STAT 992/BMI 826 University of Wisconsin-Madison

Missing Data: Theory & Methods

Chapter 2

The EM: Basic Theory and Practice

Lu Mao lmao@biostat.wisc.edu

Contents

2.1 Description & Examples

2.2 Ascent Property and Variance Estimation

2.3 Missing Data in Exponential Family

2.4 Linear Regression with Missing Covariates

Description of EM

- A basic description of the EM algorithm (Dempster et al., 1977) has been outlined in §1.4, and an example of bivariate normal with one possibly missing component is illustrated there.
- Here we recapitulate the main steps of the algorithm.
- At the $(j+1)$th iteration (with current iterate $\theta^{(j)}$ at hand),
  - E step:

$$Q(\theta|\theta^{(j)}) := E\{l_n(D;\theta)|D_{\mathsf{obs}};\theta^{(j)}\},$$

    where $l_n(D;\theta)$ is the full data log-likelihood function.
  - M step:

$$\theta^{(j+1)} := \arg\max_{\theta} Q(\theta|\theta^{(j)}).$$

Description of EM

- In practice, it is always easier to perform the M step first by setting the conditional expectation of the full-data score function to be zero.
- Denote $\dot{l}_n(D;\theta) = \frac{\partial}{\partial\theta} l_n(D;\theta)$.
- Under mild regularity conditions,

$$\frac{\partial}{\partial\theta} Q(\theta|\theta^{(j)}) = E\left\{\frac{\partial}{\partial\theta} l_n(D;\theta)\Big| D_{\mathsf{obs}};\theta^{(j)}\right\} = E\{\dot{l}_n(D;\theta)|D_{\mathsf{obs}};\theta^{(j)}\}.$$

- So $\theta^{(j+1)}$ is the root of

$$E\{\dot{l}_n(D;\theta)|D_{\mathsf{obs}};\theta^{(j)}\} = 0.$$

- Computation of conditional expectation of $\dot{l}_n(D;\theta)$ is usually easier than that of $l_n(D;\theta)$, because the score function has gotten rid of terms unrelated to θ.

Description of EM

- The idea of computing $E\{\dot{l}_n(D;\theta)|D_{\textsf{obs}};\theta^{(j)}\}$ is to replace the unobserved terms in the full-data score function with the best estimates possible given the observed data and the current "guess" of the true parameter.
- When the data consist of n independent observations, the conditional score function is

$$\sum_{i=1}^{n}\big\{R_i\dot{l}(Y_i;\theta)+(1-R_i)E[\dot{l}(Y_i;\theta)|R_i=0,Y_{\textsf{obs},i};\theta^{(j)}]\big\},$$

  where $\dot{l}(Y;\theta)$ is the score function for a single observation Y.
- When the missing mechanism is MAR, this previous display is

$$\sum_{i=1}^{n}\big\{R_i\dot{l}(Y_i;\theta)+(1-R_i)E[\dot{l}(Y_i;\theta)|Y_{\textsf{obs},i};\theta^{(j)}]\big\}.$$

Description of EM

- So that under MAR, the E step requires only the distribution of Y and has nothing to do with the missing mechanism.
- In the notation of Chapter 1, $Y_{\textsf{obs}}$ is a component of Y. However, $Y_{\textsf{obs}}$ can take any form that represents partial information contained in Y, i.e., $Y_{\textsf{obs}}$ can be any "coarsened" version of Y.
- In other words, we can write

$$Y_{\textsf{obs}}=m(Y),$$

  where m is some non-invertible function.

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- A common example of coarsened data appears frequently in population/statistical genetics.
- First, a little bit of biology.
- Characteristics of organisms, or **phenotypes**, are controlled by genes located at specific loci of chromosomes.
- Each locus can be occupied by one of several variant genes called **alleles**.
- Humans have 46 chromosomes grouped in 23 homologous pairs
- So there are two alleles at every locus, constituting the subject's **genotype** at that locus.
- The genotype is generally unobservable; what is observable is the person's phenotype.

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- Different genotypes may lead to the same phenotype, so there is a coarsening of information from genotype to phenotype.
- A simple and classic example of single-gene-controlled phenotypes is human blood type.
- The locus controlling blood type resides on chromosome 9 at band q34, and it controls blood type by determining the antigens on the surface of the red blood cells.
- There are three alleles, A, B, and O, which are responsible for generating antigen A, antigen B, and no antigen, respectively.
- So, alleles A and B are "co-dominant" and O is "recessive" (see Figure 2.1), giving rise to four blood types A, B, AB, and O.

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

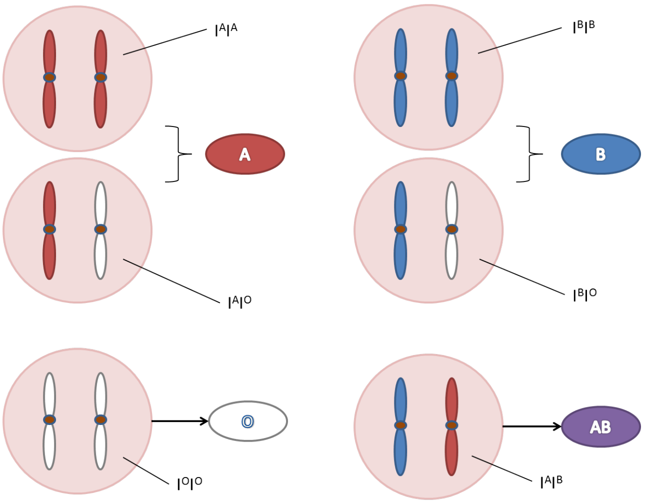

Figure 2.1: Genotypes and corresponding phenotypes of ABO gene.

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- But there are 6 genotypes, AA, AO, BB, BO, AB, OO. AA and AO lead to blood type A; BB and BO to blood type B; AB and OO correspond to blood types AB and O, respectively.
- Under the assumption of infinite population size and random mating, the occurrences of the alleles in each person can be considered independent (Hardy-Weinberg Law).
- So, if we denote the population frequency of alleles A, B, and O, as p_A, p_B, and p_O, respectively, then

$$\Pr(AA) = p_A^2, \Pr(AO) = 2p_Ap_O, \Pr(BB) = p_B^2, \Pr(BO) = 2p_Bp_O,$$

$$\Pr(AB) = 2p_Ap_B, \Pr(OO) = p_O^2.$$

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- Suppose we have a random sample of n subjects, among which n_A, n_B, n_{AB}, and n_O are observed to have blood types A, B, AB, and O, respectively.
- We want to estimate the allele frequencies $\theta := (p_A, p_B, p_O)$.
- Suppose, instead of the phenotype data $D_{\text{obs}} := (n_A, n_B, n_{AB}, n_O)$, we had observed the counts for each genotype $D := (N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}, N_{OO})$, where

$$N_{AA} + N_{AO} = n_A, N_{BB} + N_{BO} = n_B, N_{AB} = n_{AB}, N_{OO} = n_{OO}.$$

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- By the Hardy-Weinberg assumption of independent mating, the the genotype data D is equivalent to a random sample of $2n$ alleles, with observed counts $2N_{AA} + N_{AO} + N_{AB}$, $2N_{BB} + N_{BO} + N_{AB}$, and $2N_{OO} + N_{AO} + N_{BO}$. for alleles A, B, and O, respectively.
- Hence, the MLEs for θ would be

$$\widehat{p}_A = \frac{2N_{AA} + N_{AO} + N_{AB}}{2n},$$

$$\widehat{p}_B = \frac{2N_{BB} + N_{BO} + N_{AB}}{2n},$$

$$\widehat{p}_O = \frac{2N_{OO} + N_{AO} + N_{BO}}{2n}.$$

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- However, since we do not observe the Ns, we might want to replace the numbers with their corresponding estimates based on the observed phenotype data.
- For example, the population fraction of genotype AA within the population of blood type A is

$$\Pr(\text{Genotype } AA|\text{Blood type } A) = \frac{\Pr(AA)}{\Pr(AA)+\Pr(AO)} = \frac{p_A^2}{p_A^2 + 2p_Ap_O}.$$

- So we should replace N_{AA} with

$$\widehat{N}_{AA} := \frac{n_A p_A^2}{p_A^2 + 2p_Ap_O}.$$

Similarly for the other Ns.

