Example Class 9

LIU Chen

Department of Statistics and Actuarial Science,
The University of Hong Kong

November 15, 2022

Expectation–Maximization (EM) Algorithm

- Denote Y_{obs} as the observed data; Y_{mis} as the missing data
- $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ as the complete data
- $f_{\text{obs}}(Y_{\text{obs}}|\boldsymbol{\theta})$, $f_{\text{comp}}(Y|\boldsymbol{\theta})$ as the observed and complete likelihood.
- The missing data Y_{mis} follows the conditional distribution $f_{Y_{\text{mis}}|Y_{\text{obs}},\boldsymbol{\theta}}(\cdot|Y_{\text{obs}},\boldsymbol{\theta})$
- Target: find $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f_{\text{obs}}(Y_{\text{obs}}|\boldsymbol{\theta})$.
- Distinguish the missing data Y_{mis} and the observed data with missingness Y_{obs}! Ususally, some of observations Y_{obs} suffer from missingness, and Y_{mis} can be interpreted as the 'exact' value (but not observed) of these missing observations

Expectation–Maximization (EM) Algorithm

- Two step: E(xpectation) and M(aximization)
- E-step: Compute the expectation of the complete data log likelihood with respect to the conditional distribution of missing data:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\mathrm{old}}) = \int \log\left(f_{\mathrm{comp}}(Y_{\mathrm{obs}}, Y_{\mathrm{mis}}|\boldsymbol{\theta})\right) f_{\mathrm{mis}}(Y_{\mathrm{mis}}|Y_{\mathrm{obs}}, \boldsymbol{\theta}^{\mathrm{old}}) dY_{\mathrm{mis}}$$

Use the missing data Y_{obs} and then integrate them out.

- M-step: $\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\mathrm{old}})$
- EM algorithm improves $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})})$ rather than improving $\log f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}}|\boldsymbol{\theta})$

Data Augmentation: Stochastic Version of EM

- The DA algorithm consists of iterations between the imputation step (I-step) and the posterior step (P-step).
- I-step: draw $\{\theta^{(j)}\}_{j=1}^{m}$ from the current $f_k(\theta | Y_{\mathrm{obs}})$; for each $\theta^{(j)}$, draw $z^{(j)}$ from $f_{Z|Y_{\mathrm{obs}},\theta}(z | Y_{\mathrm{obs}}, \theta^{(j)})$
- P-step: Update posterior as

$$f_{k+1}(\theta | Y_{\mathrm{obs}}) = \frac{1}{m} \sum_{j=1}^{m} f(\theta | Y_{\mathrm{obs}}, z^{(j)})$$

- The produced $z^{(j)}, j = 1, \ldots, m$ are called multiple imputation.

Why EM?

- Observed log likelihood: $\ell(\boldsymbol{\theta}) = \log f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}}|\boldsymbol{\theta})$
- Ascent property: if $Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)})$, then $\ell(\boldsymbol{\theta}^{(k+1)}) \geq \ell(\boldsymbol{\theta}^{(k)})$

$$
\begin{aligned}
\ell(\boldsymbol{\theta}^{(k+1)}) &= Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) + \ell(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) \\
&= Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) + \log f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(k+1)}) - E_{\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs},\boldsymbol{\theta}^{(k)}}\left[\log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}},\boldsymbol{Y}_{\mathrm{mis}}|\boldsymbol{\theta}^{(k+1)})\right] \\
&= Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) - E_{\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs},\boldsymbol{\theta}^{(k)}}\left[\log\left\{\frac{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}},\boldsymbol{Y}_{\mathrm{mis}}|\boldsymbol{\theta}^{(k+1)})}{f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(k+1)})}\right\}\right] \\
&\geq Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) - E_{\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs},\boldsymbol{\theta}^{(k)}}\left[\log\left\{\frac{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}},\boldsymbol{Y}_{\mathrm{mis}}|\boldsymbol{\theta}^{(k+1)})}{f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(k+1)})}\right\}\right] \\
&\geq Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) - E_{\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs},\boldsymbol{\theta}^{(k)}}\left[\log\left\{\frac{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}},\boldsymbol{Y}_{\mathrm{mis}}|\boldsymbol{\theta}^{(k)})}{f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(k)})}\right\}\right] \text{(Gibbs's inequality)} \\
&= Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) + \ell(\boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) = \ell(\boldsymbol{\theta}^{(k)})
\end{aligned}
$$

