3.4 - 3.5 Monte Carlo approaches to EM

Motivation

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

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

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

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

$$Q\left(\eta \mid \eta^{(t)}\right) = E\left\{\ln L_{\text{com}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right\}$$

   - [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

$$\mathbf{y}_{\mathrm{mis}}^{*(1)}, \cdots, \mathbf{y}_{\mathrm{mis}}^{*(m)} \overset{iid}{\sim} p\left(\mathbf{y}_{\mathrm{mis}} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right) = \frac{f(\mathbf{y}, \boldsymbol{\delta}; \eta^{(t)})}{\int f(\mathbf{y}, \boldsymbol{\delta}; \eta^{(t)}) d\mathbf{y}_{\mathrm{mis}}}$$

and approximate

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

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

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

How to generate samples from (1)?

Rejection sampling method

Problem

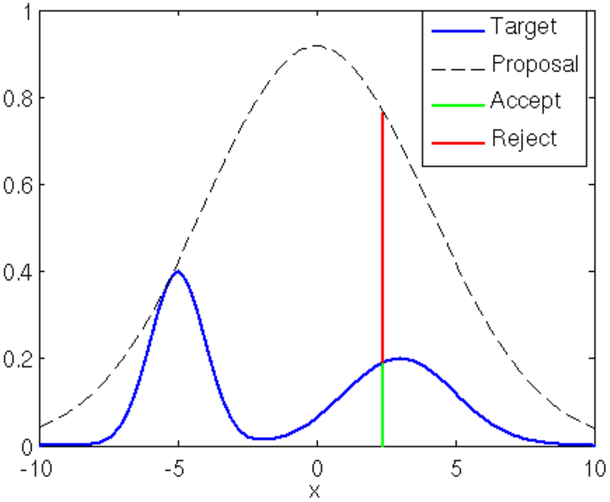
- Given a density of interest f, suppose that there exist a density g and a constant M such that
$$f(x) \leq Mg(x)$$
on the support of f.
- We are interested in generating samples from f, which is difficult. But, generating samples from g is easy.

Rejection sampling method

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.

Rejection sampling method



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.

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

Importance sampling

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 $g(x)$. The weight w_i is called important weight.

(Details are skipped. Will be covered again in Chapter 6)

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

Basic setup for MCMC

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

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

or

$$\frac{1}{N}\sum_{t=1}^{N} h\left(X^{(t)}\right) \rightarrow E_f\left[h(X)\right] = \int h(x) f(x)\, dx \tag{2}$$

as $N \rightarrow \infty$.

- A Markov chain that satisfies (2) is called ergodic.

Metropolis-Hastings algorithm: Algorithm

Starting with $X^{(0)}$ iterate for $t = 1, 2, \cdots$

1. Draw $X^* \sim q\left(\cdot \mid X^{(t-1)}\right)$.
2. Compute
$$R\left(X^*, X^{(t-1)}\right) = \frac{q\left(X^{(t-1)} \mid X^*\right)}{q\left(X^* \mid X^{(t-1)}\right)} \frac{f\left(X^*\right)}{f\left(X^{(t-1)}\right)}$$
and
$$\rho\left(X^*, X^{(t-1)}\right) = \min\{R\left(X^*, X^{(t-1)}\right), 1\}.$$
3. With probability $\rho\left(X^*, X^{(t-1)}\right)$ set $X^{(t)} = X^*$. Otherwise set $X^{(t)} = X^{(t-1)}$.

Remark

- The probability of acceptance does not depend on the normalization constant: If $f(x) = C \cdot \pi(x)$, then

$$R\left(X^* \mid X^{(t-1)}\right) = \frac{q\left(X^{(t-1)} \mid X^*\right)}{q\left(X^* \mid X^{(t-1)}\right)} \frac{\pi\left(X^*\right)}{\pi\left(X^{(t-1)}\right)}.$$

- Usually, q is chosen so that $q\left(y \mid x\right)$ is easy to sample from.
- The Markov chain is irreducible if $q(X \mid X^{(t-1)}) > 0$ for all $X, X^{(t-1)} \in \operatorname{supp}(f)$: every state can be reached in a single step.

Remark (Cont'd)

- In the independent chain where $q(X^* \mid X^{(t)}) = q(X^*)$, the Metropolis-Hastings ratio is

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

  which is the ratio of the importance weight for X^* over the importance weight for $X^{(t)}$. Thus, the Metropolis-Hastings ratio $R\left(X^*, X^{(t)}\right)$ is also called the importance ratio.
- The basic idea of the MH algorithm is
  - from the current position x, move to y according to $q\left(y \mid x\right)$ and
  - we decide to stay at y, roughly speaking, with probability $f\left(y\right)/f\left(x\right)$.
- Hence, $q\left(y \mid x\right)$ having more mass when $f\left(y\right)$ is larger is a good candidate.

