

## Chapter 2 EM Optimization Methods

## 2.1 Missing Data Problems

---

- **Missing Data Problem:** Generally, a missing/incomplete data problem can be described as follows.

- A statistical model is specified for  $X$  by the density  $f_X(x; \theta)$ , where  $\theta$  is the parameter. (Here  $X$  could be  $X_1, \dots, X_n$ .)
- We only observe  $Y = h(X)$ , where  $h$  is a **many-to-one** function, therefore we only have incomplete information about  $X$ . For example,  $X = (Y, Z)$ , but we can only observe  $Y$ .
- The MLE of  $\theta$  is derived by maximizing the observed-data likelihood as

$$\hat{\theta}_{MLE} = \arg \max_{\theta} l(\theta|Y = y) = \arg \max_{\theta} f_Y(y; \theta).$$

- In many cases, the complete-data likelihood  $l(\theta|X = x) = f_X(x; \theta)$  has a simple form (often specified by the model), but  $l(\theta|Y)$  is difficult to evaluate. In some cases,  $l(\theta|Y)$  is possible to evaluate, but hard to maximize.

## 2.1 Missing Data Problems

---

- **Example: Censored Data.** Assume that  $X_1, \dots, X_n$  are i.i.d. from the  $N(\theta, 1)$  distribution, but the data is right censored at the value  $a$ . That is, we can only observe  $Y_i = h(X_i) = \min\{X_i, a\}$ .
  - For example, we can consider  $\exp\{X_i\}$  as the survival time of individual  $i$ , and the observation is censored at  $e^a$ .
  - The distribution of  $Y_i$  consists of a continuous part and a discrete part with the PMF/PDF function

$$f_Y(y) = \begin{cases} \phi(y - \theta), & \text{if } y < a; \\ 1 - \Phi(a - \theta), & \text{if } y = a, \end{cases}$$

where  $\Phi(y)$  and  $\phi(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$  are the CDF and PDF of the  $N(0, 1)$  distribution, respectively.

## 2.1 Missing Data Problems

---

- – Suppose that the observations are  $(Y_1, \dots, Y_n) = (y_1, \dots, y_n)$ . For convenience, we re-arrange the observations and assume that  $y_i < a$  for  $i = 1, \dots, m$  and  $y_i = a$  for  $i = m + 1, \dots, n$ .
- The MLE of  $\theta$  is

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg \max_{\theta} \log l(\theta | Y_1, \dots, Y_n) \\ &= \arg \max_{\theta} \sum_{i=1}^n \log f_Y(y_i) \\ &= \arg \max_{\theta} \left[ \sum_{i=1}^m \log \phi(y_i - \theta) + (n - m) \log (1 - \Phi(a - \theta)) \right] \\ &= \arg \max_{\theta} \left[ c - \frac{1}{2} \sum_{i=1}^m (y_i - \theta)^2 + (n - m) \log (1 - \Phi(a - \theta)) \right],\end{aligned}$$

which does not have a closed-form solution.

## 2.1 Missing Data Problems

---

- **Example: Mixture Normal Distribution.** Assume that  $U_{i1}, \dots, U_{iK}$  are independent random variables with  $U_{ik} \sim N(\mu_k, \sigma_k^2)$ . Further assume that  $Z_i \in \{1, \dots, K\}$  is a discrete random variable independent of  $U_{i1}, \dots, U_{iK}$  with  $P(Z_i = k) = p_k$ . Here  $p_1 + \dots + p_K = 1$ . Define

$$Y_i = \sum_{k=1}^K I(Z_i = k)U_{ik}.$$

- The parameters are  $\mu_k, \sigma_k^2$  and  $p_k, k = 1, \dots, K$ .
- Let  $X_i = (Y_i, Z_i), i = 1, \dots, n$ . Given  $X_1, \dots, X_n$  (assume they are independent), it is easy to find the MLE of  $\theta = (\mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2, p_1, \dots, p_K)$ . (**Homework**)
- We can only observe the incomplete data  $Y_1, \dots, Y_n$ .

## 2.1 Missing Data Problems

---

- – Let  $F_Y(y; \theta)$  and  $f_Y(y; \theta)$  be the CDF and PDF of  $Y_i$ , respectively. Then

$$\begin{aligned} F_Y(y; \theta) = P(Y_i \leq y) &= \sum_{k=1}^K P(Z_i = k) P(Y_i \leq y | Z_i = k) \\ &= \sum_{k=1}^K P(Z_i = k) P(U_{ik} \leq y | Z_i = k) \\ &= \sum_{k=1}^K P(Z_i = k) P(U_{ik} \leq y) = \sum_{k=1}^K p_k \Phi\left(\frac{y - \mu_k}{\sigma_k}\right) \end{aligned}$$

and

$$f_Y(y; \theta) = \sum_{k=1}^K p_k \frac{1}{\sigma_k} \phi\left(\frac{y - \mu_k}{\sigma_k}\right)$$

where  $\Phi(y)$  and  $\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$  are the CDF and PDF of the  $N(0, 1)$  distribution, respectively. Obviously,  $\frac{1}{\sigma_k} \phi\left(\frac{y - \mu_k}{\sigma_k}\right)$  is the PDF of the  $N(\mu_k, \sigma_k^2)$  distribution. We call that  $Y_i$  follows a *mixture of normal distributions*.

