Expectation-Maximization Algorithm

Outline

1. Motivation and EM View
2. The Overview of EM Algorithm
3. Examples
4. Theoretical Issues in EM Algorithm
5. Variants of EM Algorithm

Problem

- Y : p-dimensional random vector with p.d.f. $g(y \mid \psi)$ on R^p
- Observe random variable Y modeled by $g(y \mid \psi)$, $\psi \in \Omega$.
- The likelihood function for ψ formed from the observed data y

$$L(\psi) = g(y \mid \psi)$$

- Estimate the state of ψ with maximum likelihood.

$$\psi^* = \arg\max_{\psi \in \Omega} \log L(y \mid \psi)$$

$$\frac{\partial}{\partial \psi} \log L(y \mid \psi)\Big|_{\psi=\psi^*} = 0 \qquad (1)$$

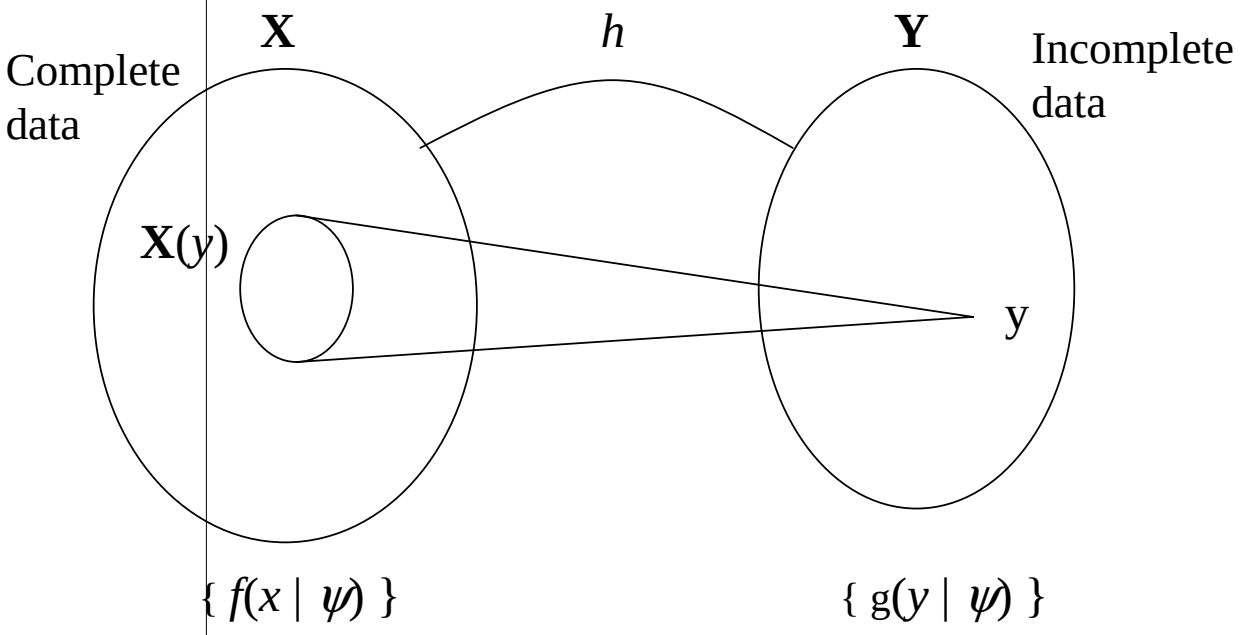
♦ Often it is difficult to solve equation (1).

Motivation

- Data sets with Missing Values
  - ♦ Censored and grouped Observation
  - ♦ Truncated Observation
- Cannot be avoided in many real-world problems

Incomplete Data Model (1)

- Image two sample space **X** and **Y**, and a map h: **X** $\rightarrow$ **Y**.