Gibbs's inequality

- Consider two distributions P and Q of random variables with densities $p(\cdot)$ and $q(\cdot)$, respectively, it holds that,

$$\int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \geq 0.$$

- Non-negativity of Kullback–Leibler divergence.
- $\because \log x \leq x - 1$ for all $x > 0$,

$$\begin{aligned} -\int p(x) \log\left(\frac{q(x)}{p(x)}\right) dx &\geq \int p(x) \left(\frac{q(x)}{p(x)} - 1\right) dx \\ &= 1 - 1 = 0 \end{aligned}$$

$$
\begin{aligned}
& -E_{\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs},\boldsymbol{\theta}^{(k)}}\left[\log\left\{\frac{f_{\text{comp}}(\boldsymbol{Y}_{\text{obs}},\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k+1)})}{f_{\text{obs}}(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta}^{(k+1)})}\right\}\right] \\
&= -\int f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)},\boldsymbol{Y}_{\text{obs}})\log(f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k+1)},\boldsymbol{Y}_{\text{obs}}))d\boldsymbol{Y}_{\text{mis}} \\
&\geq -\int f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)},\boldsymbol{Y}_{\text{obs}})\log(f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)},\boldsymbol{Y}_{\text{obs}}))d\boldsymbol{Y}_{\text{mis}} \\
&= -E_{\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs},\boldsymbol{\theta}^{(k)}}\left[\log\left\{\frac{f_{\text{comp}}(\boldsymbol{Y}_{\text{obs}},\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)})}{f_{\text{obs}}(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{\theta}^{(k)})}\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
\ell(\boldsymbol{\theta}^{(k+1)})-\ell(\boldsymbol{\theta}^{(k)}) &= Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)})-Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) \\
&\quad+\int f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)},\boldsymbol{Y}_{\text{obs}})\log(f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)},\boldsymbol{Y}_{\text{obs}}))d\boldsymbol{Y}_{\text{mis}} \\
&\quad-\int f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k)},\boldsymbol{Y}_{\text{obs}})\log(f_{\text{mis}}(\boldsymbol{Y}_{\text{mis}}|\boldsymbol{\theta}^{(k+1)},\boldsymbol{Y}_{\text{obs}}))d\boldsymbol{Y}_{\text{mis}} \\
&\geq Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)})-Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)})
\end{aligned}
$$

Choosing $\boldsymbol{\theta}^{(k+1)}$ to improve $Q(\cdot|\boldsymbol{\theta}^{(k)})$ causes $\ell(\cdot)$ to improve at least as much.

EM for Mixture Distribution

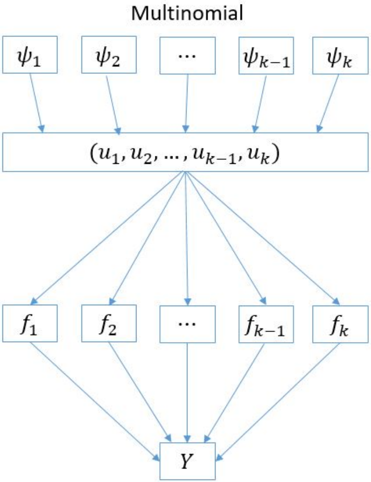
- $Y_i \sim \sum_{j=1}^{k} \psi_j f_j(y)$.
- Each f_j is a density function and $\sum_{j=1}^{k} \psi_j = 1$. At first, we assume f_j's are known.
- Take u_i as the missing data where $u_i = j$ indicates that the i-th item y_i comes from j-th component of the mixture distribution f_j.
- Complete likelihood

$$g(y_i, u_i|\boldsymbol{\psi}) = \prod_{j=1}^{k} [\psi_j f_j(y_i)]^{I(u_i=j)}$$

- Conditional distribution for u_i,

$$P(u_i = j|\boldsymbol{\psi}, y_i) = \frac{P(y_i|u_i = j, \boldsymbol{\psi})P(u_i = j|\boldsymbol{\psi})}{\sum_{l=1}^{k} P(y_i|u_i = l, \boldsymbol{\psi})P(u_i = l|\boldsymbol{\psi})} = \frac{\psi_j f_j(y_i)}{\sum_{l=1}^{k} \psi_l f_l(y_i)}$$

