Statistical Methods for Handling Incomplete Data Chapter 3: Computation

Jae-Kwang Kim

Department of Statistics, Iowa State University

Outline

1. Introduction
2. Factoring likelihood approach
3. EM algorithm
4. Monte Carlo computation
5. Monte Carlo EM

1. Introduction: Motivation

- Interested in finding the solution that

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Often the MLE can be computed from the score equation

$$S(\hat{\theta}) = 0$$

which is generally a system of nonlinear equations.

- How to solve the score equation?

Methods for solving nonlinear equations: $g(\theta) = 0$

1. Bisection method: Use the intermediate value theorem.
   "If g is continuous for all θ in the interval $g(\theta_1)\, g(\theta_2) < 0$. A root of $g(\theta) = 0$ lie in the interval (θ_1, θ_2)"
2. Method of false positions (or Secant method): Use a linear approximation

$$g(\theta) \cong g(a) + \frac{g(b) - g(a)}{b - a}(\theta - a)$$

to get

$$\theta = \frac{a g(b) - b g(a)}{g(b) - g(a)}.$$

Thus, the method of false positions can be defined as

$$\theta^{(t+2)} = \frac{\theta^{(t)} g\left(\theta^{(t+1)}\right) - \theta^{(t+1)} g\left(\theta^{(t)}\right)}{g\left(\theta^{(t+1)}\right) - g\left(\theta^{(t)}\right)}.$$

3. Newton's method (Or Newton-Raphson method): Use a linear approximation of $g(\theta)$ at $\theta^{(t)}$

$$g(\theta) \cong g\left(\theta^{(t)}\right) + \left[\frac{\partial g\left(\theta^{(t)}\right)}{\partial \theta}\right]\left(\theta - \theta^{(t)}\right).$$

Thus,

$$\theta^{(t+1)} = \theta^{(t)} - \left[\frac{\partial g\left(\theta^{(t)}\right)}{\partial \theta}\right]^{-1} g\left(\theta^{(t)}\right).$$

For score equation:

$$\theta^{(t+1)} = \theta^{(t)} + \left[I\left(\theta^{(t)}\right)\right]^{-1} S\left(\theta^{(t)}\right).$$

Other variants of Newton's method

1. Fisher scoring method: Use

$$\theta^{(t+1)} = \theta^{(t)} + \left[\mathcal{I}\left(\theta^{(t)}\right)\right]^{-1} S\left(\theta^{(t)}\right)$$

2. Ascent method:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha\left[\mathcal{I}\left(\theta^{(t)}\right)\right]^{-1} S\left(\theta^{(t)}\right)$$

for $\alpha \in (0, 1]$. If $L(\hat{\theta}^{(t+1)}) < L(\hat{\theta}^{(t)})$, then use $\alpha = \alpha/2$ and compute $\theta^{(t+1)}$ again.

3. Quasi-Newton method:

$$\theta^{(t+1)} = \theta^{(t)} - \left[M^{(t)}\right]^{-1} S\left(\theta^{(t)}\right)$$

where $M^{(t)}$ satisfies

$$S\left(\theta^{(t+1)}\right) - S\left(\theta^{(t)}\right) = M^{(t+1)}\left(\theta^{(t+1)} - \theta^{(t)}\right).$$

Example 3.1

Model

Logistic regression model

$$y_i \overset{i.i.d.}{\sim} Bernoulli\,(p_i)$$

with

$$\mathrm{logit}\,(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}.$$

Log-likelihood

$$\begin{aligned} \ln L\,(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left[y_i \ln\,(p_i) + (1-y_i)\ln\,(1-p_i)\right] \\ &= \sum_{i=1}^{n} \left[y_i\,(\mathbf{x}_i'\boldsymbol{\beta}) - \ln\,(1+\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right] \end{aligned}$$

Example 3.1 (Cont'd)

Score function

$$
\begin{aligned}
S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \{y_i - p_i(\boldsymbol{\beta})\} \mathbf{x}_i \\
I(\boldsymbol{\beta}) &= -\frac{\partial}{\partial \boldsymbol{\beta}'} S(\boldsymbol{\beta}) = \sum_{i=1}^{n} p_i(\boldsymbol{\beta})\{1 - p_i(\boldsymbol{\beta})\}\mathbf{x}_i\mathbf{x}_i'
\end{aligned}
$$

Newton-Raphson Method= Scoring method

$$
\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left[\sum_{i=1}^{n} p_i^{(t)}(1 - p_i^{(t)})\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \sum_{i=1}^{n}(y_i - p_i^{(t)})\mathbf{x}_i
$$

where

$$
p_i^{(t)} = p_i(\boldsymbol{\beta}^{(t)}).
$$

Order of convergence

Definition

Let θ^* be the unique solution to $g(\theta) = 0$. A sequence $\left\{\theta^{(t)}\right\}$ is converges to θ^* of order β if

$$\lim_{t \to \infty} \|\theta^{(t)} - \theta^*\| = 0$$

and

$$\lim_{t \to \infty} \frac{\|\theta^{(t+1)} - \theta^*\|}{\|\theta^{(t)} - \theta^*\|^{\beta}} = c$$

for some constants $c \neq 0$.

Remark

Result

Under the regularity conditions, the sequence obtained from Newton's method converges at a second order rate.

Sketched Proof:

By the second order Taylor expansion,

$$\begin{aligned} 0 &= g(\theta^*) \\ &\cong g\left(\theta^{(t)}\right) + \left\{\partial g\left(\theta^{(t)}\right)/\partial\theta\right\}\left(\theta^* - \theta^{(t)}\right) + \left\{\partial^2 g(q)/\partial\theta^2\right\}\left(\theta^* - \theta^{(t)}\right)^2/2 \end{aligned}$$

where q is between θ^* and $\theta^{(t)}$. Multiplying both sides of the above equation by $\left\{\partial g\left(\theta^{(t)}\right)/\partial\theta\right\}^{-1}$ and using the definition of the Newton method, we have

$$\frac{\theta^{(t+1)} - \theta^*}{(\theta^{(t)} - \theta^*)^2} = \frac{\partial^2 g(q)/\partial\theta^2}{2\partial g(\theta^{(t)})/\partial\theta}$$

Thus, the Lipschitz condition holds and

$$\lim_{t\to\infty} \frac{\|\theta^{(t+1)} - \theta^*\|}{\|\theta^{(t)} - \theta^*\|^\beta} = \left|\frac{g''(\theta^*)}{2g'(\theta^*)}\right| \neq 0.$$

Example 3.3 (Normal-theory random effects model)

Model

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + u_i + e_{ij}$$

where $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$, $e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_e^2)$, and e_{ij} are independent of u_i. The cluster-specific effect u_i is treated as random (Not observed).

Complete-sample likelihood

$$L_{\rm com}(\theta) = \prod_i \left[\left\{ \prod_{j=1}^{n_i} \frac{1}{\sigma_e} \phi\left(\frac{y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - u_i}{\sigma_e} \right) \right\} \frac{1}{\sigma_u} \phi\left(\frac{u_i}{\sigma_u} \right) \right]$$

where $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ is the probability density function of the standard normal distribution.