Incomplete Data Model (2)

- Occurrence of $x \in \mathbf{X} \Rightarrow$ occurrence of $y = h(x) \in \mathbf{Y}$.
- Only y can be observed and $x \in \mathbf{X}(y)$ is revealed.
- The relation between density f and g is:

$$g(y \mid \psi) = \int_{X(y)} f(x \mid \psi) dx$$

- In given problems, **Y** is fixed while **X** can be chosen.

Outline

1. Motivation and EM View
2. The Overview of EM Algorithm
3. Examples
4. Theoretical Issues in EM Algorithm
5. Variants of EM Algorithm

EM Algorithm (1)

- A Standard Tool in the Statistical Repertoire
- Basic Idea
  - To associate with the given *incomplete-data problem*, a *complete-data problem* for which ML estimation is computationally more tractable
- $\boldsymbol{Y}$: random vector corresponding to the observed data $\boldsymbol{y}$, having p.d.f. postulated as $g(\boldsymbol{y}; \boldsymbol{\psi})$ where $\boldsymbol{\psi}$ is a vector of unknown parameters with parameter space $\boldsymbol{\Omega}$
  - y : observed incomplete data
  - x : augmented complete data
  - z : additional data, referred to as the unobservable or missing data
  - $f(\boldsymbol{x}; \boldsymbol{\psi})$: p.d.f. of the random vector $\boldsymbol{X}$ corresponding to the complete-data vector $\boldsymbol{x}$
  - Complete-data log likelihood function

    $\log L_c(\boldsymbol{\psi}) = \log f(\boldsymbol{x}; \boldsymbol{\psi})$

EM Algorithm (2)

- The expectation of the complete-data log likelihood

$$Q(\psi \mid \psi') = E\left[\log f(X \mid \psi) \mid y, \psi'\right]$$

- $\psi^{(0)} \in \Omega$: any first approximation to ψ^*
- The EM algorithm consists of repeating two steps.
- On the (k+1) iteration
  - **E-step**. Calculate $Q(\psi \mid \psi^{(k)})$.
  - **M-step.** Choose $\psi^{(k+1)}$ to be any value of $\psi \in \Omega$ that maximizes $Q(\psi, \psi^{(k)})$.

$$\psi^{(k+1)} = \arg\max_{\psi \in \Omega} Q(\psi \mid \psi^{(k)})$$

Properties

- Choose $\psi^{(k+1)}$ to maximize $\log f(x \mid \psi)$, where $f(x \mid \psi)$ is unknown.
  - Maximize its expectation given the data y and the current fit $\psi^{(k)}$
- Some properties
  - The constraint $\psi \in \Omega$ is naturally incorporated in the M-step.
  - The likelihood $g(y \mid \psi^{(k)})$ is non-decreasing.
  - Simple and Stable
  - The sequence of likelihood has a finite limit L^*.
    - L^* can be the local maximum.
- Disadvantage
  - Slow linear convergence

Outline

1. Motivation and EM View
2. The Overview of EM Algorithm
3. Examples
4. Theoretical Issues in EM Algorithm
5. Variants of EM Algorithm

Multinomial Example (1)

- The observed data vector of frequencies

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4)^T$$

is postulated to arise from a multinomial distribution with four cells with cell probabilities

$$\frac{1}{2}+\frac{1}{4}\psi, \frac{1}{4}(1-\psi), \frac{1}{4}(1-\psi), \text{ and } \frac{1}{4}\psi$$

with $0 \le \boldsymbol{\psi} \le 1$.

- Example
  - $\boldsymbol{y} = (125, 18, 20, 34)^T$, $n = 197$

Multinomial Example (2)

- The probability function $g(y; \psi)$

$$g(y \mid \psi) = \frac{n!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{1}{4}\psi\right)^{y_1} \left(\frac{1}{4} - \frac{1}{4}\psi\right)^{y_2} \left(\frac{1}{4} - \frac{1}{4}\psi\right)^{y_3} \left(\frac{1}{4}\psi\right)^{y_4}$$

- Log likelihood for ψ

$$\log L(\psi) = y_1 \log(2 + \psi) + (y_2 + y_3) \log(1 - \psi) + y_4 \log \psi$$

$$\frac{\partial \log L(\psi)}{\partial \psi} = \frac{y_1}{2 + \psi} - \frac{y_2 + y_3}{1 - \psi} + \frac{y_4}{\psi}$$

  - The likelihood equation can be solved explicitly to find the MLE of ψ.

