ST509 Computational Statistics

Lecture 9: MM/EM Algorithm

Seung Jun Shin

Department of Statistics
Korea University

E-mail: sjshin@korea.ac.kr

MM algorithm I

- MM stands for
  - (Minimization) Majorization then Minimization.
  - (Maximization) Minorization then Maximization.

MM algorithm II

- A function $g(\mathbf{x} \mid \mathbf{x}_m)$ is said to majorize a function $f(\mathbf{x})$ at $\mathbf{x}_m$ provided

$$f(\mathbf{x}_m) = g(\mathbf{x}_m \mid \mathbf{x}_m)$$
$$f(\mathbf{x}) \leq g(\mathbf{x}|\mathbf{x}_m), \quad \text{for } \mathbf{x} \neq \mathbf{x}_m$$

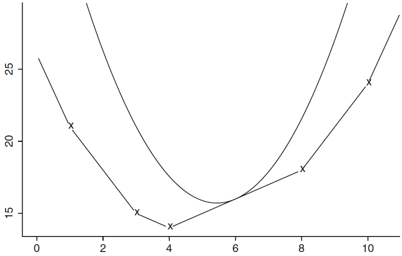

Figure: A quadratic majorizing function for the piecewise linear function $f(x) = |x-1| + |x-3| + |x-4| + |x-8| + |x-10|$ at the point $x_m = 6$.

MM algorithm III

- MM minimize the surrogate majorizing function $g(\mathbf{x} \mid \mathbf{x}_m)$ rather than the actual function $f(\mathbf{x})$.
- Let $\mathbf{x}_{m+1}$ denote the minimizer of $g(\mathbf{x}|\mathbf{x}_m)$. Then we have the following descent property:

$$f(\mathbf{x}_{m+1}) \leq g(\mathbf{x}_{m+1} \mid \mathbf{x}_m) \leq g(\mathbf{x}_m \mid \mathbf{x}_m) = f(\mathbf{x}_m)$$

MM algorithm IV

- One simple way majorizing is to use Jensen's inequality:

$$f\left(\sum_i \alpha_i t_i\right) \le \sum_i \alpha_i f(t_i)$$

- With $\alpha_i = c_i y_i / \mathbf{c}^T\mathbf{y}$ and $t_i = \mathbf{c}^T\mathbf{y} x_i / y_i$:

$$f(\mathbf{c}^T\mathbf{x}) \le \sum_i \frac{c_i y_i}{\mathbf{c}^T\mathbf{y}} f\left(\frac{\mathbf{c}^T\mathbf{y}}{y_i} x_i\right) = \sum_i \alpha_i f\left(\frac{c_i}{\alpha_i} x_i\right) = g(\mathbf{x} \mid \mathbf{y}) \qquad (1)$$

  provided $\mathbf{c}, \mathbf{x}$, and $\mathbf{y}$ are positive.

- $f(\mathbf{c}^T\mathbf{x}) = g(\mathbf{x} \mid \mathbf{y})$ when $\mathbf{x} = \mathbf{y}$.
- $g(\mathbf{x} \mid \mathbf{y})$ separates the parameters.

MM algorithm V

- To relax the positivity restrictions, we have for a convex function $f(t)$:

$$f(\mathbf{c}^T\mathbf{x}) \leq \sum_i \alpha_i f\left\{\frac{c_i}{\alpha_i}(x_i - y_i) + \mathbf{c}^T\mathbf{y}\right\} = g(\mathbf{x} \mid \mathbf{y}) \tag{2}$$

  where all $\alpha_i \geq 0, \sum_i \alpha_i = 1$, and $\alpha_i > 0$ whenever $c_i \neq 0$.
- One obvious choice of α_i is

$$\alpha_i = \frac{|c_i|^p}{\sum_j |c_j|^p}$$

  for $p \geq 0$.

MM algorithm VI

- Third method involves the linear majorization

$$f(\mathbf{x}) \leq f(\mathbf{y}) + f'(\mathbf{y})(\mathbf{x} - \mathbf{y}) = g(\mathbf{x} \mid \mathbf{y}) \tag{3}$$

  satisfied by any concave function $f(\mathbf{x})$.

