STAT 3006: Statistical Computing
Lecture 3*

22 January

3 The Expectation-Maximization (EM) Algorithm

3.1 Normal Mixture Example

Q: We collected height data from n people, but we did not record their gender (female or male). How to use the height data to cluster females into a group and cluster males into another group simultaneously?

Assume the height distribution is a mixture of two normal distribution. That is to say, female height follows a normal distribution and male height also follows a normal distribution but with a different (higher) mean.

Statistical model: assume female height follows $N(\mu_1, \sigma^2)$ and male height follows $N(\mu_2, \sigma^2)$ (notice that the two distributions have the same standard deviation). The proportion of females is p. X_i and Z_i represent the height and the gender of person i (notice that X_i is observed, but Z_i is unknown). $Z_i = 1$ if person i is female; $Z_i = 0$ if person i is male. The model is formulated as follows: for i from 1 to n,

$$
\begin{aligned}
&P(Z_i = 1) = p, \quad P(Z_i = 0) = 1 - p, \\
&X_i | Z_i = 1 \sim N(\mu_1, \sigma^2), \quad X_i | Z_i = 0 \sim N(\mu_2, \sigma^2).
\end{aligned}
$$

Based on the model, the observed likelihood function of the above model is

$$
\begin{aligned}
L_o(\mu_1, \mu_2, \sigma, p | X_1, \ldots, X_n) &= \prod_{i=1}^{n} p(X_i | \mu_1, \mu_2, \sigma, p) \\
&= \prod_{i=1}^{n} \left[p(X_i | Z_i = 1; \mu_1, \mu_2, \sigma, p) \cdot p + p(X_i | Z_i = 0; \mu_1, \mu_2, \sigma, p) \cdot (1 - p) \right] \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i - \mu_1)^2}{2\sigma^2}} \cdot p + \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i - \mu_2)^2}{2\sigma^2}} \cdot (1 - p) \right]. \qquad (3.1)
\end{aligned}
$$

*If you have any question about the note, please send an email to xyluo@link.cuhk.edu.hk

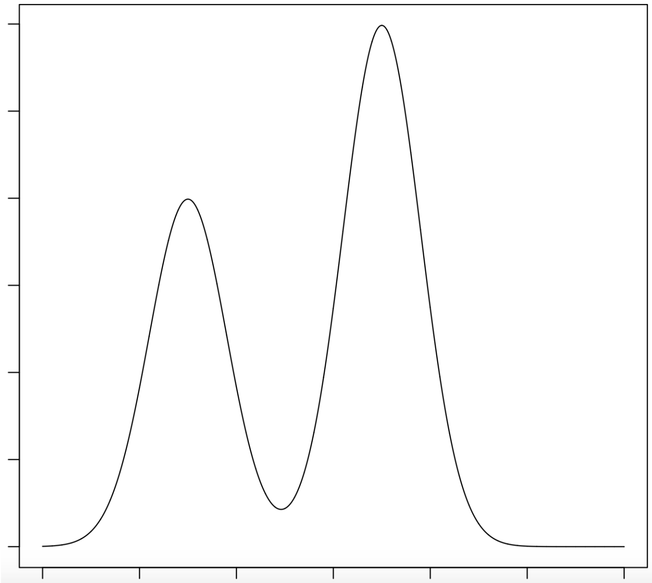

Figure 1: The pdf of a mixture of two normal distributions. The peak on the left hand side can be regarded as the average of female heights, and the peak on the right hand side can be interpreted as the average of male heights.

If we want to get MLE of μ_1, μ_2, σ, p, directly optimizing $L(\mu_1, \mu_2, \sigma, p|X_1, \ldots, X_n)$ is very difficult. Notice that $\{X_i; i = 1, \ldots, n\}$ are known and $\{Z_i; i = 1, \ldots, n\}$ are missing. When $\{Z_i; i = 1, \ldots, n\}$ are known, we call the likelihood function based on complete data $\{X_i, Z_i; i = 1, \ldots, n\}$ *complete-data likelihood function*(denoted by L_c), and call (3.1) *observed-data likelihood function*(denoted by L_o). Our idea is to use more tractable L_c to approximate the maximum of L_o.

