3 Likelihood-based Methods Under MAR

In this chapter, we provide an overview of methods for inference in the presence of missing data based on the principles of maximum likelihood when it is reasonable to assume that the missing data mechanism is MAR. Some of the material introduced here, especially that in Section 3.2, will be important in later chapters as well. The concept of ***ignorability***, discussed in Section 3.3, is central to the methods discussed in this chapter.

3.1 Review of maximum likelihood inference for full data

Before we discuss the case of missing data, we review briefly in our context and notation the basic principles of maximum likelihood.

As always, let Z denote generically the full data of interest. Suppose as discussed in Chapter 1 that we posit a ***parametric model*** for the probability density of Z in terms of a finite dimensional parameter θ, which we write as

$$p_Z(z;\theta). \tag{3.1}$$

Interest may focus on all of θ or on a function of its components.

Given a model (3.1) and a sample of data Z_i, $i = 1 \ldots, N$ (iid), the goal is to estimate θ. Define the ***(full data) score vector*** as the vector of partial derivatives of $\log\{p_Z(z;\theta)\}$ with respect to each of the elements of θ, i.e.,

$$S^F_\theta(z;\theta) = \frac{\partial}{\partial\theta}\log\{p_Z(z;\theta)\}. \tag{3.2}$$

Then the ***maximum likelihood estimator*** (MLE) for θ based on the full data, $\widehat{\theta}^F$, is the value of θ maximizing

$$\prod_{i=1}^N p_Z(Z_i;\theta) \quad \text{or equivalently} \quad \sum_{i=1}^N \log\{p_Z(Z_i;\theta)\}.$$

Under regularity conditions, the estimator $\widehat{\theta}^F$ is found by solving in θ the ***score equation***

$$\sum_{i=1}^N S^F_\theta(Z_i;\theta) = 0.$$

Assuming that the model (3.1) is ***correctly specified*** in the sense that there exists some θ_0 such that $p_Z(z;\theta_0)$ is the true distribution that generates the data, then, under regularity conditions, it is well-known that

$$\widehat{\theta}^F \xrightarrow{p} \theta_0.$$

That is, the MLE is a consistent estimator.

Moreover, under these conditions,

$$N^{1/2}(\widehat{\theta}^F - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}[0, \{\mathcal{I}^F(\theta_0)\}^{-1}], \tag{3.3}$$

where $\mathcal{I}^F(\theta_0)$ is the ***full data information matrix***, and

$$\mathcal{I}^F(\theta) = -E_\theta \left[\frac{\partial^2}{\partial\theta\,\partial\theta^T} \log\{p_Z(Z;\theta)\} \right] = E_\theta \left\{ S^F_\theta(Z;\theta) S^F_\theta(Z;\theta)^T \right\}. \tag{3.4}$$

The first expression in (3.4) is the expectation of the matrix of second partial derivatives of $\log\{p_Z(z;\theta)\}$ with respect to the elements of θ evaluated at θ_0. As in Chapter 1, this is shorthand for the fact that $N^{1/2}(\widehat{\theta}^F - \theta_0)$ converges in distribution (law) to a normal random vector with mean zero and covariance matrix $\{\mathcal{I}^F(\theta_0)\}^{-1}$.

Under regularity conditions, $\mathcal{I}^F(\widehat{\theta}^F) \xrightarrow{p} \mathcal{I}^F(\theta_0)$. Using this and (3.3), we have

$$\widehat{\theta}^F \,\dot{\sim}\, \mathcal{N}[\theta_0, N^{-1}\{\mathcal{I}^F(\widehat{\theta}^F)\}^{-1}]. \tag{3.5}$$

The expression $\mathcal{I}^F(\theta_0)$ is referred to as the ***expected information matrix*** and is an expectation for a single observation Z. The ***observed information matrix*** is a sample analog based on $\underset{\sim}{Z} = \{Z_i, i = 1 \ldots, N\}$; the usual definition is

$$I^F(\underset{\sim}{Z};\theta) = -\sum_{i=1}^N \frac{\partial^2}{\partial\theta\,\partial\theta^T} \log\{p_Z(Z_i;\theta)\}. \tag{3.6}$$

Under regularity conditions,

$$N^{-1} I^F(\underset{\sim}{Z};\widehat{\theta}^F) \xrightarrow{p} \mathcal{I}^F(\theta_0).$$

Thus, we can use this and (3.3) to conclude that

$$\widehat{\theta}^F \,\dot{\sim}\, \mathcal{N}[\theta_0, N^{-1}\{N^{-1} I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}] = \mathcal{N}[\theta_0, \{I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}]. \tag{3.7}$$

The results (3.5) and (3.7) are used in practice to derive approximate (large sample) standard errors for components of $\widehat{\theta}^F$ and confidence intervals for components of θ in the usual way.

The foregoing developments are standard and indeed fundamental results in statistical inference. The challenge in our context is that we do not observe the full data; rather, we wish carry out likelihood-based inference on θ in the model for the full data $p_Z(z;\theta)$ using the ***observed data***

$$(R, Z_{(R)}).$$

By analogy to the standard full data problem, this requires deducing the probability density of the observed data.

As a prelude to addressing this, we consider first the starting point for this and other methods we will discuss, the density of the ***ideal full data***

$$(R, Z).$$

We refer to these data as "ideal" because they are not observable in practice in general – if $R = \underset{\sim}{1}$, where $\underset{\sim}{1}$ is a K-vector of all 1s, then we do observe all of Z, but otherwise we cannot observe both R and Z. Although the ideal full data are thus not practically relevant, they are a useful device for motivating important frameworks for missing data methods, including likelihood-based methods.

3.2 Factorization of the density of (R, Z)

Consider the ideal full data (R, Z), with joint density $p_{R,Z}(r, z)$. Three different factorizations of this joint density lead to three different modeling approaches.

SELECTION MODEL FACTORIZATION: One way the joint density can be written is, in obvious notation,

$$p_{R,Z}(r, z) = p_{R|Z}(r|z)p_Z(z). \tag{3.8}$$

The factorization (3.8) involves directly the full data density $p_Z(z)$, for which the model $p_Z(z; \theta)$ has been postulated, and the missingness mechanism $p_{R|Z}(r|z)$.

As R is discrete, the probability density $p_{R|Z}(r|z)$ is a ***probability mass function*** so is analogous to the way we wrote the missingness mechanism previously, for fixed r and z, as $\text{pr}(R = r|Z = z)$. We say more about this in the next section.

This factorization suggests what is referred to as the ***selection model*** framework, which involves a full data model and a model for the missingness mechanism, which we write as

$$p_{R,Z}(r, z; \theta, \psi) = p_{R|Z}(r|z; \psi)p_Z(z; \theta). \tag{3.9}$$

In (3.9), $p_{R|Z}(r|z; \psi)$ is a parametric model for the missingness mechanism. Selection models were first used by Rubin (1976), and, according to Molenberghs and Kenward (2007, Chapter 3), the terminology was coined in the econometrics literature.

We will see in the next section how the selection model factorization is critical to formulating likelihood-based inference from the observed data.

PATTERN MIXTURE FACTORIZATION: The obvious alternative to the selection model factorization is to write the joint density as

$$p_{R,Z}(r,z) = p_{Z|R}(z|r)p_R(r). \tag{3.10}$$

The factorization (3.10) can be viewed as incorporating the density of the full data for given patterns of missingness weighted by the probability of each pattern.

The corresponding ***pattern mixture*** model framework involves modeling each of these components:

$$p_{R,Z}(r,z;\theta,\psi) = p_{Z|R}(z|r;\theta)p_R(r;\psi); \tag{3.11}$$

actually, $p_R(r;\psi)$ is simply the probability mass function for the discrete distribution of possible missingness patterns. Pattern mixture models were first proposed by Little (1993).

Note that in (3.11) θ and ψ have different interpretations from those in the selection model (3.9); we use the same symbols for convenience, but their meaning depends on the context. Thus, θ in (3.11) is not the parameter ordinarily of interest, that governing the assumed model for the full data. In fact, such a quantity is not directly represented in the pattern mixture model framework.

SHARED PARAMETER MODEL: A third type of framework is based on the same type of factorization (3.8) but conditional on an additional vector of ***random*** or ***latent effects*** b, namely

$$p_{R,Z|b}(r,z|b;\theta,\psi) = p_{Z|R,b}(z|r,b;\theta)p_{R|b}(r|b;\psi). \tag{3.12}$$

In (3.12), the random effects b are "shared" between the two components of the joint (conditional) density. The usual assumption is that Z and R are ***conditionally independent*** given b, so that (3.12) becomes

$$p_{R,Z|b}(r,z|b;\theta,\psi) = p_{Z|b}(z|b;\theta)p_{R|b}(r|b;\psi). \tag{3.13}$$

The ***shared parameter*** b can be interpreted as a latent mechanism governing both the data and missingness processes. In the shared parameter framework, one posits models depending on such a random effect; for example, if Z is a vector of longitudinal outcomes, $p_{Z|b}(z|b;\theta)$ may correspond to a hierarchical mixed effects model specified in terms of individual-specific random effects.

Note that by integrating (3.13) over the density of b, a model for the joint density of (R,Z) is obtained. Again, the interpretation of θ and ψ is different from those for selection and pattern mixture models.

We will return to pattern mixture and shared parameter models later in the course. We now demonstrate how the selection model framework forms the basis for deriving likelihood-based inference using the observed data.