## 2.1 Missing Data Problems

---

- – In this example, the logarithm of the observed-data likelihood is

$$\begin{aligned}\log l(\theta|Y_1, \dots, Y_n) &= \sum_{i=1}^n \log f_Y(Y_i; \theta) \\ &= \sum_{i=1}^n \log \left[ \sum_{k=1}^K p_k \frac{1}{\sigma_k} \phi\left(\frac{Y_i - \mu_k}{\sigma_k}\right) \right] \\ &= \sum_{i=1}^n \log \left[ \sum_{k=1}^K p_k \frac{1}{\sigma_k} \cdot \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(Y_i - \mu_k)^2}{2\sigma_k^2} \right\} \right].\end{aligned}$$

- It is not easy to maximize  $\log l(\theta|Y_1, \dots, Y_n)$  or solve the equation  $\nabla \log l(\theta|Y_1, \dots, Y_n) = 0$ .

## 2.1 Missing Data Problems

---

- **Example: Peppered Moths.** The coloring of peppered moths is determined by a gene with three possible alleles, denoted by C, I and T. Of these alleles, C is dominant to I, and I is dominant to T.
  - The genotypes CC, CI, and CT result in the *carbonaria* phenotype, which exhibits solid black coloring.
  - The genotypes II and IT produce the *insularia* phenotype, which varies widely in appearance but is generally mottled with intermediate color.
  - The genotype TT results in the *typica* phenotype, which exhibits light-colored patterned wings.
  - If the allele frequencies in the population are  $p_C$ ,  $p_I$ , and  $p_T$ , the genotype frequencies should be  $p_C^2$ ,  $2p_Cp_I$ ,  $2p_Cp_T$ ,  $p_I^2$ ,  $2p_Ip_T$ , and  $p_T^2$  for genotypes CC, CI, CT, II, IT, and TT, respectively. Here  $p_C + p_I + p_T = 1$ .
  - We want to know the allele frequencies  $\theta = (p_C, p_I, p_T)$ .



## 2.1 Missing Data Problems

---

- – Suppose we capture  $n$  moths, of which there are  $n_C$ ,  $n_I$ , and  $n_T$  of the carbonaria, insularia, and typica phenotypes, respectively. Thus,  $n = n_C + n_I + n_T$ .
- We also use  $n_{CC}$ ,  $n_{CI}$ ,  $n_{CT}$ ,  $n_{II}$ ,  $n_{IT}$ , and  $n_{TT}$  to denote counts for genotypes CC, CI, CT, II, IT, and TT, respectively. Then we have

$$n = n_{CC} + n_{CI} + n_{CT} + n_{II} + n_{IT} + n_{TT},$$

$$n_C = n_{CC} + n_{CI} + n_{CT},$$

$$n_I = n_{II} + n_{IT},$$

$$n_T = n_{TT}.$$

- The complete data are  $X = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$ , and the observed data are  $Y = (n_C, n_I, n_T)$ .

## 2.1 Missing Data Problems

---

- – **Multinomial Distribution:** Consider an experiment having  $K$  possible outcomes with  $p_i$ ,  $i = 1, \dots, K$ , as the probability for the  $i$ -th outcome, where  $p_i > 0$  and  $p_1 + \dots + p_K = 1$ .

- \* In  $n$  independent trials of this experiment, let  $X_i$ ,  $i = 1, \dots, K$ , be the number of trials resulting in the  $i$ -th outcome.

- \* The probability mass function of  $(X_1, \dots, X_K)$  is

$$f(x_1, \dots, x_K) = \frac{n!}{x_1! \dots x_K!} p_1^{x_1} \dots p_K^{x_K},$$

where  $x_i$ 's are nonnegative integers and  $x_1 + \dots + x_K = n$ .

- \* Such a distribution is called the *multinomial distribution*, denoted by  $\text{Multinomial}(n; p_1, \dots, p_K)$ .