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- The problem is that the "estimated" genotype frequencies contain the unknown parameters.
- This suggests using an iterative scheme, where at the $(j+1)$th iteration, one computes

$$\begin{aligned} p_A^{(j+1)} &= \frac{2\widehat{N}_{AA}^{(j)} + \widehat{N}_{AO}^{(j)} + n_{AB}}{2n}, \\ p_B^{(j+1)} &= \frac{2\widehat{N}_{BB}^{(j)} + \widehat{N}_{BO}^{(j)} + n_{AB}}{2n}, \\ p_O^{(j+1)} &= \frac{2n_{OO} + \widehat{N}_{AO}^{(j)} + \widehat{N}_{BO}^{(j)}}{2n}, \end{aligned}$$

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- Where

$$\widehat{N}_{AA}^{(j)} = \frac{n_A {p_A^{(j)}}^2}{{p_A^{(j)}}^2 + 2p_A^{(j)} p_O^{(j)}}, \qquad \widehat{N}_{AO}^{(j)} = \frac{n_A 2 p_A^{(j)} p_O^{(j)}}{{p_A^{(j)}}^2 + 2p_A^{(j)} p_O^{(j)}},$$

$$\widehat{N}_{BB}^{(j)} = \frac{n_B {p_B^{(j)}}^2}{{p_B^{(j)}}^2 + 2p_B^{(j)} p_O^{(j)}}, \qquad \widehat{N}_{BO}^{(j)} = \frac{n_B 2 p_B^{(j)} p_O^{(j)}}{{p_B^{(j)}}^2 + 2p_B^{(j)} p_O^{(j)}}.$$

- It is easily shown that this iterative schedule corresponds to an EM algorithm.
- Clearly, the "imputed" data

$$(\widehat{N}_{AA}^{(j)}, \widehat{N}_{AO}^{(j)}, \widehat{N}_{BB}^{(j)}, \widehat{N}_{BO}^{(j)}) = E[(N_{AA}, N_{AO}, N_{BB}, N_{BO}) | D_{\mathsf{obs}}, \theta^{(j)}].$$

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- So we only need to show that the full data score function is linear in $D = (N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}, N_{OO})$.
- This is easy to see given the multinomial distribution of D.
- The details are left as an exercise.

Examples

Example 1. Estimation of gene allele frequency based on phenotypes

- Clarke et al. (1959) considered the blood types of 521 duodenal ulcer patients, with $n_A = 186, n_B = 38, n_{AB} = 13$, and $n_O = 284$.
- Use the EM algorithm

Table 2.1: Iterations for ABO Duodenal Ulter Data

Iteration j	$p_A^{(j)}$	$p_B^{(j)}$	$p_O^{(j)}$
0	.3000	.2000	.5000
1	.2321	.0550	.7129
2	.2160	.0503	.7337
3	.2139	.0502	.7359
4	.2136	.0501	.7363
5	.2136	.0501	.7363

- It appears that convergence occurs quickly.

Examples

Example 2. Estimation of normal mixture models

- The missing data framework here applies to latent variable models, and so does the EM algorithm.
- Here we give a simple example.
- Suppose we observe a continuous outcome X, whose distribution is multi-modal.
- A popular model for multi-modal distribution is the normal mixture model.
- Suppose there is a latent categorical variable $G = 1, \cdots, K$, where K is fixed, indicating the underlying classes of normal distribution, and

$$X|G = k \sim N(\mu_k, \sigma_k^2), \qquad \Pr(G = k) = p_k, \qquad k = 1, \cdots, K,$$

where $\theta = \{(p_k, \mu_k, \sigma_k^2),\ k = 1, \cdots, K\}$ consists of unknown parameters.

Examples

Example 2. Estimation of normal mixture models

- The (marginal) density of X is

$$p_X(X;\theta) = \sum_{k=1}^{K} p_k f(X;\mu_k,\sigma_k^2),$$

  where $f(X;\mu_k,\sigma_k^2) = \sigma_k^{-1}\phi\left(\frac{X-\mu_k}{\sigma_k}\right)$.
- Given a random sample $X_1,\cdots,X_n$, it is hard to find the MLE of θ directly based on $\sum_{i=1}^n \log p_X(X_i;\theta)$.
- So we try the EM algorithm with imagined full data (X,G), whose density is

$$p_{X,G}(X,G;\theta) = \prod_{k=1}^{K} \left\{p_k f(X;\mu_k,\sigma_k^2)\right\}^{I(G=k)}.$$

Examples

Example 2. Estimation of normal mixture models

- So the log-likelihood of $\{(X_i,G_i) : i=1,\cdots,n\}$ is

$$\begin{aligned} l_n(D;\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} I(G_i=k)\log\left\{p_k f(X_i;\mu_k,\sigma_k^2)\right\} \\ &= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} I(G_i=k)\right)\log p_k \\ &\qquad + \sum_{k=1}^{K}\sum_{i=1}^{n} I(G_i=k)\log f(X_i;\mu_k,\sigma_k^2). \end{aligned}$$

- Hence, we only need to compute $E[I(G=k)|X;\theta]$ at the E step.

Examples

Example 2. Estimation of normal mixture models

- So, at the $(j+1)$th iteration,

$$Q(\theta|\theta^{(j)}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} w_{ki}^{(j)} \right) \log p_k + \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ki}^{(j)} \log f(X_i; \mu_k, \sigma_k^2),$$

  where $w_{ki}^{(j)} = E[I(G_i = k)|X_i; \theta^{(j)}]$.

- For the M step, after straightforward derivation that completely parallels that of MLEs for multinomial and Gaussian data, we have

$$p_k^{(j+1)} = n^{-1} \sum_{i=1}^{n} w_{ki}^{(j)}, \ \mu_k^{(j+1)} = \frac{\sum_{i=1}^{n} w_{ki}^{(j)} X_i}{\sum_{i=1}^{n} w_{ki}^{(j)}},$$

$$\sigma_k^{2\,(j+1)} = \frac{\sum_{i=1}^{n} w_{ki}^{(j)} (X_i - \mu_k^{(j+1)})^2}{\sum_{i=1}^{n} w_{ki}^{(j)}}$$

Examples

Example 2. Estimation of normal mixture models

- For the E step, we compute

$$\begin{aligned} w_{ki}^{(j)} = E[I(G_i = k)|X_i, \theta^{(j)}] &= \frac{p_{X,G}(X_i, G_i = k; \theta^{(j)})}{p_X(X_i; \theta^{(j)})} \\ &= \frac{p_{X,G}(X_i, G_i = k; \theta^{(j)})}{\sum_{l=1}^{K} p_{X,G}(X_i, G_i = l; \theta^{(j)})} \\ &= \frac{p_k^{(j)} f(X_i; \mu_k^{(j)}, {\sigma_k^{(j)}}^2)}{\sum_{l=1}^{K} p_l^{(j)} f(X_i; \mu_l^{(j)}, {\sigma_l^{(j)}}^2)}. \end{aligned}$$

- This wraps up the EM algorithm for normal mixture models.