- We can replace $\mathbf{x}$ with $h(\mathbf{x})$:

$$f(h(\mathbf{x})) \leq f(h(\mathbf{y})) + f'(h(\mathbf{x}))\{h(\mathbf{x}) - h(\mathbf{y})\} = g(\mathbf{x} \mid \mathbf{y}).$$

MM algorithm VII

- Assuming that $f(\mathbf{x})$ is twice differentiable, and there exist a positive-definite matrix $\mathbf{B}$ such that $\mathbf{B} - f''(\mathbf{x})$ is positive-semidefinite.

$$
\begin{aligned}
f(\mathbf{x}) =& f(\mathbf{y}) + f'(\mathbf{y})(\mathbf{x}-\mathbf{y}) \\
& + (\mathbf{x}-\mathbf{y})^T \int_0^1 f''(\mathbf{y}+t(\mathbf{x}-\mathbf{y}))(1-t)dt(\mathbf{x}-\mathbf{y}) \\
\leq & f(\mathbf{y}) + f'(\mathbf{y})(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^T \mathbf{B}(\mathbf{x}-\mathbf{y}) \\
=& g(\mathbf{x} \mid \mathbf{y}) \qquad (4)
\end{aligned}
$$

Allele Frequency Estimation I

- We are given $n = 521$, $n_A = 186$, $n_B = 38$, $n_{AB} = 13$, $n_O = 284$.
- Let us estimate p_A, p_B, and p_O via MM algorithm.
- Under the Hardy-Weinberg law of population genetics, our goal is to maximize the following multinomial likelihood:

$$\begin{aligned} f(\mathbf{p}) =& n_A \log(p_A^2 + 2p_A p_O) + n_B \log(p_B^2 + 2p_B p_O) + n_{AB} \log(2p_A p_B) \\ &+ n_O \log p_O^2 + \log \binom{n}{n_A, n_B, n_{AB}, n_O}. \end{aligned}$$

where $p_A + p_B + p_O = 1$ and $p_A, p_B, p_O \geq 0$.

Allele Frequency Estimation II

- The likelihood is not easy to maximize due to $\log(p_A^2 + 2p_Ap_O)$ and $\log(p_B^2 + 2p_Bp_O)$
- We can minorize the log function which is concave.

$$\begin{aligned}\log\left(p_A^2 + 2p_Ap_O\right) \geq & \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}} \log\left(\frac{p_{mA}^2 + 2p_{mA}p_{mO}}{p_{mA}^2} p_A^2\right) \\ & + \frac{2p_{mA}p_{mO}}{p_{mA}^2 + 2p_{mA}p_{mO}} \log\left(\frac{p_{mA}^2 + 2p_{mA}p_{mO}}{2p_{mA}p_{mO}} 2p_Ap_O\right)\end{aligned}$$

- Let

$$\begin{aligned}n_{mA/A} &= n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}} \\ n_{mA/O} &= n_A \frac{2p_{mA}p_{mO}}{p_{mA}^2 + 2p_{mA}p_{mO}}\end{aligned}$$

- A similar minorization applies to $\log(p_B^2 + 2p_Bp_O)$ and have $n_{mB/B}$ and $n_{mB/O}$.

Allele Frequency Estimation III

- Now, we have

$$g(\mathbf{p} \mid \mathbf{p}_m) = n_{mA/A} \log p_A^2 + n_{mA/O} \log(2p_A p_O) + n_{mB/B} \log p_B^2 + n_{mB/O} \log(2p_B p_O) + n_{AB} \log(2p_A p_B) + n_O \log p_O^2 + c,$$

where c represent the terms irrelevant to $\mathbf{p}$.

$$L(\mathbf{p} \mid \lambda) = g(\mathbf{p} \mid \mathbf{p}_m) + \lambda(p_A + p_B + p_O - 1).$$

- Solving stationary equations yield

$$p_{m+1,A} = \frac{2n_{mA/A} + n_{mA/O} + n_{AB}}{2n}$$
$$p_{m+1,B} = \frac{2n_{mB/B} + n_{mB/O} + n_{AB}}{2n}$$
$$p_{m+1,O} = \frac{n_{mA/O} + n_{mB/O} + 2n_O}{2n}$$

Linear Regression I

- Linear regression solves

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$.