Example 3.3 (Cont'd)

Complete-sample Score functions

$$
\begin{aligned}
S_{\mathrm{com},1}(\theta) &\equiv \partial \log\{L_{\mathrm{com}}(\theta)\}/\partial\boldsymbol{\beta} = \sum_i \sum_j \left(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - u_i\right)\mathbf{x}_{ij}/\sigma_e^2 \\
S_{\mathrm{com},2}(\theta) &\equiv \partial \log\{L_{\mathrm{com}}(\theta)\}/\partial\sigma_u^2 = \frac{1}{2\sigma_u^4}\sum_i \left(u_i^2 - \sigma_u^2\right) \\
S_{\mathrm{com},3}(\theta) &\equiv \partial \log\{L_{\mathrm{com}}(\theta)\}/\partial\sigma_e^2 = \frac{1}{2\sigma_e^4}\sum_i \sum_j \left(e_{ij}^2 - \sigma_e^2\right),
\end{aligned}
$$

where $e_{ij} = y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - u_i$.

Example 3.3 (Cont'd)

Observed likelihood

$$L_{\text{obs}}(\theta) = \prod_i \left[\int \prod_{j=1}^{n_i} \left\{ \frac{1}{\sigma_e} \phi \left(\frac{y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - u_i}{\sigma_e} \right) \right\} \frac{1}{\sigma_u} \phi \left(\frac{u_i}{\sigma_u} \right) du_i \right].$$

Predictive distribution

$$u_i \mid \mathbf{x}_i, \mathbf{y}_i \sim N\left(\tau_i \left(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta} \right), \sigma_u^2 (1 - \tau_i) \right),$$

where $\tau_i = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_i)$.

Example 3.3 (Cont'd)

Observed score equation

$$\bar{S}_1(\beta) \equiv \sum_i \sum_j \left\{ y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - \tau_i \left(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta} \right) \right\} \mathbf{x}_{ij}/\sigma_e^2 = 0.$$

$$\bar{S}_2(\sigma_u^2) \equiv \frac{1}{2\sigma_u^4} \sum_i \left\{ \tau_i^2 \left(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta} \right)^2 - \tau_i \sigma_u^2 \right\} = 0.$$

$$\bar{S}_3(\sigma_e^2) \equiv \frac{1}{2\sigma_e^4} \sum_i \sum_j \left[\left\{ y_{ij} - \tau_i \bar{y}_i - (\mathbf{x}_{ij} - \tau_i \bar{\mathbf{x}}_i)'\boldsymbol{\beta} \right\}^2 - (1 - \frac{\tau_i}{n_i})\sigma_e^2 \right] = 0.$$

The resulting $\hat{\boldsymbol{\beta}}$ is obtained by the regression of $y_{ij} - \tau_i \bar{y}_i$ on $(\mathbf{x}_{ij} - \tau_i \bar{\mathbf{x}}_i)$. Fuller and Battese (1973) obtained the same result from the estimated generalized least square method.

Example 3.4 (Bivariate Normal distribution)

- Model

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right]$$

- Observation

  r complete observations $\{(x_i, y_i); i = 1, 2, \cdots, r\}$

  $n - r$ partial observations $\{x_i; i = r+1, r+2, \cdots, n\}$

  assume missing at random.

- The observed likelihood is

$$L_{obs}(\theta) = \prod_{i=1}^{r} f(x_i, y_i; \mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy}) \times \prod_{i=r+1}^{n} f(x_i; \mu_x, \sigma_{xx})$$

  Finding the MLE using direct maximization of the observed likelihood is computationally challenging.

Factoring likelihood approach (Anderson, 1957)

Idea: Use
"Joint pdf of (x, y) = (marginal pdf of x) $\times$ (conditional pdf of y given x)"
Alternative parametrization

$$\begin{aligned} X_i &\sim N(\mu_x, \sigma_{xx}) \\ Y_i \mid X_i = x &\sim N(\beta_0 + \beta_1 x, \sigma_{ee}) \end{aligned}$$

where

$$\begin{aligned} \beta_1 &= \sigma_{xy}/\sigma_{xx} \\ \beta_0 &= \mu_y - \beta_1 \mu_x \\ \sigma_{ee} &= \sigma_{yy} - \sigma_{xy}^2/\sigma_{xx}. \end{aligned}$$

Under the new parametrization,

$$\begin{aligned} L_{obs}(\theta) &= \prod_{i=1}^{n} f(x_i; \mu_x, \sigma_{xx}) \times \prod_{i=1}^{r} f(y_i \mid x_i; \beta_0, \beta_1, \sigma_{ee}) \\ &= L_1(\mu_x, \sigma_{xx}) \times L_2(\beta_0, \beta_1, \sigma_{ee}). \end{aligned}$$

Example 3.4 (Cont'd)

- The MLEs under the new parametrization are

$$\begin{aligned}\hat{\mu}_x &= \bar{x}_n \\ \hat{\sigma}_{xx} &= S_{xxn}\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}_1 &= S_{xyr}/S_{xxr} \\ \hat{\beta}_0 &= \bar{y}_r - \hat{\beta}_1\bar{x}_r \\ \hat{\sigma}_{ee} &= S_{yyr} - S_{xyr}^2/S_{xxr},\end{aligned}$$

where the subscript r denotes that the statistics are computed from the r respondents only and subscript n denotes that the statistics are computed from the whole sample of size n.

- Thus, the MLE's for the original parametrization are

$$\begin{aligned}\hat{\mu}_y &= \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}_x = \bar{y}_r + \hat{\beta}_1\left(\hat{\mu}_x - \bar{x}_r\right) \\ \hat{\sigma}_{yy} &= S_{yyr} + \hat{\beta}_1^2\left(\hat{\sigma}_{xx} - S_{xxr}\right) \\ \hat{\sigma}_{xy} &= S_{xyr}\frac{\hat{\sigma}_{xx}}{S_{xxr}}.\end{aligned}$$

Example 3.5 (Bivariate categorical distribution)

$$(Y_1, Y_2) = \begin{cases} (1,1) & \text{with prob. } \pi_{11} \\ (1,0) & \text{with prob. } \pi_{10} \\ (0,1) & \text{with prob. } \pi_{01} \\ (0,0) & \text{with prob. } \pi_{00} \end{cases}$$

Observation
r complete observations $\{(y_{1i}, y_{2i})\,; i = 1, 2, \cdots, r\}$
$n - r$ partial observations $\{y_{1i}; i = r+1, r+2, \cdots, n\}$

Observed likelihood for $\theta_1 = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$

Example 3.5 (Cont'd)

Alternative parametrization: $\theta_2 = (\pi_{1+}, \pi_{1|1}, \pi_{1|0})$ where

$$\begin{aligned} \pi_{1+} &= Pr(Y_1 = 1) \\ \pi_{1|1} &= Pr(Y_2 = 1 \mid Y_1 = 1) \\ \pi_{1|0} &= Pr(Y_2 = 1 \mid Y_1 = 0) \end{aligned}$$

Observed likelihood for θ_2

$$L_{\rm obs}(\theta_2) = \prod_{i=1}^{n} \pi_{1+}^{y_{1i}} (1 - \pi_{1+})^{1-y_{1i}} \times \prod_{i=1}^{r} \left\{ \pi_{1|1}^{y_{2i}} \left(1 - \pi_{1|1}\right)^{1-y_{2i}} \right\}^{y_{1i}} \left\{ \pi_{1|0}^{y_{2i}} \left(1 - \pi_{1|0}\right)^{1-y_{2i}} \right\}^{1-y_{1i}} .$$

