Chapter 1
Incomplete Data and the EM Algorithm

Qing Zhou[*,†]

Contents

1. Assumptions

Reading: Schafer (1997), Section 2.1 to 2.3.
Let Y be an $n \times p$ matrix of complete data, $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, y_i be the i^{th} row of Y, $i = 1, \ldots, n$.

Example of missing data

Variables	1	2	...	p
1				
2		?		?
...				
n	?	?		

*UCLA Department of Statistics (email: zhou@stat.ucla.edu).
†I thank Elvis Cui for typesetting part of this chapter in LaTex.

Under the iid assumption, the probability density of Y

$$p(Y \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta),$$

where θ is the parameter for this data generation model.

1.1. Ignorability

Missing at random (MAR) is defined in terms of a probability model for the missingness. Let $R = (r_{ij})$ be an $n \times p$ matrix of indicator variables: $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ otherwise. We put a probability model for R, $p(R \mid Y, \xi)$, where ξ is some parameter. The MAR assumption is that

$$p(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \xi) = p(R \mid Y_{\text{obs}}, \xi), \tag{1}$$

that is, $R \perp Y_{\text{mis}} \mid Y_{\text{obs}}$. A stronger assumption is missing completely at random (MCAR): $R \perp (Y_{\text{mis}}, Y_{\text{obs}})$. If neither holds, then the data are missing not at random (MNAR): R depends on Y_{mis}.
Consider an example in Mohan and Pearl (2021): A study in a school measured age (A), gender (G), and obesity (O) for students, with missing values in O since some students fail to reveal weight.

- MCAR: some students accidentally lost questionnaires ($R \perp A, G, O$).
- MAR: some teenagers not reporting weight ($R \perp O \mid A$).
- MNAR: overweight students reluctant to report weight ($O \to R$).

Distinctness of parameters. Let θ denote the parameters of the data model, and ξ the parameters of the missingness mechanism. Then, θ and ξ are distinct if
(a) **Bayesian**: any joint prior on (θ, ξ) must factor into independent marginal priors for θ and ξ, that is:

$$\pi(\theta, \xi) = \pi_\theta(\theta)\pi_\xi(\xi).$$

(b) **Frequentist**: joint parameter space of (θ, ξ) is the Cartesian product of the individual parameter spaces for $\theta \in \Theta$ and $\xi \in \Gamma$. That is:

$$(\theta, \xi) \in \Theta \times \Gamma.$$

MAR & distinctness $\Rightarrow$ the missing-data mechanism is **ignorable**.

1.2. Observed data likelihood and posterior

$$
\begin{aligned}
\mathbb{P}(R, Y_{obs}|\theta, \xi) &= \int_{\Omega_{miss}} \mathbb{P}(R, Y|\theta, \xi) dY_{miss} \\
&= \int \mathbb{P}(R|Y, \theta, \xi)\mathbb{P}(Y|\theta, \xi) dY_{miss} \\
&= \int \mathbb{P}(R|Y, \xi)\mathbb{P}(Y|\theta) dY_{miss} \\
&= \mathbb{P}(R|Y_{obs}, \xi) \int \mathbb{P}(Y|\theta) dY_{miss} \qquad \text{by MAR} \\
&= \mathbb{P}(R|Y_{obs}, \xi)\mathbb{P}(Y_{obs}|\theta).
\end{aligned}
$$

Consider the maximum likelihood estimate (MLE) of (θ, ξ). Under distinctness,

$$
\max_{(\theta,\xi)\in\Theta\times\Gamma} \mathbb{P}(R, Y_{obs}|\theta, \xi) = \left\{ \max_{\xi\in\Gamma} \mathbb{P}(R|Y_{obs}, \xi) \right\} \left\{ \max_{\theta\in\Theta} \mathbb{P}(Y_{obs}|\theta) \right\}
$$

is separable. Define the observed-data likelihood $L(\theta|Y_{obs}) := \mathbb{P}(Y_{obs}|\theta)$. If both MAR and distinctness hold, we have the following MLE of θ:

$$
\widehat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\theta\in\Theta} \mathbb{P}(Y_{obs}|\theta) = \operatorname*{argmax}_{\theta\in\Theta} L(\theta|Y_{obs}).
$$

Now for the *posterior* distribution of the parameters:

$$
\begin{aligned}
\mathbb{P}(\theta, \xi|Y_{obs}, R) &\propto \mathbb{P}(R, Y_{obs}|\theta, \xi)\pi(\theta, \xi) \\
&=^{\text{MAR}} \mathbb{P}(R|Y_{obs}, \xi)\mathbb{P}(Y_{obs}|\theta)\pi(\theta, \xi) \\
&=^{\text{Distinctness}} \mathbb{P}(R|Y_{obs}, \xi)\mathbb{P}(Y_{obs}|\theta)\pi_\theta(\theta)\pi_\xi(\xi).
\end{aligned}
$$

Then we could derive the posterior of θ:

$$
\begin{aligned}
\mathbb{P}(\theta|Y_{obs}, R) &= \int \mathbb{P}(\theta, \xi|Y_{obs}, R) d\xi \\
&\propto \mathbb{P}(Y_{obs}|\theta)\pi_\theta(\theta) \int h(R, Y_{obs}, \xi) d\xi \\
&\propto L(\theta|Y_{obs})\pi_\theta(\theta),
\end{aligned}
$$

where $h(R, Y_{obs}, \xi)$ is a function independent of θ and $L(\theta|Y_{obs})$ is the observed data likelihood. Therefore, the observed-data posterior:

$$
\mathbb{P}(\theta|Y_{obs}, R) = \mathbb{P}(\theta|Y_{obs}) \propto \mathbb{P}(Y_{obs}|\theta)\pi_\theta(\theta).
$$

