STAT 3006: Statistical Computing
Lecture 4*

29 January

3.5 Why EM algorithm Works?

In the subsection 3.3, we have known that

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

If we can prove

$$H(\Theta^{(t+1)}|\Theta^{(t)}) \le H(\Theta^{(t)}|\Theta^{(t)}), \tag{3.1}$$

then $l_o(\Theta^{(t+1)}|\mathbf{Y}_{obs}) \ge l_o(\Theta^{(t)}|\mathbf{Y}_{obs})$, which indicates the non-decreasing property of the sequence $\{\Theta^{(t)}\}$ produced by the EM algorithm. To prove the inequality (3.1), we first review the convex function.

Definition 3.1. A function f defined on an interval $[a, b]$ is convex, if $f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2)$ for $\forall x_1, x_2 \in [a, b]$ and $\forall t \in [0, 1]$.

Intuitively, a function f is convex if the segment between points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ for $\forall x_1, x_2 \in [a, b]$ is always above or lies in the curve $\{(x, f(x)) : x_1 \le x \le x_2\}$ (illustrated in the figure 1).

There are some equivalent definitions for convexity:

1 A differentiable function f on $[a, b]$ is convex if and only if $f(x) \ge f(y) + f'(y)(x - y)$;

2 A twice differentiable function f on $[a, b]$ is convex if and only if $f''(x) \ge 0$.

Jensen's inequality: If $h(x)$ is convex and W is a random variable, then

$$E(h(W)) \ge h(EW)$$

.

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

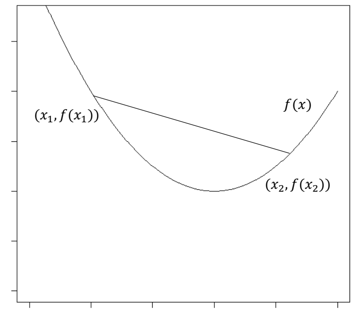

Figure 1: Figure demonstration for the convex function $f(x)$.

Proof. $h(x)$ is convex, so there exists a constant ξ such that

$$h(x) \geq h(x_0) + \xi(x - x_0) \tag{3.2}$$

for any fixed x_0. Let x_0 be $E(W)$ and x be W. The inequality (3.2) becomes $h(W) \geq h(E(W)) + \xi(W - E(W))$. Taking expectation in the inequality, we have $E(h(W)) \geq h(EW)$. As you can see, $E(h(W)) = h(E(W))$ if and only if h is a linear function.

We now come back to prove $H(\Theta^{(t+1)}|\Theta^{(t)}) \leq H(\Theta^{(t)}|\Theta^{(t)})$.

Proof. $\forall \Theta$,

$$\begin{aligned}
H(\Theta|\Theta^{(t)}) - H(\Theta^{(t)}|\Theta^{(t)}) &= \int log\frac{k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta)}{k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)})} \cdot k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)})d\mathbf{Y}_{mis} \\
&= -\int -log\frac{k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta)}{k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)})} \cdot k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)})d\mathbf{Y}_{mis} \\
&\leq (Jensen's\ Inequality)\ \ log(\int \frac{k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta)}{k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)})} \cdot k(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \Theta^{(t)})d\mathbf{Y}_{mis}) \\
&= log(1) = 0.
\end{aligned}$$

Notice that we apply Jensen inequality to the function $h(x) = -log(x)$.