Example 3.5 (Cont'd)

<u>MLE</u>

Because we can write

$$L_{\mathrm{obs}}(\pi_{1+}, \pi_{1|1}, \pi_{1|0}) = L_1(\pi_{1+})L_2(\pi_{1|1})L_3(\pi_{1|0})$$

for some $L_1(\cdot)$, $L_2(\cdot)$, and $L_3(\cdot)$, we can obtain the MLE by separately maximizing each likelihood component. Thus, we have

$$\begin{aligned} \hat{\pi}_{1+} &= \frac{1}{n}\sum_{i=1}^{n} y_{1i} \\ \hat{\pi}_{1|1} &= \frac{\sum_{i=1}^{r} y_{1i}y_{2i}}{\sum_{i=1}^{r} y_{1i}} \\ \hat{\pi}_{1|0} &= \frac{\sum_{i=1}^{r}(1-y_{1i})y_{2i}}{\sum_{i=1}^{r}(1-y_{1i})}. \end{aligned}$$

The MLE for π_{ij} can then be obtained by $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{j|i}$ for $i = 0, 1$ and $j = 0, 1$.

Remark

1. The factoring likelihood approach is particularly useful for *monotone missing patterns*, where we can relabel the variable in such a way that the set of respondents for each variable is monotonely nested:

$$R_1 \supset R_2 \supset \cdots \supset R_p$$

where R_i denotes the set of respondents for Y_i after relabeling. In this case, under MAR, the observed likelihood can be written as

$$L_{\mathrm{obs}}(\theta) = \prod_{i \in R_1} f(y_{1i}; \theta_1) \times \prod_{i \in R_2} f(y_{2i} \mid y_{1i}; \theta_2) \times \cdots \times \prod_{i \in R_p} f(y_{pi} \mid y_{p-1,i}; \theta_p)$$

and the MLE for each component of the parameters can be obtained by maximizing each component of the observed likelihood (Rubin, 1974).

2. For non-monotone missing data, we cannot directly apply the factoring likelihood method. Instead, we may use the GLS to combine the estimates.

Example (Bivariate Normal distribution with non-monotone missing pattern)

Same setup of Example 3.4, except that the missing pattern is now non-monotone.

- [Step 1] Partition the original sample into several disjoint sets according to the missing pattern.
- [Step 2] Compute the MLEs for the identified parameters separately in each partition of the sample.
- [Step 3] Combine the estimators to get a set of final estimates using a generalized least squares (GLS) form.

Step 1 Partition the original sample into H, K, L and M.

Table: **An illustration of the missing data structure under bivariate normal distribution**

Set	x	y	Sample Size	Estimable parameters
H	Observed	Observed	n_H	$\mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy}$
K	Observed	Missing	n_K	μ_x, σ_{xx}
L	Missing	Observed	n_L	μ_y, σ_{yy}
M	Missing	Missing	n_M	

Step 2: Compute MLE separately from each partition.

$$
\begin{aligned}
\hat{\theta}_H &= (\hat{\mu}_{x,H}, \hat{\mu}_{y,H}, \hat{\sigma}_{xx,H}, \hat{\sigma}_{xy,H}, \hat{\sigma}_{yy,H}) \\
\hat{\theta}_{x,K} &= (\hat{\mu}_{x,K}, \hat{\sigma}_{xx,K}) \\
\hat{\theta}_{y,L} &= (\hat{\mu}_{y,L}, \hat{\sigma}_{yy,L})
\end{aligned}
$$

Step 3 Combine using GLS

- Method 1: Use information matrix directly.

$$\hat{\theta} = \mathcal{I}_{\rm obs}^{-1} \left\{ \mathcal{I}_H \hat{\theta}_H + \mathcal{I}_K \hat{\theta}_K + \mathcal{I}_L \hat{\theta}_L \right\}$$

  where $\mathcal{I}_{\rm obs} = \mathcal{I}_H + \mathcal{I}_K + \mathcal{I}_L$, $\hat{\theta}_K = (\hat{\mu}_{x,K}, 0, \hat{\sigma}_{xx,K}, 0, 0)'$, and $\hat{\theta}_L = (0, \hat{\mu}_{y,L}, 0, 0, \hat{\sigma}_{yy,L})'$. The information matrices $\mathcal{I}_H$, $\mathcal{I}_K$, $\mathcal{I}_L$ are the expected Fisher information matrices of $\hat{\theta}_H$, $\hat{\theta}_K$, $\hat{\theta}_L$, respectively. For example, $\mathcal{I}_K = \text{diag}\left\{ n_K/\sigma_{xx}, 0, n_K/(2\sigma_{xx}^2), 0, 0 \right\}$.

- Method 2: Use Gauss-Newton method (Kim and Shin, 2012)

3.3 EM algorithm

Motivation

- Interested in finding $\hat{\eta}$ that maximizes $L_{obs}(\eta)$. The MLE can be obtained by solving $S_{obs}(\eta) = 0$, which is equivalent to solving $\bar{S}(\eta) = 0$ by Theorem 2.5.
- Computing the solution $\bar{S}(\eta) = 0$ can be challenging because it often involves computing $I_{obs}(\eta) = -\partial \bar{S}(\eta)/\partial \eta'$ in order to apply Newton method:

$$\hat{\eta}^{(t+1)} = \hat{\eta}^{(t)} + \left\{ I_{obs}(\hat{\eta}^{(t)}) \right\}^{-1} \bar{S}(\hat{\eta}^{(t)}).$$

  We may rely on Louis formula (Theorem 2.7) to compute $I_{obs}(\eta)$.

- EM algorithm provides an alternative method of solving $\bar{S}(\eta) = 0$ by writing

$$\bar{S}(\eta) = E\left\{S_{com}(\eta) \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta\right\}$$

and using the following iterative method:

$$\hat{\eta}^{(t+1)} \leftarrow \text{ solve } E\left\{S_{com}(\eta) \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \hat{\eta}^{(t)}\right\} = 0.$$

  - E-step: Compute the conditional expectation given the observed data evaluated at $\hat{\eta}^{(t)}$
  - M-step: Update the parameter by solving the above mean score equation.

EM algorithm

Definition

Let $\eta^{(t)}$ be the current value of the parameter estimate of η. The EM algorithm can be defined as iteratively carrying out the following E-step and M-steps:

- E-step: Compute

$$Q\left(\eta \mid \eta^{(t)}\right) = E\left\{\ln f\left(\mathbf{y}, \boldsymbol{\delta}; \eta\right) \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right\}$$

- M-step: Find $\eta^{(t+1)}$ that maximizes $Q(\eta \mid \eta^{(t)})$ w.r.t. η.

Theorem 3.2 (Dempster et al., 1977)

Let $L_{obs}(\eta) = \int f(\mathbf{y}, \boldsymbol{\delta}; \eta)\, d\mathbf{y}_{\rm mis}$ be the observed likelihood of η. If $Q(\eta^{(t+1)} \mid \eta^{(t)}) \geq Q(\eta^{(t)} \mid \eta^{(t)})$, then $L_{obs}(\eta^{(t+1)}) \geq L_{obs}(\eta^{(t)})$.