First, the complete-data likelihood function can be easily derived.

$$
\begin{aligned}
L_c(\mu_1, \mu_2, \sigma^2, p|X_1, \ldots, X_n, Z_1, \ldots, Z_n) &= \prod_{i=1}^{n} p(X_i, Z_i|\mu_1, \mu_2, \sigma, p) \\
&= \prod_{i=1}^{n} p(X_i|Z_i; \mu_1, \mu_2, \sigma, p) \cdot p(Z_i; p) \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu_1)^2}{2\sigma^2}} \cdot p\right]^{Z_i} \cdot \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu_2)^2}{2\sigma^2}} \cdot (1-p)\right]^{1-Z_i}.
\end{aligned}
$$

It follows that the *complete-data log likelihood function*

$$
\begin{aligned}
&l_c(\mu_1, \mu_2, \sigma^2, p|X_1, \ldots, X_n, Z_1, \ldots, Z_n) \\
&\quad = \sum_{i=1}^{n} \{ Z_i \cdot [-\frac{1}{2} log(2\pi) - \frac{1}{2} log\sigma^2 - \frac{(X_i - \mu_1)^2}{2\sigma^2} + logp] + (1 - Z_i) \cdot \\
&\qquad [-\frac{1}{2} log(2\pi) - \frac{1}{2} log\sigma^2 - \frac{(X_i - \mu_2)^2}{2\sigma^2} + log(1-p)] \} \\
&\quad = \sum_{i=1}^{n} \left\{ -\frac{1}{2} log(2\pi) - \frac{1}{2} log\sigma^2 - \frac{Z_i(X_i - \mu_1)^2}{2\sigma^2} - \frac{(1 - Z_i)(X_i - \mu_2)^2}{2\sigma^2} + Z_i logp + (1 - Z_i) log(1-p) \right\}.
\end{aligned}
\tag{3.2}
$$

Maximizing $l_c(\mu_1, \mu_2, \sigma^2, p|X_1, \ldots, X_n, Z_1, \ldots, Z_n)$ w.r.t $\mu_1, \mu_2, \sigma^2, p$ is equivalent to solving the following equation system.

$$\frac{\partial l_c}{\partial p} = 0 \tag{3.3}$$

$$\frac{\partial l_c}{\partial \mu_1} = 0 \tag{3.4}$$

$$\frac{\partial l_c}{\partial \mu_2} = 0 \tag{3.5}$$

$$\frac{\partial l_c}{\partial \sigma^2} = 0 \tag{3.6}$$

For equation (3.3),

$$
\begin{aligned}
\frac{\partial l_c}{\partial p} &= \sum_{i=1}^{n} \frac{Z_i}{p} - \frac{1 - Z_i}{1 - p} = 0 \\
\hat{p} &= \frac{\sum_{i=1}^{n} Z_i}{n}.
\end{aligned}
$$

For equation (3.4) and (3.5),

$$\frac{\partial l_c}{\partial \mu_1} = \sum_{i=1}^{n} \frac{-2Z_i(X_i - \mu_1)^2}{2\sigma^2} = 0$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} Z_i X_i}{\sum_{i=1}^{n} Z_i} \tag{3.7}$$

$$\textit{Similarly,} \quad \hat{\mu}_2 = \frac{\sum_{i=1}^{n} (1 - Z_i) X_i}{\sum_{i=1}^{n} (1 - Z_i)}. \tag{3.8}$$

The equation (3.7) and (3.8) indicate that the estimate of μ_1 is the average of heights in the female group and the estimate of μ_2 is the average of heights in the male group, respectively. Generally speaking, the estimate of μ_1 is the weighted average of heights in all people with equal weights in the female group and zero weights in the male group; the estimate of μ_2 is the weighted average of heights in all people with zero weights in the female group and equal weights in the male group.