- The surrogate function is

$$g(\boldsymbol{\beta} \mid \boldsymbol{\beta}_m) = \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_{ij} \left[y_i - \frac{x_{ij}}{\alpha_{ij}} (\beta_j - \beta_{mj}) - \mathbf{x}_i^T \boldsymbol{\beta}_m \right]^2$$

which achieves equality when $\boldsymbol{\beta} = \boldsymbol{\beta}_m$.

- Minimization of $g(\boldsymbol{\beta} \mid \boldsymbol{\beta}_m)$ yields

$$\beta_{m+1,j} = \beta_{mj} + \frac{\sum_{i=1}^{n} x_{ij} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_m)}{\sum_{i=1}^{n} x_{ij}^2 / \alpha_{ij}}$$

with $\alpha_{ij} = |x_{ij}| / \|\mathbf{x}_i\|_1$.

Linear Regression II

- Alternatively, we consider a median regression which minimizes

$$h(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mid y_i - \mathbf{x}_i^T \boldsymbol{\beta} \mid = \sum_{i=1}^{n} \mid r_i(\boldsymbol{\beta}) \mid .$$

- Notice that for the square root function, we have

$$\sqrt{u} \leq \sqrt{u_m} + \frac{u - u_m}{2\sqrt{u_m}}$$

- We find that

$$h(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sqrt{r_i(\boldsymbol{\beta})^2} \leq h(\boldsymbol{\beta}_m) + \frac{1}{2} \sum_{i=1}^{n} \frac{r_i^2(\boldsymbol{\beta}) - r_i^2(\boldsymbol{\beta}_m)}{\sqrt{r_i^2(\boldsymbol{\beta}_m)}} = g(\boldsymbol{\beta} \mid \boldsymbol{\beta}_m)$$

- Thus updating equation is

$$\boldsymbol{\beta}_{m+1} = \left[\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}_m) \mathbf{X}\right]^{-1} \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}_m) \mathbf{y}$$

where $\mathbf{W}(\boldsymbol{\beta}_m) = \mathrm{diag}\{|r_i(\boldsymbol{\beta}_m)|^{-1}\}$

Logistic Regression I

- Recall that the log-likelihood of LR is

$$\sum_{i=1}^{n} \left[y_i \log \pi_i(\boldsymbol{\beta}) + (1 - y_i) \log\{1 - \pi_i(\boldsymbol{\beta})\} \right]$$

  where

$$\pi_i(\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

- We have

$$f'(\mathbf{x}) = \sum_{i=1}^{n} \{y_i - \pi_i(\boldsymbol{\beta}\}\mathbf{x}_i^T,$$

$$f''(\mathbf{x}) = -\sum_{i=1}^{n} \pi_i(\boldsymbol{\beta})\{1 - \pi_i(\boldsymbol{\beta})\}\mathbf{x}_i\mathbf{x}_i^T$$

Logistic Regression II

- By (4), we maximize

$$g(\boldsymbol{\beta} \mid \boldsymbol{\beta}_m) = f(\boldsymbol{\beta}_m) + f'(\boldsymbol{\beta}_m)(\boldsymbol{\beta} - \boldsymbol{\beta}_m) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_m)^T \mathbf{B}(\boldsymbol{\beta} - \boldsymbol{\beta}_m)$$

  with

$$\mathbf{B} = -\frac{1}{4}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$$

- Updating equation is

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - \mathbf{B}^{-1} f'(\boldsymbol{\beta}_m)$$

  which seems quite similar to NR updating equation, but $\mathbf{B}^{-1}$ can be repeatedly used for every iteration.

EM algorithm I

- The observed $\mathbf{Y}$ is incomplete and there is missing $\mathbf{Z}$ under the presence of a latent structure or missing.
- Direct optimization of the likelihood of $\mathbf{Y}$ is often difficult.
- EM algorithm defines

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) = E_{\boldsymbol{\theta}_m} \{\log L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})|\mathbf{Y}\}$$

  where L_C denotes the complete data likelihood.