By Theorem 3.2, the sequence $\{L_{\rm obs}(\eta^{(t)})\}$ is monotone increasing and it is bounded above if the MLE exists. Thus, the sequence of $L_{\rm obs}(\eta^{(t)})$ converges to some value L^*. In most cases, L^* is a stationary value in the sense that $L^* = L_{\rm obs}(\eta^*)$ for some η^* at which $\partial L_{\rm obs}(\eta)/\partial \eta = 0$. Under fairly weak conditions, such as $Q(\eta \mid \gamma)$ satisfies

$$\partial Q(\eta \mid \gamma)/\partial \eta \text{ is continuous in } \eta \text{ and } \gamma,$$

the EM sequence $\{\eta^{(t)}\}$ converges to a stationary point η^*. (Wu, 1983)

Proof of Theorem 3.2

Writing

$$\begin{aligned} \ln L_{\text{obs}}(\boldsymbol{\eta}) &= \ln \int f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta}) d\mathbf{y}_{\text{mis}} \\ &= \ln E\left\{ \frac{f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta})}{f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta}^{(t)})} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta^{(t)} \right\} + \ln L_{\text{obs}}(\boldsymbol{\eta}^{(t)}), \end{aligned}$$

we have

$$\begin{aligned} \ln L_{\text{obs}}(\boldsymbol{\eta}) - \ln L_{\text{obs}}(\boldsymbol{\eta}^{(t)}) &= \ln E\left\{ \frac{f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta})}{f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta}^{(t)})} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta^{(t)} \right\} \\ &\geq E\left[\ln\left\{ \frac{f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta})}{f(\mathbf{y}, \boldsymbol{\delta}; \boldsymbol{\eta}^{(t)})} \right\} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta^{(t)} \right] \\ &= Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(t)}) - Q(\boldsymbol{\eta}^{(t)} \mid \boldsymbol{\eta}^{(t)}), \end{aligned}$$

where the above inequality follows from Lemma 2.1. Therefore, $Q(\eta^{(t+1)} \mid \eta^{(t)}) \geq Q(\eta^{(t)} \mid \eta^{(t)})$ implies $L_{obs}\left(\eta^{(t+1)}\right) \geq L_{obs}\left(\eta^{(t)}\right)$.

Example 3.8 (Mixture model)

- Observation

$$Y_i = (1 - W_i) Z_{1i} + W_i Z_{2i}, \qquad i = 1, 2, \cdots, n$$

where

$$\begin{aligned} Z_{1i} &\sim N\left(\mu_1, \sigma_1^2\right) \\ Z_{2i} &\sim N\left(\mu_2, \sigma_2^2\right) \\ W_i &\sim Bernoulli\,(\pi)\,. \end{aligned}$$

- Parameter of interest: $\theta = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi\right)$
- Observed likelihood

$$L_{obs}(\theta) = \prod_{i=1}^{n} \left\{ (1 - \pi)\,\phi\left(y \mid \mu_1, \sigma_1^2\right) + \pi\phi\left(y \mid \mu_2, \sigma_2^2\right) \right\}$$

where

$$\phi\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y - \mu)^2}{2\sigma^2}\right].$$

Example 3.8 (Cont'd)

Complete-sample likelihood

$$L_{com}(\theta) = \prod_{i=1}^{n} \text{pdf}(y_i, w_i \mid \theta)$$

where

$$\text{pdf}(y, w \mid \theta) = \left[\phi\left(y \mid \mu_1, \sigma_1^2\right)\right]^{1-w} \left[\phi\left(y \mid \mu_2, \sigma_2^2\right)\right]^{w} \pi^{w} (1-\pi)^{1-w}.$$

Thus,

$$\begin{aligned} \ln L_{com}(\theta) &= \sum_{i=1}^{n} \left[(1-w_i) \ln \phi\left(y_i \mid \mu_1, \sigma_1^2\right) + w_i \ln \phi\left(y_i \mid \mu_2, \sigma_2^2\right)\right] \\ &\quad + \sum_{i=1}^{n} \{w_i \ln(\pi) + (1-w_i) \ln(1-\pi)\} \end{aligned}$$

Example 3.8 (Cont'd)

[E-step]

$$
\begin{aligned}
Q\left(\theta \mid \theta^{(t)}\right) &= \sum_{i=1}^{n}\left[\left(1-r_i^{(t)}\right)\ln\phi\left(y_i \mid \mu_1,\sigma_1^2\right)+r_i^{(t)}\ln\phi\left(y_i \mid \mu_2,\sigma_2^2\right)\right] \\
&\quad +\sum_{i=1}^{n}\left\{r_i^{(t)}\ln(\pi)+\left(1-r_i^{(t)}\right)\ln(1-\pi)\right\}
\end{aligned}
$$

where $r_i^{(t)} = E\left(w_i \mid y_i, \theta^{(t)}\right)$ with

$$
E\left(w_i \mid y_i, \theta\right) = \frac{\pi\phi\left(y_i \mid \mu_2,\sigma_2^2\right)}{(1-\pi)\phi\left(y_i \mid \mu_1,\sigma_1^2\right)+\pi\phi\left(y_i \mid \mu_2,\sigma_2^2\right)}
$$

[M-step]

$$
\frac{\partial}{\partial\theta} Q\left(\theta \mid \theta^{(t)}\right) = 0.
$$

Convergence of EM algorithm is linear. It can be shown that

$$\eta^{(t+1)} - \eta^{(t)} \cong \mathcal{J}_{\rm mis}\left(\eta^{(t)} - \eta^{(t-1)}\right)$$

where $\mathcal{J}_{\rm mis} = \mathcal{I}_{\rm com}^{-1}\mathcal{I}_{\rm mis}$ is called the *fraction of missing information*. The fraction of missing information may vary across different components of $\eta^{(t)}$, suggesting that certain components of $\eta^{(t)}$ may approach η^* rapidly while other components may require many iterations. Roughly speaking, the rate of convergence of a vector sequence $\eta^{(t)}$ from the EM algorithm is given by the largest eigenvalue of the matrix $\mathcal{J}_{\rm mis}$.

Categorical Missing data

If $\mathbf{y}$ is a categorical variable that takes values in set S_y, then the E-step can be easily computed by a weighted summation

$$E\left\{\ln f(\mathbf{y}, \boldsymbol{\delta}; \eta) \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right\} = \sum_{\mathbf{y} \in S_y} P(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}) \ln f(\mathbf{y}, \boldsymbol{\delta}; \eta) \quad (1)$$

where the summation is over all possible values of $\mathbf{y}$ and $P\left(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right)$ is the conditional probability of taking $\mathbf{y}$ given $\mathbf{y}_{\mathrm{obs}}$ and $\boldsymbol{\delta}$ evaluated at $\eta^{(t)}$. The conditional probability $P\left(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}; \eta^{(t)}\right)$ can be treated as the weight assigned for the categorical variable $\mathbf{y}$. That is, if $S(\eta) = \sum_{i=1}^{n} S(\eta; \mathbf{y}_i, \delta_i)$ is the score function for η, then the EM algorithm using (1) can be obtained by solving

$$\sum_{i=1}^{n} \sum_{\mathbf{y} \in S_y} P\left(\mathbf{y}_i = \mathbf{y} \mid \mathbf{y}_{i,\mathrm{obs}}, \boldsymbol{\delta}_i, \eta^{(t)}\right) S(\eta; \mathbf{y}, \delta_i) = 0$$

for η to get $\eta^{(t+1)}$. Ibrahim (1990) called this approach *EM by weighting*.