2. The EM algorithm and its properties

Reading: Schafer (1997), Section 3.2 and 3.3. Also see Dempster, Laird and Rubin (1977) and Wu (1983).
Recall that our goal is to find:

$$
\widehat{\theta}_{\text{MLE}} = \operatorname*{argmax}_{\theta\in\Theta} \mathbb{P}(Y_{\text{obs}}|\theta) = \operatorname*{argmax}_{\theta\in\Theta} \int \mathbb{P}(Y_{\text{obs}}, Y_{\text{miss}}|\theta) dY_{\text{miss}}.
$$

2.1. The algorithm

Definition 1 (EM Algorithm)**.** First, start with an initial $\theta^{(0)}$. For the $(t+1)^{th}$ iteration:

- E-step: Calculate the expectation of complete-data log-likelihood:

$$Q(\theta|\theta^{(t)}) := \mathbb{E}[\log \mathbb{P}(Y_{obs}, Y_{miss}|\theta)|Y_{obs}, \theta^{(t)}].$$

- M-step: Find $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$:

$$\theta^{(t+1)} := \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)}).$$

Iterate the above 2 steps until convergence.

Remark 1. The expectation in the E-step is taken with respect to $\mathbb{P}(Y_{miss}|Y_{obs}, \theta^{(t)})$ (conditional distribution), but not $\mathbb{P}(Y_{miss}|\theta^{(t)})$ (marginal distribution).

Example 1 (Bivariate binary data)**.** Y_1 and Y_2 are correlated binary variables on $\{1, 2\}$. Missing values occur on either Y_1 or Y_2 in an i.i.d. sample of n units. We want to estimate $\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, where $\theta_{ij} := \mathbb{P}(Y_1 = i, Y_2 = j)$. Complete data: $X = (x_{11}, x_{12}, x_{21}, x_{22})$ (2×2 contingency table), where x_{ij} is the number of units with $Y_1 = i$ and $Y_2 = j$. Complete data log-likelihood:

$$\ell(\theta|X) = \sum_{i,j=1}^{2} x_{ij} \log \theta_{ij}.$$

According to the missingness pattern, we partition the n units into three blocks:

A: Both observed

$Y_1\backslash Y_2$	1	2	
1	x^A_{11}	x^A_{12}	x^A_{1+}
2	x^A_{21}	x^A_{22}	x^A_{2+}
	x^A_{+1}	x^A_{+2}	

B: Y_2 missing

$Y_1\backslash Y_2$	1	2	
1			x^B_{1+}
2			x^B_{2+}

C: Y_1 missing

$Y_1\backslash Y_2$	1	2
1		
2		
	x^C_{+1}	x^C_{+2}

Then we have:

$$(x^B_{i1}, x^B_{i2})|Y_{obs}, \theta^{(t)} \sim \mathcal{M}\left(x^B_{i+}, \left(\frac{\theta^{(t)}_{i1}}{\theta^{(t)}_{i+}}, \frac{\theta^{(t)}_{i2}}{\theta^{(t)}_{i+}}\right)\right), \qquad i = 1, 2.$$

$$(x_{1j}^C, x_{2j}^C)|Y_{obs}, \theta^{(t)} \sim \mathcal{M}\left(x_{+j}^C, \left(\frac{\theta_{1j}^{(t)}}{\theta_{+j}^{(t)}}, \frac{\theta_{2j}^{(t)}}{\theta_{+j}^{(t)}}\right)\right), \qquad j = 1, 2.$$

where $\theta_{i+}^{(t)} = \theta_{i1}^{(t)} + \theta_{i2}^{(t)}$, $\theta_{+j}^{(t)} = \theta_{1j}^{(t)} + \theta_{2j}^{(t)}$. Thus we derive the EM algorithm as follows:

- E-step: To calculate $\mathbb{E}[\ell(\theta|X)|Y_{obs}, \theta^{(t)}]$, let

$$x_{ij}^{(t)} := \mathbb{E}(x_{ij}|Y_{obs}, \theta^{(t)}) = x_{ij}^A + x_{i+}^B \frac{\theta_{ij}^{(t)}}{\theta_{i+}^{(t)}} + x_{+j}^C \frac{\theta_{ij}^{(t)}}{\theta_{+j}^{(t)}}, \quad 1 \le i, j \le 2.$$

  Then

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}[\ell(\theta|X)|Y_{obs}, \theta^{(t)}] = \sum_{i,j} x_{ij}^{(t)} \log \theta_{ij}.$$

- M-step: Maximizing $Q(\theta \mid \theta^{(t)})$ subject to $\sum_{i,j} \theta_{ij} = 1$, we have

$$\theta_{ij}^{(t+1)} = \frac{x_{ij}^{(t)}}{n} = \frac{1}{n}\left[x_{ij}^A + x_{i+}^B \frac{\theta_{ij}^{(t)}}{\theta_{i+}^{(t)}} + x_{+j}^C \frac{\theta_{ij}^{(t)}}{\theta_{+j}^{(t)}}\right].$$

2.2. EM as MM Algorithm

MM Algorithm: Minorization-Maximization Algorithm. It was first proposed by Professor Jan de Leeuw at UCLA.
We start with a simple identity:

$$\log \mathbb{P}(Y_{miss}, Y_{obs}|\theta) = \ell(\theta|Y_{obs}) + \log \mathbb{P}(Y_{miss}|Y_{obs}, \theta).$$

