The Expectation Maximization (EM) Algorithm

gustavoc@msu.edu

The Expectation-Maximization (EM) algorithm is often used to obtained maximum likelihood (ML) estimates in cases where the likelihood function does not have a closed form. The algorithm can also be used for estimation of the posterior mode in Bayesian models; however, we will be focused on ML estimation. The algorithm was formalized by Dempster, Laird and Rubin (1977)– earlier versions appeared in Baum, Petrie Soules and Weiss (1970). The EM algorithm formalizes ideas earlier used to deal with missing values.

The ***likelihood function*** is the joint distribution of the data $(\boldsymbol{y}_o)$, given the parameters $(\boldsymbol{\theta})$, viewed as a function of the model parameters. In many cases the likelihood function has a closed form; therefore, we can obtain ML estimates using standard analytical or numerical approaches such as the ones discussed in the first half of the course. In other cases, the likelihood of the observed data may not have a closed form–it exactly in these cases where the EM-algorithm becomes useful. The nature of the 'unobserved' random variables (or missing data) may include random effects, censored data points or latent variables (e.g., hidden states in a hidden Markov model).

The ***complete-data likelihood*** is the joint distribution of the observed and the missing data given the parameters, viewed as a function of the parameters, that is, $L_C(\boldsymbol{\theta}) = p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\theta})$. Since $\boldsymbol{y}_m$ is unobserved, we cannot evaluate the complete likelihood directly. However, as we will see, this object will become useful in one of the steps of the EM-algorithm.

The ***(marginal) likelihood*** is the marginal distribution of the observed data given the parameters, viewed as a function of the parameters, that is

$$\widehat{\boldsymbol{\theta}}\overset{=}{_{argmax}}\quad L(\boldsymbol{\theta};\boldsymbol{y}_o) = p(\boldsymbol{y}_o|\boldsymbol{\theta}) = \int \quad p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\theta})d\,\boldsymbol{y}_m$$

As noted, in many cases, the integral in the right-hand side does not have closed form.

The EM-algorithm is an iterative procedure; at each iteration, we complete two steps: The E (or 'expectation') step, followed by the M (or maximization) step.

Outline of the EM-algorithm

- **Initialize**: set $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ where $\boldsymbol{\theta}^{(0)}$ is a value within the parameter space,
- **Iterate** until convergence the following steps:
  - **E-step**: find $Q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{y}_o) = E_{\boldsymbol{y}_m|\boldsymbol{y}_o,\boldsymbol{\theta}_{t-1}}\{\log(\boldsymbol{y}_o, \boldsymbol{y}_m^t|\boldsymbol{\theta}_{t-1}\ )\}$.
  - **M-step**: maximize $Q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{y}_o)$ with respect to $\boldsymbol{\theta}_t$ and set

$$\widehat{\boldsymbol{\theta}}^{(t)}\overset{=}{_{argmax}}\quad Q(\boldsymbol{\theta}_t)$$

The expression involved in the E-step, is the expected value of the complete-data log-likelihood given the observed data and the current values of the parameters. The expectation is taken with respect to the conditional distribution of the missing data; the missing data is 'integrated out', leading a function, $Q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{y}_o)$, which only depend on the parameters and the observed data. In the M-step we maximize this function with respect to the parameters.

In many cases (notably when the distribution of the data belongs to the exponential family) the E-step reduces to 'impute' the missing data from its conditional expectation. In the next section we discuss a few implementations of the EM-algorithm in cases where the data follows a distribution from the exponential family.

Example 1: Maximum likelihood estimation with right-censored data

Suppose that $y_i \sim Exponential(\lambda)$ is a time-to-event variable following a exponential distribution. We have a sample consisting of n pairs of the form (y_i, d_i) where y_i is time to event or time to censoring and d_i is a dummy variable indicating whether y_i is time-to-event $(d_i = 1)$ or time to censoring $(d_i = 0)$. Technically, the time-to-event variable is missing for all the data points with $d_i = 0$. Ignoring the right-censored data will lead to bias missing data points are not 'uniformative': we know that for right-censored points the un-observed time-to-event $(\tilde{y}_i)$ variable is greater than the censoring time, that is $\tilde{y}_i > y_i$. The following scripts simulates right-censored exponential data.

```
 set.seed(195021)
 n=100
 y=rexp(n=n,rate=4)
 # let's consider fixed censoring time
 d=y<0.3 # TRUE here indicate event and FALSE right-censored

 yCen=y; yCen[!d]=0.3 # this is the data we observe
```