Multinomial Example (3)

- Suppose that the first of the original four multinomial cells could be split into two subcells having probabilities ½ and ¼ψ.
- MLE of ψ on the basis of this split

$$(y_{12} + y_4) / (y_{12} + y_2 + y_3 + y_4)$$

- View y as being incomplete
- Complete-data vector

$$\boldsymbol{x} = (y_{11}, y_{12}, y_2, y_3, y_4)^T$$

- The cell frequencies in $\boldsymbol{x}$:

$$\frac{1}{2}, \frac{1}{4}\psi, \frac{1}{4}(1 - \psi), \frac{1}{4}(1 - \psi), \text{ and } \frac{1}{4}\psi$$

- y_{11} and y_{12} : unobservable and missing data

Multinomial Example (4)

- Over all values of $\boldsymbol{x}$ such that $y_{11} + y_{12} = y_1$,

$$g(y \mid \psi) = \sum f(x \mid \psi)$$

$$f(x \mid \psi) = C(x)\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{1}{4}\psi\right)^{y_{12}}\left(\frac{1}{4}-\frac{1}{4}\psi\right)^{y_2}\left(\frac{1}{4}-\frac{1}{4}\psi\right)^{y_3}\left(\frac{1}{4}\psi\right)^{y_4}$$

$$C(x) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!}$$

- Complete-data log likelihood

$$\log L_c(\psi) = (y_{12} + y_4)\log\psi + (y_2 + y_3)\log(1-\psi)$$

Multinomial Example (5)

- Let $\psi^{(0)}$ be the initial value for ψ.
- E-step:
  - Filling in for unobservable data by averaging the complete-data log likelihood over conditional distribution given the observed data y
  - Conditional Expectation of log $L_c(\psi)$ given y

$$Q(\psi \mid \psi^{(0)}) = E[\log L_c(\psi) \mid y, \psi^{(0)}]$$

Multinomial Example (6)

- Replace y_{11} and y_{12} by their conditional expectations given the observed data $\boldsymbol{y}$
  - Y_{11} : random variable corresponding to y_{11}
  - Y_{11} has a binomial distribution with sample size y_1 and probability parameter $\frac{1}{2}\Big/\left(\frac{1}{2}+\frac{1}{4}\psi^{(0)}\right)$
  - Initial conditional expectation of Y_{11} given y_1

$$E_{\psi^{(0)}}\left(Y_{11} \mid y_1\right) = y_{11}^{(0)}$$

$$y_{11}^{(0)} = \frac{1}{2} y_1 \Big/ \left(\frac{1}{2}+\frac{1}{4}\psi^{(0)}\right)$$

$$y_{12}^{(0)} = y_1 - y_{11}^{(0)} = \frac{1}{4} y_1 \psi^{(0)} \Big/ \left(\frac{1}{2}+\frac{1}{4}\psi^{(0)}\right)$$

Multinomial Example (7)

- M-step
  - Choose $\psi^{(1)}$ to be the value of ψ that maximizes $Q(\psi;\psi^{(0)})$ w.r.t ψ.

$$\begin{aligned}\psi^{(1)} &= \left(y_{12}^{(0)} + y_4\right)\Big/\left(y_{12}^{(0)} + y_2 + y_3 + y_4\right) \\ &= \left(y_{12}^{(0)} + y_4\right)\Big/\left(n - y_{11}^{(0)}\right)\end{aligned}$$