- \* The marginal distribution of each  $X_i$  is the binomial distribution  $Bi(n; p_i)$ , that is,

$$P(X_i = x_i) = \frac{n!}{x_i!(n - x_i)!} p_i^{x_i} (1 - p_i)^{n - x_i} \quad \text{for } x_i = 0, 1, \dots, n.$$

## 2.1 Missing Data Problems

---

- – If we have the complete data  $X = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$ , the log-likelihood function is

$$\begin{aligned}\log l(\theta|X) = c_1 &+ n_{CC} \log p_C^2 + n_{CI} \log(2p_C p_I) + n_{CT} \log(2p_C p_T) \\ &+ n_{II} \log p_I^2 + n_{IT} \log(2p_I p_T) + n_{TT} \log p_T^2,\end{aligned}$$

where  $c_1$  is a constant. **Note that**  $p_C + p_I + p_T = 1$ . The MLE of  $\theta = (p_C, p_I, p_T)$  is

$$\begin{aligned}\hat{p}_{C,X} &= \frac{2n_{CC} + n_{CI} + n_{CT}}{2n}, \\ \hat{p}_{I,X} &= \frac{n_{CI} + 2n_{II} + n_{IT}}{2n}, \\ \hat{p}_{T,X} &= \frac{n_{CT} + n_{IT} + 2n_{TT}}{2n}.\end{aligned}$$

## 2.1 Missing Data Problems

---

- – We can only observe  $Y = (n_C, n_I, n_T)$ . The logarithm of the observed-data likelihood is

$$\begin{aligned}\log l(\theta|Y) = c_2 + (n_{CC} + n_{CI} + n_{CT}) \log(p_C^2 + 2p_Cp_I + 2p_Cp_T) \\ + (n_{II} + n_{IT}) \log(p_I^2 + 2p_Ip_T) + n_{TT} \log p_T^2,\end{aligned}$$

which does not have a closed-form solution.

## 2.2 The EM Algorithm

---

- **EM Algorithm:** Let  $X$  and  $Y$  be the complete data and observed data, respectively. Define

$$Q(\theta, \theta') = E_{\theta'} [\log l(\theta|X)|Y = y] = \int [\log f_X(x; \theta)] f_{X|Y}(x|y; \theta') dx$$

Suppose at iteration  $t$ , we have  $\theta^{(t)}$ . We update  $\theta^{(t)}$  as follows.

- The *E-Step (Expectation Step)*: Calculate

$$Q(\theta, \theta^{(t)}) = E_{\theta^{(t)}} [\log l(\theta|X)|Y = y].$$

Note that given  $\theta^{(t)}$  and  $Y = y$ ,  $Q(\theta, \theta^{(t)})$  is a function of  $\theta$ .

- The *M-Step (Maximization Step)*: Let

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{(t)}) = \arg \max_{\theta} E_{\theta^{(t)}} [\log l(\theta|X)|Y = y].$$

## 2.2 The EM Algorithm

---

- **Remarks:**

- When  $X = (Y, Z)$ , the EM algorithm has the update equation

$$\theta^{(t+1)} = \arg \max_{\theta} \int [\log f_{YZ}(y, z; \theta)] f_{Z|Y}(z|y; \theta^{(t)}) dz.$$

The standard MLE can be written as

$$\hat{\theta}_{MLE} = \arg \max_{\theta} \left[ \log \int f_{YZ}(y, z; \theta) dz \right].$$

In many cases, it is easier to maximize  $\int [\log f_{YZ}(y, z; \theta)] f_{Z|Y}(z|y; \theta^{(t)}) dz$  than  $\log \int f_{YZ}(y, z; \theta) dz$ .

- The EM algorithm needs to update  $\theta^{(t)}$  iteratively.
- We can stop the algorithm when  $\theta^{(t+1)} - \theta^{(t)}$  is close to 0.

## 2.2 The EM Algorithm

---

- **Convergence of EM Algorithm:**

- Note that  $f_Y(y; \theta)f_{X|Y}(x|y; \theta) = f_{XY}(x, y; \theta) = f_X(x; \theta)$ , then

$$\log l(\theta|Y = y) = \log f_Y(y; \theta) = \log f_X(x; \theta) - \log f_{X|Y}(x|y; \theta).$$

- Multiply both sides of the equation by  $f_{X|Y}(x|y; \theta')$ , and take integration with respect to  $x$ , we have

$$\begin{aligned}\log l(\theta|Y = y) &= \int [\log f_X(x; \theta)] f_{X|Y}(x|y; \theta') dx \\ &\quad - \int [\log f_{X|Y}(x|y; \theta)] f_{X|Y}(x|y; \theta') dx \\ &= Q(\theta, \theta') - H(\theta, \theta')\end{aligned}$$

for any  $\theta' \in \Theta$ . Here  $H(\theta, \theta') = \int [\log f_{X|Y}(x|y; \theta)] f_{X|Y}(x|y; \theta') dx$ .

- Using the Jensen's inequality, we can prove that

$$H(\theta, \theta') \leq H(\theta', \theta')$$

for any  $\theta, \theta' \in \Theta$ . (**How to prove it?**)

## 2.2 The EM Algorithm

---

- – We have

$$\begin{aligned}\log l(\theta^{(t+1)}|Y = y) &= Q(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) \\ &\geq Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) \\ &\geq Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) \\ &= \log l(\theta^{(t)}|Y = y).\end{aligned}$$

- So  $\log l(\theta^{(0)}|Y = y), \log l(\theta^{(1)}|Y = y), \dots, \log l(\theta^{(t)}|Y = y), \dots$  forms a non-decreasing sequence, which has a limit when  $t$  goes to infinity.
- Under certain conditions, it can be shown that  $\theta^{(t)}$  converges to a stationary point of the likelihood function.