Example 3.14 (Normal-Cauchy model)

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

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)} \tag{3}$$

- 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 (3).
  3. Accept θ^* as $\theta^{(t+1)}$ with probability $\rho(\theta^{(t)}, \theta^*) = \min\left\{R\left(\theta^{(t)}, \theta^*\right), 1\right\}$.

Random-walk Metropolis: Idea

- In the Metropolis-Hastings algorithm the proposal is from $X \sim q\left(\cdot \mid X^{(t-1)}\right)$.
- A popular choice for the proposal is $q(X \mid X^{(t-1)}) = g(X - X^{(t-1)})$ with g being a symmetric distribution. That is,

$$X = X^{(t-1)} + \epsilon, \quad \epsilon \sim g.$$

- The Metropolis-Hastings ratio becomes

$$R\left(X^*, X^{(t)}\right) = \frac{g\left(X^{(t-1)} - X^*\right)}{g\left(X^* - X^{(t-1)}\right)} \frac{f\left(X^*\right)}{f\left(X^{(t-1)}\right)} = \frac{f\left(X^*\right)}{f\left(X^{(t-1)}\right)}.$$

- We accept
    - every move to a more probable state with probability 1.
    - Moves to less probable states with a probability $f(X^*)/f(X^{(t-1)}) < 1$.

Example 3.18 (GLMM)

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

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

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

- To apply EM algorithm, note that a_i are observed in the complete sample log-likelihood

$$\ell_{\text{com}}(\beta, \sigma^2) = \ell_{\text{com},1}(\beta) + \ell_{\text{com},2}(\sigma^2)$$

where

$$\ell_{\text{com},1}(\beta) = \sum_i \sum_j \left[y_{ij} \log p_{ij}(\beta) + (1 - y_{ij}) \log\{1 - p_{ij}(\beta)\} \right]$$

and

$$\ell_{\text{com},2}(\sigma^2) = -\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} a_i^2$$

- Thus, in the E-step, we need to compute

$$Q(\theta \mid \theta^{(t)}) = E\left\{ \ell_{\text{com}}(\beta, \sigma^2) \mid \text{data}, \theta^{(t)} \right\}$$

where the conditional expectation is with respect to $f(a_i \mid \mathbf{x}_i, \mathbf{y}_i; \theta^{(t)})$.

- MCEM approach: Target distribution is

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

- M-H algorithm 1: Generate a_i^* from $f_2\left(a_i; \hat{\sigma}\right)$. Then, we accept a_i^* with probability

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

M-H algorithm 2: Random Walk Metropolis algorithm

1. Starting with any $a_i^{(0)}$, iterate for $t = 1, 2 \cdots$
2. Draw $\epsilon_i \sim N(0, 0.1)$ and set
$$a_i^* = a_i^{(t-1)} + \epsilon_i$$
3. Compute
$$R(a_i^*, a_i^{(t-1)}) = \frac{f(a_i^* \mid \mathbf{x}_i, \mathbf{y}_i; \hat{\boldsymbol{\beta}}, \hat{\sigma})}{f(a_i^{(t-1)} \mid \mathbf{x}_i, \mathbf{y}_i; \hat{\boldsymbol{\beta}}, \hat{\sigma})}$$
and
$$\rho(a_i^*, a_i^{(t-1)}) = \min\{R(a_i^*, a_i^{(t+1)}), 1\}.$$
4. Set $a_i^{(t)} = a_i^*$ with probability $\rho(a_i^*, a_i^{(t+1)})$. Otherwise, set $a_i^{(t)} = a_i^{(t-1)}$.

Summary

- Monte Carlo methods can be used to compute the E-step of the EM algorithm.
- Sometimes, MCMC algorithm is used for the E-step and so it is an MCMC-EM algorithm.
- Because of the nature of MC algorithm, the convergence is not guaranteed for a fixed MC sample size. Booth and Hobert (1999) discussed some convergence criteria for MCEM.

REFERENCES

Booth, J. G. and J. P. Hobert (1999), 'Maximizing generalized linear models with an automated Monte Carlo EM algorithm', *Journal of the Royal Statistical Society: Series B* **61**, 625–685.

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.