3.3 Observed data likelihood and ignorability

Consider a model for the joint density of (R, Z) formulated according to the ***selection model*** factorization (3.9), namely

$$p_{R,Z}(r, z; \theta, \psi) = p_{R|Z}(r|z; \psi) p_Z(z; \theta). \tag{3.14}$$

SEPARABILITY CONDITION: It is conventional to assume that the parameters θ and ψ in (3.14) are ***variation independent***. If θ is $(p \times 1)$ and ψ is $(q \times 1)$, this states that the possible values of the $(p + q \times 1)$ vector $(\theta^T, \psi^T)^T$ lie in a rectangle of dimension $(p \times q)$; that is, if the parameter spaces of θ and ψ are Θ and Ψ, say, then the parameter space of $(\theta^T, \psi^T)^T$ is $\Theta \times \Psi$. This implies that the range of θ is the same for all possible values of ψ and vice versa. Thus, intuitively, knowing the value of ψ, say, provides no information on that of θ, and vice versa. This is often referred to as the ***separability condition***.

JOINT DENSITY OF $(R, Z_{(R)})$***:*** Recall that, for a specific value r of R, we write $Z_{(r)}$ to denote the part of Z that is observed for missingness pattern r and $Z_{(\bar{r})}$ to denote the part that is missing. Thus, we can think of Z as being partitioned as $Z = (Z_{(r)}, Z_{(\bar{r})})$ when $R = r$. Under (3.14), we can write the joint density of $(R, Z_{(R)})$ evaluated at r and $z_{(r)}$ as

$$\begin{aligned} p_{R,Z_{(R)})}(r, z_{(r)}; \theta, \psi) &= \int p_{R|Z}(r|z; \psi)\, p_Z(z; \theta)\, d\nu(z_{(\bar{r})}) \\ &= \int p_{R|Z}\{r|(z_{(r)}, z_{(\bar{r})}); \psi\}\, p_Z\{(z_{(r)}, z_{(\bar{r})}); \theta\}\, d\nu(z_{(\bar{r})}) \end{aligned} \tag{3.15}$$

- In (3.15), $\nu(\cdot)$ is a ***dominating measure***, which is Lebesgue measure for continuous random variables and the counting measure for discrete random variables. This is a technicality that allows us to write probability densities as we have been doing regardless of whether the random variables are continuous or discrete; a probability mass function is a probability density with respect to the counting measure. We can thus write expressions like (3.15) as integrals regardless of the nature of the random variables in question (instead of having to use summations explicitly when random variables are discrete.) (If you take ST 779, you will learn more about this.)

- Note that the interpretation of (3.15) is that we are integrating (summing) the joint density $p_{R,Z}(r, z; \theta, \psi)$ over the part of the full data that is missing, so that the resulting expression involves only $z_{(r)}$.

JOINT DENSITY UNDER MAR: The integral in (3.15) seems fairly complicated and implies that, even though interest focuses on θ, one must contend with ψ as well.

When the missingness mechanism is MAR, however, (3.15) takes on a simpler form. In particular, under MAR,

$$\text{pr}(R = r|Z) = \text{pr}(R = r|Z_{(r)}),$$

which under (3.14) is analogous to writing

$$p_{R|Z}(r|z;\psi) = p_{R|Z_{(r)}}(r|z_{(r)};\psi);$$

that is, $p_{R|Z}(r|z;\psi)$ depends only on $z_{(r)}$.

Under this condition, because $p_{R|Z}(r|z_{(r)};\psi)$ does not depend on $z_{(\bar{r})}$, it is not involved in the integral, and we can rewrite (3.15) as

$$p_{R,Z_{(R)}}(r, z_{(r)};\theta,\psi) = p_{R|Z_{(r)}}(r|z_{(r)};\psi) \int p_Z(z;\theta)\, d\nu(z_{(\bar{r})}). \tag{3.16}$$

Note further that

$$p_{Z_{(r)})}(z_{(r)};\theta) = \int p_Z(z;\theta)\, d\nu(z_{(\bar{r})}); \tag{3.17}$$

that is, (3.17) is the density for the part of Z that is observed for fixed r, as it is simply the density for the full data with the "missing part" integrated out.

Thus, combining (3.16) and (3.17), we have that the joint density of $(R, Z_{(R)})$ evaluated at r and $z_{(r)}$ can be written as

$$p_{R,Z_{(R)})}(r, z_{(r)};\theta,\psi) = p_{R|Z_{(r)}}(r|z_{(r)};\psi)\, p_{Z_{(r)}}(z_{(r)};\theta). \tag{3.18}$$

Some observations are immediate:

- Unlike in (3.15), where the involvement of θ and ψ is linked within the integral, the involvement of θ and ψ in (3.16) and (3.18) (so under MAR) is distinct.
- This feature is the basis for ***ignorability***, which we discuss next.

LIKELIHOOD: Given sample observed data $(R_i, Z_{(R_i)i})$, $i = 1, \dots, N$, we can use the foregoing results to write the likelihood for θ and ψ as follows.

Note that for an individual i for whom the full data are observed, so that $R_i = r = \underset{\sim}{1}$, in fact $Z_{(r)i} = Z_i$. With this convention, it should be clear from (3.16) and (3.18) that the contribution to the likelihood for a subject i with $R_i = r$ is

$$\begin{aligned} p_{R,Z_{(R)}}(r, Z_{(r)i}; \theta, \psi) &= p_{R|Z_{(r)}}(r|Z_{(r)i}; \psi) \int p_Z\{(Z_{(r)i}, z_{(\bar{r})}); \theta\}\, d\nu(z_{(\bar{r})}) \\ &= p_{R|Z_{(r)}}(r|Z_{(r)i}; \psi)\, p_{Z_{(r)}}(Z_{(r)i}; \theta) \qquad (3.19) \end{aligned}$$

Thus, using (3.19), the contribution to the likelihood for the ith individual is

$$\begin{aligned} \prod_r p_{R,Z_{(R)}}(r, Z_{(r)i}; \theta, \psi)^{I(R_i=r)} &= \prod_r \left\{ p_{R|Z_{(r)}}(r|Z_{(r)i}; \psi)\, p_{Z_{(r)}}(Z_{(r)i}; \theta) \right\}^{I(R_i=r)} \\ &= \prod_r p_{R|Z_{(r)}}(r|Z_{(r)i}; \psi)^{I(R_i=r)} \prod_r p_{Z_{(r)}}(Z_{(r)i}; \theta)^{I(R_i=r)}, \qquad (3.20) \end{aligned}$$

where the products in (3.20) are over all possible values of r.

IGNORABILITY: Under the ***separability condition***, which says that there is no information about θ in ψ, it is clear from (3.20) that only the second term is relevant to obtaining the MLE $\widehat{\theta}$ for θ. Accordingly, we need only maximize in θ

$$\prod_{i=1}^N \prod_r p_{Z_{(r)}}(Z_{(r)i}; \theta)^{I(R_i=r)} \text{ or equivalently } \quad \sum_{i=1}^N \sum_r I(R_i = r) \log\left\{ p_{Z_{(r)}}(Z_{(r)i}; \theta) \right\} \qquad (3.21)$$

- That is, we can ***ignore*** the missingness mechanism and need not even model it for the purpose of calculating the MLE. All we require is a model for the full data density.
- This feature is commonly referred to as ***ignorability*** of the missingness mechanism.
- Under ignorability and the separability condition, it is commonplace in the literature on missing data to refer to the first expression in (3.21), viewed as a function of θ for fixed data, as the ***observed data likelihood***. However, strictly speaking, this term should be used to refer to the product of (3.20) over $i = 1, \dots, N$,

To demonstrate how these general results are used in practice, we consider two examples.

EXAMPLE 1: Estimation of the mean. Recall the situation in ***EXAMPLE 1*** of Chapter 1 in which the full data are $Z = (Z_1, Z_2) = (Y, V)$. For simplicity, we take both Y and V to be continuous here. Interest focuses on estimation of

$$\mu = E(Y).$$

As before, V is a set of variables comprising ***auxiliary information*** that may be useful in justifying the MAR assumption.

Suppose that Y may be missing, but V is ***always*** observed. As in Chapter 1, $R = (R_1, R_2)$ can take on the two possible values $(1, 1)$ and $(0, 1)$. Assume that

$$\text{pr}\{R = (1,1)|Y, V\} = \text{pr}\{R = (1,1)|V\} = \pi(V),$$

which, as in Chapter 1, implies that $\text{pr}\{R = (0,1)|Y, V\} = 1 - \pi(V)$, so that $R \perp\!\!\!\perp Y|V$, and the missingness mechanism is MAR.

A likelihood-based approach to estimation of μ would proceed as follows. Suppose that we posit a model for the density of the full data of the form

$$p_Z(z;\theta) = p_{Y|V}(y|v;\theta_1)\, p_V(v;\theta_2), \quad \theta = (\theta_1^T, \theta_2^T)^T,$$

where θ_1 and θ_2 are ***variation independent***. Then we can write

$$\mu = E(Y) = E\{E(Y|V)\} = \int y\, p_{Y|V}(y|v;\theta_1)\, p_V(v;\theta_2)\, dy\, dv. \tag{3.22}$$

From (3.22), if we were to obtain estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for θ_1 and θ_2, we could estimate μ by

$$\widehat{\mu} = \int y\, p_{Y|V}(y|v;\widehat{\theta}_1)\, p_V(v;\widehat{\theta}_2)\, dy\, dv. \tag{3.23}$$

Consider the MLEs $\widehat{\theta}_1$ and $\widehat{\theta}_2$ based on the observed data. To find these, we must maximize the likelihood based on the observed data. Under MAR, from (3.20), if we were to posit a model for the missingness mechanism in terms of parameter ψ, the likelihood for θ and ψ is

$$\begin{aligned}
&\prod_{i=1}^N \left\{ \prod_r p_{R|Z_{(r)}}(r|Z_{(r)i};\psi)^{I(R_i=r)} \right\} \prod_r p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)} \\
&\qquad = \prod_{i=1}^N \left\{ \prod_r \text{pr}(R_i = r|Z_{(r)i};\psi)^{I(R_i=r)} \right\} \prod_r p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)}.
\end{aligned} \tag{3.24}$$

Under the ***separability condition***, from (3.21), we may ***ignore*** the first product over r in braces in (3.24) and need only maximize in θ

$$\prod_{i=1}^N \prod_r p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)}. \tag{3.25}$$