EM for Mixture Distribution



EM for Mixture Distribution

$$
\begin{aligned}
Q(\psi|\psi^{(\text{old})}) &= E_{\boldsymbol{u}|\psi^{(\text{old})},\boldsymbol{y}}[\log \prod_{i=1}^{n} g(y_i, u_i|\psi)] \\
&= E_{\boldsymbol{u}|\psi^{(\text{old})},\boldsymbol{y}} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} \{(\log \psi_j + \log f_j(y_i)) I(u_i = j)\} \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} P(u_i = j|\psi^{(\text{old})}, \boldsymbol{y}) \log \psi_j + \text{unrelated terms} \\
&= \sum_{j=1}^{k} \log \psi_j \sum_{i=1}^{n} P(u_i = j|\psi^{(\text{old})}, \boldsymbol{y})
\end{aligned}
$$

Multinomial likelihood: $\ell(\boldsymbol{p}|\boldsymbol{n}) = \prod_{j=1}^{k} p_j^{n_j}$, $\log \ell(\boldsymbol{p}|\boldsymbol{n}) = \sum_{j=1}^{k} n_j \log p_j$, $\hat{p}_j = \frac{n_j}{\sum_{l=1}^{k} n_l}$

EM for Mixture Distribution

$$
\begin{aligned}
\psi_j^{(\text{new})} &= \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(\text{old})}) = \frac{\sum_{i=1}^{n} P(u_i = j|\boldsymbol{\psi}^{(\text{old})}, \boldsymbol{y})}{n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\psi_j^{(\text{old})} f_j(y_i)}{\sum_{l=1}^{k} \psi_l^{(\text{old})} f_l(y_i)}
\end{aligned}
$$

EM for Mixture Distribution (Parametric)

- What if f_j's are distributions with unknown parameters?
- $f_j \sim N(\mu_j, \sigma^2)$, σ^2 is known.

$$Q(\psi, \boldsymbol{\mu}|\psi^{(\text{old})}, \boldsymbol{\mu}^{(\text{old})}) = E_{\boldsymbol{u}|\boldsymbol{\psi}^{(\text{old})}, \boldsymbol{\mu}^{\text{old}}, \boldsymbol{y}}[\log \prod_{i=1}^{n} g(y_i, u_i|\psi, \boldsymbol{\mu})]$$

$$= E_{\boldsymbol{u}|\boldsymbol{\psi}^{(\text{old})}, \boldsymbol{\mu}^{(\text{old})}, \boldsymbol{y}} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} \{(\log \psi_j + \log f_j(y_i)) I(u_i = j)\} \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} P(u_i = j|\boldsymbol{\psi}^{(\text{old})}, \boldsymbol{\mu}^{(\text{old})}, \boldsymbol{y}) \left\{ \log \psi_j - \frac{(y_i - \mu_j)^2}{2\sigma^2} \right\}$$

EM for Mixture Distribution (Parametric)