- E-step:

$$\bar{S}_1\left(\beta \mid \beta^{(t)}, \phi^{(t)}\right) = \sum_{\delta_i=1}\left\{y_i - p_i(\beta)\right\}\mathbf{x}_i + \sum_{\delta_i=0}\sum_{j=0}^{1} w_{ij(t)}\left\{j - p_i(\beta)\right\}\mathbf{x}_i,$$

where

$$\begin{aligned} w_{ij(t)} &= Pr(Y_i = j \mid \mathbf{x}_i, \delta_i = 0; \beta^{(t)}, \phi^{(t)}) \\ &= \frac{Pr(Y_i = j \mid \mathbf{x}_i; \beta^{(t)})Pr(\delta_i = 0 \mid \mathbf{x}_i, j; \phi^{(t)})}{\sum_{y=0}^{1} Pr(Y_i = y \mid \mathbf{x}_i; \beta^{(t)})Pr(\delta_i = 0 \mid \mathbf{x}_i, y; \phi^{(t)})} \end{aligned}$$

and

$$\begin{aligned} \bar{S}_2\left(\phi \mid \beta^{(t)}, \phi^{(t)}\right) &= \sum_{\delta_i=1}\left\{\delta_i - \pi\left(\mathbf{x}_i, y_i; \phi\right)\right\}\left(\mathbf{x}_i', y_i\right)' \\ &\quad + \sum_{\delta_i=0}\sum_{j=0}^{1} w_{ij(t)}\left\{\delta_i - \pi_i\left(\mathbf{x}_i, j; \phi\right)\right\}\left(\mathbf{x}_i', j\right)'. \end{aligned}$$

- M-step:

  The parameter estimates are updated by solving

$$\left[\bar{S}_1\left(\beta \mid \beta^{(t)}, \phi^{(t)}\right), \bar{S}_2\left(\phi \mid \beta^{(t)}, \phi^{(t)}\right)\right] = (0, 0)$$

  for β and ϕ.

- Thus, the conditional expectation in the E-step can be computed using the weighted mean with weights $w_{ij(t)}$.
- Observed information matrix can also be obtained by the Louis formula (in Theorem 2.7) using the weighted mean in the E-step.

EM in the exponential family

Under MAR and for the exponential family of the distribution of the form

$$f(\mathbf{y};\theta) = b(\mathbf{y})\,exp\left\{\theta'\mathbf{T}(\mathbf{y}) - A(\theta)\right\}.$$

Under MAR, the E-step of the EM algorithm is

$$Q\left(\theta \mid \theta^{(t)}\right) = \text{ constant } + \theta' E\left\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\mathrm{obs}}, \theta^{(t)}\right\} - A(\theta) \tag{2}$$

and the M-step is

$$\frac{\partial}{\partial\theta} Q\left(\theta \mid \theta^{(t)}\right) = 0 \quad \Longleftrightarrow \quad E\left\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\mathrm{obs}}, \theta^{(t)}\right\} = \frac{\partial}{\partial\theta} A(\theta).$$

Because $\int f(\mathbf{y};\theta)\,d\mathbf{y} = 1$, we have

$$\frac{\partial}{\partial\theta} A(\theta) = E\left\{\mathbf{T}(\mathbf{y})\,;\theta\right\}.$$

Therefore, the M-step reduces to finding $\theta^{(t+1)}$ as a solution to

$$E\left\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\mathrm{obs}}, \theta^{(t)}\right\} = E\left\{\mathbf{T}(\mathbf{y}) \mid \theta\right\}. \tag{3}$$

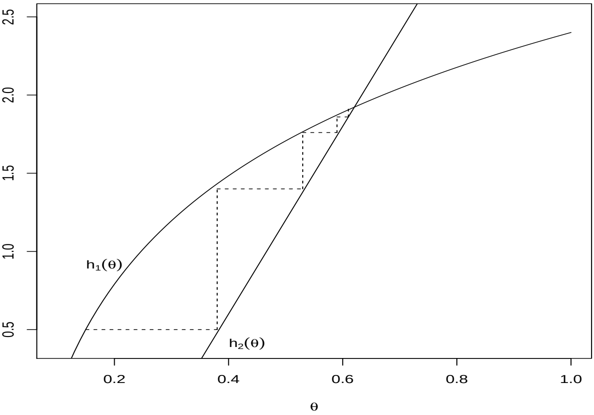

Figure: Illustration of EM algorithm for exponential family ($h_1(\theta) = E\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\rm obs}, \theta\}$, $h_2(\theta) = E\{\mathbf{T}(\mathbf{y}) \mid \theta\}$)

Example 3.9 (Bivariate Normal distribution)

- Model

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right]$$

- Sufficient statistics

$$S = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n y_i, \sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i y_i, \sum_{i=1}^n y_i^2\right)$$

- The EM algorithm reduces to solving

$$\sum_{i=1}^n E\left\{\left(x_i, y_i, x_i^2, x_i y_i, y_i^2\right) \mid \delta_i^{(x)}, \delta_i^{(y)}, \delta_i^{(x)} x_i, \delta_i^{(y)} y_i; \theta^{(t)}\right\}$$
$$= \sum_{i=1}^n E\left\{\left(x_i, y_i, x_i^2, x_i y_i, y_i^2\right); \theta\right\}$$

for θ. Under MAR, the above conditional expectation can be obtained using the usual conditional expectation under normality.

Back to Example 3.6

Table: **A 2 × 2 table with supplemental margins for both variables**

Set	y_1	y_2	Count
H	1	1	100
	1	2	50
	2	1	75
	2	2	75
K	1		30
	2		60
L		1	28
		2	60

Example 3.6 (Cont'd)

- The parameters of interest are $\pi_{ij} = P\left(Y_1 = i,\ Y_2 = j\right),\ i = 1, 2, j = 1, 2.$
- The sufficient statistics for the parameters are $n_{ij}, i = 1, 2; j = 1, 2$, where n_{ij} is the sample size for the set with $Y_1 = i$ and $Y_2 = j$.
- The E-step computes the conditional expectation of the sufficient statistics. This gives

$$n_{ij}^{(t)} = E\left(n_{ij} \mid \text{data}, \pi_{ij}^{(t)}\right) = n_{ij,H} + n_{i+,K}\frac{\pi_{ij}^{(t)}}{\pi_{i+}^{(t)}} + n_{+j,L}\frac{\pi_{ij}^{(t)}}{\pi_{+j}^{(t)}},$$

  for $i = 1, 2; j = 1, 2$.
- In the M-step, the parameters are updated by $\pi_{ij}^{(t+1)} = n_{ij}^{(t)}/n$.