Now denote by F any distribution for Y_{miss}. Then re-arrange the above equation to get

$$\ell(\theta|Y_{obs}) = \log \mathbb{P}(Y_{miss}, Y_{obs}|\theta) - \log F(Y_{miss}) + \log \frac{F(Y_{miss})}{\mathbb{P}(Y_{miss}|Y_{obs}, \theta)}.$$

Take expectation on both sides w.r.t. F (L.H.S. is a constant since it does not involve Y_{miss}):

$$\ell(\theta|Y_{obs}) = \mathbb{E}_F[\log \mathbb{P}(Y_{miss}, Y_{obs}|\theta)] + H(F) + D(F\|\mathbb{P}(Y_{miss}|Y_{obs}, \theta)),$$

where $H(F)$ denotes the entropy of distribution F and $D(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. Since $D(\cdot\|\cdot) \ge 0$, thus for any F we have:

$$\ell(\theta|Y_{obs}) \ge \mathbb{E}_F[\log \mathbb{P}(Y_{miss}, Y_{obs}|\theta)] + H(F) := L(\theta, F),$$

and equality holds when $F = \mathbb{P}(Y_{miss}|Y_{obs}, \theta)$. Let $F^{(t)} = \mathbb{P}(Y_{miss}|Y_{obs}, \theta^{(t)})$. Then $L(\theta, F^{(t)})$, called a minorization function of $\ell(\theta|Y_{obs})$, satisfies the following two conditions:

(i) $\ell(\theta|Y_{obs}) \geq L(\theta, F^{(t)})$ for any θ;
(ii) $\ell(\theta^{(t)}|Y_{obs}) = L(\theta^{(t)}, F^{(t)})$.

EM iterates between two steps:

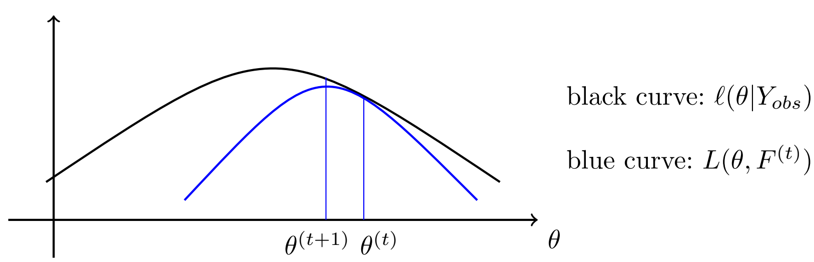
1. Minorization (E-step): Find $L(\theta, F^{(t)})$ by calculating

$$\mathbb{E}_{F^{(t)}}[\log \mathbb{P}(Y_{miss}, Y_{obs}|\theta)] = Q(\theta|\theta^{(t)}).$$

   Note that $L(\theta, F^{(t)}) = Q(\theta|\theta^{(t)}) + H(F^{(t)})$, where $H(F^{(t)})$ is a constant w.r.t θ and thus can be omitted.
2. Maximization (M-step): $\max_\theta L(\theta, F^{(t)}) \Leftrightarrow \max_\theta Q(\theta|\theta^{(t)})$ to obtain $\theta^{(t+1)}$.

Then, we can show the ascent property (Proposition 1) of the EM:

$$\begin{aligned}
\ell(\theta^{(t+1)}|Y_{obs}) &\geq L(\theta^{(t+1)}, F^{(t)}) && \text{by (i)} \\
&\geq L(\theta^{(t)}, F^{(t)}) && \text{M-step} \\
&= \ell(\theta^{(t)}|Y_{obs}). && \text{by (ii)}
\end{aligned}$$



2.3. Properties of the EM

To establish the ascent property of the EM algorithm, we need the following inequality:

Lemma 1 (Jensen's inequality)**.** *Assume that a random variable W is defined in the interval (a, b). If $h(W)$ is convex on (a, b), then*

$$\mathbb{E}[h(W)] \geq h[\mathbb{E}(W)],$$

provided that both expectations exist. For a strictly convex function, equality hold iff $W = \mathbb{E}(W)$ a.s.

Proof. Use the supporting hyperplane theorem. Denote $g(W)$ as the supporting hyperplane of $h(W)$ at point $w_0 = \mathbb{E}(W)$. By convexity, we have $h(w) \geq g(w)$ $\forall w \in (a, b)$, and thus,

$$\mathbb{E}[h(W)] \geq \mathbb{E}[g(W)] = g[\mathbb{E}(W)] = h[\mathbb{E}(W)].$$

The second equality is due to the linearity of $\mathbb{E}(\cdot)$ and $g(\cdot)$. □

Proposition 1 (Ascent property of the EM)**.** *Let $\ell(\theta|Y_{obs}) := \log \mathbb{P}(Y_{obs}|\theta)$, which is the observed-data log-likelihood. Then the EM iterations satisfy*

$$\ell(\theta^{(t+1)}|Y_{obs}) \geq \ell(\theta^{(t)}|Y_{obs}).$$

Proof. There are three crucial steps. First, write

$$\ell(\theta|Y_{obs}) = \log \mathbb{P}(Y_{obs}|\theta) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}),$$

where

$$H(\theta|\theta^{(t)}) = \int [\log \mathbb{P}(Y_{miss}|Y_{obs},\theta)]\, \mathbb{P}(Y_{miss}|Y_{obs},\theta^{(t)}) dY_{mis}.$$