## 2.2 The EM Algorithm

---

- **Example: Censored Data.** Assume that  $X_1, \dots, X_n$  are i.i.d. from the  $N(\theta, 1)$  distribution, but the data is right censored at the value  $a$ . That is, we can only observe  $Y_i = h(x_i) = \min\{X_i, a\}$ .

- For convenience, we re-arrange the observations and assume that  $y_i < a$  for  $i = 1, \dots, m$  and  $y_i = a$  for  $i = m + 1, \dots, n$ .
- Th E-step:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E_{\theta^{(t)}} \left[ \sum_{i=1}^n \log \phi(X_i - \theta) \middle| Y_1 = y_1, \dots, Y_n = y_n \right] \\ &= \sum_{i=1}^m E_{\theta^{(t)}} \left[ \log \phi(X_i - \theta) \middle| Y_i = y_i \right] + \sum_{i=m+1}^n E_{\theta^{(t)}} \left[ \log \phi(X_i - \theta) \middle| Y_i = a \right] \\ &= c - \frac{1}{2} \sum_{i=1}^m E_{\theta^{(t)}} \left[ (X_i - \theta)^2 \middle| X_i = y_i \right] - \frac{1}{2} \sum_{i=m+1}^n E_{\theta^{(t)}} \left[ (X_i - \theta)^2 \middle| X_i > a \right]. \end{aligned}$$

## 2.2 The EM Algorithm

---

- – Obviously,

$$E_{\theta^{(t)}} \left[ (X_i - \theta)^2 \middle| X_i = y_i \right] = (y_i - \theta)^2.$$

- We also have

$$E_{\theta^{(t)}} \left[ (X_i - \theta)^2 \middle| X_i > a \right] = \int_a^\infty (x - \theta)^2 f_X(x | X_i > a; \theta^{(t)}) dx$$

where  $f_X(x | X_i > a; \theta^{(t)})$  is the density function of  $X_i$  conditional on  $X_i > a$  when  $X_i \sim N(\theta^{(t)}, 1)$ . It is easy to show that

$$f_X(x | X_i > a; \theta^{(t)}) = \frac{f_X(x; \theta^{(t)})}{P(X_i > a; \theta^{(t)})} = \frac{\phi(x - \theta^{(t)})}{1 - \Phi(a - \theta^{(t)})}, \quad x > a.$$

- Therefore,

$$Q(\theta, \theta^{(t)}) = c - \frac{1}{2} \sum_{i=1}^m (y_i - \theta)^2 - \frac{1}{2} \sum_{i=m+1}^n \int_a^\infty \frac{(x - \theta)^2 \phi(x - \theta^{(t)})}{1 - \Phi(a - \theta^{(t)})} dx.$$

## 2.2 The EM Algorithm

---

- – Note that

$$\begin{aligned} Q'(\theta, \theta^{(t)}) &= \sum_{i=1}^m (y_i - \theta) + (n - m) \int_a^\infty \frac{(x - \theta^{(t)} + \theta^{(t)} - \theta) \phi(x - \theta^{(t)})}{1 - \Phi(a - \theta^{(t)})} dx \\ &= \sum_{i=1}^m y_i - m\theta + (n - m)\theta^{(t)} - (n - m)\theta \\ &\quad + \frac{n - m}{1 - \Phi(a - \theta^{(t)})} \int_a^\infty (x - \theta^{(t)}) \phi(x - \theta^{(t)}) dx \\ &= \sum_{i=1}^m y_i + (n - m)\theta^{(t)} - n\theta + \frac{(n - m)}{1 - \Phi(a - \theta^{(t)})} \int_{a - \theta^{(t)}}^\infty u \phi(u) du \\ &= \sum_{i=1}^m y_i + (n - m)\theta^{(t)} - n\theta + \frac{(n - m)}{1 - \Phi(a - \theta^{(t)})} \int_{a - \theta^{(t)}}^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}} d\frac{u^2}{2} \\ &= \sum_{i=1}^m y_i + (n - m)\theta^{(t)} - n\theta + \frac{(n - m)}{1 - \Phi(a - \theta^{(t)})} \cdot \frac{e^{-(a - \theta^{(t)})^2/2}}{\sqrt{2\pi}}. \end{aligned}$$

– The M-step:

$$\begin{aligned}\theta^{(t+1)} &= \frac{1}{n} \left[ \sum_{i=1}^m y_i + (n-m)\theta^{(t)} + \frac{(n-m)}{1 - \Phi(a - \theta^{(t)})} \cdot \frac{e^{-(a-\theta^{(t)})^2/2}}{\sqrt{2\pi}} \right] \\ &= \frac{1}{n} \left[ \sum_{i=1}^m y_i + (n-m)\theta^{(t)} + \frac{(n-m)\phi(a - \theta^{(t)})}{1 - \Phi(a - \theta^{(t)})} \right].\end{aligned}$$

