i\right)\right\} \\
& +\sum_{i=1}^{n}\left\{\left(1-E\left[B_i|\widehat{\theta}, Y_i\right]\right)\ln\left(1-p\right)+E\left[B_i|\widehat{\theta}, Y_i\right]\ln p\right\} \\
= & \sum_{i=1}^{n}\left\{\left(1-\gamma_i\left(\widehat{\theta}\right)\right)\ln\phi_{\theta_1}\left(y_i\right)+\gamma_i\left(\widehat{\theta}\right)\ln\phi_{\theta_2}\left(y_i\right)\right\} \\
& +\sum_{i=1}^{n}\left\{\left(1-\gamma_i\left(\widehat{\theta}\right)\right)\ln\left(1-p\right)+\gamma_i\left(\widehat{\theta}\right)\ln p\right\}
\end{aligned}
$$

EM Algorithm

- **Remark:** The EM algorithm iterates back and forth between an expectation step and a maximization step. Under the expectation step, we do a soft assignment of each observation to each model, i.e., the current estimates of the parameters are used to assign responsibilities according to the relative density (under each model) of the sample points. Under the maximization step these responsibities are used to construct a weighted log-likelihood, which we then maximize to update our estimates of the parameters.

EM Algorithm

- More precisely, the EM algorithm goes as follows:
  Step 1: Take initial estimates of the parameters $\widehat{\mu}_{1,0}, \widehat{\sigma}^2_{1,0}, \widehat{\mu}_{2,0}, \widehat{\sigma}^2_{2,0}, \widehat{p}_0$ (to be specified below).
  Step 2: **(**Expectation Step) In the k^{th} step, compute the responsibilities

$$\widehat{\gamma}_{i,k} = \frac{\widehat{p}_{k-1}\phi_{\widehat{\theta}_{2,k-1}}(y_i)}{(1-\widehat{p}_{k-1})\,\phi_{\widehat{\theta}_{1,k-1}}(y_i) + \widehat{p}_{k-1}\phi_{\widehat{\theta}_{2,k-1}}(y_i)} \text{ for } i = 1, ..., n;$$

where $\widehat{\theta}_{1,k-1} = \left(\widehat{\mu}_{1,k-1}, \widehat{\sigma}^2_{1,k-1}\right)'$ and $\widehat{\theta}_{2,k-1} = \left(\widehat{\mu}_{2,k-1}, \widehat{\sigma}^2_{2,k-1}\right)$.

EM Algorithm

- **Step 3:** (Maximization Step) Compute the weighted means and variances for the $(k+1)^{th}$ step as

$$\widehat{\mu}_{1,k} = \frac{\sum_{i=1}^{n}\left(1-\widehat{\gamma}_{i,k}\right)y_i}{\sum_{i=1}^{n}\left(1-\widehat{\gamma}_{i,k}\right)}, \widehat{\sigma}^2_{1,k} = \frac{\sum_{i=1}^{n}\left(1-\widehat{\gamma}_{i,k}\right)\left(y_i-\widehat{\mu}_{1,k}\right)^2}{\sum_{i=1}^{n}\left(1-\widehat{\gamma}_{i,k}\right)}$$

$$\widehat{\mu}_{2,k} = \frac{\sum_{i=1}^{n}\widehat{\gamma}_{i,k}y_i}{\sum_{i=1}^{n}\widehat{\gamma}_{i,k}}, \qquad \widehat{\sigma}^2_{2,k} = \frac{\sum_{i=1}^{n}\left(1-\widehat{\gamma}_{i,k}\right)\left(y_i-\widehat{\mu}_{2,k}\right)^2}{\sum_{i=1}^{n}\left(1-\widehat{\gamma}_{i,k}\right)}$$

Also, compute the mixing probability as

$$\widehat{p}_{k+1} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\gamma}_{i,k}.$$

Step 4: Iterate steps 2 and 3 until convergence.

EM Algorithm

- **Remark:** Initial estimates $\widehat{\mu}_{1,0}$ and $\widehat{\mu}_{2,0}$ could be made by simply choosing two of the y_i's. $\widehat{\sigma}^2_{1,0}$ and $\widehat{\sigma}^2_{2,0}$ could both be set equal to the overall sample variance

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

and the initial mixing proportion $\widehat{p}_0$ can be set to 0.5.