Contents

2.1 Description & Examples

2.2 Ascent Property and Variance Estimation

2.3 Missing Data in Exponential Family

2.4 Linear Regression with Missing Covariates

The Ascent Property of EM

- The essence of the EM algorithm is that maximizing $Q(\theta|\theta^{(j)})$ leads to an increase in the observed-data log-likelihood $l_n(D_{\text{obs}};\theta)$.
- This assertion is proved in the this theoretical section.
- The entropy (or information) inequality at the heart of the EM algorithm is a consequence of Jensen's inequality.

Proposition 2.1 (Jensen's Inequality)

Let Y be a random variable with values in a possibly infinite interval (a,b). If $h(Y)$ is convex on (a,b), then $E[h(Y)] \geq h[E(Y)]$. For a strictly convex function, equality holds in Jensen's inequality if and only if $Y = E(Y)$ almost surely.

The Ascent Property of EM

Proposition 2.2 (Entropy Inequality)

Let f and g be probability densities with respect to a measure μ. If E_f denotes expectation under density f, then,

$$E_f \log\left(\frac{g}{f}\right) \leq 0,$$

with equality only if $f = g$ almost everywhere relative to μ.

Proof. The function $h(x) = -\log x$ is strictly convex on $(0, \infty)$. We apply Proposition 2.1 to the random variable g/f:

$$-E_f \log\left(\frac{g}{f}\right) \geq -\log E_f\left(\frac{g}{f}\right) = -\log \int \frac{g}{f} f d\mu = -\log 1 = 0.$$

□

The Ascent Property of EM

- Denote $L(D;\theta)$ and $L(D_{\textsf{obs}};\theta)$ as the full data likelihood and the observed data likelihood, respectively.
- We want to show that each EM step increase the log-likelihood of the observed data, i.e.,

$$\log L(D_{\textsf{obs}};\theta^{(j+1)}) \geq \log L(D_{\textsf{obs}};\theta^{(j)}).$$

- Recall that

$$Q(\theta|\theta^{(j)}) = E[\log L(D;\theta)|D_{\textsf{obs}}, \theta^{(j)}].$$

So,

$$Q(\theta|\theta^{(j)}) - \log L(D_{\textsf{obs}};\theta) = E\left[\log \frac{L(D;\theta)}{L(D_{\textsf{obs}};\theta)} \Big| D_{\textsf{obs}}, \theta^{(j)}\right].$$

The Ascent Property of EM

- Note that
$$L(D|D_{\textsf{obs}};\theta) := \frac{L(D;\theta)}{L(D_{\textsf{obs}};\theta)}$$
is the conditional density of D given $D_{\textsf{obs}}$ under θ.
- By Proposition 2.2,
$$\begin{aligned} Q(\theta|\theta^{(j)}) - \log L(D_{\textsf{obs}};\theta) &= E\left[\log L(D|D_{\textsf{obs}};\theta)\middle|D_{\textsf{obs}},\theta^{(j)}\right] \\ &\le E\left[\log L(D|D_{\textsf{obs}};\theta^{(j)})\middle|D_{\textsf{obs}},\theta^{(j)}\right] \\ &= E\left[\log \frac{L(D;\theta^{(j)})}{L(D_{\textsf{obs}};\theta^{(j)})}\middle|D_{\textsf{obs}},\theta^{(j)}\right] \\ &= Q(\theta^{(j)}|\theta^{(j)}) - \log L(D_{\textsf{obs}};\theta^{(j)}). \qquad (2.1) \end{aligned}$$

The Ascent Property of EM

- Since inequality (2.1) holds for all θ, take $\theta = \theta^{(j+1)}$ and we have
$$Q(\theta^{(j+1)}|\theta^{(j)}) - \log L(D_{\textsf{obs}};\theta^{(j+1)}) \le Q(\theta^{(j)}|\theta^{(j)}) - \log L(D_{\textsf{obs}};\theta^{(j)}).$$
- Because by definition $Q(\theta^{(j+1)}|\theta^{(j)}) \ge Q(\theta^{(j)}|\theta^{(j)})$, we have the result
$$\log L(D_{\textsf{obs}};\theta^{(j+1)}) \ge \log L(D_{\textsf{obs}};\theta^{(j)}).$$

Variance Estimation after EM

- One initial criticism of the EM algorithm was that it does not automatically provide an estimate of the covariance matrix of the MLE, as do some other methods, such as Newton-Raphson type methods.
- We know from asymptotic likelihood theory, that when the sample size n is large

  $$\widehat{\theta} \sim N(\theta_0, \mathcal{I}(\theta)^{-1}),$$

  where $\mathcal{I}(\theta)$ is the Fisher information.
- The Fisher information is the expectation of the observed information, which is the negative quadrature of the log-likelihood.
- Because the EM algorithm does not directly provide the correct information matrix at convergence, a variety of methods have been proposed for obtaining the asymptotic covariance matrix of the MLE.

Variance Estimation after EM

- The observed information matrix in missing data problems, $I(\theta; D_{\text{obs}})$, can be found directly by differentiating the log-likelihood $\log L(D_{\text{obs}}; \theta)$ twice w.r.t. θ.
- However, this approach is not always practical since the likelihood of the observed data may be difficult to find.
- Louis (1982) proposed a method of obtaining the observed information matrix directly from quantities calculated in the EM algorithm.
- For simplicity, denote the missing part of the data as D_{mis}. So, $D = (D_{\text{obs}}, D_{\text{mis}})$.

Variance Estimation after EM

- Observe that the full data likelihood can be written as

$$L(D;\theta) = L(D_{\text{mis}}|D_{\text{obs}};\theta)L(D_{\text{obs}};\theta).$$

So,

$$\log L(D;\theta) = \log L(D_{\text{mis}}|D_{\text{obs}};\theta) + \log L(D_{\text{obs}};\theta).$$

- Differentiating both sides of the equation twice with respect to θ yields

$$I(D_{\text{obs}};\theta) = I(D;\theta) + \frac{\partial^2}{\partial\theta^{\otimes 2}} \log L(D_{\text{mis}}|D_{\text{obs}};\theta),$$

where $I(D;\theta)$ is the full-data information, and $-\frac{\partial^2}{\partial\theta^{\otimes 2}} \log L(D_{\text{mis}}|D_{\text{obs}};\theta)$ represents missing-data information.

Variance Estimation after EM

- So, the above equation can be interpreted as

Available information = Full information − Missing information.

- Taking conditional expectation of both sides given D_{obs}, we have

$$I(D_{\text{obs}};\theta) = E[I(D;\theta)|D_{\text{obs}};\theta] + E\left[\frac{\partial^2}{\partial\theta^{\otimes 2}} \log L(D_{\text{mis}}|D_{\text{obs}};\theta)\Big|D_{\text{obs}};\theta\right]. \tag{2.2}$$

- The first term on the right hand side of (2.2) can be estimated by

$$\begin{aligned}
E[I(D;\widehat{\theta})|D_{\text{obs}};\widehat{\theta}] &= -E\left[\frac{\partial^2}{\partial\theta^{\otimes 2}} \log L(D;\theta)\Big|D_{\text{obs}};\widehat{\theta}\right]\Bigg|_{\theta=\widehat{\theta}} \\
&= -\frac{\partial^2}{\partial\theta^{\otimes 2}} E\left[\log L(D;\theta)|D_{\text{obs}};\widehat{\theta}\right]\Big|_{\theta=\widehat{\theta}} \\
&= -\frac{\partial^2}{\partial\theta^{\otimes 2}} Q(\theta|\widehat{\theta})\Big|_{\theta=\widehat{\theta}}.
\end{aligned}$$

Variance Estimation after EM

- Louis (1982) showed that the missing information, represented by the negative of the second term on the right hand side of (2.2), can be estimated by

$$E\left[\dot{l}_n(D;\widehat{\theta})^{\otimes 2}\Big|D_{\text{obs}};\widehat{\theta}\right]-\dot{l}_n(D_{\text{obs}};\widehat{\theta})^{\otimes 2},$$

  where $\dot{l}_n(D_{\text{obs}};\theta)$ is the observed-data score function. At the MLE $\widehat{\theta}$, $\dot{l}_n(D_{\text{obs}};\widehat{\theta})=0$.