## 2.2 The EM Algorithm

---

- **Example: Mixture Normal Distribution.** Assume that  $U_i = (U_{i1}, \dots, U_{iK})$ , where  $U_{i1}, \dots, U_{iK}$  are independent with  $U_{ik} \sim N(\mu_k, \sigma_k^2)$ . Further assume that  $Z_i \in \{1, \dots, K\}$  is a discrete random variable independent of  $U_i$  with  $P(Z_i = k) = p_k$ . Here  $p_1 + \dots + p_K = 1$ . Define

$$Y_i = \sum_{k=1}^K I(Z_i = k)U_{ik}$$

and  $X_i = (Y_i, Z_i)$ ,  $i = 1, \dots, n$ . The parameters of interests are  $\mu_k$ ,  $\sigma_k^2$  and  $p_k$ ,  $k = 1, \dots, K$ .

- The logarithm of the complete-data likelihood is

$$\begin{aligned} \log l(\theta | X_1, \dots, X_n) &= \sum_{i=1}^n \log f(X_i; \theta) = \sum_{i=1}^n \log [f(Y_i | Z_i; \theta) f(Z_i; \theta)] \\ &= \sum_{i=1}^n \sum_{k=1}^K I(Z_i = k) \left[ \log p_k - \log \sigma_k + \log \phi\left(\frac{Y_i - \mu_k}{\sigma_k}\right) \right]. \end{aligned}$$

## 2.2 The EM Algorithm

---

- – Given  $Y_i = y_i$ , we have

$$\begin{aligned} E_{\theta^{(t)}} [I(Z_i = k) | Y_i = y_i] &= P(Z_i = k | Y_i = y_i; \theta^{(t)}) \\ &= \frac{P(Z_i = k, Y_i = y_i; \theta^{(t)})}{P(Y_i = y_i; \theta^{(t)})} \\ &= \frac{P(Z_i = k; \theta^{(t)}) P(Y_i = y_i | Z_i = k; \theta^{(t)})}{\sum_{s=1}^K P(Z_i = s; \theta^{(t)}) P(Y_i = y_i | Z_i = s; \theta^{(t)})} \\ &= \frac{p_k^{(t)} \cdot \frac{1}{\sigma_k^{(t)}} \phi\left(\frac{y_i - \mu_k^{(t)}}{\sigma_k^{(t)}}\right)}{\sum_{s=1}^K p_s^{(t)} \cdot \frac{1}{\sigma_s^{(t)}} \phi\left(\frac{y_i - \mu_s^{(t)}}{\sigma_s^{(t)}}\right)} := q_{k,i}^{(t)}. \end{aligned}$$

- The E-step:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E_{\theta^{(t)}} \left[ \log l(\theta | X_1, \dots, X_n) \middle| Y_1 = y_1, \dots, Y_n = y_n \right] \\ &= \sum_{i=1}^n \sum_{k=1}^K q_{k,i}^{(t)} \left[ \log p_k - \frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + c \right]. \end{aligned}$$

## 2.2 The EM Algorithm

---

- – The M-step: Solving the equation  $\nabla Q(\theta, \theta^{(t)}) = \mathbf{0}$ , we obtain

$$\begin{aligned} p_k^{(t+1)} &= \frac{\sum_{i=1}^n q_{k,i}^{(t)}}{\sum_{i=1}^n \sum_{s=1}^K q_{s,i}^{(t)}} = \frac{1}{n} \sum_{i=1}^n q_{k,i}^{(t)}, \\ \mu_k^{(t+1)} &= \frac{\sum_{i=1}^n q_{k,i}^{(t)} y_i}{\sum_{i=1}^n q_{k,i}^{(t)}}, \\ (\sigma_k^2)^{(t+1)} &= \frac{\sum_{i=1}^n q_{k,i}^{(t)} (y_i - \mu_k^{(t+1)})^2}{\sum_{i=1}^n q_{k,i}^{(t)}}. \end{aligned}$$

(Note that  $p_1 + \cdots + p_K = 1$  and  $q_{1,i}^{(t)} + \cdots + q_{K,i}^{(t)} = 1$ .)

## 2.2 The EM Algorithm

---

- **Example: Peppered Moths.** Suppose we capture  $n$  moths, of which there are  $n_C$ ,  $n_I$ , and  $n_T$  of the carbonaria, insularia, and typica phenotypes, respectively. We also use  $n_{CC}$ ,  $n_{CI}$ ,  $n_{CT}$ ,  $n_{II}$ ,  $n_{IT}$ , and  $n_{TT}$  to denote counts for genotypes CC, CI, CT, II, IT, and TT, respectively. We want to estimate the parameter  $\theta = (p_C, p_I, p_N)$ .