For equation (3.6),

$$\frac{\partial l}{\partial \sigma^2} = \sum_{i=1}^{n} \left[-\frac{1}{2\sigma^2} + \frac{Z_i(X_i - \hat{\mu}_1)^2}{2\sigma^4} + \frac{(1-Z_i)(X_i - \hat{\mu}_2)^2}{2\sigma^4} \right] = 0$$

$$\hat{\sigma}^2 = \frac{1}{n} \left[\sum_{i=1}^{n} Z_i(X_i - \hat{\mu}_1)^2 + (1-Z_i)(X_i - \hat{\mu}_2)^2 \right]. \tag{3.9}$$

As you can see, the estimate of σ^2 is the weighted average of two sample variances with the weight proportional to the female number in the female group and to the male number in the male group.

However, $\{Z_i; i = 1, \ldots, n\}$ are unknown, so how to approximate Z_i? In the EM algorithm, we replace Z_i by the *conditional expectation* (E step) $E(Z_i|X_i; \mu_1^{(t)}, \mu_2^{(t)}, \sigma^{2(t)}, p^{(t)})$ in the *complete-data log likelihood function*l_c, where $\mu_1^{(t)}, \mu_2^{(t)}, \sigma^{2(t)}, p^{(t)}$ are estimates in the current iteration, and then maximize l_c (M step), which is usually much easier than directly maximizing L_o. We alternate E step and M step enough times (with good initial values), and the final esimates of parameters are the maximum points of L_o.

3.2 Calculate Conditional Expectation

Definition 3.1. A random variable $X \sim f(x)$, where $f(x)$ is the probability density function or the probability mass function. The expectation of X, $E(X)$, is defined as $\int x f(x) dx$. Specifically, $E(X) = \int_{-\infty}^{\infty} x f(x) dx$ (continuous case); $E(X) = \sum_{i=-\infty}^{\infty} x_k f(x_k) = \sum_{i=-\infty}^{\infty} x_k p_k$ (discrete case), where we let p_k be $f(x_k)$.

Example: flip a biased coin twice, what is the expected number of observed head?

Suppose the probability to observe one head in one trial is p and X represents the number of heads. It follows that $X \sim Binomial(2, p)$, $E(X) = 0 \cdot (1-p)^2 + 1 \cdot 2(1-p)p + 2 \cdot p^2 = 2p$.

Definition 3.2. We have two random variables X and Y. We also know the conditional density (or mass) function of X given $Y = y$ is $f_X(x|y)$. The conditional expectation of X given $Y = y$, $E(X|Y = y)$, is defined as $\int x f_X(x|y) dx$.

Remark 1. when the joint density (or mass) function of X and Y $(f(x, y))$ and the marginal density (or mass) function of $Y(f_Y(y))$ is known, the conditional density (or mass) function $f_X(x|y) = \frac{f(x,y)}{f_Y(y)}$.

Keeping these definitions in mind, we calculate the E step, which is calculating $E(Z_i|X_i; \mu_1^{(t)}, \mu_2^{(t)}, \sigma^{2(t)}, p^{(t)})$.

We denote the conditional expectation by w_{it}.

$$\begin{aligned}
w_{it} &= E(Z_i|X_i;\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)}) \\
&= 0\cdot p(Z_i=0|X_i;\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)}) + 1\cdot p(Z_i=1|X_i,\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)}) \\
&= p(Z_i=1|X_i,\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)}) \\
&= \frac{p(Z_i=1,X_i;\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)})}{p(Z_i=0,X_i;\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)}) + p(Z_i=1,X_i;\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)})} \\
&= \frac{p^{(t)}\cdot\frac{1}{\sqrt{2\pi}\sigma^{(t)}}e^{-\frac{(X_i-\mu_1^{(t)})^2}{2\sigma^{2(t)}}}}{p^{(t)}\cdot\frac{1}{\sqrt{2\pi}\sigma^{(t)}}e^{-\frac{(X_i-\mu_1^{(t)})^2}{2\sigma^{2(t)}}} + (1-p^{(t)})\cdot\frac{1}{\sqrt{2\pi}\sigma^{(t)}}e^{-\frac{(X_i-\mu_2^{(t)})^2}{2\sigma^{2(t)}}}}.
\end{aligned} \tag{3.10}$$

The EM algorithm for maximizing L_o in the equation (3.1) is as follows.