- Therefore, the observed-data information can be estimated by

$$I(D_{\text{obs}};\widehat{\theta})=E[I(D;\widehat{\theta})|D_{\text{obs}},\widehat{\theta}]-E\left[\dot{l}_n(D;\widehat{\theta})^{\otimes 2}\Big|D_{\text{obs}};\widehat{\theta}\right],$$

  where

$$E[I(D;\widehat{\theta})|D_{\text{obs}},\widehat{\theta}]=-\frac{\partial^2}{\partial\theta^{\otimes 2}}Q(\theta|\widehat{\theta})\Big|_{\theta=\widehat{\theta}}.$$

Variance Estimation after EM

- So the first term of $I(D_{\text{obs}};\widehat{\theta})$ can be computed based on the Q function. The second term is not a byproduct of the EM algorithm.
- In the iid case,

$$\dot{l}_n(D;\widehat{\theta})^{\otimes 2}=\sum_{i=1}^{n}\dot{l}(Y_i;\widehat{\theta})^{\otimes 2}+\sum\sum_{i\neq j}\dot{l}(Y_i;\widehat{\theta})\dot{l}(Y_j;\widehat{\theta})^{\mathrm{T}}.$$

  So,

$$E\left[\dot{l}_n(D;\widehat{\theta})^{\otimes 2}\Big|D_{\text{obs}};\widehat{\theta}\right]=\sum_{i=1}^{n}\widehat{\dot{l}_i^{\otimes 2}}+\sum\sum_{i\neq j}\widehat{\dot{l}}_i\widehat{\dot{l}}_j^{\mathrm{T}},$$

  where

$$\widehat{\dot{l}_i^{\otimes 2}}=E\left[\dot{l}(Y_i;\widehat{\theta})^{\otimes 2}\Big|D_{\text{obs}};\widehat{\theta}\right],$$

$$\widehat{\dot{l}}_i=E\left[\dot{l}(Y_i;\widehat{\theta})\Big|D_{\text{obs}};\widehat{\theta}\right]$$

Variance Estimation after EM

- Note that

$$\sum_{i\neq j}\sum \widehat{\dot{l}}_i \widehat{\dot{l}}_j^{\mathrm{T}} = \sum_{i=1}^{n} \widehat{\dot{l}}_i \sum_{j\neq i} \widehat{\dot{l}}_j^{\mathrm{T}} = -\sum_{i=1}^{n} \widehat{\dot{l}}_i^{\otimes 2},$$

since $\sum_{i=1}^{n} \widehat{\dot{l}}_i = \sum_{i=1}^{n} E\left[\dot{l}(Y_i;\widehat{\theta})\middle| D_{\textsf{obs}};\widehat{\theta}\right] = 0$.

- In conclusion,

$$\begin{aligned}\widehat{I}(D_{\textsf{obs}};\widehat{\theta}) = &-\sum_{i=1}^{n} E\left[\ddot{l}(Y_i;\widehat{\theta})\middle| D_{\textsf{obs}};\widehat{\theta}\right] - \sum_{i=1}^{n} E\left[\dot{l}(Y_i;\widehat{\theta})^{\otimes 2}\middle| D_{\textsf{obs}};\widehat{\theta}\right] \\ &+ \sum_{i=1}^{n} E\left[\dot{l}(Y_i;\widehat{\theta})\middle| D_{\textsf{obs}};\widehat{\theta}\right]^{\otimes 2}. \qquad (2.3)\end{aligned}$$

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

- To put this example into iid framework, let $Y \in \{AA, AO, BB, BO, AB, OO\}$ be the full genotype data for one individual.
- The full data log-likelihood for $\theta = (p_A, p_B)^{\mathrm{T}}$ is

$$\begin{aligned} l(Y;\theta) = I(Y = AA) f_{AA}(\theta) &+ I(Y = AO) f_{AO}(\theta) \\ + I(Y = BB) f_{BB}(\theta) &+ I(Y = BO) f_{BO}(\theta) \\ + I(Y = AB) f_{AB}(\theta) &+ I(Y = OO) f_{OO}(\theta), \end{aligned}$$

where $f_{AA}(\theta) = \log p_A^2$, $f_{AO}(\theta) = \log 2p_A(1 - p_A - p_B)$, $f_{BB}(\theta) = \log p_B^2$, $f_{BO} = \log 2p_B(1 - p_A - p_B)$, $f_{AB}(\theta) = \log 2p_A p_B$, and $f_{OO}(\theta) = \log(1 - p_A - p_B)^2$.

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

- So,

$$\begin{aligned}\dot{l}(Y;\theta) = I(Y=AA)\dot{f}_{AA}(\theta) + I(Y=AO)\dot{f}_{AO}(\theta) \\ + I(Y=BB)\dot{f}_{BB}(\theta) + I(Y=BO)\dot{f}_{BO}(\theta) \\ + I(Y=AB)\dot{f}_{AB}(\theta) + I(Y=OO)\dot{f}_{OO}(\theta).\end{aligned}$$

- Note that the square of $\dot{l}(Y;\theta)$ has a simple form

$$\begin{aligned}\dot{l}(Y;\theta)^{\otimes 2} = I(Y=AA)\dot{f}_{AA}(\theta)^{\otimes 2} + I(Y=AO)\dot{f}_{AO}(\theta)^{\otimes 2} \\ + I(Y=BB)\dot{f}_{BB}(\theta)^{\otimes 2} + I(Y=BO)\dot{f}_{BO}(\theta)^{\otimes 2} \\ + I(Y=AB)\dot{f}_{AB}(\theta)^{\otimes 2} + I(Y=OO)\dot{f}_{OO}(\theta)^{\otimes 2}.\end{aligned}$$

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

- Let $Y_{\textsf{obs}} \in \{A, B, AB, O\}$ denote the observed blood type data.
- In the E step, we have calculated the conditional probabilities of Y given $Y_{\textsf{obs}}$. In particular, define $w_A(\theta) = \frac{p_A^2}{p_A^2 + 2p_A p_O}$, and $w_B(\theta) = \frac{p_B^2}{p_B^2 + 2p_B p_O}$.
- Then, we have

$$\begin{aligned}& -\sum_{i=1}^{n} E\left[\ddot{l}(Y_i;\widehat{\theta})\middle| D_{\textsf{obs}};\widehat{\theta}\right] \\ = & -\sum_{i=1}^{n}\Bigg(I(Y_{\textsf{obs},i}=A)\Big[w_A(\widehat{\theta})\ddot{f}_{AA}(\widehat{\theta}) + \{1-w_A(\widehat{\theta})\}\ddot{f}_{AO}(\widehat{\theta})\Big] \\ & + I(Y_{\textsf{obs},i}=B)\Big[w_B(\widehat{\theta})\ddot{f}_{BB}(\widehat{\theta}) + \{1-w_B(\widehat{\theta})\}\ddot{f}_{BO}(\widehat{\theta})\Big] \\ & + I(Y_{\textsf{obs},i}=AB)\ddot{f}_{AB}(\widehat{\theta}) + I(Y_{\textsf{obs},i}=O)\ddot{f}_{OO}(\widehat{\theta})\Bigg).\end{aligned}$$