Although, under ***ignorability***, the first product in braces in (3.24) is ***irrelevant*** for the purpose of estimating θ, for illustration we present the forms of both products over r in (3.24) in this specific case. Here, the two possible values of r are $r = (1,1)$ and $r = (0,1)$, and $Z_{(r)} = (Y, V)$ when $r = (1,1)$ and $Z_{(r)} = V$ when $r = (0,1)$. Thus, when $r = (1,1)$,

$$p_{R|Z_{(r)}}(r|Z_{(r)i};\psi) = \text{pr}\{R_i = (1,1)|Y_i, V_i;\psi\} = \text{pr}\{R_i = (1,1)|V_i;\psi\} = \pi(V_i;\psi),$$

and when $r = (0,1)$,

$$p_{R|Z_{(r)}}(r|Z_{(r)i};\psi) = \text{pr}\{R_i = (0,1)|V_i;\psi\} = 1 - \pi(V_i;\psi),$$

where $\pi(v;\psi)$ is a parametric model for $\pi(v)$ depending on ψ. Thus, the first product becomes

$$\prod_r p_{R|Z_{(r)}}(r|Z_{(r)i};\psi)^{I(R_i=r)} = \{\pi(V_i;\psi)\}^{I\{R_i=(1,1)\}}\{1 - \pi(V_i;\psi)\}^{I\{R_i=(0,1)\}}.$$

Similarly, for the second product in (3.24), i.e., that in (3.25), when $r = (1,1)$, $Z_{(\bar{r})}$ is null, and thus

$$p_{Z_{(r)}}(Z_{(r)i};\theta) = \int p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}\, d\nu(z_{(\bar{r})}) = p_{Y|V}(Y_i|V_i;\theta_1)\, p_V(V_i;\theta_2);$$

when $r = (0,1)$, $Z_{(\bar{r})} = Y$, and

$$p_{Z_{(r)}}(Z_{(r)i};\theta) = \int p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}\, d\nu(z_{(\bar{r})}) = \int p_{Y|V}(y|V_i;\theta_1)\, p_V(V_i;\theta_2)\, dy = p_V(V_i;\theta_2).$$

Thus, the second product in (3.24) is

$$\begin{aligned}\prod_r p_{Z_{(r)}}(Z_{(r)i};\theta) &= \{p_{Y|V}(Y_i|V_i;\theta_1)\, p_V(V_i;\theta_2)\}^{I\{R_i=(1,1)\}}\,\{p_V(V_i;\theta_2)\}^{I\{R_i=(0,1)\}} \\ &= \{p_{Y|V}(Y_i|V_i;\theta_1)\}^{I\{R_i=(1,1)\}}\, p_V(V_i;\theta_2).\end{aligned}$$

(Why?) The ***observed data likelihood*** (3.25) is then

$$\prod_{i=1}^N \{p_{Y|V}(Y_i|V_i;\theta_1)\}^{I\{R_i=(1,1)\}}\, p_V(V_i;\theta_2) = \left\{\prod_{i=1}^N \{p_{Y|V}(Y_i|V_i;\theta_1)\}^{I\{R_i=(1,1)\}}\right\}\left\{\prod_{i=1}^N p_V(V_i;\theta_2)\right\}. \quad (3.26)$$

The MLEs $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are found by maximizing (3.26). Note that, under the assumption that θ_1 and θ_2 are ***variation independent***, the MLE $\widehat{\theta}_1$ is found by maximizing the product of the $p_{Y|V}(Y_i|V_i;\theta_1)$ over the ***complete cases*** for whom $R_i = (1,1)$, and $\widehat{\theta}_2$ is found by maximizing the product of $p_V(V_i;\theta_2)$ over all individuals (recalling that V is observed for all individuals). The MLEs are substituted into (3.23) to obtain $\widehat{\mu}$.

EXAMPLE 2: Bivariate normal model. Suppose that $Z = (Y_1, Y_2)$, where Y_1, Y_2 are scalar outcomes, and define $Y = (Y_1, Y_2)^T$. Assume that Y is bivariate normal, which we write as

$$Y \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_1, \mu_2)^T, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Then the full data model is

$$\begin{aligned} p_Z(z;\theta) &= p_Y(y;\theta) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\{-(y-\mu)^T\Sigma^{-1}(y-\mu)/2\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{y_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho}{\sigma_1\sigma_2}(y_1-\mu_1)(y_2-\mu_2) + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right\}\right] \end{aligned} \tag{3.27}$$

where $\sigma_{12} = \rho\sigma_1\sigma_2$, and $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \sigma_{12})^T$.

Define $R = (R_1, R_2)$ as usual, and assume that there are three possible values r of R:

(i) Y_1, Y_2 are both observed, $r = (1, 1)$, $Z_{(r)} = Y$

(ii) Y_1 is observed, Y_2 is missing, $r = (1, 0)$, $Z_{(r)} = Y_1$

(iii) Y_1 is missing, Y_2 is observed, $r = (0, 1)$, $Z_{(r)} = Y_2$.

Interest focuses on estimation of θ based on the observed data.

Assuming the separability condition, under ignorability, the MLE $\widehat{\theta}$ is found by maximizing the ***observed data likelihood*** in (3.21). We thus characterize $p_{Z_{(r)}}(z_{(r)};\theta)$ under each of (i) – (iii) above:

(i) $r = (1, 1)$, $Z_{(r)} = Y$, and hence $p_{Z_{(r)}}(z_{(r)};\theta) = p_Y(y;\theta)$ given in (3.27).

(ii) $r = (1, 0)$, $Z_{(r)} = Y_1$. For the multivariate normal distribution, it is well known that the distribution of each component is normal with the corresponding mean and variance. Accordingly,

$$p_{z_{(r)}}(Z_{(r)};\theta) = p_{Y_1}(y_1;\theta_1),$$

where $p_{Y_1}(y_1;\theta_1)$ is the $\mathcal{N}(\mu_1, \sigma_1^2)$ density depending on $\theta_1 = (\mu_1, \sigma_1^2)^T$.

(iii) $r = (0, 1)$, $Z_{(r)} = Y_2$. Similarly,

$$p_{z_{(r)}}(Z_{(r)};\theta) = p_{Y_2}(y_2;\theta_2), \quad \theta_2 = (\mu_2, \sigma_2^2)^T,$$

where $p_{Y_2}(y_2;\theta_2)$ is the $\mathcal{N}(\mu_2, \sigma_2^2)$ density.

Combining, the observed data likelihood in (3.21) is given by

$$\left\{\prod_{i=1}^{N}\{p_Y(Y_i;\theta)\}^{I\{R_i=(1,1)\}}\right\}\left\{\prod_{i=1}^{N}\{p_{Y_1}(Y_{i1};\theta_1)\}^{I\{R_i=(1,0)\}}\right\}\left\{\prod_{i=1}^{N}\{p_{Y_2}(Y_{i2};\theta_2)\}^{I\{R_i=(0,1)\}}\right\}, \tag{3.28}$$

where $\theta = (\theta_1^T, \theta_2^T, \sigma_{12})^T$.

PROPERTIES OF THE MLE: It is important to recognize that, although calculation of the MLE $\widehat{\theta}$ itself does not require specification of a model for the missingness mechanism, the large sample properties of the estimator ***do depend*** on the missingness mechanism, as we now demonstrate.

We continue to assume MAR, so that the missingness mechanism is ignorable. Under regularity conditions, as in Section 3.1, with the full data model correctly specified,

$$N^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}[0, \{\mathcal{I}(\theta_0)\}^{-1}], \tag{3.29}$$

where $\mathcal{I}(\theta_0)$ is the expected information matrix, and

$$\begin{aligned}\mathcal{I}(\theta) &= -E_\theta\left(\frac{\partial^2}{\partial\theta\partial\theta^T}\left[\sum_r I(R=r)\log\{p_{Z_{(r)}}(Z_{(r)};\theta)\}\right]\right)\\ &= E_\theta\Big\{S_\theta(R,Z_{(R)};\theta)S_\theta(R,Z_{(R)};\theta)^T\Big\}. \qquad (3.30)\end{aligned}$$

In (3.30), the ***observed data score vector*** is, from (3.20),

$$S_\theta(R,Z_{(R)};\theta) = \frac{\partial}{\partial\theta}\left[\sum_r I(R=r)\log\{p_{Z_{(r)}}(Z_{(r)};\theta)\}\right]. \tag{3.31}$$

In fact, (3.31) may be shown to be equivalent to

$$\begin{aligned}S_\theta(R,Z_{(R)};\theta) &= \sum_r I(R=r)E_\theta\Big\{S_\theta^F(Z;\theta)|Z_{(r)}\Big\} \qquad (3.32)\\ &= \sum_r I(R=r)E_\theta\Big\{S_\theta^F(Z;\theta)|R=r,Z_{(r)}\Big\} = E_\theta\Big\{S_\theta^F(Z;\theta)|R,Z_{(R)}\Big\}. \qquad (3.33)\end{aligned}$$

We first show (3.32). From (3.19), we have

$$\log\{p_{Z_{(r)}}(Z_{(r)};\theta)\} = \log\left\{\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})\right\},$$

so that the score vector is

$$S_\theta(R,Z_{(R)};\theta) = \sum_r I(R=r)\frac{\partial}{\partial\theta}\left[\log\left\{\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})\right\}\right].$$

This may be rewritten, interchanging the order of differentiation and integration, as

$$S_\theta(R,Z_{(R)};\theta) = \sum_r I(R=r)\,\frac{\displaystyle\int \frac{\partial}{\partial\theta}\left[p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\right]\,d\nu(z_{(\bar r)})}{\displaystyle\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}.$$