  - In general,

$$\psi^{(k+1)} = \left(y_{12}^{(k)} + y_4\right)\Big/\left(n - y_{11}^{(k)}\right)$$

$$y_{11}^{(k)} = \frac{1}{2} y_1 \Big/ \left(\frac{1}{2}+\frac{1}{4}\psi^{(k)}\right)$$

$$y_{12}^{(k)} = y_1 - y_{11}^{(k)}$$

Multinomial Example (8)

Iteration	$\Psi^{(k)}$	$\Psi^{(k)} - \hat{\Psi}$	$r^{(k)}$	$\log L(\Psi^{(k)})$
0	0.500000000	0.126821498	—	64.62974
1	0.608247423	0.018574075	0.1465	67.32017
2	0.624321051	0.002500447	0.1346	67.38292
3	0.626488879	0.000332619	0.1330	67.38408
4	0.626777323	0.000044176	0.1328	67.38410
5	0.626815632	0.000005866	0.1328	67.38410
6	0.626820719	0.000000779	0.1328	67.38410
7	0.626821395	0.000000104	0.1328	67.38410
8	0.626821484	0.000000014	—	67.38410

- $r(k) = (\psi^{(k+1)} - \psi^{(k)}) / (\psi^{(k)} - \psi^{(k-1)})$
- Linear convergence equal to 0.1328

Maximum Likelihood Estimator

- Likelihood $P(D|M)$: probability of data D being generated from model M.
- Learning (induction): inference process of deriving a parameterized model $M=M(w)$ from data D

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)} = P(M)\frac{P(D \mid M)}{P(D)}$$

- Maximum Likelihood Hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

- Maximum Likelihood Estimator

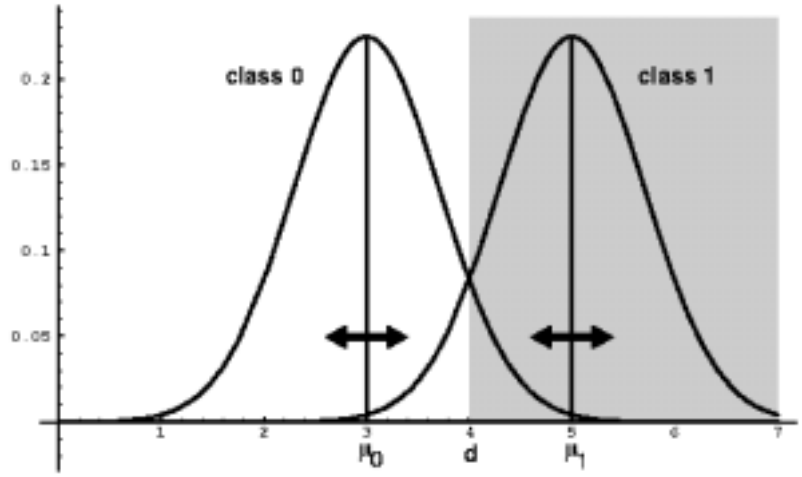
$$\theta^* = \arg\max_{\theta \in H} L(\theta) = \arg\max_{\theta \in H} P(D \mid \theta)$$

Estimating Mixing Proportions (1)

- Suppose that the p.d.f. of a random vector W has a g-component mixture form

$$f(w \mid \psi) = \sum_{i=1}^{g} \pi_i f_i(w)$$

  - $\psi = (\pi_1, \ldots, \pi_{g-1})^T$: g - 1 mixing proportions



Estimation of Mixing Proportions (2)

- Observed random sample obtained from mixture density

$$\boldsymbol{y} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_n^T)^T,$$

  where

$$w_i = (x, f(x \mid \psi)) = (x, \sum_{i=1}^{g} \pi_i f_i(x))$$

- We can observe $f_i(w)$'s, but cannot observe $\pi_i f_i(w)$'s.
  - We do not know π_i.