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

- So,

$$
\begin{aligned}
-\sum_{i=1}^{n} E\left[\ddot{l}(Y_i;\widehat{\theta})\Big|D_{\mathsf{obs}};\widehat{\theta}\right] = &- n_A\left[w_A(\widehat{\theta})\ddot{f}_{AA}(\widehat{\theta}) + \{1-w_A(\widehat{\theta})\}\ddot{f}_{AO}(\widehat{\theta})\right] \\
&- n_B\left[w_B(\widehat{\theta})\ddot{f}_{BB}(\widehat{\theta}) + \{1-w_B(\widehat{\theta})\}\ddot{f}_{BO}(\widehat{\theta})\right] \\
&- n_{AB}\ddot{f}_{AB}(\widehat{\theta}) - n_O\ddot{f}_{OO}(\widehat{\theta}).
\end{aligned}
$$

- Similarly,

$$
\begin{aligned}
\sum_{i=1}^{n} E\left[\dot{l}(Y_i;\widehat{\theta})^{\otimes 2}\Big|D_{\mathsf{obs}};\widehat{\theta}\right] = &\, n_A\left[w_A(\widehat{\theta})\dot{f}_{AA}(\widehat{\theta})^{\otimes 2} + \{1-w_A(\widehat{\theta})\}\dot{f}_{AO}(\widehat{\theta})^{\otimes 2}\right] \\
&+ n_B\left[w_B(\widehat{\theta})\dot{f}_{BB}(\widehat{\theta})^{\otimes 2} + \{1-w_B(\widehat{\theta})\}\dot{f}_{BO}(\widehat{\theta})^{\otimes 2}\right] \\
&+ n_{AB}\dot{f}_{AB}(\widehat{\theta})^{\otimes 2} + n_O\dot{f}_{OO}(\widehat{\theta})^{\otimes 2}.
\end{aligned}
$$

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

- And,

$$
\begin{aligned}
\sum_{i=1}^{n} E\left[\dot{l}(Y_i;\widehat{\theta})\Big|D_{\mathsf{obs}};\widehat{\theta}\right]^{\otimes 2} = &\, n_A\left[w_A(\widehat{\theta})\dot{f}_{AA}(\widehat{\theta}) + \{1-w_A(\widehat{\theta})\}\dot{f}_{AO}(\widehat{\theta})\right]^{\otimes 2} \\
&+ n_B\left[w_B(\widehat{\theta})\dot{f}_{BB}(\widehat{\theta}) + \{1-w_B(\widehat{\theta})\}\dot{f}_{BO}(\widehat{\theta})\right]^{\otimes 2} \\
&+ n_{AB}\dot{f}_{AB}(\widehat{\theta})^{\otimes 2} + n_O\dot{f}_{OO}(\widehat{\theta})^{\otimes 2}.
\end{aligned}
$$

- The variance of $\widehat{\theta}$ can be computed using the Louis formula (2.3). See A2.1 for details.

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

- We perform some numerical studies to assess the EM and Louis formula.
- Set $p_A = 0.7$, $p_B = 0.2$, and $p_O = 0.1$, and generate data under the Hardy-Weinberg Law.

Table 2.2: Simulation of allele frequency estimation by EM and Louis formula.

	p_A				p_B			
n	Est	SE	SEE	CP	Est	SE	SEE	CP
100	0.705	0.045	0.044	0.948	0.201	0.029	0.030	0.948
200	0.702	0.032	0.031	0.939	0.200	0.021	0.021	0.946
500	0.701	0.020	0.020	0.938	0.200	0.013	0.013	0.948

Est and SE are the empirical means and standard error of the parameter estimator; SEE is the empirical average of the standard error estimator; CP is the empirical coverage probability of the 95% confidence interval. Each entry is based on 2,000 replicates.

Variance Estimation after EM

Example 1. Estimation of gene allele frequency based on phenotypes

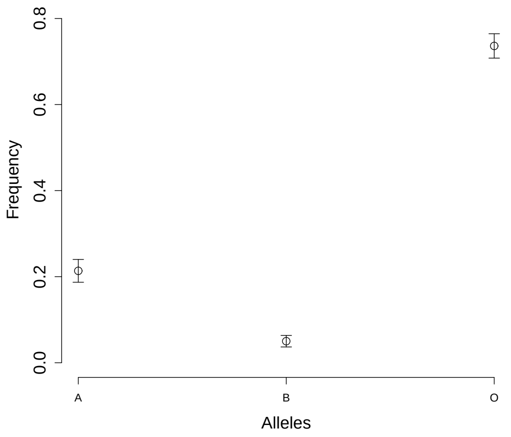

Figure 2.2: Estimates of ABO allele frequencies in the duodenal ulter data (Clarke et al., 1959) and their 95% confidence intervals produced by the Louis formula.

Contents

2.1 Description & Examples

2.2 Ascent Property and Variance Estimation

2.3 Missing Data in Exponential Family

2.4 Linear Regression with Missing Covariates

Missing Data in Exponential Family

- To make our framework more general so as to allow for coarsening of data (Example 1) as well as actually missing values, we introduce, for the ith subject, the (random) M_i representing the "coarsening" pattern, so that the coarsened data is

$$Y_{i,\mathsf{obs}} = M_i(Y_i).$$

- For example, if full data $Y_i = (Y_{i1}, Y_{i2})^{\mathrm{T}}$, and the observed data are (R_i, Y_{i1}, R_i), then

$$M_i = R_i f_1 + (1 - R_i) f_0,$$

where $f_1(y) = y$ and $f_0(y) = y_1$ for $y = (y_1, y_2)$. So there are two possible coarsening patterns f_1 and f_0, which represent no coarsening at all and extracting the first element, respectively.

Missing Data in Exponential Family

- In Example 1, M_i is the (fixed) function

$$m(Y_i) = \begin{cases} A, & \text{if } Y_i \in \{AA, AO\} \\ B, & \text{if } Y_i \in \{BB, BO\} \\ AB, & \text{if } Y_i = AB \\ O, & \text{if } Y_i = OO \end{cases}$$

- The representation of **observed data** as $\{M_i, Y_{i,\text{obs}} := M_i(Y_i)\}$ also allows each subject to have a different coarsening pattern.

Missing Data in Exponential Family

- Under this framework, suppose there are K levels of coarsening pattern, i.e., $M \in \{m_1, \cdots, m_K\}$.
- Denote the coarsening mechanism by

$$\pi_k(y) = \Pr(M = m_k | Y = y), \quad k = 1, \cdots, K.$$

- The familiar classifications of missing data based on missingness mechanism can be easily extend to the coarsened data as follows:
  - **MCAR**: $M \perp\!\!\!\perp Y$, i.e., the $\pi_k(y)$ are constants

$$\pi_k(y) = \pi_{0k}, \quad k = 1, \cdots, K.$$

  - **MAR**: $M \perp\!\!\!\perp Y | M(Y)$, i.e., $\pi_k(y)$ is only a function of $m_k(y)$

$$\pi_k(y) = \widetilde{\pi}_k\Big(m_k(y)\Big), \quad k = 1, \cdots, K.$$

  - **NMAR**: M is not independent of Y even given $M(Y)$, i.e., at least one $\pi_k(y)$ depends on elements of y other than $m_k(y)$.