Dividing and multiplying by $p_Z\{(Z_{(r)}, z_{(\bar{r})});\theta\}$ in the integrand of the numerator yields

$$\begin{aligned} S_\theta(R, Z_{(R)};\theta) &= \sum_r I(R=r)\,\frac{\displaystyle\int S^F_\theta\{(Z_{(r)}, z_{(\bar{r})});\theta\} p_Z\{(Z_{(r)}, z_{(\bar{r})});\theta\}\, d\nu(z_{(\bar{r})})}{\displaystyle\int p_Z\{(Z_{(r)}, z_{(\bar{r})});\theta\}\, d\nu(z_{(\bar{r})})} \qquad (3.34)\\ &= \sum_r I(R=r) \int S^F_\theta\{(Z_{(r)}, z_{(\bar{r})});\theta\}\, p_{Z|Z_{(r)}}\{(Z_{(r)}, z_{(\bar{r})})|Z_{(r)};\theta\}\, d\nu(z_{(\bar{r})})\\ &= \sum_r I(R=r) \int S^F_\theta\{(Z_{(r)}, z_{(\bar{r})});\theta\}\, p_{Z_{(\bar{r})}|Z_{(r)}}\{z_{(\bar{r})}|Z_{(r)};\theta\}\, d\nu(z_{(\bar{r})}). \qquad (3.35) \end{aligned}$$

The equality in (3.35) is a consequence of the fact that

$$p_{Z|Z_{(r)}}\{(Z_{(r)}, z_{(\bar{r})})|Z_{(r)};\theta\} = \frac{p_Z\{(Z_{(r)}, z_{(\bar{r})});\theta\}}{p_{Z_{(r)}}(Z_{(r)};\theta)} = p_{Z_{(\bar{r})}|Z_{(r)}}\{z_{(\bar{r})}|Z_{(r)};\theta\},$$

where we are regarding $Z_{(r)}$ as fixed in these calculations. The expression in (3.35) is equivalent to (3.32). Note that nowhere in this argument did we use the MAR assumption.

We now assume MAR and demonstrate the first equality in (3.33); the second equality is by definition. To show this, it suffices to show that, for fixed r,

$$p_{Z|R,Z_{(R)}}(z|r, z_{(r)};\theta,\psi) = p_{Z|R,Z_{(R)}}(z|r, z_{(r)};\theta) = p_{Z|Z_{(r)}}(z|z_{(r)};\theta).$$

This is true because, when $Z_{(r)} = z_{(r)}$,

$$\begin{aligned} p_{Z|R,Z_{(R)}}(z|r, z_{(r)};\theta,\psi) &= \frac{p_{R,Z}(r, z;\theta,\psi)}{\displaystyle\int p_{R,Z}(r, z;\theta,\psi)\, d\nu(z_{(\bar{r})})} = \frac{p_{R|Z}(r|z;\psi)p_Z(z;\theta)}{\displaystyle\int p_{R|Z}(r|z;\psi)p_Z(z;\theta)\, d\nu(z_{(\bar{r})})}\\ &= \frac{p_{R|Z_{(r)}}(r|z_{(r)};\psi)p_Z(z;\theta)}{p_{R|Z_{(r)}}(r|z_{(r)};\psi)\displaystyle\int p_Z(z;\theta)\, d\nu(z_{(\bar{r})})} \qquad (3.36)\\ &= \frac{p_Z(z;\theta)}{\displaystyle\int p_Z(z;\theta)\, d\nu(z_{(\bar{r})})} = p_{Z|Z_{(r)}}(z|z_{(r)};\theta), \end{aligned}$$

where the equality in (3.36) follows by MAR.

Analogous to (3.6) in the full data case, the ***observed information matrix*** for the observed data is

$$I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta) = -\sum_{i=1}^N \frac{\partial^2}{\partial\theta\partial\theta^T}\left[\sum_r I(R_i = r)\log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}\right]. \qquad (3.37)$$

Taking (3.29) together with (3.30) and (3.31) makes clear that the properties of $\widehat{\theta}$ depend on the distribution of (R, Z). We discuss methods for computation of standard errors for $\widehat{\theta}$ later in this chapter.

Note also that the MLE $\widehat{\theta}$ may be equivalently expressed as the solution to the score equation

$$\sum_{i=1}^N S_\theta(R_i, Z_{(R_i)i};\theta) = 0. \qquad (3.38)$$

3.4 Expectation-Maximization (EM) algorithm

The foregoing results demonstrate that, under the assumption of MAR, the MLE for θ in a posited model for the full data $p_Z(z;\theta)$ can be found in principle by maximizing either of the expressions in (3.21). As the derivation of this result suggests, however, actual implementation in practice may be ***computationally challenging*** because of the need to calculate the integrals involved, which in most problems are likely to be analytically intractable.

EXPECTATION-MAXIMIZATION ALGORITHM: The Expectation-Maximization (EM) algorithm is an iterative numerical technique for maximizing an objective function whose formulation involves unobservable latent quantities. Here, the latent quantities are the missing components of Z over which integration is performed to arrive at the observed data expressions in (3.21).

We first present the generic EM algorithm in the context of maximizing (3.21) and then discuss practical implications.

Consider the first expression in (3.21),

$$\prod_{i=1}^{N} \prod_{r} p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)}. \tag{3.39}$$

The goal is to maximize (3.39) in θ. Note that (3.39) can be written as

$$\prod_{i=1}^{N} \prod_{r} p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)} = \prod_{i=1}^{N} \prod_{r} \left\{ \frac{p_Z(Z_i;\theta)}{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)} \right\}^{I(R_i=r)}.$$

Taking logarithms yields

$$\sum_{i=1}^{N} \sum_{r} I(R_i=r)\log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\} = \sum_{i=1}^{N} \log\{p_Z(Z_i;\theta)\} - \sum_{i=1}^{N} \sum_{r} I(R_i=r)\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}, \tag{3.40}$$

and we equivalently wish to maximize the objective function (3.40).

Now take the conditional expectation of both sides of (3.40) given the observed sample data $(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})$. Because the sample data are iid across i, this is equivalent to taking the conditional expectation of the ith summand with respect to $(R_i, Z_{(R_i)i})$ for $i = 1, \ldots, N$. We consider this conditional expectation for the ith summand in each term in (3.40).

For the left hand side of (3.40), we have

$$
\begin{aligned}
&E_{\theta'}\Big[\sum_r I(R_i = r)\ \log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}|R_i, Z_{(R_i)i}\Big]\\
&\quad = \sum_s\sum_r I(R_i = r)I(R_i = s)\ E_{\theta'}\Big[\log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}|R_i = s, Z_{(s)i}\Big]\\
&\quad = \sum_r I(R_i = r)E_{\theta'}\Big[\log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}|R_i = r, Z_{(r)i}\Big] = \sum_r I(R_i = r)\log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}, \qquad (3.41)
\end{aligned}
$$

where (3.41) follows because all terms in the double sum are equal to zero except when $r = s$, and $\log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}$ is function of the observed data on i. In (3.41), we emphasize that this expectation is taken with respect to the distribution of Z evaluated at some value θ' that may be ***different from*** the value θ in $p_{Z_{(r)}}(Z_{(r)i};\theta)$; the reason for this will be clear shortly.

Now consider the first term on the right hand side of (3.40). It is straightforward that

$$
\begin{aligned}
E_{\theta'}\Big[\log\{p_Z(Z_i;\theta)\}|R_i, Z_{(R_i)i}\Big] &= \sum_r I(R_i = r)\ E_{\theta'}\Big[\log\{p_Z(Z_i;\theta)\}|R_i = r, Z_{(r)i}\Big]\\
&= \sum_r I(R_i = r)\ E_{\theta'}\Big[\log\{p_Z(Z_i;\theta)\}|Z_{(r)i}\Big], \qquad (3.42)
\end{aligned}
$$

where the equality in (3.42) follows as a result of the MAR assumption by an argument similar to that leading to the first equality in (3.33) (try showing this).

Finally, for a summand in the second term on the right hand side of (3.40), we have

$$
\begin{aligned}
&E_{\theta'}\Big[\sum_r I(R_i = r)\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}|R_i, Z_{(R_i)i}\Big]\\
&\quad = \sum_s\sum_r I(R_i = r)I(R_i = s)E_{\theta'}\Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}|R_i = s, Z_{(s)i}\Big]\\
&\quad = \sum_s\sum_r I(R_i = r)I(R_i = s)E_{\theta'}\Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}|Z_{(s)i}\Big] \qquad (3.43)\\
&\quad = \sum_r I(R_i = r)E_{\theta'}\Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}|Z_{(r)i}\Big], \qquad (3.44)
\end{aligned}
$$

where the equality in (3.43) is a consequence of MAR, and that in (3.44) follows because all terms in the double sum are equal to zero except when $r = s$.

Using all of (3.41), (3.42), and (3.44), we thus have from (3.40) that

$$
\begin{aligned}
&\sum_{i=1}^N\sum_r I(R_i = r)\ \log\{p_{Z_{(r)}}(Z_{(r)i};\theta)\}\\
&\quad = \sum_{i=1}^N\sum_r I(R_i = r)\ E_{\theta'}\Big[\log\{p_Z(Z_i;\theta)\}|Z_{(r)i}\Big] - \sum_{i=1}^N\sum_r I(R_i = r)E_{\theta'}\Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}|Z_{(r)i}\Big]\\
&\quad = Q(\theta|\theta') - H(\theta|\theta'), \qquad (3.45)
\end{aligned}
$$

say. That is, the objective function to be maximized in θ to obtain the MLE can be expressed as in (3.45) for any fixed value θ' of θ.

The result in (3.45) is the basis for the iterative procedure known as the EM algorithm.

If we index iterations by t and let $\theta^{(t)}$ be the tth iterate, starting from some initial value $\theta^{(0)}$, the $(t+1)$th iteration of the algorithm involves the following two steps:

E-step. Calculate the conditional ***E***xpectation

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{r} I(R_i = r)E_{\theta^{(t)}}\Big[\log\{p_Z(Z_i;\theta)\}|Z_{(r)i}\Big]. \tag{3.46}$$

M-step. Maximize $Q(\theta|\theta^{(t)})$ in θ to obtain $\theta^{(t+1)}$; that is,

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

Iteration continues until some convergence criterion is satisfied. The value of θ at convergence can be taken to be the MLE $\widehat{\theta}$.