Note that $-H(\theta^{(t)}|\theta^{(t)})$ is the entropy of the distribution $[Y_{miss}|Y_{obs},\theta^{(t)}]$. Second, we have

$$Q(\theta^{(t)}|\theta^{(t)}) \leq Q(\theta^{(t+1)}|\theta^{(t)})$$

since $\theta^{(t+1)}$ is a maximizer of $Q(\bullet|\theta^{(t)})$. Third, note that by Jensen's inequality and convexity of $-\log(\cdot)$:

$$H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) = \mathbb{E}\left\{ \log \frac{\mathbb{P}(Y_{miss}|Y_{obs},\theta^{(t)})}{\mathbb{P}(Y_{miss}|Y_{obs},\theta^{(t+1)})} \middle| Y_{obs},\theta^{(t)} \right\} \geq 0.$$

Therefore,

$$\begin{aligned}\ell(\theta^{(t)}|Y_{obs}) &= Q(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \\ &\leq Q(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) = \ell(\theta^{(t+1)}|Y_{obs}).\end{aligned}$$

□

Theorem 1 (Convergence property of the EM)**.** *Under some conditions, the sequence $\{\theta^{(t)}\}$ defined by the EM iterations converges to a stationary point of the observed-data log-likelihood $\ell(\theta|Y_{obs})$.*

2.4. Missing information and convergence rate

Recall that $Q(\theta|\theta) = \ell(\theta|Y_{obs}) + H(\theta|\theta)$. Taking second derivatives on both sides:

$$\underbrace{-\frac{\partial^2}{\partial\theta^2} Q(\theta|\theta)}_{\mathcal{I}_C(\theta)} = \underbrace{-\frac{\partial^2}{\partial\theta^2}\ell(\theta|Y_{obs})}_{\mathcal{I}_O(\theta)} + \underbrace{(-\frac{\partial^2}{\partial\theta^2} H(\theta|\theta))}_{\mathcal{I}_M(\theta)}.$$

Thus, $\mathcal{I}_C(\theta) = \mathcal{I}_O(\theta) + \mathcal{I}_M(\theta)$. This is called **missing information principle**. For regular problems where $\theta^{(t+1)} \Leftarrow \frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta} = 0$, we have

$$(\theta^{(t+1)} - \widehat{\theta}) \doteq D(\theta^{(t)} - \widehat{\theta}),$$

when $\theta^{(t)}$ is close to the MLE $\widehat{\theta} = \operatorname{argmax}_\theta \ell(\theta|Y_{obs})$. Here, $D = \mathcal{I}_C(\widehat{\theta})^{-1}\mathcal{I}_M(\widehat{\theta})$ is called <u>the fraction of missing information</u>. Therefore after r iterations,

$$(\theta^{(t+r)} - \widehat{\theta}) \doteq D^r(\theta^{(t)} - \widehat{\theta}),$$

which shows that the convergence rate of EM is governed by the largest eigenvalue of D.

2.5. Another example

Example 2. Multinomial distribution with cell probabilities

$$(\pi_1, \pi_2, \pi_3, \pi_4) = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right),$$

where $\theta \in (0,1)$ is the only unknown parameter. Given observations

$$y = (y_1, y_2, y_3, y_4), \qquad \sum_{i=1}^{4} y_i = n,$$

we want to find the MLE of θ.

We could directly maximize the likelihood via numerical optimization, but we could also use EM algorithm, i.e., treat this as a missing data problem. Split the first category $\pi_1 = \pi_{11} + \pi_{12}, \pi_{11} = \frac{1}{2}, \pi_{12} = \frac{\theta}{4}$. Therefore, the complete data is $y_{cmp} = (y_{11}, y_{12}, y_2, y_3, y_4)$. The complete data log-likelihood is:

$$\begin{aligned}\ell(\theta|y_{cmp}) &= y_{11} \log \frac{1}{2} + (y_{12} + y_4) \log \frac{\theta}{4} + (y_2 + y_3) \log \frac{1-\theta}{4} \\ &= (y_{12} + y_4) \log \theta + (y_2 + y_3) \log(1-\theta) + \text{constant}.\end{aligned}$$

EM algorithm:

- E-step: Calculate

  $$\mathbb{E}(y_{12} \mid y, \theta^{(t)}) = y_1 \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4} := y_{12}^{(t)}.$$

  Then

  $$\begin{aligned}Q(\theta \mid \theta^{(t)}) = \mathbb{E}[\ell(\theta|y_{cmp})|y, \theta^{(t)}] &= (y_{12}^{(t)} + y_4) \log \theta + (y_2 + y_3) \log(1-\theta) \\ &\quad + \text{constant}.\end{aligned}$$

- M-step: Maximizing $Q(\theta \mid \theta^{(t)})$ (binomial log-likelihood),

  $$\theta^{(t+1)} = \frac{y_{12}^{(t)} + y_4}{y_{12}^{(t)} + y_4 + y_2 + y_3}.$$

3. EM for exponential families

3.1. Exponential families

Definition 2. A family of pdfs or pmfs is called an exponential family (EF) if it can be expressed as

$$f(x \mid \theta) = h(x)c(\theta) \exp\left[\phi(\theta)^{\mathsf{T}} t(x)\right], \tag{2}$$

where $\theta = (\theta_m)_{1:d} \in \mathbb{R}^d$, $\phi(\theta) = (\phi_j(\theta))_{1:k} \in \mathbb{R}^k$, $t(x) = (t_j(x))_{1:k} \in \mathbb{R}^k$ and $d \leq k$. If $d < k$, the family is called a curved exponential family.