Missing Data in Exponential Family

- The **exponential family** of statistical models encompass the most commonly used models such as the normal, binomial/multinomial, Poission, and etc..
- A model belongs to the exponential family if its desnity can be parameterized as

$$p_Y(Y;\theta) = \exp\Big\{\theta^{\mathrm{T}} t(Y) + b(Y) - a(\theta)\Big\},$$

  where θ is called the **canonical parameter**, $t(\cdot)$ and $b(\cdot)$ are functions of Y, and $a(\theta)$ is the normalizing constant for density under θ.
- So, the score function is

$$\dot{l}(Y;\theta) = t(Y) - \dot{a}(\theta).$$

Missing Data in Exponential Family

- Based on an iid sample of n subjects, the score function is

$$\dot{l}_n(D;\theta) = \sum_i^n t(Y_i) - n\dot{a}(\theta).$$

- So, inference of θ is solely based on $\sum_i^n t(Y_i)$, which is hence termed the **sufficient statistic**.
- With full data, the MLE solves

$$\sum_i^n t(Y_i) - n\dot{a}(\theta) = 0,$$

  using possibly the Newton-Raphson algorithm.
- With coarsened data $D_{\textsf{obs}} = \{(M_i, M_i(Y_i)),\, i = 1, \cdots, n\}$, the M step for the EM algorithm closely mimics the MLE for full data.

Missing Data in Exponential Family

- Observe that the full data score function is linear in the sufficient statistic, so, at the $(j+1)$th iteration, the M step solves

$$\sum_{i}^{n} \widehat{t}_i^{(j)} - n\dot{a}(\theta) = 0,$$

where in general

$$\widehat{t}_i^{(j)} = E[t(Y_i)|M_i, M_i(Y_i); \theta^{(j)}],$$

and under MAR

$$\widehat{t}_i^{(j)} = E[t(Y_i)|M_i(Y_i); \theta^{(j)}].$$

- We can also derive a general expression for the variance estimator of the MLE $\widehat{\theta}$ with coarsened data using Louis formula.

Missing Data in Exponential Family

- Assuming MAR, denote

$$\widehat{t}_i = E[t(Y_i)|M_i(Y_i); \widehat{\theta}], \quad \text{and } \widehat{t_i^{\otimes 2}} = E[t(Y_i)^{\otimes 2}|M_i(Y_i); \widehat{\theta}].$$

- Using the Louis formula (2.3), an estimator for the observed data information is

$$\begin{aligned}\widehat{I}(D_{\mathsf{obs}}; \widehat{\theta}) &= \sum_{i=1}^{n} \Bigg\{ \ddot{a}(\widehat{\theta}) - \Big(\widehat{t_i^{\otimes 2}} - 2\dot{a}(\widehat{\theta})\widehat{t}_i + \dot{a}(\widehat{\theta})^{\otimes 2} \Big) \\ &\qquad + \widehat{t}_i^{\otimes 2} - 2\dot{a}(\widehat{\theta})\widehat{t}_i + \dot{a}(\widehat{\theta})^{\otimes 2} \Bigg\} \\ &= \sum_{i=1}^{n} \Big\{ \ddot{a}(\widehat{\theta}) - \big(\widehat{t_i^{\otimes 2}} - \widehat{t}_i^{\otimes 2} \big) \Big\}\end{aligned}$$

Missing Data in Exponential Family

- In fact, it is easily seen that the missing information for the ith subject

$$\widehat{t_i^{\otimes 2}} - \widehat{t}_i^{\otimes 2} = \widehat{\mathsf{Var}}\big(t(Y_i)|M_i(Y_i);\widehat{\theta}\big).$$

- So the missing information is due to the extra variability of the sufficient statistic $T(Y)$ unaccounted for by $M(Y)$.
- It may very well happen that the canonical parameter is not directly in the form of the parameter of interest.
- If the parameter of interest is ψ and the canonical parameter θ can be expressed as $\theta(\psi)$. Then the information of ψ can be estimated by

$$\widehat{I}(D_{\mathsf{obs}};\widehat{\psi}) = \dot{\theta}(\widehat{\psi})^{\mathrm{T}}\widehat{I}\big(D_{\mathsf{obs}};\theta(\widehat{\psi})\big)\dot{\theta}(\widehat{\psi})$$

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- Consider the multinomial data $Y = k$ with probability p_k, $k = 1, \cdots, K$, where $\sum_{k=1}^{K} p_k = 1$.
- The log-likelihood of Y is

$$\begin{aligned}\sum_{k=1}^{K} I(Y=k)\log p_k &= \sum_{k=1}^{K-1} I(Y=k)\log p_k + \left(1 - \sum_{k=1}^{K-1} I(Y=k)\right)\log p_K \\ &= \sum_{k=1}^{K-1} I(Y=k)\log\left(p_k/p_K\right) + \log p_K \\ &=: \theta^{\mathrm{T}}T - \log\left(1 + \sum_{k=1}^{K-1} e^{\theta_k}\right),\end{aligned}$$

where $\theta = (\theta_1, \cdots, \theta_{K-1})^{\mathrm{T}}$, $\theta_k = \log(p_k/p_K)$, and $T = (I(Y=1), \cdots, I(Y=K-1))^{\mathrm{T}}$.

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- To get back to $p := (p_1, \cdots, p_{K-1})^{\mathrm{T}}$ from θ, note that

$$p_k = \frac{e^{\theta_k}}{1 + \sum_{k'=1}^{K-1} e^{\theta_{k'}}}, \qquad k = 1, \cdots, K-1,$$

  and $p_K = \left(1 + \sum_{k=1}^{K-1} e^{\theta_k}\right)^{-1}$.

- In this case,

$$a(\theta) = \log\left(1 + \sum_{k=1}^{K-1} e^{\theta_k}\right),$$

  and it is easily shown that

$$\dot{a}(\theta) = \frac{e^{\theta}}{1 + \sum_{k=1}^{K-1} e^{\theta_k}} = p,$$

  where $e^{\theta} = (e^{\theta_1}, \cdots, e^{\theta_{K-1}})$.

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- Hence, the full data score function is

$$\dot{l}_n(Y;\theta) = \sum_{i=1}^{n} (T_i - p).$$

- So, the M step for the $(j+1)$th iteration of EM algorithm with incomplete data is as follows:

$$p^{(j+1)} = n^{-1} \sum_{i=1}^{n} \widehat{T}_i^{(j)},$$

  where

$$\widehat{T}_i^{(j)} = E[T_i | M_i, M_i(Y_i); \theta^{(j)}].$$

- The E step (computation of the $\widehat{T}_i^{(j)}$), of course, depends on the specific coarsening patterns and mechanism.

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- Here we consider a grouping scenario that frequently occurs to multinomial data.
- Assume that the K-level data Y are grouped into L categories, where $L < K$.
- Specifically, define a many-to-one function $m : \{1, \cdots, K\} \to \{1, \cdots, L\}$ representing the grouping pattern. For example, Levels 1, 2, and 3 are grouped into Group 1; Levels 4 and 5 to Group 2, and so on.
- Assume that this pattern occurs at random (MAR), that is, the probability of being coarsened (grouped) depends only on the group. In other words, all levels within a group have the same probability of being grouped.

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- Further, in order to identify the full $(K-1)$ vector p, we have to assume that for each level, there is a positive probability of not being grouped.
- If otherwise, say, levels within $m^{-1}(l_0) := \{k : m(k) = l_0, k = 1, \cdots, K\}$ are always grouped (into the new level l_0), then we can only identify the probability of Group l_0, which is $\sum_{k \in m^{-1}(l_0)} p_k$, not those of the individual levels, if more than one, within it.
- These assumptions guarantee that vector p is identifiable by Proposition 1.1.