REMARKS:

- The basic idea behind the EM algorithm appeared in applications in the statistical literature for much of the twentieth century, but it was not until 1977 and a landmark paper (Dempster, Laird, and Rubin, 1977) that the idea was formalized and presented in generality.

- Note that the ***E-step*** can be written as

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{r} I(R_i = r)\int \log[p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}]\,p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}\,d\nu(z_{(\bar{r})})\\
&\quad= \sum_{i=1}^{N}\sum_{r} I(R_i = r)\int \log[p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}]\,p_{Z_{(\bar{r})}|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}\,d\nu(z_{(\bar{r})}) \qquad (3.47)\\
&\quad= \sum_{i=1}^{N}\sum_{r} I(R_i = r)\int \log[p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}]\,p_{Z|R,Z_{(R)}}\{(Z_{(r)i}, z_{(\bar{r})})|R_i = r, Z_{(r)i};\theta^{(t)}\}\,d\nu(z_{(\bar{r})}).
\end{aligned}$$

  This makes explicit that this step can be viewed as averaging the loglikelihood of the full data over the (frequentist) ***predictive distribution*** of the missing part of Z given the observed, evaluated at the current iterate $\theta^{(t)}$. Under MAR, this predictive distribution for fixed θ and r is

$$p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta\} = p_{Z|R,Z_{(R)}}\{(Z_{(r)i}, z_{(\bar{r})})|R_i = r, Z_{(r)i};\theta\} = p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i};\theta). \tag{3.48}$$

  The last expression is the one that is useful in practice, as it makes clear the interpretation of predicting missing components of Z from those observed. The predictive distribution, both from a frequentist and Bayesian point of view, will play a major role in our discussion of multiple imputation in Chapter 4.

ALTERNATIVE NOTATION: In the foregoing developments, we have used the notation introduced in Chapter 1, which makes precise which parts of the full data Z are missing and observed under a given pattern of missingness. In the literature on missing data, it is customary to present these arguments using the ***informal notation***

$$Z = (Z^{obs}, Z^{mis}),$$

also introduced in Chapter 1. As noted there, it is important to be aware that, in these more informal arguments, care must be taken to recognize that the dimensions of the quantities Z^{obs} and Z^{mis} for any individual are ***not fixed*** and may be ***different*** for different individuals indexed by i. Moreover, Z^{obs} by itself ***does not*** fully characterize the observed data.

RATIONALE FOR THE EM ALGORITHM: We now sketch the usual argument justifying the EM algorithm. The argument shows that the value of the objective function

$$\sum_{i=1}^{N} \sum_{r} I(R_i = r) \log\{p_{Z_{(r)}}(Z_{(r)i}; \theta)\}$$

increases at each iteration of the algorithm.

Consider the difference in the objective function at two successive iterates, which, using (3.45), is given by

$$\sum_{i=1}^{N} \sum_{r} I(R_i = r) \log\{p_{Z_{(r)}}(Z_{(r)i}; \theta^{(t+1)})\} - \sum_{i=1}^{N} \sum_{r} I(R_i = r) \log\{p_{Z_{(r)}}(Z_{(r)i}; \theta^{(t)})\}$$
$$= Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) - \left\{ H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \right\} \tag{3.49}$$

Trivially, by the definition of $\theta^{(t+1)}$, the first term in (3.49) is such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) \geq 0, \tag{3.50}$$

as $\theta^{(t+1)}$ maximizes $Q(\theta|\theta^{(t)})$ in θ.

Moreover, we can show that

$$H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \leq 0. \tag{3.51}$$

Combining (3.50) and (3.51) then shows that the difference (3.49) is nonnegative, from which it follows that the objective function increases at each iteration.

To show that (3.51) holds, recall that

$$H(\theta|\theta^{(t)}) = \sum_{i=1}^{N} \sum_{r} I(R_i = r) E_{\theta^{(t)}} \Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta)\}|Z_{(r)i} \Big],$$

so that (3.51) can be written as

$$\sum_{i=1}^{N} \sum_{r} I(R_i = r) \Big(E_{\theta^{(t)}} \Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta^{(t+1)})\}|Z_{(r)i} \Big] - E_{\theta^{(t)}} \Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta^{(t)})\}|Z_{(r)i} \Big] \Big). \quad (3.52)$$

Consider the difference of expectations in the summand of this expression for any i. We have

$$\begin{aligned} & E_{\theta^{(t)}} \Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta^{(t+1)})\}|Z_{(r)i} \Big] - E_{\theta^{(t)}} \Big[\log\{p_{Z|Z_{(r)}}(Z_i|Z_{(r)i};\theta^{(t)})\}|Z_{(r)i} \Big] \\ &\qquad = \int \log[p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t+1)}\}]\, p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}\, d\nu(z_{(\bar{r})}) \\ &\qquad - \int \log[p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}]\, p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}\, d\nu(z_{(\bar{r})}) \\ &\qquad = \int \log[p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t+1)}\}]\, p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i};\theta^{(t)})\, d\nu(z_{(\bar{r})}) \\ &\qquad - \int \log[p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}]\, p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i};\theta^{(t)})\, d\nu(z_{(\bar{r})}). \end{aligned}$$

Note that in these calculations we are treating $Z_{(r)i}$ as fixed.

This is of the form

$$\int \log\{p_2(z)\}\, p_1(z)\, dz - \int \log\{p_1(z)\}\, p_1(z)\, dz, \quad (3.53)$$

identifying $p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t+1)}\}$ with $p_2(z)$ and $p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}$ with $p_1(z)$ and treating the argument of integration as continuous.

Note that (3.53) can be rewritten as

$$\int \log\left\{\frac{p_2(z)}{p_1(z)}\right\} p_1(z)\, dz,$$

so that the result follows if we can show that this expression is ≤ 0. Because the logarithm is a ***concave function***, by ***Jensen's inequality***, we have

$$\int \log\left\{\frac{p_2(z)}{p_1(z)}\right\} p_1(z)\, dz \leq \log\left\{\int \frac{p_2(z)}{p_1(z)} p_1(z)\, dz\right\} = \log\left\{\int p_2(z)\, dz\right\} = 0.$$

Thus, each summand in (3.52) is ≤ 0, and the result follows.

REMARKS:

- Under suitable regularity conditions, in general, the EM algorithm is guaranteed to converge to a ***stationary point*** of the objective function (observed data loglikelihood), which could be a local or global maximum of the objective function or a saddle point.

- Dempster et al. (1977) and Wu (1983) demonstrated that if the objective function has a unique maximum in θ in the interior of the parameter space, the EM algorithm ***will converge*** as $t \to \infty$ to the (unique global) maximum.

- In less well-behaved problems, the algorithm ***may not*** converge to a unique maximum; situations like this are discussed by Schafer (1997, Chapter 3).

- In practice, it is advisable to carry out the algorithm starting from ***several different initial values*** $\theta^{(0)}$. If the algorithm converges to different values from different starting points, this may indicate the presence of local maxima. If the algorithm converges to the same value from different starting points, this may engender confidence that a unique global maximum has been found.

- It may be shown (see, for example, Molenberghs and Kenward, 2007, Section 8.4 or Schafer, 1997, Section 3.3.2) that the EM algorithm ***converges linearly*** and that the rate of convergence is related to the amount of missing information. In practice, the convergence can be extremely slow. In contrast, the convergence of optimization methods such as the ***Newton-Raphson algorithm*** can be shown to be ***quadratic***, so that if one is able to maximize the observed data likelihood directly via such techniques, this is likely to be more computationally efficient. There is thus a ***tradeoff*** between the ease of implementation of the EM algorithm and performance.

- A ***disadvantage*** is that ***standard error estimates*** are not readily available at the conclusion of the algorithm as they would be, for example, from direct maximization of the observed data likelihood using optimization techniques such as the Newton-Raphson algorithm, in which the ***observed information matrix*** is calculated at each internal iteration.

  We discuss approaches to obtaining standard errors following the EM algorithm shortly.

PRACTICAL ADVANTAGE: The EM algorithm is particularly attractive in problems where

- $Q(\theta|\theta^{(t)})$ is relatively ***easy to compute*** (***E-step***)
- $Q(\theta|\theta^{(t)})$ is relatively ***easy to maximize*** in θ (***M-step***).

For a large class of problems, this is the case. From (3.46),

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N} \sum_{r} I(R_i = r) E_{\theta^{(t)}} \Big[\log\{p_Z(Z_i;\theta)\} | Z_{(r)i} \Big] = \sum_{i=1}^{N} E_{\theta^{(t)}} \Big[\log\{p_Z(Z_i;\theta)\} | R_i, Z_{(R_i)i} \Big].$$

From (3.42), and using the fact that Z_i, $i = 1, \ldots, N$, and $(R_i, Z_{(R_i)i})$, $i = 1, \ldots, N$, are iid, we can write

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}} \Big[\sum_{i=1}^{N} \log\{p_Z(Z_i;\theta)\} | \underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})} \Big] = E_{\theta^{(t)}} \Big[\log\{p_{\underset{\sim}{Z}}(\underset{\sim}{Z};\theta) | \underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})} \Big],$$

where

$$p_{\underset{\sim}{Z}}(\underset{\sim}{Z};\theta) = \prod_{i=1}^{N} p_Z(Z_i;\theta)$$

is the joint density of the full sample data.