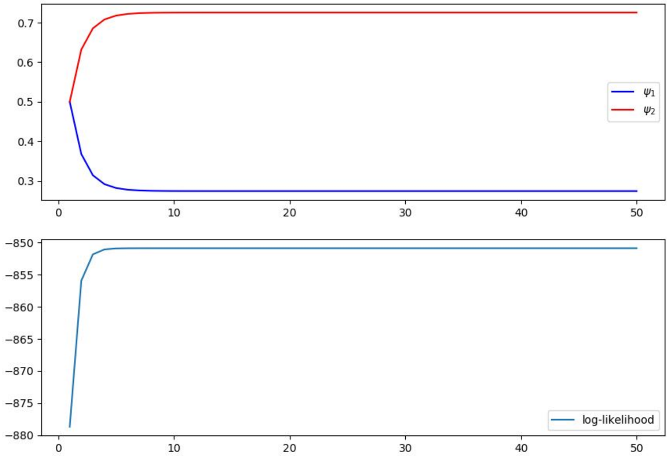
$$
\begin{aligned}
\psi_j^{(\mathrm{new})} &= \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}, \boldsymbol{\mu} | \boldsymbol{\psi}^{(\mathrm{old})}, \boldsymbol{\mu}^{(\mathrm{old})}) = \frac{\sum_{i=1}^n P(u_i = j | \boldsymbol{\psi}^{(\mathrm{old})}, \boldsymbol{\mu}^{(\mathrm{old})}, \boldsymbol{y})}{n} \\
&= \frac{1}{n} \sum_{i=1}^n \frac{\psi_j^{(\mathrm{old})} f_j(y_i | \boldsymbol{\mu}^{(\mathrm{old})})}{\sum_{l=1}^k \psi_l^{(\mathrm{old})} f_l(y_i | \boldsymbol{\mu}^{(\mathrm{old})})}
\end{aligned}
$$

$$
\frac{\partial Q(\boldsymbol{\psi}, \boldsymbol{\mu} | \boldsymbol{\psi}^{(\mathrm{old})}, \boldsymbol{\mu}^{(\mathrm{old})})}{\partial \mu_j} = -\frac{1}{\sigma^2} \sum_{i=1}^n P(u_i = j | \boldsymbol{\psi}^{(\mathrm{old})}, \boldsymbol{\mu}^{(\mathrm{old})}, \boldsymbol{y})(y_i - \mu_j) = 0
$$

$$
\to \mu_j^{(\mathrm{new})} = \frac{\sum_{i=1}^n y_i P(u_i = j | \boldsymbol{\psi}^{(\mathrm{old})}, \boldsymbol{\mu}^{(\mathrm{old})}, \boldsymbol{y})}{\sum_{i=1}^n P(u_i = j | \boldsymbol{\psi}^{(\mathrm{old})}, \boldsymbol{\mu}^{(\mathrm{old})}, \boldsymbol{y})}
$$

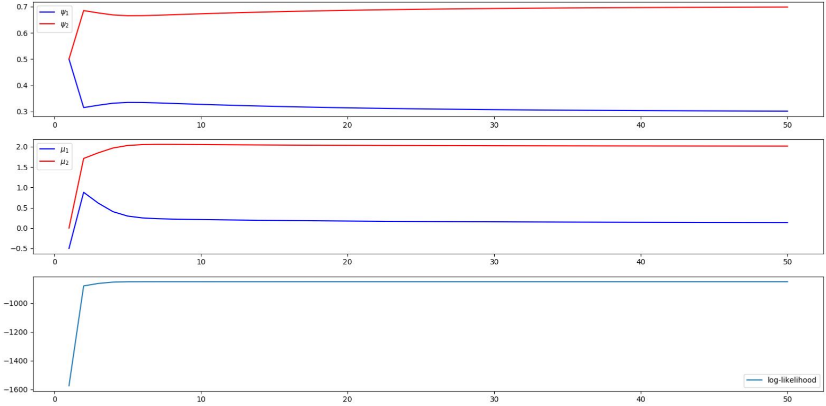
Example (Mixture Distribution)

- $y_i \sim 0.3N(0,1) + 0.7N(2,1), n = 500$



Example (Mixture Distribution, Parametric)

- $y_i \sim 0.3N(\mu_1, 1) + 0.7N(\mu_2, 1)$, $n = 500$, $\boldsymbol{\mu}_{\text{true}} = (0, 2)^T$



Monte Carlo EM

- In the EM algorithm, each E-step requires the computation of an expectation

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) = E_{\boldsymbol{Y}_{\mathrm{mis}}|\boldsymbol{\theta}^{(\mathrm{old})}, \boldsymbol{Y}_{\mathrm{obs}}}(\log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{Y}_{\mathrm{mis}}|\boldsymbol{\theta}))$$

- However, in some cases, the E-step might be complex and does not admit a closed form solution. That is, the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})})$ function cannot be computed explicitly.
- Solution: evaluate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})})$ by Monte Carlo methods $\longrightarrow$ MCEM algorithm.