Estimation of Mixing Proportions (3)

- Log likelihood function for ψ

$$\log L(\psi) = \sum_{j=1}^{n} \log f(w_j \mid \psi)$$

$$= \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i f_i(w_j) \right\}$$

  - ♦ Differentiating with respect to π_i and equating the result to zero

$$\sum_{j=1}^{n} \left\{ \frac{f_i(w_j)}{f(w_j \mid \hat{\psi})} - \frac{f_g(w_j)}{f(w_j \mid \hat{\psi})} \right\} = 0 \quad (i = 1, \ldots, g-1)$$

    - ■ Does not yield am explicit solution for $\hat{\psi}$

Estimation of Mixing Proportions (4)

- To pose this problem as an incomplete-data one
  - ♦ Introduce as the unobservable or missing data the vector

$$\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$$

  - ♦ $\boldsymbol{z}_i$: g-dimensional vector of zero-one indicator variables
  - ♦ $z_{ij} = (z_j)_i$: whether $\boldsymbol{w}_j$ arose from the ith component
- MLE of π_I

$$\sum_{j=1}^{n} z_{ij} \Big/ n$$

- Complete-data vector $\boldsymbol{x}$

$$x = (y^T, z^T)^T$$

Estimation of Mixing Proportion (5)

- Complete-data has multinomial form.

$$x_j \sim \text{Mult}_g(1, \{\pi_i f_i(w)\})$$

$$\Pr(\mathbf{X}_j = x_j) = \{\pi_{1i} f_1(w)\}^{z_{1j}} \{\pi_{2i} f_2(w)\}^{z_{2j}} \cdots \{\pi_{gi} f_g(w)\}^{z_{gj}}$$

- Complete-data log likelihood for ψ

$$\log L_c(\psi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log \pi_i + \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log f_i(w_j)$$

Estimation of Mixing Proportions (6)

- E-step
  - Calculation of the current conditional expectation of Z_{ij} given the observed data y

$$E(Z_{ij} \mid y, \psi^{(k)}) = \text{pr}\{Z_{ij} = 1 \mid y, \psi^{(k)}\}$$

$$= z_{ij}^{(k)}$$

$$z_{ij}^{(k)} = \frac{\pi_i^{(k)} f_i(w_j)}{f(w_j \mid \psi^{(k)})} = \frac{\pi_i^{(k)} f_i(w_j)}{\sum_{k=1}^{g} \pi_k^{(k)} f_k(w_j)}$$

- M-step

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} z_{ij}^{(k)} / n$$

Outline

1. Motivation and EM View
2. The Overview of EM Algorithm
3. Examples
4. Theoretical Issues in EM Algorithm
5. Variants of EM Algorithm

Non-Decreasing Likelihood (1)

- The conditional density of **X** given **Y** = y

$$k(x \mid y, \psi) = \frac{f(x \mid \psi)}{\int_{X(y)} f(x \mid \psi) dx} = \frac{f(x \mid \psi)}{g(y \mid \psi)} \quad x \in X(y)$$

- The log likelihood

$$\begin{aligned} \log L(\psi) &= \log g(y \mid \psi) \\ &= \log f(x \mid \psi) - \log k(x \mid y, \psi) \\ &= \log L_c(\psi) - \log k(x \mid y, \psi) \end{aligned}$$

- Taking the conditional expectation

$$\begin{aligned} \log L(\psi^{(k)}) &= E[\log L_C(\psi) \mid y, \psi^{(k)}] - E[\log k(X \mid y, \psi) \mid y, \psi^{(k)}] \\ &= Q(\psi \mid \psi^{(k)}) - H(\psi \mid \psi^{(k)}) \end{aligned}$$

Non-Decreasing Likelihood (2)

- What we want show

$$\log L(\psi^{(k+1)}) - \log L(\psi^{(k)})$$
$$= Q(\psi^{(k+1)} \mid \psi^{(k)}) - Q(\psi^{(k)} \mid \psi^{(k)}) - \{H(\psi^{(k+1)} \mid \psi^{(k)}) - H(\psi^{(k)} \mid \psi^{(k)})\}$$
$$\geq 0$$