The case where $\log\{p_{\underset{\sim}{Z}}(\underset{\sim}{Z};\theta)\}$ is ***linear*** in ***sufficient statistics*** enjoys these advantages. For example, suppose that $p_Z(z;\theta)$ belongs to an ***exponential family***, where θ is $(p \times 1)$, so that, expressed in ***natural or canonical form***,

$$p_{\underset{\sim}{Z}}(\underset{\sim}{Z};\theta) = b(\underset{\sim}{Z}) \exp\left\{ \sum_{\ell=1}^{p} T_\ell(\underset{\sim}{Z}) \eta_\ell(\theta) - a(\theta) \right\},$$

for known functions $b(\,\cdot\,)$ and $a(\,\cdot\,)$, where $\eta_\ell(\theta)$, $\ell = 1, \ldots, p$, are the ***natural parameters***, and $T_\ell(\underset{\sim}{Z})$, $\ell = 1, \ldots, p$, are the associated ***sufficient statistics***. Then

$$\log\{p_{\underset{\sim}{Z}}(\underset{\sim}{Z};\theta)\} = \log\{b(\underset{\sim}{Z})\} + \sum_{\ell=1}^{p} T_\ell(\underset{\sim}{Z}) \eta_\ell(\theta) - a(\theta). \tag{3.54}$$

Taking derivatives of (3.54) with respect to the elements of θ and setting equal to zero yields the MLE. If we define

$$q_\ell(\theta) = E_\theta\{T_\ell(\underset{\sim}{Z})\}$$

then it is well known (see ST 522) that this results in solving the p moment equations

$$q_\ell(\theta) = T_\ell(\underset{\sim}{Z}), \quad \ell = 1, \ldots, p, \tag{3.55}$$

which is a set of p equations in p unknowns (elements of θ).

In this situation, the E- and M-steps are straightforward. At the $(t+1)$th iteration, the E-step involves finding

$$E_{\theta^{(t)}}\Big[\log\{b(\underset{\sim}{Z})\} + \sum_{\ell=1}^{p} T_\ell(\underset{\sim}{Z})\eta_\ell(\theta) - a(\theta)|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\Big], \tag{3.56}$$

and then the M-step involves maximizing (3.56) in θ_ℓ, $\ell = 1, \dots, p$. But from (3.55), this maximization will involve ***only***

$$E_{\theta^{(t)}}\{T_\ell(\underset{\sim}{Z})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}, \quad \ell = 1, \dots, p.$$

Thus, the algorithm at the $(t+1)$th iteration reduces to the following:

E-step. Find $\quad E_{\theta^{(t)}}\{T_\ell(\underset{\sim}{Z})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}, \quad \ell = 1, \dots, p.$

M-step. Find $\theta^{(t+1)}$ as the solution to the p equations

$$q_\ell(\theta) = E_{\theta^{(t)}}\{T_\ell(\underset{\sim}{Z})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}, \quad \ell = 1, \dots, p.$$

We now demonstrate how this works in two examples.

EXAMPLE 2, CONTINUED: Bivariate normal model. The case where $p_Z(z;\theta)$ is bivariate normal as in (3.27) is an example where the EM algorithm is straightforward. Recall $Z = (Y_1, Y_2)$, $Y = (Y_1, Y_2)^T$, $R = (R_1, R_2)$, and the three possible situations are

(i) Y_1, Y_2 are both observed, $r = (1,1)$, $Z_{(r)} = Y$

(ii) Y_1 is observed, Y_2 is missing, $r = (1,0)$, $Z_{(r)} = Y_1$

(iii) Y_1 is missing, Y_2 is observed, $r = (0,1)$, $Z_{(r)} = Y_2$.

The full data density is again

$$p_Z(z;\theta) = p_Y(y;\theta) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\{-(y-\mu)^T\Sigma^{-1}(y-\mu)/2\},$$

$\mu = (\mu_1, \mu_2)^T$, $\text{vech}(\Sigma) = (\sigma_1^2, \sigma_{12}, \sigma_2^2)^T$, where $\text{vech}(A)$ is the vector of distinct elements of a symmetric matrix A, so that $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \sigma_{12})^T$, and $p = 5$.

By the calculations above, it can be shown (try it) that the $p = 5$ sufficient statistics are

$$T_1(\underset{\sim}{Z}) = \sum_{i=1}^{N} Y_{i1}, \quad T_2(\underset{\sim}{Z}) = \sum_{i=1}^{N} Y_{i1}^2, \quad T_3(\underset{\sim}{Z}) = \sum_{i=1}^{N} Y_{i2}, \quad T_4(\underset{\sim}{Z}) = \sum_{i=1}^{N} Y_{i2}^2, \quad T_5(\underset{\sim}{Z}) = \sum_{i=1}^{N} Y_{i1}Y_{i2}. \tag{3.57}$$

Thus, the moment equations (3.55) yield

$$\mu_1 = N^{-1}\sum_{i=1}^N Y_{i1},\quad \sigma_1^2 = N^{-1}\sum_{i=1}^N Y_{i1}^2 - (\mu_1)^2,\quad \mu_2 = N^{-1}\sum_{i=1}^N Y_{i2},\quad \sigma_2^2 = N^{-1}\sum_{i=1}^N Y_{i2}^2 - (\mu_2)^2,$$

$$\sigma_{12} = N^{-1}\sum_{i=1}^N Y_{i1}Y_{i2} - (\mu_1\mu_2)^2, \tag{3.58}$$

Thus, from (3.57) and (3.58), recalling that Z_i, $i = 1, \ldots, N$, and $(R_i, Z_{(R_i)i})$, $i = 1, \ldots, N$, are iid, the E-step involves finding

$$E(Y_1|Z_{(r)}),\quad E(Y_1^2|Z_{(r)}),\quad E(Y_2|Z_{(r)}),\quad E(Y_2^2|Z_{(r)}),\quad E(Y_1Z_2|Z_{(r)}), \tag{3.59}$$

for the three cases (i) – (iii) above. Trivially, in case (i), $r = (1, 1)$, the conditional expectations (3.59) are equal to the observed values Y_1, Y_1^2, Y_2, Y_2^2, Y_1Y_2. For (ii), $r = (1, 0)$, $Z_{(r)} = Y_1$,

$$E(Y_1|Z_{(r)}) = Y_1,\quad E(Y_1^2|Z_{(r)}) = Y_1^2,$$

while, by standard properties of the multivariate normal distribution,

$$E(Y_2|Z_{(r)}) = E(Y_2|Y_1) = \mu_2 + \sigma_{12}(Y_1 - \mu_1)/\sigma_1^2,$$

$$E(Y_2^2|Z_{(r)}) = E(Y_2^2|Y_1) = \sigma_2^2 - \sigma_{12}^2/\sigma_1^2 + \{E(Y_2|Y_1)\}^2,$$

$$E(Y_1Y_2|Z_{(r)}) = E(Y_1Y_2|Y_1) = Y_1E(Y_2|Y_1).$$

For (iii), $r = (0, 1)$, reverse the roles of Y_1 and Y_2 in these expressions.

Combining, the EM-algorithm takes the following form. At the $(t + 1)$ iteration

E-step. With $\theta^{(t)} = (\mu_1^{(t)}, \sigma_1^{2(t)}, \mu_2^{(t)}, \sigma_2^{2(t)}, \sigma_{12}^{(t)})^T$, calculate

$$T_1^{(t)} = \sum_{i=1}^N\sum_r I(R_i = r)E_{\theta^{(t)}}(Y_{i1}|Z_{(r)i}),\quad T_2^{(t)} = \sum_{i=1}^N\sum_r I(R_i = r)E_{\theta^{(t)}}(Y_{i1}^2|Z_{(r)i}),$$

$$T_3^{(t)} = \sum_{i=1}^N\sum_r I(R_i = r)E_{\theta^{(t)}}(Y_{i2}|Z_{(r)i}),\quad T_4^{(t)} = \sum_{i=1}^N\sum_r I(R_i = r)E_{\theta^{(t)}}(Y_{i2}^2|Z_{(r)i}),$$

$$T_5^{(t)} = \sum_{i=1}^N\sum_r I(R_i = r)E_{\theta^{(t)}}(Y_{i1}Y_{i2}|Z_{(r)i}).$$

M-step. Update $\theta^{(t+1)}$ as

$$\mu_1^{(t+1)} = T_1^{(t)}/N,\quad \sigma_1^{2(t+1)} = T_2^{(t)}/N - \left(\mu_1^{(t+1)}\right)^2,\quad \mu_2^{(t+1)} = T_3^{(t)}/N,\quad \sigma_2^{2(t+1)} = T_4^{(t)}/N - \left(\mu_2^{(t+1)}\right)^2,$$

$$\sigma_{12}^{(t+1)} = T_5^{(t)}/N - \left(\mu_1^{(t+1)}\mu_2^{(t+1)}\right).$$

EXAMPLE 3. Multivariate normal model with monotone missingness (dropout). Consider again the simple longitudinal situation of ***EXAMPLE 3*** of Chapter 1, in which a scalar outcome is collected at each of T times $t_1, \dots, t_T$, so that the full data are $Z = (Y_1, \dots, Y_T)$. Let $Y = (Y_1, \dots, Y_T)^T$, and suppose that we assume that Y is $\mathcal{N}(\mu, \Sigma)$ for $(T \times 1)$ mean vector μ and $(T \times T)$ covariance matrix Σ. Thus,

$$\theta = \{\mu^T, \text{vech}(\Sigma)^T\}^T,$$

and the full data density $p_Z(z;\theta)$ is

$$p_Z(z;\theta) = p_Y(y;\theta) = \frac{1}{(2\pi)^T |\Sigma|^{T/2}} \exp\{-(y-\mu)^T \Sigma^{-1}(y-\mu)/2\}.$$

It is straightforward to deduce that the ***sufficient statistics*** may be summarized as $\sum_{i=1}^N Y_i$ and $\sum_{i=1}^N Y_i Y_i^T$ and that the ***moment equations*** (3.55) are

$$\mu = N^{-1}\sum_{i=1}^N Y_i, \qquad \Sigma = N^{-1}\sum_{i=1}^N Y_i Y_i^T - \mu\mu^T. \tag{3.60}$$

Suppose that individuals ***drop out***, so that there is ***monotone missingness***. Then with $R = (R_1, \dots, R_T)$, R can take on T possible values r given by

$$r^{(0)} = (0,0,\dots,0), \;\; r^{(1)} = (1,0,\dots,0), \;\; \dots, \;\; r^{(T)} = (1,1,\dots,1).$$

It is convenient to use the ***dropout notation*** defined in Chapter 1; i.e.,

$$D = 1 + \sum_{j=1}^T R_j,$$

so that $R = r^{(j)}$ corresponds to $D = j+1$. Thus, when $D = j+1$, write $Z_{(j+1)} = (Y_1, \dots, Y_j)$ to denote the part of Z that is observed. Note that $Z_{(1)}$ is thus null, as no part of Y is observed, and $Z_{(T+1)} = (Y_1, \dots, Y_T)$.