R program for EM algorithm in Example 3.6

```
> setH=matrix(c(100,75,50,75),2,2)
> setK=c(30,60)
> setL=c(28,60)
> th=prop.table(setH) #initial estimates of pi from setH
> round(th,3)
      [,1]  [,2]
[1,] 0.333 0.167
[2,] 0.250 0.250
> nij=matrix(nrow=2,ncol=2)
> repeat{
+ th0=th
+ #E-step
+ for(i in 1:2){
+ for(j in 1:2){
+ nij[i,j]=setH[i,j]+setK[i]*th[i,j]/sum(th[i,])+setL[j]*th[i,j]/sum(th[,j])
+ }}
+ #M-step
+ th=nij/n
+ dif=sum((th0-th)^2)
+ if(dif<1e-8) break}
> round(th,3)
      [,1]  [,2]
[1,] 0.279 0.174
[2,] 0.239 0.308
```

Example 3.12

- Model: $x_i = \mu + \sigma e_i$ with $e_i \overset{indep}{\sim} t(\nu)$, ν: known.
- Missing data setup:
$$e_i = u_i / \sqrt{w_i}$$
where
$$x_i \mid w_i \sim N\left(\mu, \sigma^2 / w_i\right), \quad w_i \sim \chi^2_\nu / \nu.$$
- (x_i, w_i): complete data
- x_i always observed, w_i always missing
- Parameter: $\theta = (\mu, \sigma)$

Example 3.12 (Cont'd)

E-step: Find the conditional distribution of w_i given x_i. By Bayes theorem,

$$\begin{aligned} f(w_i \mid x_i) &\propto f(w_i) f(x_i \mid w_i) \\ &\propto (w_i\nu)^{\frac{\nu}{2}-1} \exp\left(-\frac{w_i\nu}{2}\right) \times \left(\sigma^2/w_i\right)^{-1/2} \exp\left\{-\frac{w_i}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right\} \\ &\sim \text{Gamma}\left[\frac{\nu+1}{2}, 2\left\{\nu + \left(\frac{x_i-\mu}{\sigma}\right)^2\right\}^{-1}\right]. \end{aligned}$$

Thus, the E-step of EM algorithm can be written as

$$E(w_i \mid x_i, \theta^{(t)}) = \frac{\nu+1}{\nu + \left(d_i^{(t)}\right)^2},$$

where $d_i^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)}$.

Example 3.12 (Cont'd)

M-step:

$$\begin{aligned}
\mu^{(t+1)} &= \frac{\sum_{i=1}^{n} w_i^{(t)} x_i}{\sum_{i=1}^{n} w_i^{(t)}} \\
\sigma^{2(t+1)} &= \frac{1}{n} \sum_{i=1}^{n} w_i^{(t)} \left(x_i - \mu^{(t+1)} \right)^2
\end{aligned}$$

where $w_i^{(t)} = E(w_i \mid x_i, \theta^{(t)})$.

Return to Example 3.3

- Consider the setup of Example 3.3, random effect model,

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + a_i + e_{ij}, \quad i = 1, \cdots, n_1, j = 1, \cdots, n_2,$$

  where $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.
- Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_2})'$ be observed, but a_i is never observed.
- The joint density of $(\mathbf{y}_i, a_i)$ is

$$f(\mathbf{y}_i, a_i; \theta) = f_1(\mathbf{y}_i \mid a_i; \boldsymbol{\beta}, \sigma_e^2) f_2(a_i; \sigma_a^2)$$

  where

$$\begin{aligned} f_1(\mathbf{y}_i \mid a_i; \boldsymbol{\beta}, \sigma_e^2) &= (2\pi\sigma_e^2)^{-n_2/2} \exp\left\{ -\frac{1}{2\sigma_e^2} \sum_j (y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - a_i)^2 \right\} \\ f_2(a_i; \sigma_a^2) &= (2\pi\sigma_a^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_a^2} a_i^2 \right\}. \end{aligned}$$

Return to Example 3.3 (Cont'd)

- EM algorithm:
  - E-step: Compute the conditional expectation of the score functions given the observed data:

    $$E\{S(\theta) \mid \mathbf{y}; \hat{\theta}^{(t)}\}.$$

    When both f_1 and f_2 are normal, then the above conditional distribution is also normal

    $$a_i \mid \mathbf{y}_i \sim N\left(\tau_i\left(\bar{y}_i - \bar{\mathbf{x}}_i'\beta\right), \sigma_u^2(1-\tau_i)\right), \tag{4}$$

    where $\tau_i = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_i)$.
  - M-step: Update the parameter by solving

    $$E\{S(\theta) \mid \mathbf{y}; \hat{\theta}^{(t)}\} = 0$$

    for θ, where the conditional expectation is computed from the E-step.
- If the normality does not hold either in f_1 or in f_2, then (4) is not necessarily normal. In this case, E-step may involve Monte Carlo approximation.

3.4 Monte Carlo method

Basic Setup

We want to compute the expectation of a function of a (continuous) random variable, say $\theta \equiv E\{h(X)\} = \int h(x) f(x)\, dx$, where $f(x)$ is the pdf of a random variable X. The Monte Carlo approximation of θ is to use

$$\hat{\theta}_{MC} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

where $X_1, \cdots, X_n$ are IID with pdf $f(x)$.

Properties

1. $\hat{\theta}_{MC}$ converges to θ as $n \to \infty$.
2. The variance of $\hat{\theta}_{MC}$ is $n^{-1}\sigma^2$ where $\sigma^2 = Var\left\{h\left(X\right)\right\}$.
3. Thus, $\hat{\theta}_{MC} - \theta = O_p\left(n^{-1/2}\right)$.

Approaches for Monte Carlo simulation

1. Probability integral transformation approach:
   For any continuous distribution function F, if $U \sim \mathit{Unif}\,(0,1)$, then $X = F^{-1}(U)$ has cdf equal to F.
2. Rejection sampling method
3. Importance sampling

Rejection sampling method

Given a density of interest f, suppose that there exist a density g and a constant M such that

$$f(x) \leq Mg(x)$$

on the support of f. The rejection sampling method (or accept-rejection method) is

1. Sample $Y \sim g$ and $U \sim Unif(0, 1)$.
2. Reject Y if
$$U > \frac{f(Y)}{Mg(Y)}.$$
In this case, do not record the value of Y as an element in the target random sample. Instead, return to step 1.
3. Otherwise, keep the value of Y. Set $X = Y$, and consider X to be an element of the target random sample.

Write $\theta \equiv \int h(x) f(x)\, dx = \int h(x) \frac{f(x)}{g(x)} g(x)\, dx$ for some density $g(x)$ and approximate θ by

$$\hat{\theta} = \sum_{i=1}^{n} w_i h(X_i)$$

where

$$w_i = \frac{f(X_i)/g(X_i)}{\sum_{j=1}^{n} f(X_j)/g(X_j)}$$

and $X_1, \cdots, X_n$ are IID with pdf $g(x)$.

Remark

1. In the rejection sampling method,

$$
\begin{aligned}
P(Y \le y) &= P\left[X \le y \mid U \le \frac{f(X)}{Mg(X)}\right] \\
&= \frac{\int_{-\infty}^{y} \int_{0}^{f(x)/Mg(x)} du g(x)\, dx}{\int_{-\infty}^{\infty} \int_{0}^{f(x)/Mg(x)} du g(x)\, dx} \\
&= \frac{\int_{-\infty}^{y} f(x)\, dx}{\int_{-\infty}^{\infty} f(x)\, dx}
\end{aligned}
$$

2. The rejection sampling method can be applicable when the density f is known up to a multiplicative factor.

Markov Chain Monte Carlo (MCMC) method

What is MCMC ?