  1. E-step: Calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y})$.
  2. M-step: Calculate $\boldsymbol{\theta}_{m+1}$ that maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y})$ w.r.t $\boldsymbol{\theta}$.

- EM is an MM algorithm as shown in the following.

EM algorithm II

- Let $P(\boldsymbol{\theta}|\boldsymbol{\theta}_m) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) + \log L(\boldsymbol{\theta}_m \mid \mathbf{Y}) - Q(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m, \mathbf{Y})$
- We can show that $P(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ minorizes $\log L(\boldsymbol{\theta} \mid \mathbf{Y})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_m$.
- First, we have

$$\begin{aligned} P(\boldsymbol{\theta}_m|\boldsymbol{\theta}_m) &= Q(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m, \mathbf{Y}) + \log L(\boldsymbol{\theta}_m \mid \mathbf{Y}) - Q(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m, \mathbf{Y}) \\ &= \log L(\boldsymbol{\theta}_m \mid \mathbf{Y}) \end{aligned}$$

- Next, we have

$$P(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) \leq \log L(\boldsymbol{\theta} \mid \mathbf{Y})$$

which is equivalent to

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) - \log L(\boldsymbol{\theta} \mid \mathbf{Y}) \leq Q(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m, \mathbf{Y}) - \log L(\boldsymbol{\theta}_m \mid \mathbf{Y})$$

EM algorithm III

This is because

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) - \log L(\boldsymbol{\theta} \mid \mathbf{Y}) &= E_{\boldsymbol{\theta}_m}\left\{\log \frac{L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})}{L(\boldsymbol{\theta} \mid \mathbf{Y})} \;\middle|\; \mathbf{Y}\right\} \\
&\leq E_{\boldsymbol{\theta}_m}\left\{\log \frac{L_C(\boldsymbol{\theta}_m|\mathbf{Y}, \mathbf{Z})}{L(\boldsymbol{\theta}_m|\mathbf{Y})} \;\middle|\; \mathbf{Y}\right\} \\
&= Q(\boldsymbol{\theta}_m, \boldsymbol{\theta}_m, \mathbf{Y}) - \log L(\boldsymbol{\theta}_m \mid \mathbf{Y})
\end{aligned}$$

The inequality holds since

$$E_f(\log f) \geq E_f(\log g)$$

where f and g are densities. (Information Inequality)

- Finally,

$$\operatorname*{argmin}_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\boldsymbol{\theta}_m) = \operatorname*{argmin}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}).$$

Gaussian Mixture I

- Suppose that $Y_1, \cdots, Y_n$ are iid from the Gaussian mixture density:

$$f(y; \boldsymbol{\theta}) = p\phi(y; \mu_1, \sigma_1^2) + (1-p)\phi(y; \mu_2, \sigma_2^2)$$

  where $\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$, $\phi(y; \mu, \sigma)$ denotes the normal density with mean μ and variance σ^2, and $p \in [0, 1]$ is a mixture proportion.

- The log likelihood is

$$\log L(\boldsymbol{\theta} \mid \mathbf{Y}) = \sum_{i=1}^{n} \log \left\{ p\phi(y; \mu_1, \sigma_1^2) + (1-p)\phi(y; \mu_2, \sigma_2^2) \right\}$$

  which is simple to write down, but not so simple to maximize.

Gaussian Mixture II

- In order to apply EM, we introduce independent random variables:
  - $Z_i \sim \text{Bernoulli}(p)$
  - $X_{i1} \sim N(\mu_1, \sigma_1^2)$ and $X_{i2} \sim N(\mu_2, \sigma_2^2)$
- We can represent Y_i as

$$Y_i = Z_i X_{i1} + (1 - Z_i) X_{i2}.$$

- The joint likelihood of the complete data (Y_i, Z_i) is

$$\{p\phi(y_i; \mu_1, \sigma_1^2)\}^{z_i} \{(1-p)\phi(y_i; \mu_2, \sigma_2^2)\}^{1-z_i}$$