Theorem 2. *Suppose that $f(x \mid \theta)$ and its partial derivatives $\partial f(x \mid \theta)/\partial\theta_m$ are continuous in x and θ. If X is a random variable with density $f(x \mid \theta)$, then*

$$\mathbb{E}\left[\sum_{j=1}^{k} \frac{\partial \phi_j(\theta)}{\partial \theta_m} t_j(X)\right] = -\frac{\partial \log c(\theta)}{\partial \theta_m} \quad \textit{for } m = 1, \ldots, d.$$

Theorem 3 (Sufficient statistic)**.** *Let $Y_1, \ldots, Y_n$ be an iid sample of size n from $f(\cdot \mid \theta)$. Then*

$$T(Y_1, \ldots, Y_n) = \left(\sum_{i=1}^{n} t_1(Y_i), \ldots, \sum_{i=1}^{n} t_k(Y_i)\right) := \sum_{i=1}^{n} t(Y_i)$$

is a sufficient statistic for θ.

Proof. Let $Y = (Y_1, \ldots, Y_n)$ and y_i be the observed value of Y_i. Then

$$f(y \mid \theta) = f(y_1, \ldots, y_n \mid \theta) = \left[\prod_{i=1}^{n} h(y_i)\right] [c(\theta)]^n \exp\left[\phi(\theta)^{\mathsf{T}} \sum_{i=1}^{n} t(y_i)\right].$$

Suppose $\sum_{i=1}^{n} t(y_i) = t^*$. The conditional distribution $[Y \mid T(Y) = t^*, \theta]$ is given by

$$\begin{aligned} p(y \mid t^*, \theta) &\propto f(y \mid \theta) \cdot I(T(y) = t^*) \\ &= \prod_{i=1}^{n} h(y_i) \cdot I(T(y) = t^*) \cdot [c(\theta)]^n \exp\left[\phi(\theta)^{\mathsf{T}} t^*\right] \\ &\propto \prod_{i=1}^{n} h(y_i) \cdot I(T(y) = t^*), \end{aligned}$$

which is independent of θ. □

3.2. MLE for complete data

Let $T_j(y) = \sum_{i=1}^{n} t_j(y_i)$, $j = 1, \ldots, k$. The log-likelihood given complete data

$$\begin{aligned} \ell(\theta \mid y) &= n \log c(\theta) + \phi(\theta)^{\mathsf{T}} \sum_{i=1}^{n} t(y_i) \\ &= n \log c(\theta) + \sum_{j=1}^{k} \phi_j(\theta) T_j(y). \end{aligned} \tag{3}$$

The MLE is given by the solution to

$$\frac{\partial \ell(\theta \mid y)}{\partial \theta_m} = n \frac{\partial \log c(\theta)}{\partial \theta_m} + \sum_{j=1}^{k} \frac{\partial \phi_j(\theta)}{\partial \theta_m} T_j(y) = 0, \quad m = 1, \ldots, d.$$

From Theorem 2 and that $Y_i \sim f(\cdot \mid \theta)$, we have

$$n\frac{\partial \log c(\theta)}{\partial \theta_m} = -n\mathbb{E}\left[\sum_{j=1}^{k} \frac{\partial \phi_j(\theta)}{\partial \theta_m} t_j(Y_1)\right],$$

and therefore, the MLE is given by the solution to

$$\sum_{j=1}^{k} \frac{\partial \phi_j(\theta)}{\partial \theta_m} T_j(y) = n\sum_{j=1}^{k} \frac{\partial \phi_j(\theta)}{\partial \theta_m} \mathbb{E}\left[t_j(Y_1)\right], \quad m = 1, \ldots, d.$$

Assume that $d = k$ and the matrix

$$\frac{\partial \phi}{\partial \theta} = \left(\frac{\partial \phi_j(\theta)}{\partial \theta_m}\right)_{k\times k}$$

is invertible, where $\partial \phi_j(\theta)/\partial \theta_m$ is the $(m, j)^{\text{th}}$ element. Then the MLE $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is the solution to

$$\begin{aligned} \frac{\partial \phi}{\partial \theta}\begin{pmatrix} T_1(y) \\ \vdots \\ T_k(y) \end{pmatrix} &= n\frac{\partial \phi}{\partial \theta}\begin{pmatrix} \mathbb{E}t_1(Y_1) \\ \vdots \\ \mathbb{E}t_k(Y_1) \end{pmatrix} \\ \iff \qquad T_j(y) &= n\mathbb{E}_\theta[t_j(Y_1)], \quad j = 1, \ldots, k. \end{aligned}$$

That is,

$$\sum_{i=1}^{n} t_j(y_i) = n\mathbb{E}_\theta[t_j(Y_1)] = \mathbb{E}_\theta\left[\sum_{i=1}^{n} t_j(Y_i)\right], \quad j = 1, \ldots, k.$$

Note that the left-hand side is the observed value of the sufficient statistic and the right-hand side the expectation which depends on θ.

Example 3. $\mathcal{N}(\mu, \sigma^2)$ and $\text{Bin}(n, p)$.

3.3. EM for incomplete data

Let y_{obs} be the observed data.