- Markov Chain Monte Carlo: A body of methods for generating pseudorandom draws from probability distributions via Markov chains
- Markov chain: A sequence of random variables in which the distribution of each element depends only the previous one:

$$\{X_t; t = 1, 2, \cdots\}$$

where

$$P\left(X_t \mid X_0, X_1, \cdots, X_{t-1}\right) = P\left(X_t \mid X_{t-1}\right).$$

- "Today is the tomorrow of yesterday".

History of MCMC

1. Metropolis et al (1953): algorithm for indirect simulation of energy distributions
2. Hastings (1970): extension of Metropolis to a non-symmetric jumping distributions
3. Geman and Geman (1984): the "Gibbs sampler" for Bayesian image reconstruction
4. Tanner and Wong (1987): data augmentation for Bayesian inference in generic missing-data problems
5. Gelfand and Smith (1990): simulation of marginal distributions by repeated draws from conditionals

Use of MCMC

- Enables simulation of distributions that are known up to proportionality constant but are otherwise intractable
- Especially useful in Bayesian statistics, where information about parameters is summarized in a posterior distribution

$$P(\theta \mid \text{ data }) \propto P(\theta)\, P(\text{ data } \mid \theta)$$

- Helpful in generic missing-data problems: closely resembles EM

Basic setup for MCMC

- Z: generic random vector with density $f(Z)$
- $f(Z)$: difficult to simulate directly
- Idea: construct a Markov chain $\left\{Z^{(t)}; t = 1, 2\cdots\right\}$ with f as its stationary distribution,

$$P\left(Z^{(t)}\right) \to f \quad \text{as } t \to \infty$$

or

$$\frac{1}{N}\sum_{t=1}^{N} h\left(Z^{(t)}\right) \to E_f\left[h(Z)\right] = \int h(z)\, f(z)\, dz \qquad (*)$$

as $N \to \infty$.

- A Markov chain that satisfies (*) is called ergodic.

Gibbs sampling

Idea: Sample from conditional distributions

Given $Z^{(t)} = \left(Z_1^{(t)}, Z_2^{(t)}, \cdots, Z_J^{(t)}\right)$, draw $Z^{(t+1)}$ by sampling from the full conditionals of f,

$$\begin{aligned}
Z_1^{(t+1)} &\sim P\left(Z_1 \mid Z_2^{(t)}, Z_3^{(t)}, \cdots, Z_J^{(t)}\right) \\
Z_2^{(t+1)} &\sim P\left(Z_2 \mid Z_1^{(t)}, Z_3^{(t)}, \cdots, Z_J^{(t)}\right) \\
&\vdots \\
Z_J^{(t+1)} &\sim P\left(Z_J \mid Z_1^{(t)}, Z_2^{(t)}, \cdots, Z_{J-1}^{(t)}\right).
\end{aligned}$$

Under mild regularity conditions, $P\left(Z^{(t)}\right) \to f$ as $t \to \infty$.

Example

Suppose $Z = (Z_1, Z_2)'$ is bivariate normal,

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

The Gibbs sampler would be

$$\begin{aligned} Z_1 &\sim N\left(\rho Z_2, 1-\rho^2\right) \\ Z_2 &\sim N\left(\rho Z_1, 1-\rho^2\right). \end{aligned}$$

After a suitably large "burn-in period" we would find that

$$\begin{aligned} Z_1^{(t+1)}, Z_1^{(t+2)}, \cdots, Z_1^{(t+n)} &\sim N(0,1) \\ Z_2^{(t+1)}, Z_2^{(t+2)}, \cdots, Z_2^{(t+n)} &\sim N(0,1). \end{aligned}$$

But, if $\rho \neq 0$ the samples are dependent.

Metropolis-Hastings algorithm

Basic setup:

- Let $f(Z)$ be a distribution on R^k known except for the normalizing constant.
- The aim is to generate $Z \sim f$
- Direct generation from f is difficult but generating from $q(Z \mid Z^{(t-1)})$ is easy.

Metropolis-Hastings algorithm: Algorithm

1. Choose a transition function $q(y \mid x)$ of a certain Markov chain.
2. Initialize $Z^{(0)}$.
3. For $i = 1$ to N
   1. Generate $\tilde{Z} \sim q\left(Z \mid Z^{(i-1)}\right)$
   2. With probability

$$\rho\left(Z^{(i-1)}, \tilde{Z}\right) = \min\left\{\frac{q\left(Z^{(i-1)} \mid \tilde{Z}\right)}{q\left(\tilde{Z} \mid Z^{(i-1)}\right)} \frac{f\left(\tilde{Z}\right)}{f\left(Z^{(i-1)}\right)}, 1\right\},$$

set

$$Z^{(i)} = \tilde{Z} \quad \text{(accept)}$$

else set

$$Z^{(i)} = Z^{(i-1)} \quad \text{(reject)}.$$

Remark

- The normalizing constant in $f(Z)$ is not required in the MH algorithm.
- If $q(y \mid x) = f(y)$, then we obtain independent samples.
- Usually, q is chosen so that $q(y \mid x)$ is easy to sample from. (Theoretically any density $q(\cdot \mid x)$ having the same support as $f(\cdot)$ should work.)
- In the independent chain where $q(Z^* \mid Z^{(t)}) = q(Z^*)$, the Metropolis-Hastings ratio is

$$R\left(Z^*, Z^{(t)}\right) = \frac{f(Z^*)/q(Z^*)}{f(Z^{(t)})/q(Z^{(t)})},$$

  which is the ratio of the importance weight for Z^* over the importance weight for $Z^{(t)}$. Thus, the Metropolis-Hastings ratio $R\left(Z^*, Z^{(t)}\right)$ is also called the importance ratio.

Remark (Cont'd)

- The basic idea of the MH algorithm is
  - from the current position x, move to y according to $q(y \mid x)$ and
  - we decide to stay at y, roughly speaking, with probability $f(y)/f(x)$.
- Hence, $q(y \mid x)$ having more mass when $f(y)$ is larger is a good candidate.

Example 3.13 (Normal-Cauchy model)

- Let $Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} N(\theta, 1)$
- Prior: Cauchy distribution

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)} \tag{5}$$

- Posterior

$$\begin{aligned} \pi(\theta \mid y) &\propto exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - \theta)^2}{2} \right\} \times \frac{1}{1+\theta^2} \\ &\propto exp\left\{ -\frac{n(\theta - \bar{y})^2}{2} \right\} \times \frac{1}{1+\theta^2} \end{aligned}$$

- We want to generate $\theta \sim \pi(\theta \mid y)$.

Example 3.13 (Normal-Cauchy model)

- MH algorithm

1. Generate θ^* from Cauchy (0,1).
2. Given $y_1, \cdots, y_n$, compute the importance ratio

$$R\left(\theta^*, \theta^{(t)}\right) = \frac{\pi(\theta^* \mid y)/\pi(\theta^*)}{\pi(\theta^{(t)} \mid y)/\pi(\theta^{(t)})} = \frac{f\left(y \mid \theta^*\right)}{f\left(y \mid \theta^{(t)}\right)}$$

   where $f(y \mid \theta) = C \exp\left\{-n\left(\theta - \bar{y}\right)^2/2\right\}$ and $\pi(\theta)$ is defined in (5).
3. Accept θ^* as $\theta^{(t+1)}$ with probability $\rho(\theta^{(t)}, \theta^*) = \min\left\{R\left(\theta^{(t)}, \theta^*\right), 1\right\}$.