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- Under the current formulation, there are two coarsening patterns, one is no grouping, the other grouping by the rule $m(\cdot)$.
- So we use a binary random variable $R \in \{0,1\}$ define the random coarsening pattern
$$M = Rm_0 + (1-R)m,$$
where $m_0(y) = y$ is the identity function.
- Thus, the kth component of $\widehat{T}_i^{(j)}$ is
$$E[I(Y_i = k)|M_i(Y_i); \theta^{(j)}] = R_i I(Y_i = k) + (1-R_i)\frac{I(m(Y_i) = m(k))p_k^{(j)}}{\sum_{k' \in \mathcal{C}_{m(k)}} p_{k'}^{(j)}},$$
where $\mathcal{C}_l = m^{-1}(l), l = 1, \cdots, L$.

Missing Data in Exponential Family

Example 3. Grouped multinomial data

- Now, define
$$N_k = \sum_{i=1}^{n} R_i I(Y_i = k)$$
as the count in the kth level among the ungrouped, and
$$\widetilde{N}_l = \sum_{i=1}^{n} (1-R_i) I(m(Y_i) = l)$$
as the count in the lth group among the grouped.
- Then, the M step can be expressed as
$$p_k^{(j+1)} = n^{-1}\left(N_k + \widetilde{N}_{m(k)}\frac{p_k^{(j)}}{\sum_{k' \in \mathcal{C}_{m(k)}} p_{k'}^{(j)}}\right).$$
- Information calculation by the Louis formula is relegated to A2.2.

Contents

2.1 Description & Examples

2.2 Ascent Property and Variance Estimation

2.3 Missing Data in Exponential Family

2.4 Linear Regression with Missing Covariates

Linear Regression with Missing Covariates

- Consider the usual linear regression model:

$$Y = \beta_1 + \beta_2^{\mathrm{T}} Z + \epsilon,$$

where $\beta = (\beta_1, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ are regression parameters, $Z = (Z_1, \cdots, Z_p)^{\mathrm{T}}$ are the covariates, and $\epsilon \sim N(0, \sigma^2)$ is the random error independent of Z.

- Suppose the full data are $(Y_i, Z_i), i = 1, \cdots, n$. In the observed data, some components of the Z_i are missing at random, so we observe $(M_i, M_i(Z_i))$
- The full data likelihood can be expressed as

$$p(Y|Z; \beta, \sigma^2)\eta(Z),$$

where $p(Y|Z; \beta, \sigma^2)$ is the conditional density of Y given Z and $\eta(Z)$ is the density of Z.

Linear Regression with Missing Covariates

- Because of factorization, inference on (β, σ^2) can be based on

$$\prod_{i=1}^{n} p(Y_i | Z_i; \beta, \sigma^2),$$

regardless of η, which can thus be left totally nonparametric.

- However, the density of observed data with missing covariates is (proportional to)

$$\int_{M(z)=M(Z)} p(Y|z; \beta, \sigma^2)\eta(z)dz.$$

So η gets entangled with the parameters of interest.

- In order to do MLE, we have to build a model for Z.
- We discuss two scenarios: (1) Z is multivariate normal; (2) Z is categorical (i.e., consists of binary indicators).

Linear Regression with Missing Covariates

Case 1. Continuous covariates

- Let $Z \sim N(\mu, \Sigma)$. Clearly, $(Y, Z^{\mathrm{T}})^{\mathrm{T}}$ is multivariate normal:

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N \left\{ \begin{pmatrix} \beta_1 + \beta_2^T \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 + \beta_2^T \Sigma \beta_2 & \beta_2^T \Sigma \\ \Sigma \beta_2 & \Sigma \end{pmatrix} \right\}.$$

- So this time the conditional expectation, conditional variance and covariance of any components of Z given Y and other elements of Z have closed-form expressions (cf. A1.2.3).
- We obtain the MLE for β and σ^2 based on the incomplete data $(Y_i, M_i(Z_i)), i = 1, \cdots, n$.

Linear Regression with Missing Covariates

Case 1. Continuous covariates

- For simplicity of notation, re-define Z to include the intercept, so $Z = (Z_0, Z_1, \cdots, Z_p)^{\mathrm{T}}$, where $Z_0 = 1$.
- The full data log-likelihood is

$$
\begin{aligned}
l_n(D;\theta) = & -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta^{\mathrm{T}} Z_i)^2 - \frac{n}{2}\log\sigma^2 \\
& -\frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{p}\sum_{k=1}^{p}(Z_{il}-\mu_l)\sigma_z^{l,k}(Z_{ik}-\mu_k) - \frac{n}{2}\log\det\Sigma,
\end{aligned}
$$

where $\sigma_z^{l,k}$ is the (l,k)th element of Σ^{-1}.

Linear Regression with Missing Covariates

Case 1. Continuous covariates

- At the $(j+1)$th iteration, we compute

$$
\begin{aligned}
Q(\theta|\theta^{(j)}) = & \left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n} E\left[(Y_i - \beta^{\mathrm{T}} Z_i)^2 | D_{\textsf{obs}}; \theta^{(j)}\right] - \frac{n}{2}\log\sigma^2 \right\} \\
& + \left\{ -\frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{p}\sum_{k=1}^{p}\sigma_z^{l,k} E[(Z_{il}-\mu_l)(Z_{ik}-\mu_k)|D_{\textsf{obs}};\theta^{(j)}] \right. \\
& \left. - \frac{n}{2}\log\det\Sigma \right\} \\
=: & \; Q_1(\theta|\theta^{(j)}) + Q_2(\theta|\theta^{(j)}).
\end{aligned}
$$

- The conditional expectations of the square terms can be calculated by partitioning into (conditional) variance and bias.

Linear Regression with Missing Covariates

Case 1. Continuous covariates

- So,

$$Q_1(\theta|\theta^{(j)}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big((Y_i - \beta^{\mathrm{T}}\widehat{Z}_i^{(j)})^2 + \beta^{\mathrm{T}}\widehat{C}_i^{(j)}\beta\Big) - \frac{n}{2}\log\sigma^2,$$

  where

$$\widehat{Z}_i^{(j)} = E[Z_i|Y_i, M_i(Z_i);\theta^{(j)}], \text{ and } \widehat{C}_i^{(j)} = \mathrm{Var}\left[Z_i\middle|Y_i, M_i(Z_i);\theta^{(j)}\right].$$

- Thus, the E step fills in the missing components of Z_i using conditional expectation and adds an additional term to account for the filling in.

Linear Regression with Missing Covariates

Case 1. Continuous covariates

- Similarly,

$$\begin{aligned} Q_2(\theta|\theta^{(j)}) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{p}\sum_{k=1}^{p}\sigma_z^{l,k}\Big((\widehat{Z}_{il}-\mu_l)(\widehat{Z}_{ik}-\mu_k) + \widehat{c}_{ikl}^{(j)}\Big) - \frac{n}{2}\log\det\Sigma \\ &= -\frac{1}{2}tr\left\{\sum_{i=1}^{n}\Big((\widehat{Z}_i^{(j)}-\mu)^{\otimes 2} + \widehat{C}_i^{(j)}\Big)\Sigma^{-1}\right\} - \frac{n}{2}\log\det\Sigma, \end{aligned}$$

  where $\widehat{c}_{ikl}^{(j)}$ is the (k,l)th element of $\widehat{C}_i^{(j)}$.

- In the M step, we can maximize Q_1 and Q_2 separately since the parameters (β, σ^2) are distinct from (μ, Σ). The maximizations are similar to the full data cases.