- The complete data are  $X = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$ , and the observed data are  $Y = (n_C, n_I, n_T)$ .
- The log-likelihood function of complete data is

$$\begin{aligned}\log l(\theta|X) = & c_1 + n_{CC} \log p_C^2 + n_{CI} \log(2p_C p_I) + n_{CT} \log(2p_C p_T) \\ & + n_{II} \log p_I^2 + n_{IT} \log(2p_I p_T) + n_{TT} \log p_T^2.\end{aligned}$$

- We want to calculate

$$Q(\theta, \theta^{(t)}) = E_{\theta^{(t)}} \left[ \log l(\theta|X) \middle| Y = (n_C, n_I, n_T) \right].$$



## 2.2 The EM Algorithm

---

- Note that  $n_C = n_{CC} + n_{CI} + n_{CT}$ . It is easy to know given  $\theta^{(t)}$  and  $n_C$ ,  $(n_{CC}, n_{CI}, n_{CT})$  follows a multinomial distribution with cell probabilities proportional to  $((p_C^{(t)})^2, 2p_C^{(t)}p_I^{(t)}, 2p_C^{(t)}p_T^{(t)})$ . Thus,

$$E_{\theta^{(t)}}[n_{CC}|n_C] = n_C \cdot \frac{(p_C^{(t)})^2}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}} := n_{CC}^{(t)},$$

$$E_{\theta^{(t)}}[n_{CI}|n_C] = n_C \cdot \frac{2p_C^{(t)}p_I^{(t)}}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}} := n_{CI}^{(t)},$$

$$E_{\theta^{(t)}}[n_{CT}|n_C] = n_C \cdot \frac{2p_C^{(t)}p_T^{(t)}}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}} := n_{CT}^{(t)}.$$

– Similarly,

$$E_{\theta^{(t)}}[n_{II}|n_I] = n_I \cdot \frac{(p_I^{(t)})^2}{(p_I^{(t)})^2 + 2p_I^{(t)}p_T^{(t)}} := n_{II}^{(t)},$$

$$E_{\theta^{(t)}}[n_{IT}|n_I] = n_I \cdot \frac{2p_I^{(t)}p_T^{(t)}}{(p_I^{(t)})^2 + 2p_I^{(t)}p_T^{(t)}} := n_{IT}^{(t)}.$$

## 2.2 The EM Algorithm

---

- – Obviously,  $E_{\theta^{(t)}}[n_{TT}|n_T] = n_{TT}$ .
- The E-step:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E_{\theta^{(t)}} \left[ \log l(\theta|X) \middle| Y = (n_C, n_I, n_T) \right] \\ &= c_1 + n_{CC}^{(t)} \log p_C^2 + n_{CI}^{(t)} \log(2p_C p_I) + n_{CT}^{(t)} \log(2p_C p_T) \\ &\quad + n_{II}^{(t)} \log p_I^2 + n_{IT}^{(t)} \log(2p_I p_T) + n_{TT} \log p_T^2. \end{aligned}$$

- The M-step: Solving the equation  $\nabla Q(\theta, \theta^{(t)}) = 0$ , we obtain

$$\begin{aligned} p_C^{(t+1)} &= \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}, \\ p_I^{(t+1)} &= \frac{n_{CI}^{(t)} + 2n_{II}^{(t)} + n_{IT}^{(t)}}{2n}, \\ p_T^{(t+1)} &= \frac{n_{CT}^{(t)} + n_{IT}^{(t)} + 2n_{TT}}{2n}. \end{aligned}$$

## 2.3 EM Variants

---

- **Monte Carlo EM:** When  $Q(\theta, \theta^{(t)}) = E_{\theta^{(t)}} [\log l(\theta|X)|Y = y]$  is difficult to compute analytically, it can be approximated via the Monte Carlo method. At each iteration  $t$ ,
  - Draw samples  $X_1^{(t)}, \dots, X_m^{(t)}$  from the conditional distribution  $f_{X|Y}(x|y; \theta^{(t)})$ .
  - The E-step: Calculate

$$\hat{Q}(\theta, \theta^{(t)}) = \frac{1}{m} \sum_{i=1}^m \log l(\theta|X_i^{(t)}) \xrightarrow{a.s.} E_{\theta^{(t)}} [\log l(\theta|X)|Y = y].$$

- The M-step: let

$$\theta^{(t+1)} = \arg \max_{\theta} \hat{Q}(\theta, \theta^{(t)}).$$

## 2.3 EM Variants

---

- **EM Gradient Algorithm:** In some cases, it may not be easy to find  $\theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{(t)})$ .

– Actually, if  $\theta^{(t+1)}$  satisfies  $Q(\theta^{(t+1)}, \theta^{(t)}) > Q(\theta^{(t)}, \theta^{(t)})$ , then

$$\begin{aligned} \log l(\theta^{(t+1)} | Y = y) &= Q(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) \\ &> Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) \\ &\geq Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) \\ &= \log l(\theta^{(t)} | Y = y). \end{aligned}$$

– We let  $g(\theta) = Q(\theta, \theta^{(t)})$  and apply the Newton's iteration once with  $\theta^{(t)}$  as the initial value, that is,

$$\begin{aligned} \theta^{(t+1)} &= \theta^{(t)} - [\nabla^2 g(\theta^{(t)})]^{-1} \nabla g(\theta^{(t)}) \\ &= \theta^{(t)} - [\nabla_{\theta}^2 Q(\theta, \theta^{(t)})]^{-1} \big|_{\theta=\theta^{(t)}} \nabla_{\theta} Q(\theta, \theta^{(t)}) \big|_{\theta=\theta^{(t)}}. \end{aligned}$$

## 2.4 The Minorization-Maximization (MM) Algorithm

---

- **Problem:** We want to maximize an objective function  $g(\theta)$  for  $\theta \in \Theta$ .