3.5 Monte Carlo EM

Motivation

1. In the mean score approach, the MLE can be found by solving

$$E\left\{S\left(\eta\right) \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}\right\} = 0$$

which requires the knowledge of the conditional distribution of $\mathbf{y}_{mis}$ given $\mathbf{y}_{obs}$ and $\boldsymbol{\delta}$.

2. In the EM algorithm defined by
   - [E-step] Compute

$$Q\left(\eta \mid \eta^{(t)}\right) = E\left\{\ln f\left(\mathbf{y}, \boldsymbol{\delta}; \eta\right) \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right\}$$

   - [M-step] Find $\eta^{(t+1)}$ that maximizes $Q\left(\eta \mid \eta^{(t)}\right)$,

E-step is computationally cumbersome because it involves integral.

Monte Carlo EM (MCEM) method (Wei and Tanner, 1990)

In the E-step, first draw

$$\mathbf{y}_1, \cdots, \mathbf{y}_m \overset{i.i.d.}{\sim} p\left(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right)$$

and approximate

$$Q\left(\eta \mid \eta^{(t)}\right) \cong \frac{1}{m} \sum_{j=1}^{m} \ln f\left(\mathbf{y}_j, \boldsymbol{\delta}; \eta\right).$$

Example 3.14 (Nonignorable missing)

$$y_i \sim f\left(y_i \mid x_i; \theta\right)$$

Assume that x_i is always observed but we observe y_i only when $\delta_i = 1$ where $\delta_i \sim Bernoulli\left[\pi_i\left(\phi\right)\right]$ and

$$\pi_i\left(\phi\right) = \frac{\exp\left(\phi_0 + \phi_1 x_i + \phi_2 y_i\right)}{1 + \exp\left(\phi_0 + \phi_1 x_i + \phi_2 y_i\right)}.$$

To implement the MCEM method, we need to generate samples from

$$f\left(y_i \mid x_i, \delta_i = 0; \hat{\theta}, \hat{\phi}\right) = \frac{f\left(y_i \mid x_i; \hat{\theta}\right)\left[1 - \pi_i\left(\hat{\phi}\right)\right]}{\int f\left(y_i \mid x_i; \hat{\theta}\right)\left[1 - \pi_i\left(\hat{\phi}\right)\right] dy_i}$$

Example 3.14 (Cont'd)

We can use the following rejection method to generate m Monte Carlo samples from $f(y_i \mid x_i, \delta_i = 0; \hat{\theta}, \hat{\phi})$:

1. Generate y_i^* from $f(y_i \mid x_i; \hat{\theta})$.
2. Using y_i^*, compute
$$\pi_i^*\left(\hat{\phi}\right) = \frac{\exp\left(\hat{\phi}_0 + \hat{\phi}_1 x_i + \hat{\phi}_2 y_i^*\right)}{1 + \exp\left(\hat{\phi}_0 + \hat{\phi}_1 x_i + \hat{\phi}_2 y_i^*\right)}.$$
3. Accept y_i^* with probability $1 - \pi_i^*(\hat{\phi})$. Otherwise, goto Step 1.

Example 3.14 (Cont'd)

M-step: Update the parameters by solving

$$\sum_{i=1}^{n} \sum_{j=1}^{m} S\left(\theta; x_i, y_i^{*(j)}\right) = 0$$

and

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \left\{\delta_i - \pi(\phi; x_i, y_i^{*(j)})\right\} \left(1, x_i, y_i^{*(j)}\right) = 0,$$

where $S(\theta; x_i, y_i) = \partial \log f(y_i \mid x_i; \theta)/\partial\theta$.

Example 3.18 (GLMM)

- Basic Setup: Let y_{ij} be a binary random variable (that takes 0 or 1) with probability $p_{ij} = Pr\left(y_{ij} = 1 \mid x_{ij}, a_i\right)$ and we assume that

$$\text{logit}\left(p_{ij}\right) = \mathbf{x}_{ij}'\boldsymbol{\beta} + a_i$$

where $\mathbf{x}_{ij}$ is a p-dimensional covariate associate with j-th repetition of unit i, $\boldsymbol{\beta}$ is the parameter of interest that can represent the treatment effect due to $\mathbf{x}$, and a_i represents the random effect associate with unit i. We assume that a_i are iid with $N\left(0, \sigma^2\right)$.

- Missing data : a_i

- Observed likelihood:

$$L_{obs}\left(\boldsymbol{\beta}, \sigma^2\right) = \prod_i \int \left\{ \prod_j p(x_{ij}, a_i; \boldsymbol{\beta})^{y_{ij}} \left[1 - p(x_{ij}, a_i; \boldsymbol{\beta})\right]^{1-y_{ij}} \right\} \frac{1}{\sigma}\phi\left(\frac{a_i}{\sigma}\right) da_i$$

where $\phi\left(\cdot\right)$ is the pdf of the standard normal distribution.

Example 3.18 (GLMM)

- MCEM approach: generate a_i^* from

$$f(a_i \mid \mathbf{x}_i, \mathbf{y}_i; \hat{\boldsymbol{\beta}}, \hat{\sigma}) \propto f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i; \hat{\boldsymbol{\beta}}) f_2\left(a_i; \hat{\sigma}\right).$$

  To do this, we first generate a_i^* from $f_2\left(a_i; \hat{\sigma}\right)$ and then accept it with probability proportional to $f_1\left(\mathbf{y}_i \mid \mathbf{x}_i, a_i^*; \hat{\boldsymbol{\beta}}\right)$.

- M-H algorithm: Choose $q\left(a_i \mid a_i^{(t-1)}\right) = f_2\left(a_i; \hat{\sigma}\right)$. Then, we accept $\tilde{a}_i$ from $f_2\left(a_i; \hat{\sigma}\right)$ with probability

$$\rho\left(a_i^{(t-1)}, \tilde{a}_i\right) = \min\left\{\frac{f_1\left(\mathbf{y}_i \mid \mathbf{x}_i, \tilde{a}_i; \hat{\boldsymbol{\beta}}\right)}{f_1\left(\mathbf{y}_i \mid \mathbf{x}_i, a_i^{(t-1)}; \hat{\boldsymbol{\beta}}\right)}, 1\right\}.$$

REFERENCES

Anderson, T. W. (1957), 'Maximum likelihood estimates for the multivariate normal distribution when some observations are missing', *Journal of the American Statistical Association* **52**, 200–203.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B* **39**, 1–38.

Fuller, W. A. and G. E. Battese (1973), 'Transformations for estimation of linear models with nested-error structure', *Journal of the American Statistical Association* **68**, 626–632.

Ibrahim, J. G. (1990), 'Incomplete data in generalized linear models', *Journal of the American Statistical Association* **85**, 765–769.

Kim, J. K. and D. W. Shin (2012), 'The factoring likelihood method for non-monotone missing data', *Journal of the Korean Statistical Society* **41**, 375–386.

Rubin, D. B. (1974), 'Characterizing the estimation of parameters in incomplete data problems', *Journal of the American Statistical Association* **69**, 467–474.

Wei, G. C. and M. A. Tanner (1990), 'A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms', *Journal of the American Statistical Association* **85**, 699–704.

Wu, C. F. J. (1983), 'On the convergence properties of the EM algorithm', *The Annals of Statistics* **11**, 95–103.