The mean of an exponential random variable with rate λ is $E[y_i] = \frac{1}{\lambda}$, in this case 1/4. Ignoring the right-censored data leads to serious downward-bias.

```

# estimate with the 'complete' data
   mean(y)
[1] 0.2193996

# estimate ignoring censoring
  mean(yCen[d])
[1] 0.08742043
```

The exponential distribution has the following (memoryless) property $E[y_i|\tau] = \tau + 1/\lambda$. Furthermore, for a complete-data likelihood the maximum likelihood estimate of the rate can be shown to be: $\hat{\lambda} = \frac{1}{\bar{y}}$. We use these two results in the EM-algorithm below

```
lambda=rep(NA,10) # a vector to store estimates iterations
lambda[1]=1/mean(y[d]) # initial value (estimate ignoring censoring)
completeData=y  # this vector stores the 'complete' data
for(i in 2:length(lambda)){
   # E-step
   completeData[!d]=y[!d]+1/lambda[i-1]
   # M-step
   lambda[i]=1/mean(completeData)
}
round(1/lambda,3)

[1] 0.087 0.247 0.299 0.315 0.320 0.322 0.322 0.323 0.323 0.323
```

[See in-class assignment]

Example 2: Mixed Models

Consider a **linear mixed effects model** of the form

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{b} + u_i + \varepsilon_{ij}$$

where $\boldsymbol{x}_{ij}$ is a vector of covariates (including the constant '1' for the intercept) $\boldsymbol{b}$ is a vector of effects, u_i (*i=1,…,q*) is the effect of the *ith* level of the grouping factor and ε_{ij} (j=1,….,n$_i$) are model residuals. The standard mixed effects model assumes $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.

Stacking all the data equations we have:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad [1]$$

where $\boldsymbol{y} = \{y_{ij}\}$, $\boldsymbol{u} = (u_1, \dots, u_q)'$ and $\boldsymbol{\varepsilon} = \{\varepsilon_{ij}\}$ are the response, random effects, and error vectors, respectively, and $\boldsymbol{X}$ is an incidence matrix whose rows are $\boldsymbol{x}'_{ij}$.

Since the random effects and the errors are normal, the joint distribution of the data and the random effects is multivariate normal:

$$\begin{matrix}\boldsymbol{y}\\ \boldsymbol{u}\end{matrix}|\boldsymbol{b},\sigma_\varepsilon^2,\sigma_u^2\sim MVN\left\{\begin{matrix}\boldsymbol{Xb}\\ \boldsymbol{0}\end{matrix}\ ,\quad \begin{matrix}\boldsymbol{ZZ'}\sigma_u^2+\boldsymbol{I}\sigma_\varepsilon^2 & \boldsymbol{Z}\sigma_u^2\\ \boldsymbol{Z'}\sigma_u^2 & \boldsymbol{I}\sigma_u^2\end{matrix}\right\}\quad [2]$$

Next, we describe the M and E-steps that will form a EM-algorithm to estimate the model parameters $\boldsymbol{\theta}=\{\boldsymbol{b},\sigma_u^2,\sigma_\varepsilon^2\}$. In the formulation we will regard the random effects as the 'missing data'.

M-step

The **complete data likelihood** is as follows

$$\begin{aligned}p(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{b},\sigma_u^2,\sigma_\varepsilon^2)&=p(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{b},\sigma_\varepsilon^2)p(\boldsymbol{u}|\sigma_u^2)\\&=\left\{\prod_{i=1}^{q}\prod_{j=1}^{n_i}(\sigma_\varepsilon^2)^{-\frac{1}{2}}e^{-\frac{\left(y_{ij}-\boldsymbol{x}'_{ij}\boldsymbol{b}-u_i\right)^2}{2\sigma_\varepsilon^2}}\right\}\times(\sigma_u^2)^{-q/2}e^{-\frac{\sum_{i=1}^{q}u_i^2}{2\sigma_u^2}}\end{aligned}$$

In the complete-data likelihood we regard both ***y*** and ***u*** as observed data, thus, we can re-write the above expression as follows

$$L_C(\boldsymbol{b},\sigma_u^2,\sigma_\varepsilon^2)=p(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{b},\sigma_u^2,\sigma_\varepsilon^2)=(\sigma_\varepsilon^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{ij}\left(\tilde{y}_{ij}-\boldsymbol{x}'_{ij}\boldsymbol{b}\right)^2}\times(\sigma_u^2)^{-q/2}e^{-\frac{\sum_{i=1}^{q}u_i^2}{2\sigma_u^2}}$$

where $\tilde{y}_{ij}=y_{ij}-u_i$. The logarithm of the complete data likelihood is

$$l_C(\boldsymbol{b},\sigma_u^2,\sigma_\varepsilon^2)=-\frac{n}{2}log(\sigma_\varepsilon^2)-\frac{1}{2\sigma_\varepsilon^2}\sum_{ij}\left(\tilde{y}_{ij}-\boldsymbol{x}'_{ij}\boldsymbol{b}\right)^2-\frac{q}{2}\log(\sigma_u^2)-\frac{1}{2\sigma_u^2}\sum_{i=1}^{q}u_i^2\quad[3]$$