Algorithm: EM algorithm for the normal mixture example.

Input: $\{X_i; i=1,\ldots,n\}$,$p^{(0)}$, $\mu_1^{(0)}$, $\mu_2^{(0)}$, $\sigma^{2(0)}$, total interation number T.
Initialize: $p^{(0)}$, $\mu_1^{(0)}$, $\mu_2^{(0)}$, $\sigma^{2(0)}$, $t \leftarrow 0$.
Repeat
(E step) calculate w_{it} for $i=1,\ldots,n$ based on the equation (3.10);
(M step) maximize l_c in the equation (3.2) with Z_i being replaced by w_{it}:
$p^{(t+1)} = \sum_{i=1}^n w_{it}/n$;
$\mu_1^{(t+1)} = \sum_{i=1}^n w_{it}X_i/\sum_{i=1}^n w_{it}$;
$\mu_2^{(t+1)} = \sum_{i=1}^n (1-w_{it})X_i/\sum_{i=1}^n (1-w_{it})$;
$\sigma^{2(t+1)} = 1/n\cdot[\sum_{i=1}^n w_{it}(X_i-\mu_1^{(t+1)})^2 + (1-w_{it})(X_i-\mu_2^{(t+1)})^2]$.
$t \leftarrow t+1$;
Until $t == T$.
Output:$\mu_1^{(t)},\mu_2^{(t)},\sigma^{2(t)},p^{(t)}$ are the MLE of L_o in the equation (3.1) .

3.3 The General Case

In this subsection, we will talk about the EM algorithm to deal with general problems with unknown data. The data $\mathbf{Y}$ has two parts. One is observed data $\mathbf{Y}_{obs}$. The other is unknown (missing) data $\mathbf{Y}_{mis}$. That is to say, $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. Assume Θ are the parameters of our interest, $f(\mathbf{Y}|\Theta)$ is the complete-data likelihood function, $g(\mathbf{Y}_{obs}|\Theta)$ is the observed-data likelihood function, and $k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta)$ is the conditional density function of $\mathbf{Y}_{mis}$ given $\mathbf{Y}_{obs}$. We have the following derivation.

$$
\begin{aligned}
f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}|\Theta) &= g(\mathbf{Y}_{obs}|\Theta) \cdot k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta) \\
(Taking\ \ logarithm)\ l_c(\Theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) &= l_o(\Theta|\mathbf{Y}_{obs}) + logk(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta) \\
l_o(\Theta|\mathbf{Y}_{obs}) &= l_c(\Theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) - logk(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta)
\end{aligned}
$$

Our target is to find $\hat{\Theta} = \arg\max_{\Theta} l_o(\Theta|\mathbf{Y}_{obs})$ assuming it is more convenient to work with $l_c(\Theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$.

Given $\Theta^{(t)}$,

$$
\begin{aligned}
l_o(\Theta|\mathbf{Y}_{obs}) = & \int l_c(\Theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) \cdot k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)}) d\mathbf{Y}_{mis} - \\
& \int logk(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta) \cdot k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)}) d\mathbf{Y}_{mis} \\
:= & \, Q(\Theta|\Theta^{(t)}) - H(\Theta|\Theta^{(t)}).
\end{aligned}
$$

By calculating (E step) and maximizing (M step) $Q(\Theta|\Theta^{(t)})$, we get $\Theta^{(t+1)} = \arg\max Q(\Theta|\Theta^{(t)})$. In addition, it can be proved that $H(\Theta^{(t+1)}|\Theta^{(t)}) \leq H(\Theta^{(t)}|\Theta^{(t)})$ by the Jensen inequality. It follows that $l_o(\Theta^{(t+1)}|\mathbf{Y}_{obs}) \geq l_o(\Theta^{(t)}|\mathbf{Y}_{obs})$. The inequality indicates that after each iteration of the EM algorithm, the obtained $\Theta^{(t+1)}$ always make the observed likelihood increasing.