- E-step:

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \mathbb{E}\left[\ell(\theta \mid Y) \mid y_{\text{obs}}, \theta^{(t)}\right] \\ &= n\log c(\theta) + \sum_{j=1}^{k} \phi_j(\theta)\mathbb{E}\left[T_j(Y) \mid y_{\text{obs}}, \theta^{(t)}\right] \quad \text{(due to (3))} \\ &= n\log c(\theta) + \sum_{j=1}^{k} \phi_j(\theta)\mathbb{E}\left[\sum_{i=1}^{n} t_j(Y_i) \,\middle|\, y_{\text{obs}}, \theta^{(t)}\right]. \end{aligned}$$

- M-step: $\theta^{(t+1)}$ is the solution to

$$\mathbb{E}\left[\sum_{i=1}^{n} t_j(Y_i) \,\middle|\, y_{\text{obs}}, \theta^{(t)}\right] = n\mathbb{E}_\theta[t_j(Y_1)], \quad j = 1, \dots, k.$$

Example 4. Let $y_1, \dots, y_n$ be iid observations from $\mathcal{N}(\mu, 1)$, but only $\text{sgn}(y_i)$ are observed for $i = 1, \dots, k$. Find the MLE of μ.
Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the pdf and cdf of $\mathcal{N}(0, 1)$, respectively. Suppose that $\text{sgn}(y_i) = 1$ for $i = 1, \dots, k_1$ and $\text{sgn}(y_i) = -1$ for $i = k_1 + 1, \dots, k_1 + k_2 = k$.

$$(\underbrace{\underbrace{+ \dots +}_{k_1} \mid \underbrace{- \dots -}_{k_2}}_{k} \mid y_{k+1}, \dots, y_n)$$

(1) By EM: Regard $y_1, \dots, y_k$ as missing. Sufficient statistic for μ is $T = \sum_{i=1}^n Y_i$. In E-step, calculate $\mathbb{E}(T \mid y_{\text{obs}}, \mu^{(t)}) = \sum_i \mathbb{E}(Y_i \mid y_{\text{obs}}, \mu^{(t)})$.

(a) For $i > k$, $\mathbb{E}(Y_i \mid y_{\text{obs}}, \mu^{(t)}) = y_i$.
(b) For $i = 1, \dots, k_1$,

$$\mathbb{E}(Y_i \mid y_{\text{obs}}, \mu^{(t)}) = \mathbb{E}(Y_i \mid Y_i > 0, \mu^{(t)}) = \mu^{(t)} + \frac{\phi(\mu^{(t)})}{\Phi(\mu^{(t)})}.$$

(c) For $i = k_1 + 1, \dots, k$,

$$\mathbb{E}(Y_i \mid y_{\text{obs}}, \mu^{(t)}) = \mathbb{E}(Y_i \mid Y_i < 0, \mu^{(t)}) = \mu^{(t)} - \frac{\phi(\mu^{(t)})}{1 - \Phi(\mu^{(t)})}.$$

M-step: Solve $\mathbb{E}(T \mid y_{\text{obs}}, \mu^{(t)}) = n\mu (= \mathbb{E}_\mu(T))$ to obtain

$$\mu^{(t+1)} = \frac{1}{n}\left[\sum_{i>k} y_i + k\mu^{(t)} + \left(\frac{k_1}{\Phi(\mu^{(t)})} - \frac{k_2}{1 - \Phi(\mu^{(t)})}\right)\phi(\mu^{(t)})\right]. \tag{4}$$

(2) Direct approach: Since $P(Y_i > 0) = \Phi(\mu)$ and $P(Y_i < 0) = 1 - \Phi(\mu)$,

$$p(y_{\text{obs}} \mid \mu) \propto [\Phi(\mu)]^{k_1}[1 - \Phi(\mu)]^{k_2} \exp\left[-\frac{1}{2}\sum_{i>k}(y_i - \mu)^2\right].$$

Thus, observed data log-likelihood

$$\ell(\mu \mid y_{\text{obs}}) = k_1 \log\Phi(\mu) + k_2 \log[1 - \Phi(\mu)] - \frac{1}{2}\sum_{i>k}(\mu - y_i)^2.$$

Therefore, setting

$$\frac{\partial \ell(\mu \mid y_{\text{obs}})}{\partial \mu} = \frac{k_1\phi(\mu)}{\Phi(\mu)} - \frac{k_2\phi(\mu)}{1 - \Phi(\mu)} - (n - k)\mu + \sum_{i>k} y_i = 0$$

shows that MLE $\hat{\mu}$ satisfies

$$\hat{\mu} = \frac{1}{n}\left[\sum_{i>k} y_i + k\hat{\mu} + \left(\frac{k_1}{\Phi(\hat{\mu})} - \frac{k_2}{1-\Phi(\hat{\mu})}\right)\phi(\hat{\mu})\right]. \tag{5}$$

Compare (4) and (5): $\hat{\mu}$ is a fixed point of the EM iteration, i.e., $\mu^{(t+1)} = \hat{\mu}$ if $\mu^{(t)} = \hat{\mu}$.

4. Incomplete normal data

4.1. The complete-data model

Complete data: $Y = (y_{ij})_{n\times p}$, $y_i = (y_{i1}, y_{i2}, \ldots, y_{ip}) \in \mathbb{R}^p$, and

$$y_i|\theta \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma), \quad i = 1, \ldots, n.$$

Put $\theta = (\mu, \Sigma)$. Complete-data likelihood is

$$L(\theta|Y) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^{\mathsf{T}}\Sigma^{-1}(y_i-\mu)\right\}.$$

Let $S := \sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^{\mathsf{T}} \in \mathbb{R}^{p\times p}$. The exponent

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\mu)^{\mathsf{T}}\Sigma^{-1}(y_i-\mu) &= \operatorname{tr}\left[\sum_i (y_i-\mu)^{\mathsf{T}}\Sigma^{-1}(y_i-\mu)\right] \\
&= \operatorname{tr}\left[\sum_i \Sigma^{-1}(y_i-\mu)(y_i-\mu)^{\mathsf{T}}\right] \\
&= \operatorname{tr}(\Sigma^{-1}S) + \operatorname{tr}[\Sigma^{-1}n(\bar{y}-\mu)(\bar{y}-\mu)^{\mathsf{T}}] \\
&= \operatorname{tr}[\Sigma^{-1}S] + n(\bar{y}-\mu)^{\mathsf{T}}\Sigma^{-1}(\bar{y}-\mu).
\end{aligned}$$