The E-step thus involves finding for any $j = 0, \dots, T$

$$E(Y|Z_{(j+1)}), \qquad E(YY^T|Z_{(j+1)}),$$

Define $Y_{(j)} = (Y_1, \dots, Y_j)^T$, $j = 1, \dots, T$, so that, for any $j = 1, \dots, T$, Y can be partitioned as

$$Y = \begin{pmatrix} Y_{(j)} \\ Y_{(\bar{j})} \end{pmatrix}.$$

When $j = 0$, so that $D = 1$, there is no contribution to the likelihood, so consider $j = 1, \dots, T$. For fixed j, partition μ and Σ analogous to Y as

$$\mu = \begin{pmatrix} \mu_{(j)} \\ \mu_{(\bar{j})} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{(jj)} & \Sigma_{(j\bar{j})} \\ \Sigma_{(\bar{j}j)} & \Sigma_{(\bar{j}\bar{j})} \end{pmatrix}.$$

Then it is straightforward to derive that (try it)

$$E(Y|Z_{(j+1)}) = E(Y|Y_{(j)}) = \begin{pmatrix} Y_{(j)} \\ \mu_{(\bar{j})} + \Sigma_{(\bar{j}j)}\Sigma^{-1}_{(j\bar{j})}(Y_{(j)} - \mu_{(j)}) \end{pmatrix},$$

$$E(YY^T|Z_{(j+1)}) = E(YY^T|Y_{(j)}) = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{(\bar{j}\bar{j})} - \Sigma_{(\bar{j}j)}\Sigma^{-1}_{(jj)}\Sigma_{(j\bar{j})} \end{pmatrix} + E(Y|Y_{(j)})E(Y|Y_{(j)})^T.$$

Thus, the $(t+1)$th iterate of the EM algorithm is as follows.

E-step. With $\theta^{(t)} = \{\mu^{(t)\,T}, \text{vech}(\Sigma)^{(t)\,T}\}^T$, define

$$T_1^{(t)} = \sum_{i=1}^{N}\sum_{j=1}^{T} I(D_i = j+1)E_{\theta^{(t)}}(Y_i|Y_{(j)i}), \qquad T_2^{(t)} = \sum_{i=1}^{N}\sum_{j=1}^{T} I(D_i = j+1)E_{\theta^{(t)}}(Y_i Y_i^T|Y_{(j)i}).$$

M-step. Update $\theta^{(t+1)}$ as

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

REMARKS:

- Schafer (1997, Section 5.3) presents a detailed derivation of the EM algorithm for the situation in ***EXAMPLE 3*** in the more general case of ***nonmonotone*** (or ***arbitrary patterns*** of) missingness.
- In fact, the EM algorithm for the case where the full data Z are a vector of outcomes Y that are distributed as multivariate normal, interest focuses on estimation of the mean and covariance matrix, and arbitrary patterns of missingness is implemented in the SAS procedure `proc mi`.
- These examples are cases where the E-step is straightforward and the solution to the maximization problem in the M-step is available in a closed form. This is generally ***not*** the case in most problems.

MONTE CARLO EM ALGORITHM: When interest focuses on more complex statistical models for the full data, it is often the case that the E-step may not be carried out in a ***closed form***. That is, it is not possible to calculate the summand in

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N} \sum_{r} I(R_i = r) E_{\theta^{(t)}} \Big[\log\{p_Z(Z_i;\theta)\} | Z_{(r)i} \Big] = \sum_{i=1}^{N} E_{\theta^{(t)}} \Big[\log\{p_Z(Z_i;\theta)\} | R_i, Z_{(R_i)i} \Big].$$

directly. In particular, recall from (3.47) that this involves calculation of the integrals in

$$\begin{aligned} &\sum_{i=1}^{N} \sum_{r} I(R_i = r) \int \log[p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}]\, p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}\, d\nu(z_{(\bar{r})}) \\ &\quad = \sum_{i=1}^{N} \sum_{r} I(R_i = r) \int \log[p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}]\, p_{Z_{(\bar{r})}|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\}\, d\nu(z_{(\bar{r})}) \\ &\quad = \sum_{i=1}^{N} \sum_{r} I(R_i = r) \int \log[p_Z\{(Z_{(r)i}, z_{(\bar{r})});\theta\}]\, p_{Z|R,Z_{(R)}}\{(Z_{(r)i}, z_{(\bar{r})})|R_i = r, Z_{(r)i};\theta^{(t)}\}\, d\nu(z_{(\bar{r})}). \end{aligned}$$

for each i and r for which $R_i = r$.

One practical approach is to carry out these integrations numerically. The idea is to make a large number S of random draws from the (frequentist) predictive distribution (3.48),

$$p_{Z_{(\bar{r})}|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta^{(t)}\},$$

for each individual i and value r where $R_i = r$, $z^{(1)}_{(\bar{r})i}, \ldots, z^{(S)}_{(\bar{r})i}$, say, and approximate the integrals for each i by

$$S^{-1} \sum_{s=1}^{S} \log[p_Z\{(Z_{(r)i}, z^{(s)}_{(\bar{r})i});\theta\}]. \tag{3.61}$$

Then $Q(\theta|\theta^{(t)})$ can be approximated numerically and the numerical approximation maximized at the M-step.

This general approach is referred to as a ***Monte Carlo (MCEM) EM algorithm***, as the E-step is carried out by Monte Carlo integration as in (3.61), and was first proposed by Wei and Tanner (1990). Of course, a requirement is a way to generate random draws from the predictive distribution.

Given that the EM algorithm is already ***slow*** to converge, incorporating potentially computationally intensive Monte Carlo integration into each E-step can make the algorithm even more unwieldy. Moreover, because ***Monte Carlo error*** is introduced into the E-step, it is no longer guaranteed that the likelihood need increase at every iteration.

These features make the choice of S and monitoring convergence critical. Wei and Tanner (1990) recommend using small values of S in the initial stages and be increasing S as the algorithm moves closer to convergence. They also recommend monitoring convergence by plotting successive iterates $\theta^{(t)}$ against t and, when the values appear to stabilize, the process may be terminated or continued with a larger value of S.

Variations on this theme have been proposed.

STANDARD ERRORS: As noted above, the observed information matrix (3.37) is not readily available as a by-product of the EM algorithm. Accordingly, standard errors for $\widehat{\theta}$ must be calculated explicitly. Louis (1982) describes how to compute the observed information matrix, as we now demonstrate.

From (3.37), the observed information matrix is

$$I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \theta) = -\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta\partial\theta^T} \left[\sum_r I(R_i = r) \log\{p_{Z_{(r)}}(Z_{(r)i}; \theta)\} \right],$$

which, from (3.31), can be written equivalently as

$$-\sum_{i=1}^{N} \frac{\partial}{\partial\theta^T} \{S_\theta(R_i, Z_{(R_i)i}; \theta)\}. \tag{3.62}$$

Recall from (3.2) that

$$S_\theta^F(Z; \theta) = \frac{\partial}{\partial\theta} \log\{p_Z(Z; \theta)\},$$

and define similarly

$$B_\theta^F(Z; \theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log\{p_Z(Z; \theta)\}.$$

From (3.34), we have

$$S_\theta(R, Z_{(R)}; \theta) = \sum_r I(R = r) \, \frac{\displaystyle\int S_\theta^F\{(Z_{(r)}, z_{(\bar{r})}); \theta\} p_Z\{(Z_{(r)}, z_{(\bar{r})}); \theta\} \, d\nu(z_{(\bar{r})})}{\displaystyle\int p_Z\{(Z_{(r)}, z_{(\bar{r})}); \theta\} \, d\nu(z_{(\bar{r})})}. \tag{3.63}$$

Thus, to find (3.62), we must take the partial derivative of (3.63) with respect to θ.

The derivative of the quotient in the summand in (3.63) for a fixed r may be found by using the "quotient rule." We obtain

$$
\begin{aligned}
&\frac{\partial}{\partial\theta^T}\{S_\theta(R,Z_{(R)};\theta)\}\\
&= \sum_r I(R=r)\left(\frac{-\int B_\theta^F\{(Z_{(r)},z_{(\bar r)});\theta\}p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}{\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}\right.\\
&+\frac{\int S_\theta^F\{(Z_{(r)},z_{(\bar r)});\theta\}S_\theta^F\{(Z_{(r)},z_{(\bar r)});\theta\}^T p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}{\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}\\
&-\left.\left[\frac{\int S_\theta^F\{(Z_{(r)},z_{(\bar r)});\theta\}p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}{\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}\right]\left[\frac{\int S_\theta^F\{(Z_{(r)},z_{(\bar r)});\theta\}^T p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}{\int p_Z\{(Z_{(r)},z_{(\bar r)});\theta\}\,d\nu(z_{(\bar r)})}\right]\right)\\
&= \sum_r I(R=r)\Big[-E_\theta\{B_\theta^F(Z;\theta)|Z_{(r)}\}+E_\theta\{S_\theta^F(Z;\theta)S_\theta^F(Z;\theta)^T|Z_{(r)}\}\\
&\qquad -E_\theta\{S_\theta^F(Z;\theta)|Z_{(r)}\}E_\theta\{S_\theta^F(Z;\theta)^T|Z_{(r)}\}\Big].
\end{aligned}
$$

Thus, the observed information matrix can be written as

$$
\begin{aligned}
I(\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})};\theta) &= -\sum_{i=1}^N\sum_r I(R_i=r)\Big[-E_\theta\{B_\theta^F(Z_i;\theta)|Z_{(r)i}\}+E_\theta\{S_\theta^F(Z_i;\theta)S_\theta^F(Z_i;\theta)^T|Z_{(r)i}\}\\
&\qquad -E_\theta\{S_\theta^F(Z_i;\theta)|Z_{(r)i}\}E_\theta\{S_\theta^F(Z_i;\theta)^T|Z_{(r)i}\}\Big]. \qquad (3.64)
\end{aligned}
$$

REMARKS:

- Although (3.64) does provide an expression for the observed information matrix, it still involves derivation of some rather complicated quantities. Consequently, although it is discussed in books on missing data methods, it is not widely used nowadays in practice.