Algorithm: EM algorithm for the general case.

Input: $\mathbf{Y}_{obs}$, $\Theta^{(0)}$, total interation number T.
Initialize: $\Theta^{(0)}$, $t \leftarrow 0$.
Repeat
  (E step) calculate the conditional expectation $Q(\Theta|\Theta^{(t)})$;
  (M step) maximize $Q(\Theta|\Theta^{(t)})$ w.r.t Θ;
  $\Theta^{(t+1)} = \arg\max Q(\Theta|\Theta^{(t)})$;
  $t \leftarrow t + 1$;
Until $t == T$.
Output:$\Theta^{(t+1)}$ is an approaximate value to the MLE of observed likelihood function .

3.4 Example: Blood Type

Q: There are n people. n_A people are observed to have blood type A; n_B people are observed to have blood type B; n_{AB} people are observed to have blood type B; n_O poeple are observed to have blood type O. What is the frequency of allele A, B, O (p_A, p_B, p_O) in the population ?

Interpretation: n_A people have genotype AA or AO; n_B people have genotype BB or BO; n_{AB} people have genotype AB; n_O people have genotype OO. The frequencey of AA, AO, BB, BO, AB and OO is p_A^2, $2p_Ap_O$, p_B^2, $2p_Bp_O$, p_Ap_B and p_O^2, respectively. Moreover, $n_A = n_{AA} + n_{AO}$, $n_B = n_{BB} + n_{BO}$, $n_{AB} = n_{AB}$, and $n_O = n_{OO}$. Complete data is $\{n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO}\}$, and the observed data is $\{n_A, n_B, n_{AB}, n_O\}$.

The complete-data likelihood can be derived as follows:

$$\begin{aligned} &L(p_A, p_B, p_O|n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO}) \\ &\qquad = \frac{n!}{n_{AA}!n_{AO}!n_{BB}!n_{BO}!n_{AB}!n_{OO}!}(p_A^2)^{n_{AA}}(2p_Ap_O)^{n_{AO}}(p_B^2)^{n_{BB}}(2p_Bp_O)^{n_{BO}}(2p_Ap_B)^{n_{AB}}(p_O^2)^{n_{OO}}. \end{aligned} \tag{3.11}$$

Taking logarithm,

$$\begin{aligned} &l(p_A, p_B, p_O|n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO}) \\ &\qquad = C + (n_{AO} + n_{BO} + n_{AB})log2 + logp_A(2n_{AA} + n_{AO} + n_{AB}) + \\ &\qquad\quad logp_B(2n_{BB} + n_{BO} + n_{AB}) + logp_0(2n_{OO} + n_{AO} + n_{BO}). \end{aligned}$$

In the E step, we calculate

$$\begin{aligned} n_{AA}^{(t)} &:= E[n_{AA}|n_A; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] = n_A \frac{p_A^{(t)}}{p_A^{(t)} + 2p_O^{(t)}} \\ n_{AO}^{(t)} &:= E[n_{AO}|n_A; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] = n_A \frac{2p_O^{(t)}}{p_A^{(t)} + 2p_O^{(t)}} \\ n_{BB}^{(t)} &:= E[n_{BB}|n_B; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] = n_B \frac{p_B^{(t)}}{p_B^{(t)} + 2p_O^{(t)}} \\ n_{BO}^{(t)} &:= E[n_{BO}|n_B; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] = n_B \frac{2p_B^{(t)}}{p_B^{(t)} + 2p_O^{(t)}}. \end{aligned} \tag{3.12}$$

In the M step, we calculate

$$\begin{aligned} p_A^{(t)} &= \frac{2n_{AA}^{(t)} + n_{AO}^{(t)} + n_{AB}}{2n} \\ p_B^{(t)} &= \frac{2n_{BB}^{(t)} + n_{BO}^{(t)} + n_{AB}}{2n}. \end{aligned}$$