EM Algorithm - General Formulation

- More generally, EM algorithms are useful for problems for which direct maximization of the likelihood function is difficult, but could be made easier by enlarging the sample with latent (unobserved) data. Suppose we have

$$
\begin{aligned}
Y &- \text{ observed data,} \\
\theta &- \text{ parameter vector,} \\
\ell(\theta, Y) &\text{ - log-likelihood associated with observed data}
\end{aligned}
$$

and

$$
\begin{aligned}
& & Z &- \text{ latent or missing data,} \\
W &= & (Y, Z) &- \text{ complete data} \\
& & \ell_0(\theta; W) &- \text{log-likelihood associated with complete data}
\end{aligned}
$$

EM Algorithm - General Formulation

- The more general EM algorithm goes as follows:
- **Step 1:** Take initial estimate of the parameter vector $\widehat{\theta}^{(0)}$
  Step 2: (Expectation Step) In the k^{th} step, compute the responsibilities

$$Q\left(\theta', \widehat{\theta}^{(k-1)}\right) = E\left[\ell_0\left(\theta', W\right) | Y, \widehat{\theta}^{(k-1)}\right].$$

  as a function of the dummy argument θ'.

  Step 3: (Maximization Step) Determine $\widehat{\theta}^{(k)}$ as

$$\widehat{\theta}^{(k)} = \arg\max_{\theta'} Q\left(\theta', \widehat{\theta}^{(k-1)}\right)$$

  Step 4: Iterate steps 2 and 3 until convergence.

EM Algorithm - General Formulation

- **Remark:** To see why the EM algorithm works, note that

$$f\left(Z|Y,\theta'\right)=\frac{f\left(Z,Y|\theta'\right)}{f\left(Y|\theta'\right)}=\frac{f\left(W|\theta'\right)}{f\left(Y|\theta'\right)},$$

so that

$$f\left(Y|\theta'\right)=\frac{f\left(W|\theta'\right)}{f\left(Z|Y,\theta'\right)}$$

In terms of log-likelihoods, this implies that

$$\begin{aligned}\ell\left(\theta';Y\right) &= \ell_0\left(\theta';W\right)-\ln f\left(Z|Y,\theta'\right)\\ &= \ell_0\left(\theta';W\right)-\ell_1\left(Y,\theta'|Z\right)\end{aligned}$$

Taking the conditional expectation with respect to the distribution of $W|Y$ evaluated at the parameter value $\widehat{\theta}$, we get

$$\begin{aligned}\ell\left(\theta';Y\right) &= E\left[\ell\left(\theta';Y\right)|Y,\widehat{\theta}\right]\\ &= E\left[\ell_0\left(\theta';W\right)|Y,\widehat{\theta}\right]-E\left[\ell_1\left(Y,\theta'|Z\right)|Y,\widehat{\theta}\right].\end{aligned}$$

EM Algorithm - General Formulation

- or

$$\ell\left(\theta';Y\right)=Q\left(\theta',\widehat{\theta}\right)-R\left(\theta',\widehat{\theta}\right).$$

- **Claim:** $R\left(\theta',\widehat{\theta}\right)$ is maximized at $R\left(\widehat{\theta},\widehat{\theta}\right)$.
- **Proof of Claim:** Consider

$$\begin{aligned}
& R\left(\theta',\widehat{\theta}\right)-R\left(\widehat{\theta},\widehat{\theta}\right) \\
= \;& E\left[\ell_1\left(Y,\theta'|Z\right)|Y,\widehat{\theta}\right]-E\left[\ell_1\left(Y,\widehat{\theta}|Z\right)|Y,\widehat{\theta}\right] \\
= \;& E\left(\ln\left[\frac{f\left(Z|Y,\theta'\right)}{f\left(Z|Y,\widehat{\theta}\right)}\right]|Y,\widehat{\theta}\right) \\
\leq \;& \ln\left(E\left[\frac{f\left(Z|Y,\theta'\right)}{f\left(Z|Y,\widehat{\theta}\right)}\right]|Y,\widehat{\theta}\right) \\
= \;& \ln\int f\left(Z|Y,\theta'\right)dZ=\ln 1=0.
\end{aligned}$$

EM Algorithm - General Formulation

- Using this result, it follows that if $\widehat{\theta}^{(k)}$ maximizes $Q\left(\theta', \widehat{\theta}^{(k-1)}\right)$; then,

$$
\begin{aligned}
& \ell\left(\theta'; Y\right)\Big|_{\theta'=\widehat{\theta}^{(k)}} - \ell\left(\theta'; Y\right)\Big|_{\theta'=\widehat{\theta}^{(k-1)}} \\
= \; & E\left[\ell\left(\theta'; Y\right) | Y, \widehat{\theta}^{(k-1)}\right]\Big|_{\theta'=\widehat{\theta}^{(k)}} - E\left[\ell\left(\theta'; Y\right) | Y, \widehat{\theta}^{(k-1)}\right]\Big|_{\theta'=\widehat{\theta}^{(k-1)}} \\
= \; & Q\left(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\right) - Q\left(\widehat{\theta}^{(k-1)}, \widehat{\theta}^{(k-1)}\right) \\
& - \left[R\left(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\right) - R\left(\widehat{\theta}^{(k-1)}, \widehat{\theta}^{(k-1)}\right)\right] \\
\geq \; & 0.
\end{aligned}
$$