  - The first term

$$Q(\psi^{(k+1)} \mid \psi^{(k)}) - Q(\psi \mid \psi^{(k)}) \geq 0$$

    - $\psi^{(k+1)}$ is chosen so that for all $\psi \in \Omega$

$$Q(\psi^{(k+1)} \mid \psi^{(k)}) \geq Q(\psi \mid \psi^{(k)})$$

  - The second term

$$H(\psi^{(k+1)} \mid \psi^{(k)}) - H(\psi^{(k)} \mid \psi^{(k)}) \leq 0 \qquad (?)$$

Non-Decreasing Likelihood (3)

- By Jensen's inequality and the concavity of the logarithmic function

$$H(\psi^{(k+1)} \mid \psi^{(k)}) - H(\psi^{(k)} \mid \psi^{(k)})$$
$$= E[\log\{k(X \mid y, \psi^{(k+1)}) / k(X, y, \psi^{(k)})\} \mid y, \psi^{(k)}]$$
$$\leq \log\left[E\{k(X \mid y, \psi^{(k+1)}) / k(X, y, \psi^{(k)})\} \mid y, \psi^{(k)}\right]$$
$$= \log \int_{X(y)} k(x \mid y, \psi^{(k+1)}) dx$$
$$= 0$$

- For a bounded sequence of likelihood value $\{L(\psi^{(k)})\}$, $L(\psi^{(k)})$ converges monotonically to some L^*.

Rate of Convergence (1)

- The EM algorithm implicitly defines a mapping $\boldsymbol{\psi} \rightarrow \boldsymbol{M}(\boldsymbol{\psi})$

$$\boldsymbol{\psi}^{(k+1)} = \boldsymbol{M}(\boldsymbol{\psi}^{(k)}) \qquad (k = 0, 1, 2, \ldots)$$

- If $\boldsymbol{\psi}^{(k)}$ converges to to some point $\boldsymbol{\psi}^*$,

$$\boldsymbol{\psi}^* = \boldsymbol{M}(\boldsymbol{\psi}^*)$$

- In a neighborhood of $\boldsymbol{\psi}^*$,

$$\boldsymbol{\psi}^{(k+1)} - \boldsymbol{\psi}^* \approx \boldsymbol{J}(\boldsymbol{\psi}^*)(\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}^*)$$

  where $\boldsymbol{J}(\boldsymbol{\psi})$ is Jacobian matrix for $\boldsymbol{M}(\boldsymbol{\psi})$.

- Thus, $$\frac{\boldsymbol{\psi}^{(k+1)} - \boldsymbol{\psi}^*}{\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}^*} \approx J(\boldsymbol{\psi}^*)$$

Rate of Convergence (2)

- Near $\boldsymbol{\psi}^*$, the EM algorithm is a linear iteration with $\boldsymbol{J}(\boldsymbol{\psi}^*)$.
  - $\boldsymbol{J}(\boldsymbol{\psi}^*)$: rate of convergence
- Global rate of convergence

$$r = \lim_{k \rightarrow \infty} \frac{\| \boldsymbol{\psi}^{(k+1)} - \boldsymbol{\psi}^* \|}{\| \boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}^* \|}$$

  - Under certain regularity conditions,

$$r = \lambda_{\max} = \text{the largest eigenvalue of } \boldsymbol{J}(\boldsymbol{\psi}^*)$$

- Global speed of convergence

$$s = 1 - r$$

  - s is the smallest eigenvalue of $\boldsymbol{S} = \boldsymbol{I}_d - \boldsymbol{J}(\boldsymbol{\psi}^*)$

Outline

1. Motivation and EM View
2. The Overview of EM Algorithm
3. Examples
4. Theoretical Issues in EM Algorithm
5. Variants of EM Algorithm

Kinds of EM Variants

1. Generalized EM Algorithm
2. Modification for Maximum *a Posteriori* Estimation
3. ECM Algorithm
4. ECME Algorithm
5. MCEM Algorithm