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) = \frac{1}{M}\sum_{m=1}^{M} \log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{z}_m|\boldsymbol{\theta})$$

$$\boldsymbol{z}_m \sim f_{mis}(\cdot|\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{\theta}^{(\mathrm{old})}), m = 1, \ldots, M$$

Monte Carlo EM

- What if we cannot draw samples from $f_{mis}(\cdot|\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{\theta}^{(\mathrm{old})})$ directly?
- You may draw $z_1, \ldots, z_M$ from $f_{mis}(\cdot|\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{\theta}^{(\mathrm{old})})$ at each iteration, but it would cost much computation power.
- Application of importance sampling: with an initial value $\psi^{(0)}$ of ψ,

$$\boldsymbol{u}_m \sim f_{mis}(\cdot|\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{\theta}^{(0)})$$

$$\hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) = \sum_{m=1}^{M} \omega_m \log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m|\boldsymbol{\theta}) / \sum_{m=1}^{M} \omega_m$$

$$\omega_m = \frac{f_{mis}(\boldsymbol{u}_m|\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{\theta}^{(\mathrm{old})})}{f_{mis}(\boldsymbol{u}_m|\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{\theta}^{(0)})}$$

- The cost in obtaining the weights ω_m is less than obtaining a new sample.

Monte Carlo EM

$$
\begin{aligned}
\omega_m &= \frac{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m | \boldsymbol{\theta}^{(\mathrm{old})}) / f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}} | \boldsymbol{\theta}^{(\mathrm{old})})}{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m | \boldsymbol{\theta}^{(0)}) / f_{\mathrm{obs}}(\boldsymbol{Y}_{\mathrm{obs}} | \boldsymbol{\theta}^{(0)})} \\
\hat{Q}(\boldsymbol{\theta} | \boldsymbol{\theta}^{(\mathrm{old})}) &= \sum_{m=1}^{M} \omega'_m \log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m | \boldsymbol{\theta}) / \sum_{m=1}^{M} \omega'_m \\
\omega'_m &= \frac{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m | \boldsymbol{\theta}^{(\mathrm{old})})}{f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m | \boldsymbol{\theta}^{(0)})}
\end{aligned}
$$

Logistic Regression with Random Effects

- For $i = 1, \ldots, I$ and $j = 1, \ldots, J$,

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) &= \beta x_{ij} + u_i, \\ u_i &\sim N(0, \sigma_u^2). \end{aligned}$$

- Observations: $\{Y_{ij}, x_{ij}\}, i = 1, \ldots, I, j = 1 \ldots, J$.
- Unknown parameters: $\boldsymbol{\theta} = (\beta, \sigma_u^2)^T$.

Logistic Regression with Random Effects

- Complete-data likelihood:

$$
\begin{aligned}
p(\boldsymbol{Y}, \boldsymbol{u}|\boldsymbol{\theta}) &= \prod_{i=1}^{I} p(u_i) \prod_{j=1}^{J} p(y_{ij}|u_i) \\
&= \prod_{i=1}^{I} \left\{ \frac{1}{\sqrt{2\pi}\sigma_u} \exp\left(-\frac{u_i^2}{2\sigma_u^2}\right) \prod_{j=1}^{J} \frac{\exp\{y_{ij}(\beta x_{ij} + u_i)\}}{1 + \exp\{\beta x_{ij} + u_i\}} \right\}
\end{aligned}
$$

$$
\begin{aligned}
\ell(\boldsymbol{Y}, \boldsymbol{u}|\boldsymbol{\theta}) = \sum_{i=1}^{I} \Bigg\{ & -\frac{1}{2} \log \sigma_u^2 - \frac{u_i^2}{2\sigma_u^2} \\
& + \sum_{j=1}^{J} [y_{ij}(\beta x_{ij} + u_i) - \log(1 + \exp\{\beta x_{ij} + u_i\})] \Bigg\}
\end{aligned}
$$