- The complete log likelihood is

$$\log L_C(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{n} \Big\{ z_i \log \phi(y_i; \mu_1, \sigma_1^2) + (1 - z_i) \log \phi(y_i; \mu_2, \sigma_2^2) \\ + z_i \log p + (1 - z_i) \log(1 - p) \Big\}.$$

Gaussian Mixture III

- The conditional expectation of the E-step is given by

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) =& E_{\boldsymbol{\theta}_m} \{\log L_C(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}\} \\
=& \sum_{i=1}^{n} \Big[w_{im} \log \phi(y_i; \mu_1, \sigma_1^2) + (1 - w_{im}) \log \phi(y_i; \mu_2, \sigma_2^2) \\
& + w_{im} \log p + (1 - w_{im}) \log(1 - p) \Big] \\
=& \sum_{i=1}^{n} \left[w_{im} \left\{ -\frac{1}{2} \log \sigma_1^2 - \frac{(y_i - \mu_1)^2}{2\sigma_1^2} \right\} \right. \\
& + (1 - w_{im}) \left\{ -\frac{1}{2} \log \sigma_2^2 - \frac{(y_i - \mu_2)^2}{2\sigma_2^2} \right\} \\
& \left. + w_{im} \log p + (1 - w_{im}) \log(1 - p) \right] \qquad (5)
\end{aligned}
$$

where

$$
w_{im} = E_{\boldsymbol{\theta}_m}(Z_i \mid y_i) = \frac{p_m \phi(y_i; \mu_{1m}, \sigma_{1m}^2)}{p_m \phi(y_i; \mu_{1m}, \sigma_{1m}^2) + (1 - p_m) \phi(y_i; \mu_{2m}, \sigma_{2m}^2)}.
$$

Gaussian Mixture IV

- Taking derivative (5) with respect to μ_1 and σ_1^2, we have

$$
\begin{aligned}
\frac{\partial}{\partial \mu_1} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) = & \quad \textstyle\sum_{i=1}^n w_{im} \frac{(y_i - \mu_1)}{\sigma_1^2} & = 0 \\
\frac{\partial}{\partial \mu_2} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) = & \quad \textstyle\sum_{i=1}^n (1 - w_{im}) \frac{(y_i - \mu_2)}{\sigma_2^2} & = 0 \\
\frac{\partial}{\partial \sigma_1^2} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) = & \quad \textstyle\sum_{i=1}^n w_{im} \left\{ \frac{1}{\sigma_1^2} - \frac{(y_i - \mu_1)^2}{\sigma_1^4} \right\} & = 0 \\
\frac{\partial}{\partial \sigma_2^2} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) = & \quad \textstyle\sum_{i=1}^n (1 - w_{im}) \left\{ \frac{1}{\sigma_2^2} - \frac{(y_i - \mu_2)^2}{\sigma_2^4} \right\} & = 0 \\
\frac{\partial}{\partial p} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_m, \mathbf{Y}) = & \quad \textstyle\sum_{i=1}^n \left\{ \frac{w_{im}}{p} + \frac{1 - w_{im}}{1 - p} \right\} & = 0
\end{aligned}
$$

Gaussian Mixture V

- EM algorithm updates $\boldsymbol{\theta}$ until conergenge:

$$\mu_{1,m+1} = \frac{\sum_{i=1}^{n} w_{im} y_i}{\sum_{i=1}^{n} w_{im}}$$
$$\mu_{2,m+1} = \frac{\sum_{i=1}^{n} (1 - w_{im}) y_i}{\sum_{i=1}^{n} (1 - w_{im})}$$
$$\sigma_{1,m+1}^2 = \frac{\sum_{i=1}^{n} w_{im} (y_i - \mu_{1,m+1})^2}{\sum_{i=1}^{n} w_{im}}$$
$$\sigma_{2,m+1}^2 = \frac{\sum_{i=1}^{n} (1 - w_{im}) (y_i - \mu_{2,m+1})^2}{\sum_{i=1}^{n} (1 - w_{im})}$$
$$p_{m+1} = \frac{1}{n} \sum_{i=1}^{n} w_{im}$$

and then $w_{i,m+1}$.