Generalized EM Algorithm (1)

- M-step of the EM Algorithm

$$\psi^{(k+1)} = \arg\max_{\psi \in \Omega} Q(\psi \mid \psi^{(k)})$$

  - That is, for all $\psi \in \Omega$,

$$Q(\psi^{(k+1)} \mid \psi^{(k)}) \geq Q(\psi \mid \psi^{(k)})$$

- M-step of the GEM Algorithm
  - Choose $\psi^{(k+1)}$ such that

$$Q(\psi^{(k+1)} \mid \psi^{(k)}) \geq Q(\psi^{(k)} \mid \psi^{(k)})$$

Generalized EM Algorithm (2)

- Monotonicity of GEM Algorithm

$$\log L(\psi^{(k+1)}) - \log L(\psi^{(k)}) \geq 0$$

  - In the prove of EM Algorithm, only

$$Q(\psi^{(k+1)} \mid \psi^{(k)}) - Q(\psi \mid \psi^{(k)}) \geq 0$$

    is changed into

$$Q(\psi^{(k+1)} \mid \psi^{(k)}) - Q(\psi^{(k)} \mid \psi^{(k)}) \geq 0.$$

For MAP Estimate (1)

- MAP Estimate

$$\psi^* = \arg\max_{\psi \in \Omega} P(\psi \mid y)$$

$$= \arg\max_{\psi \in \Omega} P(y \mid \psi) P(\psi)$$

$$\therefore \log P(\psi \mid y) = \log P(y \mid \psi) + \log P(\psi)$$

$$= \log L(\psi) + \log P(\psi)$$

For MAP Estimate (2)

- E-step
  - Calculate

$$E\left[\log p(\psi \mid x) \mid y, \psi^{(k)}\right] = Q(\psi \mid \psi^{(k)}) + \log p(\psi)$$

- M-step
  - Choose $\psi^{(k+1)}$ to maximize the above equation over $\psi \in \Omega$.

ECM Algorithm (1)

- Motivation
  - ♦ when complete-data ML estimation is rather complicated
  - ♦ Replace a complicated M-step with simpler CM-steps
- CM-steps
  - ♦ Conditional Maximization steps
  - ♦ Maximize the conditional expectation of complete-data log likelihood function found in the preceding E-step subject to constraints on ψ
  - ♦ May require iteration
    - ■ Maximization over smaller dimensional spaces
    - ■ Simpler, faster, and more stable

ECM Algorithm (2)

- M-step is replaced by $S > 1$ steps.
  - ♦ $\psi^{(k+s/S)}$: the value of ψ on the sth CM-step of $(k + 1)$th iteration
  - ♦ Choose $\psi^{(k+s/S)}$ to maximize $Q(\psi|\psi^{(k)})$ subject to the constraint
    $$g_s(\psi) = g_s(\psi^{(k+(s-1)/S)})$$
  - ♦ $C = \{g_s(\psi), s = 1, \ldots, S\}$: a set of S preselected (vector) functions
- $\psi^{(k+s/S)}$ satisfies
  $$Q(\psi^{(k+s/S)}|\psi^{(k)}) \geq Q(\psi|\psi^{(k)}) \text{ for all } \psi \in \Omega_s(\psi^{(k+(s-1)/S)})$$
  where $\Omega_s(\psi^{(k+(s-1)/S)}) \equiv \{\psi \in \Omega : g_s(\psi) = g_s(\psi^{(k+(s-1)/S)})\}$
- The value of ψ on the final CM-step
  $$\psi^{(k+S/S)} = \psi^{(k+1)}$$

ECM Algorithm (3)

- ECM algorithm is a GEM algorithm.

$$\begin{aligned} Q(\psi^{(k+1)} \mid \psi^{(k)}) &\geq Q(\psi^{(k+(S-1)/S)} \mid \psi^{(k)}) \\ &\geq Q(\psi^{(k+(S-2)/S)} \mid \psi^{(k)}) \\ &\vdots \\ &\geq Q(\psi^{(k)} \mid \psi^{(k)}) \end{aligned}$$