Logistic Regression with Random Effects

- The conditional density of the latent variable u_i's is

$$
\begin{aligned}
p(u_i|\boldsymbol{Y}, \boldsymbol{\theta}^{(\text{old})}) &\propto \prod_{j=1}^{J} p(y_{ij}|u_i, \boldsymbol{\theta}^{(\text{old})}) p(u_i|\boldsymbol{\theta}^{(\text{old})}) \\
&= \frac{1}{\sqrt{2\pi}\sigma_u^{(0)}} \exp\left(-\frac{u_i^2}{2(\sigma_u^{(0)})^2}\right) \\
&\times \prod_{j=1}^{J} \frac{\exp\{y_{ij}(\beta^{(0)} x_{ij} + u_i)\}}{1 + \exp\{\beta^{(0)} x_{ij} + u_i\}}
\end{aligned}
$$

Logistic Regression with Random Effects

- Q-function:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) &= \int \ell(\boldsymbol{Y}, \boldsymbol{u}|\boldsymbol{\theta}) p(\boldsymbol{u}|\boldsymbol{Y}, \boldsymbol{\theta}^{(\mathrm{old})}) d\boldsymbol{u} \\ &= \int \sum_{i=1}^{I} \Bigg\{ -\log \sigma_u - \frac{u_i^2}{2\sigma_u^2} \\ &\quad + \sum_{j=1}^{J} [y_{ij}(\beta x_{ij} + u_i) - \log(1 + \exp\{\beta x_{ij} + u_i\})] \Bigg\} \\ &\quad \times \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi}\sigma_u^{(0)}} \exp\left(-\frac{u_i^2}{2(\sigma_u^{(0)})^2} \right) \prod_{j=1}^{J} \frac{\exp\{y_{ij}(\beta^{(0)} x_{ij} + u_i)\}}{1 + \exp\{\beta^{(0)} x_{ij} + u_i\}} \{d\mu_i\}_{i=1}^{I} \end{aligned}$$

Logistic Regression with Random Effects

$$\hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) = \frac{\sum_{m=1}^{M} \omega'_m \log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m|\boldsymbol{\theta}^{(\mathrm{old})})}{\sum_{m=1}^{M} \omega'_m}$$

$$\begin{aligned}\log f_{\mathrm{comp}}(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{u}_m|\boldsymbol{\theta}^{(\mathrm{old})}) = \sum_{i=1}^{I} \Bigg\{ & -\log \sigma_u - \frac{u_{i,m}^2}{2\sigma_u^2} \\ & + \sum_{j=1}^{J} \left[y_{ij}(\beta x_{ij} + u_{i,m}) - \log(1 + \exp\{\beta x_{ij} + u_{i,m}\}) \right] \Bigg\}\end{aligned}$$

$$(\sigma^2)^{(\mathrm{new})} = \arg\max_{\sigma^2} \hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) = \frac{\sum_{m=1}^{M} \omega'_m \sum_{i=1}^{I} u_{i,m}^2}{I \sum_{m=1}^{M} \omega'_m}$$

$$\beta^{(\mathrm{new})} = \arg\max_{\beta} \hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\mathrm{old})}) = \frac{\sum_{m=1}^{M} \omega'_m \sum_i \sum_j \{y_{ij}\beta x_{ij} - \log(1 + \exp(\beta x_{ij} + u_{i,m}))\}}{\sum_{m=1}^{M} \omega'_m}$$

Logistic Regression with Random Effects

- We can update $\beta^{(\text{new})}$ by the iteratively reweighted least squares method (which is equivalent to one-step Newton–Raphson method).

$$\boldsymbol{\mu}_i^T(\beta, \boldsymbol{u}) = \left(\frac{1}{1 + \exp(-\beta x_{ij} - u_j)}\right)_{j=1}^{J}$$

$$\boldsymbol{W} = \operatorname{diag}\left[\operatorname{vec}(\boldsymbol{\mu}_i, i = 1 \ldots, I)\right]$$

$$\beta^{(\text{new})} = \beta^{(\text{old})} + \hat{E}[\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}]^{-1} \boldsymbol{X}^T (\boldsymbol{Y} - \hat{E}(\boldsymbol{\mu}_i^T(\beta^{(\text{old})}, \boldsymbol{u}), i = 1, \ldots, I$$

MCEM without importance sampling