Therefore,

$$\ell(\theta|Y) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}[\Sigma^{-1}S] - \frac{1}{2}n(\bar{y}-\mu)^{\mathsf{T}}\Sigma^{-1}(\bar{y}-\mu).$$

This gives us the maximum likelihood estimate of θ:

$$\widehat{\mu}_{\text{MLE}} = \bar{y}, \quad \widehat{\Sigma}_{\text{MLE}} = \frac{1}{n}S.$$

4.2. Sufficient statistics and conditional distributions

We start with the log-likelihood given complete data,

$$\begin{aligned}
\ell(\theta|Y) &= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left(\mu^{\mathsf{T}}\Sigma^{-1}\mu - 2\mu^{\mathsf{T}}\Sigma^{-1}y_i + y_i^{\mathsf{T}}\Sigma^{-1}y_i\right) \\
&= -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mu^{\mathsf{T}}\Sigma^{-1}\mu + \mu^{\mathsf{T}}\Sigma^{-1}\sum_i y_i - \frac{1}{2}\sum_i y_i^{\mathsf{T}}\Sigma^{-1}y_i.
\end{aligned}$$

Using properties of trace,

$$\sum_i y_i^\top \Sigma^{-1} y_i = \sum_i \operatorname{tr}\left(y_i^\top \Sigma^{-1} y_i\right) = \sum_i \operatorname{tr}\left(\Sigma^{-1} y_i y_i^\top\right) = \operatorname{tr}\left(\Sigma^{-1} \sum_i y_i y_i^\top\right).$$

Letting

$$T_1 := \sum_{i=1}^n y_i = n\bar{y}, \qquad T_2 := \sum_{i=1}^n y_i y_i^\top = Y^\top Y,$$

we arrive at

$$\ell(\theta|Y) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mu^\top\Sigma^{-1}\mu + \mu^\top\Sigma^{-1}T_1 - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}T_2) \tag{6}$$

Note that

$$\begin{aligned} \mu^\top\Sigma^{-1}T_1 &= \langle \Sigma^{-1}\mu, T_1\rangle, \\ \operatorname{tr}(\Sigma^{-1}T_2) &= \langle \operatorname{vec}(\Sigma^{-1}), \operatorname{vec}(T_2)\rangle. \end{aligned}$$

Therefore, (i) $\mathcal{N}(\mu,\Sigma)$ is an exponential family and (ii) (T_1,T_2) is a sufficient statistic for $\theta = (\mu,\Sigma)$. Also we have the following facts:

- $\mathbb{E}_\theta(T_1) = n\mu$;
- $\mathbb{E}_\theta(T_2) = n(\Sigma + \mu\mu^\top)$.

Now we can find the MLE by solving

$$\begin{aligned} \sum_{i=1}^n y_i &= n\mu, \\ \sum_{i=1}^n y_i y_i^\top &= n(\Sigma + \mu\mu^\top), \end{aligned}$$

which leads to

$$\widehat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n y_i = \bar{y}, \qquad \widehat{\Sigma}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n y_i y_i^\top - \bar{y}\bar{y}^\top = \frac{1}{n}S. \tag{7}$$

Theorem 4 (Conditional distributions)**.** *Suppose* $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2) \sim \mathcal{N}(\mu,\Sigma)$, *where* $\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$. *Then*

$$\mathbf{x}_1|\mathbf{x}_2 \sim \mathcal{N}(\mu_{1|2}(\mathbf{x}_2), \Sigma_{1|2}),$$

where $\mu_{1|2}(\mathbf{x}_2) := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)$ *and* $\Sigma_{1|2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

4.3. EM algorithm for incomplete normal data

Illustration of missing values

Variables	1	2	3	4	...	p
y_i	✓	✓	✓	?	?	?

Let $O(i)$ index the observed data in i^{th} observation, and $M(i)$ index the missing data in i^{th} observation. By Theorem 4,

$$y_{i,M(i)} \mid y_{i,O(i)} \sim \mathcal{N}\left(\mu_{M(i)|O(i)}(y_{i,O(i)}), \Sigma_{M(i)|O(i)}\right),$$

which will be used in the E-step.

- E-step:

$$\mathbb{E}[\ell(\theta|Y)|Y_{obs}, \theta^{(t)}] = \mu^{\mathsf{T}}\Sigma^{-1} \underbrace{\mathbb{E}(T_1|Y_{obs}, \theta^{(t)})}_{*} - \frac{1}{2}\operatorname{tr}[\Sigma^{-1} \underbrace{\mathbb{E}(T_2|Y_{obs}, \theta^{(t)})}_{\boxtimes}] \\ - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\mu^{\mathsf{T}}\Sigma^{-1}\mu. \qquad (8)$$

1) $* = \sum_i \mathbb{E}(y_i|Y_{obs}, \theta^{(t)})$ and

$$\mathbb{E}(y_{ij}|Y_{obs}, \theta^{(t)}) = \begin{cases} y_{ij} & \text{if } j \in O(i) \\ y^*_{ij} & \text{if } j \in M(i) \end{cases},$$

where $y^*_{i,M(i)} := \mathbb{E}(y_{i,M(i)}|y_{i,O(i)}, \theta^{(t)}) = \mu^{(t)}_{M(i)|O(i)}(y_{i,O(i)})$.
2) $\boxtimes = \sum_i \mathbb{E}(y_i y_i^{\mathsf{T}}|Y_{obs}, \theta^{(t)})$. Note that

$$\mathbb{E}(y_i y_i^{\mathsf{T}}|Y_{obs}, \theta^{(t)}) = [\mathbb{E}(y_{ij}y_{ik}|Y_{obs}, \theta^{(t)})]_{p\times p}.$$