Variance of the random effect: Differentiating [3] with respect to σ_u^2 leads

$$\frac{\partial l_C(\boldsymbol{b},\sigma_u^2,\sigma_\varepsilon^2)}{\partial\sigma_u^2}=-\frac{1}{2}\left(\frac{q}{\sigma_u^2}-\frac{1}{{\sigma_u^2}^2}\sum_{i=1}^{q}u_i^2\right)$$

Thus,

$$\begin{aligned}\frac{q}{\hat{\sigma}_u^2}&=\frac{1}{{\hat{\sigma}_u^2}^2}\sum_{i=1}^{q}u_i^2\\\Leftrightarrow q&=\frac{1}{\hat{\sigma}_u^2}\sum_{i=1}^{q}u_i^2\\\Leftrightarrow\hat{\sigma}_u^2&=\frac{\sum_{i=1}^{q}u_i^2}{q}\end{aligned}$$

Fixed effects: Inspection of the complete-data log-likelihood shows that the value of the fixed effects ($\boldsymbol{b}$) that maximizes [3] is the least-square estimate derived using $\tilde{y}_{ij}$ as data, that is

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widetilde{\boldsymbol{y}}$$

where $\boldsymbol{X}$ is a matrix whose rows are $\boldsymbol{x}'_{ij}$ and $\widetilde{\boldsymbol{y}}$ is a vector whose entries are $\tilde{y}_{ij} = y_{ij} - u_i$.

Error variance: Finally, differentiation [3] with respect to the error variance we get

$$\frac{\partial l_C\left(\boldsymbol{b}, \sigma_u^2, \sigma_\varepsilon^2\right)}{\partial \sigma_\varepsilon^2} = -\frac{1}{2}\left\{\frac{n}{\sigma_\varepsilon^2} - \frac{1}{\left(\sigma_\varepsilon^2\right)^2} RSS(\boldsymbol{b}, \widetilde{\boldsymbol{y}})\right\}$$

Therefore,

$$\hat{\sigma}_\varepsilon^2 = \left\{\frac{RSS(\boldsymbol{b}, \widetilde{\boldsymbol{y}})}{n}\right\}$$

Where $RSS(\boldsymbol{b}, \widetilde{\boldsymbol{y}}) = \sum_{ij}\left(\tilde{y}_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{b}\right)^2$.

Collecting the above results, we have that the M-step can be carried out using:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{q} u_i^2}{q} \quad ; \quad \hat{\sigma}_\varepsilon^2 = \left\{\frac{RSS(\boldsymbol{b}, \widetilde{\boldsymbol{y}})}{n}\right\} \quad ; \quad \widehat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widetilde{\boldsymbol{y}} \qquad [4]$$

E-Step

The normal distribution belongs to the exponential family; therefore, the E-step can be carried out by imputing the missing data using the conditional expectation of the missing data given the observed data and the parameters. Using the properties of the multivariate normal density [2], we have that

$$\begin{aligned} E(\boldsymbol{u} | \boldsymbol{y}, \boldsymbol{b}, \sigma_u^2, \sigma_\varepsilon^2) &= Cov(\boldsymbol{u}, \boldsymbol{y}')Var(\boldsymbol{y})^{-\mathbf{1}}\{\boldsymbol{y} - E(\boldsymbol{y})\} \\ &= \boldsymbol{Z}'\sigma_u^2[\boldsymbol{Z}\boldsymbol{Z}'\sigma_u^2 + \boldsymbol{I}\sigma_\varepsilon^2][\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}] \\ &= \boldsymbol{Z}'[\boldsymbol{Z}\boldsymbol{Z}' + \boldsymbol{I}\sigma_\varepsilon^2\sigma_u^{-2}][\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}] \qquad [5] \end{aligned}$$

The above expression are the so-called BLUPs (Best Linear Unbiased Predictor) of the random effects.

An implementation of the EM-algorithm for the mixed effects model is then as follows:

- Initialize
- Iterate using [5] for the E-step, followed by [4] for the M-step.

(See example in Github).