- Space-Filling Condition

$$\bigcup_{s=1}^{S} Gs(\psi^{(k)}) = \{\mathbf{0}\} \quad \text{for all } k$$

  - $Gs(\psi)$: column space of $\nabla \boldsymbol{g}_s(\psi)$
  - Complement
    - Convex hull of all feasible direction by $\boldsymbol{\Omega}_s(\psi^{(k+(s-1)/S)})$ is the whole Euclidian space $\mathbf{R}^d$.
    - Resulting maximization is over the whole parameter space $\boldsymbol{\Omega}$.

ECME Algorithm (4)

- ECME (ECM Either) Algorithm
  - *either* : with this extension, some or all of the CM-steps of the ECM algorithm are replaced by steps that computationally maximize the incomplete-data log likelihood, log $L(\psi)$, and not the Q-function.
  - Faster convergence.
- CM steps that act on the Q-function must be performed *before* those that act on the actual log likelihood.
- There are situations where the speed of convergence of the ECME algorithm is less than that of the ECM algorithm.

MCEM Algorithm (1)

- When E-step is complex and does not admit a closed-form solution to the computation of $Q(\psi|\psi^{(k)})$.
  - E-step by a Monte Carlo process

$$Q(\psi|\psi^{(k)}) = E[\log p(\psi|x)|y]$$
$$= \int_Z \log p(\psi|x)p(z|y,\psi^{(k)})dz$$

  - Z : sample space of the missing data
  - $p(z|y, \psi)$: conditional density of $\mathbf{Z}$ given y

MCEM Algorithm (2)

- MCEM Algorithm
  - Monte Carlo E-step
    - On the kth iteration, draw $z^{(1k)}, \ldots, z^{(Mk)}$ from $p(z|y, \psi^{(k)})$

$$Q(\psi|\psi^{(k)}) = \frac{1}{M}\sum_{m=1}^{M} \log p(\psi|z^{(mk)}, y)$$

  - M-step
    - Maximize $Q(\psi|\psi^{(k)})$ over ψ to obtain $\psi^{(k+1)}$.

MCEM Algorithm (3)

- An Example
  - Single unknown parameter ψ
  - Single missing variable $z = y_{12}$

$$\text{binomial}\left(y_1, \frac{\frac{1}{4}\psi^{(k)}}{\frac{1}{2}+\frac{1}{4}\psi^{(k)}}\right)$$

  - Complete-data log likelihood

$$\log L_c(\psi) = (y_{12} + y_4)\log\psi + (y_2 + y_3)\log(1-\psi)$$

  - Q-function in MCEM Algorithm

$$Q(\psi \mid \psi^{(k)}) = (\bar{z}^{(k)} + y_4)\log\psi + (y_2 + y_3)\log(1-\psi)$$

$$\psi^{(k+1)} = \frac{\bar{z}^{(k)} + y_4}{\bar{z} + y_4 + y_2 + y_3}$$

MCEM Algorithm (4)

- Difficult problems
  - Monotonicity property is lost.
    - Monitoring of convergence
    - Plot $\psi^{(k)}$ against k.
    - Stabilization of the process indicates the convergence.
    - With fluctuations, the process is terminated or continued with a larger value of M.
  - Choice of M
    - Small values of M in initial stage
    - Increased as the algorithm moves closer to convergence

References

- A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM algorithm,” *Jour. Roy. Stat. Soc*., B39, pp.1-38, 1977.
- G. J. McLachlan and T. Krishnan, “*The EM Algorithm and Extensions*,” A Wiley-Interscience Publication, 1997.
- G. McLachlan and D. Peel, “*Finite Mixture Models*,” A Wiley-Interscience Publication, 2000.
- V. Cherkassky and F. Mulier, “*Learning from Data: Concepts, Theory, and Methods*,” A Wiley-Interscience Publication, 1998.