We have

$$\mathbb{E}(y_{ij}y_{ik}|Y_{obs}, \theta^{(t)}) = \begin{cases} y_{ij}y_{ik} & \text{if } j,k \in O(i) \\ y_{ij}y^*_{ik} & \text{if } j \in O(i), k \in M(i) \\ y^*_{ij}y_{ik} & \text{if } j \in M(i), k \in O(i) \\ \left(\Sigma^{(t)}_{M(i)|O(i)}\right)_{jk} + y^*_{ij}y^*_{ik} & \text{if } j,k \in M(i) \end{cases}.$$

The last case, i.e. $j,k \in M(i)$, is due to

$$\operatorname{Cov}(y_{ij}, y_{ik}|y_{i,O(i)}, \theta^{(t)}) = \mathbb{E}(y_{ij}y_{ik}|y_{i,O(i)}, \theta^{(t)}) - y^*_{ij}y^*_{ik}.$$

- M-step:
Let $T_1^{(t)} := \mathbb{E}(T_1|Y_{obs}, \theta^{(t)}), T_2^{(t)} := \mathbb{E}(T_2|Y_{obs}, \theta^{(t)})$. Max (8) over $\theta = (\mu, \Sigma)$ or solve the following equations for (μ, Σ)

$$T_1^{(t)} = \mathbb{E}_\theta(T_1) = n\mu$$
$$T_2^{(t)} = \mathbb{E}_\theta(T_2) = n(\Sigma + \mu\mu^{\mathsf{T}})$$

to update:

$$\mu^{(t+1)} = \frac{1}{n} T_1^{(t)}, \qquad \Sigma^{(t+1)} = \frac{1}{n} T_2^{(t)} - (\mu^{(t+1)})(\mu^{(t+1)})^{\mathsf{T}}.$$

Compare to (7).

5. Problem set

1. (a) Let $f(x)$ and $g(x)$ be probability densities defined on R^n. Suppose $f(x) > 0$ and $g(x) > 0$ for all x. Show that $\mathbb{E}_f(\log f) \geq \mathbb{E}_f(\log g)$ using Jensen's inequality, where $\mathbb{E}_f(h) = \int h(x)f(x)dx$.

   (b) The entropy of a probability distribution $p(x)$ on R^n is

   $$H(p) := -\mathbb{E}_p(\log p) = -\int p(x) \log p(x) dx.$$

   Among all distributions with mean $\mu = \int x p(x) dx$ and covariance matrix $\Sigma = \int (x-\mu)(x-\mu)^{\mathsf{T}} p(x) dx$, prove that the multivariate normal distribution has the maximum entropy.

   Hint: In fact, (b) is a special case of a more general result: Consider the Boltzmann distribution

   $$p_\beta(x) \propto \exp[-\beta h(x)]$$

   with energy function $h(x)$ at inverse temperature $\beta > 0$. Define the average energy of a distribution $q(x)$ by $\mathbb{E}_q(h) = \int h(x)q(x)dx$. Let $U(\beta)$ be the average energy of p_β. Then among all distributions with average energy $U(\beta)$, the Boltzmann distribution p_β has the maximum entropy.
   Proof outline: First show that the cross-entropy $-\mathbb{E}_q(\log p_\beta)$ is a constant depending on β for any q with average energy $U(\beta)$. Then apply (a).
2. In a genetic linkage experiment, 197 animals are randomly assigned to four categories according to the multinomial distribution with cell probabilities $\pi_1 = \frac{1}{2} + \frac{\theta}{4}$, $\pi_2 = \frac{1-\theta}{4}$, $\pi_3 = \frac{1-\theta}{4}$, and $\pi_4 = \frac{\theta}{4}$. The corresponding observations are $y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$.

   (a) Derive and implement an EM algorithm to estimate θ.

   (b) Plot the observed data log-likelihood function $\ell(\theta \mid y)$ for $\theta \in (0, 1)$. Compare the maximum of this function with your EM estimate.
3. Consider an i.i.d. sample drawn from a bivariate normal distribution with mean $\mu = (\mu_1, \mu_2)$ and covariance matrix

   $$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

   Suppose that the first k observations are missing their first component, the next m observations are missing their second component, and the last r observations are complete. Derive an EM algorithm for estimating the mean assuming that the covariance matrix Σ is known.

4. Prove the following propositions.
   (a) If $Y \sim \mathcal{N}(\mu, 1)$, then $\mathbb{E}(Y \mid Y > 0) = \mu + \phi(\mu)/\Phi(\mu)$.
   (b) Under the assumptions of Theorem 2, if X is a random variable with pdf in an exponential family, then

$$\mathbb{E}\left[\sum_{j=1}^{k} \frac{\partial \phi_j(\theta)}{\partial \theta_m} t_j(X)\right] = -\frac{\partial \log c(\theta)}{\partial \theta_m} \quad \text{for } m = 1, \ldots, d.$$

   Hint: Start from the equality $\int f(x \mid \theta)dx = 1$ and differentiate both sides.

References

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

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association* **116** 1023–1037.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, 1st ed. Chapman & Hall/CRC.

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