- **The MM algorithm:**

- Suppose that there exists a function  $M(\theta, \theta^*)$  satisfying

$$M(\theta, \theta^*) \leq g(\theta) \quad \text{and} \quad M(\theta^*, \theta^*) = g(\theta^*)$$

for all  $\theta, \theta^* \in \Theta$ , it is said that  $M(\theta, \theta^*)$  *minorizes*  $g(\theta)$ .

- For each given  $\theta^*$ ,  $m(\theta) := M(\theta, \theta^*)$  is an “optimal” lower bound of  $g(\theta)$ .

- We iteratively maximize the minorizing function by letting

$$\theta^{(t+1)} = \arg \max_{\theta \in \Theta} M(\theta, \theta^{(t)}).$$

for  $t = 0, 1, 2, \dots$ .

## 2.4 The Minorization-Maximization (MM) Algorithm

---

- **Remarks:**

- It can be seen that

$$g(\theta^{(t)}) = M(\theta^{(t)}, \theta^{(t)}) \leq M(\theta^{(t+1)}, \theta^{(t)}) \leq g(\theta^{(t+1)}).$$

- $\theta^{(t+1)}$  need not to be the maximum point of  $M(\theta, \theta^{(t)})$ . It is suffice to find  $\theta^{(t+1)}$  such that

$$M(\theta^{(t+1)}, \theta^{(t)}) > M(\theta^{(t)}, \theta^{(t)}).$$

- The EM algorithm is a special case of the MM algorithm.

- \* Let  $X$  and  $Y$  be the complete data and observed data, respectively.

- \* Define

$$Q(\theta, \theta^*) := E_{\theta^*} [\log l(\theta|X)|Y = y] = \int [\log f_X(x; \theta)] f_{X|Y}(x|y; \theta^*) dx$$


and

$$H(\theta, \theta^*) := \int [\log f_{X|Y}(x|y; \theta)] f_{X|Y}(x|y; \theta^*) dx.$$

## 2.4 The Minorization-Maximization (MM) Algorithm

---

- – \* We have


$$\begin{aligned}\log l(\theta|Y = y) &= \log f_Y(y; \theta) \\ &= \log f_X(x; \theta) - \log f_{X|Y}(x|y; \theta) \\ &= \int [\log f_X(x; \theta)] f_{X|Y}(x|y; \theta^*) dx \\ &\quad - \int [\log f_{X|Y}(x|y; \theta)] f_{X|Y}(x|y; \theta^*) dx \\ &= Q(\theta, \theta^*) - H(\theta, \theta^*)\end{aligned}$$

for any  $\theta, \theta^* \in \Theta$ .

- \* Using the Jensen's inequality, we also know that

$$H(\theta, \theta^*) \leq H(\theta^*, \theta^*)$$

## 2.4 The Minorization-Maximization (MM) Algorithm

---

- – \* Let  $M(\theta, \theta^*) = Q(\theta, \theta^*) - H(\theta^*, \theta^*)$ . Then

$$\begin{aligned} M(\theta, \theta^*) &= Q(\theta, \theta^*) - H(\theta^*, \theta^*) \\ &\leq Q(\theta, \theta^*) - H(\theta, \theta^*) = \log l(\theta|Y = y) \end{aligned}$$

and

$$\begin{aligned} M(\theta^*, \theta^*) &= Q(\theta^*, \theta^*) - H(\theta^*, \theta^*) \\ &= \log l(\theta^*|Y = y). \end{aligned}$$

So  $M(\theta, \theta^*)$  minorize  $\log l(\theta|Y = y)$ .

- \* The EM algorithm calculates

$$\begin{aligned} \theta^{(t+1)} &= \arg \max_{\theta \in \Theta} Q(\theta, \theta^{(t)}) \\ &= \arg \max_{\theta \in \Theta} \{Q(\theta, \theta^{(t)}) - -H(\theta^{(t)}, \theta^{(t)})\} = \arg \max_{\theta \in \Theta} M(\theta, \theta^{(t)}). \end{aligned}$$



## 2.4 The Minorization-Maximization (MM) Algorithm

---

- How to find the minorizing function  $M(\theta, \theta^*)$ ?
- One way is to use the second order Taylor series expansion.

– Note that

$$g(\theta) = g(\theta^*) + (\theta - \theta^*)^T \nabla g(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T \nabla^2 g(\bar{\theta})(\theta - \theta^*),$$

where  $\bar{\theta}$  is a point between  $\theta$  and  $\theta^*$ .