Linear Regression with Missing Covariates

Case 1. Continuous covariates

- Maximizing Q_1 w.r.t. (β, σ^2) leads to

$$\beta^{(j+1)} = \left(\sum_{i=1}^{n} \left(\widehat{Z}_i^{(j)\otimes 2} + \widehat{C}_i^{(j)} \right) \right)^{-1} \sum_{i=1}^{n} \widehat{Z}_i^{(j)} Y_i,$$

$$\sigma^{2(j+1)} = n^{-1} \sum_{i=1}^{n} \Big((Y_i - \beta^{(j+1)\mathrm{T}} \widehat{Z}_i^{(j)})^2 + \beta^{(j+1)\mathrm{T}} \widehat{C}_i^{(j)} \beta^{(j+1)} \Big).$$

- Maximizing Q_2 w.r.t. (μ, Σ) leads to

$$\mu^{(j+1)} = n^{-1} \sum_{i=1}^{n} \widehat{Z}_i^{(j)},$$

$$\Sigma^{(j+1)} = n^{-1} \sum_{i=1}^{n} \Big((\widehat{Z}_i^{(j)} - \mu^{(j+1)})^{\otimes 2} + \widehat{C}_i^{(j)} \Big).$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- We then consider the linear regression model where the covariates are binary indicators.
- Let $p(z|\alpha)$ denote the covariate distribution of Z. For example, for the saturated model, Z can be viewed as multinomial with 2^p categories.
- The full data log-likelihood is thus

$$l_n(D; \theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta^{\mathrm{T}} Z_i)^2 - \frac{n}{2} \log \sigma^2 + \sum_{i=1}^{n} \log p(Z_i; \alpha).$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- Denote $\mathcal{Z} = \{z_k : k = 1, \cdots, m\}$ as the set of possible values of the binary vector Z. Of course, $m \le 2^p$.
- Let $\mathcal{C}_i$ denote the set of all possible z_ks that are "compatible" with the observed values $M_i(Z_i)$ of the ith subject. That is

$$\mathcal{C}_i = \{z \in \mathcal{Z} : M_i(z) = M_i(Z_i)\}.$$

- So, $\mathcal{C}_i$ contains all possibilities for the full covariate vector Z_i given the observed part $M_i(Z_i)$.
- At the $(j+1)$th iteration, we compute

$$Q(\theta|\theta^{(j)}) = -\frac{1}{2\sigma^2}\sum_{i=1}^n E\left[(Y_i - \beta^{\mathrm{T}} Z_i)^2 \Big| Y_i, M_i(Z_i); \theta^{(j)}\right] - \frac{n}{2}\log\sigma^2$$
$$+ \sum_{i=1}^n E\left[\log p(Z_i;\alpha) \Big| Y_i, M_i(Z_i); \theta^{(j)}\right].$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- Note that because Z is categorical, the conditional expectations can be written as a weighted sum of all possible values of the expectands.
- So,

$$Q(\theta|\theta^{(j)}) = -\frac{1}{2\sigma^2}\sum_{i=1}^n \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} (Y_i - \beta^{\mathrm{T}} z_k)^2 - \frac{n}{2}\log\sigma^2$$
$$+ \sum_{i=1}^n \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} \log p(z_k;\alpha),$$

where

$$w_{ik}^{(j)} = \frac{f(Y_i|z_k; \beta^{(j)}, \sigma^{2(j)}) p(z_k; \alpha^{(j)})}{\sum_{k' \in \mathcal{C}_i} f(Y_i|z_{k'}; \beta^{(j)}, \sigma^{2(j)}) p(z_{k'}; \alpha^{(j)})}.$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- In fact, the E-step for any regression model with missing categorical covariates can be written as a weighted sum of the full data log-likelihood (Ibrahim, 1990).
- The form of $Q(\theta|\theta^{(j)})$ gives a weighted least squares problem for the M step.
- It is easy to obtain

$$\beta^{(j+1)} = \left(\sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} z_k^{\otimes 2} \right)^{-1} \left(\sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} z_k Y_i \right),$$

and

$$\sigma^{2(j+1)} = n^{-1} \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} (Y_i - \beta^{(j+1)^{\mathrm{T}}} z_k)^2.$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- The M step for α solves the following weighted score function

$$n^{-1} \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} \dot{l}_\alpha(z_k; \alpha) = 0,$$

where $\dot{l}_\alpha(z;\alpha) = \frac{\partial}{\partial \alpha} \log p(z;\alpha)$.

- In the case of a saturated model for all 2^p categories of Z, there is a closed-form solution for the M step of α.
- Let $\alpha_k = \Pr(Z = k)$. Then,

$$\alpha_k^{(j+1)} = n^{-1} \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)}.$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- To find a variance estimator for the MLE, Ibrahim (1990) recommends the Louis formula.
- For ease of competing the derivatives, we make the transformation $\tau = 1/\sigma^2$. So $\theta = (\beta^{\mathrm{T}}, \tau, \alpha^{\mathrm{T}})^{\mathrm{T}}$.
- For a single observation of full data,

$$\dot{l}(Y, Z; \theta) = \left(-\tau(\beta^{\mathrm{T}} Z^{\otimes 2} - Z^T Y), -\frac{1}{2}(Y - \beta^{\mathrm{T}} Z)^{\otimes 2} + \frac{1}{2\tau}, \dot{l}_\alpha(Z; \alpha)\right)^{\mathrm{T}},$$

and

$$-\ddot{l}(Y, Z; \theta) = \begin{pmatrix} \tau Z^{\otimes 2} & \beta^{\mathrm{T}} Z^{\otimes 2} - Z^T Y & 0 \\ Z^{\otimes 2}\beta - YZ & \frac{1}{2\tau^2} & 0 \\ 0 & 0 & -\ddot{l}_\alpha(Z; \alpha) \end{pmatrix}.$$

Linear Regression with Missing Covariates

Case 2. Binary covariates

- Computation of conditional expectations is facilitated by observing the fact that, for any function $g(Y_i, Z_i; \widehat{\theta})$,

$$E[g(Y_i, Z_i; \widehat{\theta}) | Y_i, M_i(Z_i); \widehat{\theta}] = \sum_{z_k \in \mathcal{C}_i} \widehat{w}_{ik} g(Y_i, z_k; \widehat{\theta}),$$

where $\widehat{w}_{ik}$ is the expression of $w_{ik}^{(j)}$ with $\theta^{(j)}$ replaced by $\widehat{\theta}$.

- The information can thus be estimated by the Louis formula using (2.3).

Linear Regression with Missing Covariates

- It also frequently occurs that the covariates are a mixed of continuous and categorical variables.
- Let Z be a vector of continuous variables and $\widetilde{Z}$ be a vector of binary variables. The linear regression model is

$$Y = \beta^{\mathrm{T}} Z + \widetilde{\beta}^{\mathrm{T}} \widetilde{Z} + \epsilon.$$

- When components of both Z and $\widetilde{Z}$ are subject to missingness, we need to have a joint model for $(Z, \widetilde{Z})$ in order to compute the MLE for β and $\widetilde{\beta}$.
- In such cases, it is convenient to model the joint distribution of $(Z, \widetilde{Z})$ by a normal mixture:

$$p(Z, \widetilde{Z}; \alpha) = p(Z|\widetilde{Z}; \alpha_1) p(\widetilde{Z}; \alpha_2),$$

where $Z|\widetilde{Z} = z_k \sim N(\mu_k, \Sigma_k)$.

Linear Regression with Missing Covariates

- Clearly, $(Y, Z^{\mathrm{T}})^{\mathrm{T}}$ is multivariate normal conditioning on each category of $\widetilde{Z}$.
- It is easy to use the techniques in the previous two cases to derive the EM algorithm for the case of mixed covariates, where both the E and M steps also have closed-form solutions.
- The details are left as an exercise.

References

- Clarke, C. A. (1959). Distribution of ABO blood groups and the secretor status in duodenal ulcer families. Digestion, 92, 99-103.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1-38.
- Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association, 85, 765-769.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society. Series B, 226-233.