– Suppose that we can find a **positive definite matrix**  $B$  so that  $\nabla^2 g(\theta) - B$  is nonnegative definite for all  $\theta$ , denoted by  $B \leq \nabla^2 g(\theta)$ .

– Define

$$M(\theta, \theta^*) := g(\theta^*) + (\theta - \theta^*)^T \nabla g(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T B(\theta - \theta^*).$$

Then  $M(\theta, \theta^*)$  minorizes  $g(\theta)$ .

– The MM algorithm let  $\theta^{(t+1)} = \theta^{(t)} - B^{-1} \nabla g(\theta^{(t)})$ .

## 2.4 The Minorization-Maximization (MM) Algorithm

---

- **Example: Logistic Regression.** Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be independent and identically distributed random vectors, where  $Y_i \in \{0, 1\}$  and  $X_i$ ,  $i = 1, \dots, n$ , are  $d$ -dimensional random vectors. Consider a logistic regression model

$$P(Y_i = 1 \mid X_i = x_i, \beta) = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}.$$

Given observations  $(x_1, y_1), \dots, (x_n, y_n)$ , we want to estimate  $\beta = (\beta_1, \dots, \beta_d)^T$ .

- The log-likelihood function is

$$\log l(\beta) = c + \sum_{i=1}^n y_i \cdot x_i^T \beta - \sum_{i=1}^n \log [1 + \exp\{x_i^T \beta\}].$$

## 2.4 The Minorization-Maximization (MM) Algorithm

---

- – The gradient and Hessian of  $\log l(\beta)$  are

$$\nabla \log l(\beta) = \sum_{i=1}^n y_i \cdot x_i - \sum_{i=1}^n \frac{1}{1 + \exp\{-x_i^T \beta\}} \cdot x_i$$

and

$$\nabla^2 \log l(\beta) = - \sum_{i=1}^n \frac{\exp\{-x_i^T \beta\}}{[1 + \exp\{-x_i^T \beta\}]^2} \cdot x_i x_i^T.$$

- Note that

$$-\frac{1}{4} \sum_{i=1}^n x_i x_i^T \leq - \sum_{i=1}^n \frac{\exp\{-x_i^T \beta\}}{[1 + \exp\{-x_i^T \beta\}]^2} \cdot x_i x_i^T.$$

- The MM algorithm let

$$\beta^{(t+1)} = \beta^{(t)} + \left[ \frac{1}{4} \sum_{i=1}^n x_i x_i^T \right]^{-1} \nabla \log l(\beta^{(t)}).$$

We don't need to calculate  $\nabla^2 \log l(\beta^{(t)})$  for each iteration.

## Homework

---

1. Assume that  $U_i = (U_{i1}, \dots, U_{iK})$ , where  $U_{i1}, \dots, U_{iK}$  are independent with  $U_{ik} \sim N(\mu_k, \sigma_k^2)$ . Further assume that  $Z_i \in \{1, \dots, K\}$  is a discrete random variable independent of  $U_i$  with  $P(Z_i = k) = p_k$ , where  $p_1 + \dots + p_K = 1$ . Define

$$Y_i = \sum_{k=1}^K I(Z_i = k) U_{ik}$$

and  $X_i = (Y_i, Z_i)$ ,  $i = 1, \dots, n$ . Given  $X_1, \dots, X_n$  (assume they are independent), find the MLE of  $\theta = (\mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2, p_1, \dots, p_K)$ .

## Homework

---

2. Assume  $X = (X_1, X_2, X_3)$  follows the Multinomial( $n; p_1, p_2, p_3$ ) distribution, where  $p_i > 0$  and  $p_1 + p_2 + p_3 = 1$ .

(a) Use the probability mass function of the multinomial distribution to prove that  $X_1 + X_2$  follows a Binomial( $n; p_1 + p_2$ ) distribution.

(b) Prove that given  $X_1 + X_2 = m$ ,  $m \leq n$ ,  $(X_1, X_2)$  follows the distribution

$$\text{Multinomial} \left( m; \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \right).$$

Equivalently, given  $X_1 + X_2 = m$ ,  $X_1$  follows a Binomial( $m; p_1/(p_1 + p_2)$ ) distribution.

## Homework

---

3. Consider the peppered moth example with  $n_C = 85$ ,  $n_I = 196$ , and  $n_T = 341$ . Suppose the sample collected by the researchers actually included  $n_U = 578$  **more moths** that were known to be insularia or typica but whose exact phenotypes could not be determined.
- (a) Derive the EM algorithm for maximum likelihood estimation of  $p_C$ ,  $p_I$ , and  $p_T$  for this modified problem having observed data  $n_C$ ,  $n_I$ ,  $n_T$ , and  $n_U$  as given above.
- (b) Apply the algorithm to find the MLE of  $p_C$ ,  $p_I$ , and  $p_T$ .