- A host of other methods has been proposed. For example, Meng and Rubin (1991) propose what they call the "supplemental" EM (SEM) algorithm, which allows standard errors to be calculated given code for computing the full data information matrix.

- With the advent of modern computing power, use of the ***bootstrap*** has been proposed. Here, one constructs the bth bootstrap data set by sampling N individuals with replacement from the original data set then runs the EM algorithm on the sampled data set to obtain an estimate $\widehat{\theta}^b$, say. This is repeated B times, and the covariance matrix is approximated by the sample covariance matrix of $\widehat{\theta}^b$, $b = 1, \ldots, B$. Taking B to be 100 to 250 is generally sufficient. Obviously, this is quite computationally intensive and is only feasible when the EM algorithm runs fairly quickly, but if it is, it is straightforward.

MISSING INFORMATION PRINCIPLE: The result (3.64) is often interpreted as demonstrating a concept referred to as the ***missing information principle***. The first term on the right hand side of (3.64) is

$$-\sum_{i=1}^{N}\sum_{r} I(R_i = r)\ E_\theta\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log\{p_Z(Z_i;\theta)\}|Z_{(r)i}\right] = -\sum_{i=1}^{N} E_\theta\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log\{p_Z(Z_i;\theta)\}|R_i, Z_{(R_i)i}\right]. \tag{3.65}$$

This can be interpreted as the expectation of the full data observed information given the observed data, which we denote as $I^F(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta)$. It is straightforward to see that

$$E_\theta\{I^F(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta)\} = N\mathcal{I}^F(\theta).$$

The second and third terms on the right hand side of (3.64), taken together, can likewise be seen to be

$$I^M(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta) = \sum_{i=1}^{N} \text{var}_\theta\{S_\theta^F(Z_i;\theta)|R_i, Z_{(R_i)i}\}, \tag{3.66}$$

where for each i the summand is the covariance matrix of the full data score vector $S_\theta^F(Z_i;\theta)$ given the observed data on i. Defining $\mathcal{I}^M(\theta) = E_\theta[\text{var}_\theta\{S_\theta^F(Z;\theta)|R, Z_{(R)}\}]$, note that

$$E_\theta\{I^M(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta)\} = N\mathcal{I}^M(\theta).$$

Combining (3.64)-(3.66), we have

$$I^F(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta) = I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta) + I^M(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta) \tag{3.67}$$

The interpretation is that the expectation of the full data observed information, conditional on the observed data, is equal to the observed information plus an additional term representing how they differ. Consequently, from the point of view of (3.67), $I^M(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta)$ in (3.66) has been interpreted as the "missing information" or "lost information" resulting from not observing the full data. Thus, (3.67) has been referred to as the missing information principle.

Taking the expectation of both sides of (3.67) and recalling that $N^{-1} E_\theta\{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})};\theta)\} = \mathcal{I}(\theta)$, yields the unconditional statement

$$\mathcal{I}^F(\theta) = \mathcal{I}(\theta) + \mathcal{I}^M(\theta). \tag{3.68}$$

The expression (3.68) has also been referred to as the missing information principle; in fact, this is the result that was derived in the original paper by Orchard and Woodbury (1972) in which this concept was first described. They did this directly by noting that, from (3.4), $\mathcal{I}^F(\theta) = \text{var}_\theta\{S^F_\theta(Z;\theta)\}$, and then writing this as the sum of two terms using the ***law of total variance*** (also known as ***Eve's law*** or the conditional variance formula). This derivation is left as an exercise for the diligent student.

3.5 Inference in practice

The foregoing results demonstrate that, for a likelihood-based analysis under the assumption of MAR, implemented via direct maximization of the observed data likelihood or using the EM algorithm, achieving valid assessments of uncertainty (e.g., standard errors) for estimators for components of θ of interest in the full data model can be ***challenging***. As noted in the last section, one approach in either case is to use a ***nonparametric bootstrap*** however, this can be computationally intensive.

APPROXIMATE PRACTICAL APPROACH: An alternative, approximate approach is common in practice in the context of ***longitudinal data analysis*** via popular models; e.g., for continuous longitudinal outcomes, the ***linear mixed effects model***. Here, the ***full data model*** boils down to assuming that the intended full longitudinal response vector $Y = (Y_1, \dots, Y_T)^T$ is distributed as ***multivariate normal*** (possibly conditional on covariates), with posited structures for the population mean vector and covariance matrix of Y depending on elements of the parameter θ.

Under the conditions that

(i) the assumptions of multivariate normality and the forms of the population mean and covariance matrix are correct, so that the full data model is ***correctly specified***, with true value θ_0 of θ, and

- (ii) the assumptions of separability and MAR are also ***correct***,

then, analogous to the demonstration in ***EXAMPLE 2*** in the simplest case bivariate Y (so $T = 2$), as in (3.28), the observed data likelihood (3.21) is ***also*** that of a multivariate normal.

- Specifically, the observed data likelihood (3.21),
$$\prod_{i=1}^{N} \prod_{r} p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)}, \tag{3.69}$$
is exactly that one would obtain by treating the observed data as if they were the ***intended*** data; that is, if the numbers of components in Y_i that are observed under the MAR mechanism for each individual i had been ***fixed in advanced***.

- Accordingly, under MAR, the estimator obtained by maximizing the "usual" likelihood treating the lengths of the response vectors Y_i as if they were fixed in advance is the same as that maximizing the observed data likelihood, $\widehat{\theta}$, and thus, by standard likelihood theory, should be a ***consistent estimator*** for θ_0.

- In fact, a little thought reveals that, in general, ***likelihood ratio test statistics*** for comparing, for example, ***nested models*** for the full data where the null hypothesis sets some components of θ to zero, based on observed data likelihoods (3.69) for competing "full" and "reduced" models will yield ***valid inferences***. This is because in the "actual" observed data likelihood based on (3.20), namely,
$$\prod_{i=1}^{N} \prod_{r} p_{R|Z_{(r)}}(r|Z_{(r)i};\psi)^{I(R_i=r)} \prod_{r} p_{Z_{(r)}}(Z_{(r)i};\theta)^{I(R_i=r)}$$
including a model for the missingness mechanism, the missingness model will be estimated identically under the "full" and "reduced" full data models. This model thus appears in both the numerator and denominator of the likelihood ratio for each i and cancels, leaving the likelihood ratio statistic to depend only on (3.69) under the "full" and "reduced" models.

- Although the developments in the last section suggest otherwise, one would thus also hope that standard errors for the components of $\widehat{\theta}$ can be obtained from the expected information matrix for the "usual" likelihood treating the lengths of the Y_i as fixed in advance. As we now demonstrate in a very simple example, this is ***not*** the case.

EXAMPLE 4: Consider again the bivariate full data model in ***EXAMPLE 2*** with $Z = (Y_1, Y_2)$ and
$$Y = (Y_1, Y_2)^T \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_1, \mu_2)^T, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}, \tag{3.70}$$
$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_{12}, \sigma_2^2)^T$. Suppose we are interested in inference on $\mu = (\mu_1, \mu_2)^T$. Under the assumption that this model is correctly specified, there are true values μ_0 and Σ_0 that correspond to the true bivariate normal distribution generating the data.

Suppose that Y_1 is ***always observed*** and only Y_2 is possibly missing. Thus, R can take on two possible values, $(1,1)$, and $(1,0)$. Let $C = 1$ if $R = (1,1)$ and $C = 0$ if $R = (1,0)$. The observed data likelihood is then

$$\left\{ \prod_{i=1}^{N} \{p_Y(Y_i;\theta)\}^{I\{C_i=1\}} \right\} \left\{ \prod_{i=1}^{N} \{p_{Y_1}(Y_{i1};\theta_1)\}^{I\{C_i=0\}} \right\} \tag{3.71}$$

where $\theta_1 = (\mu_1, \sigma_1^2)^T$, $p_Y(y;\theta)$ is the $\mathcal{N}(\mu,\Sigma)$ density, and $p_{Y_1}(y_1;\theta_1)$ is the $\mathcal{N}(\mu_1,\sigma_1^2)$ density.

As above, the observed data likelihood (3.71) is exactly that obtained by treating the observed part of Y_i for each i as if it were ***planned in advance***. Thus, if for individual i only Y_{i1} is observed, i's contribution to (3.71) is the same as it would be if it were planned in advance to collect only Y_1 on this individual.

The loglikelihood corresponding to (3.71) is

$$\ell = (1/2) \sum_{i=1}^{N} \Big[I(C_i = 1)\{-\log|\Sigma| - (Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)\} + I(C_i = 0)\{-\log\sigma_1^2 - (Y_{i1} - \mu_1)^2/\sigma_1^2\} \Big]. \tag{3.72}$$

Write the estimators maximizing (3.72) as $\widehat{\mu}$ and $\widehat{\Sigma}$.

It is straightforward (and left as an exercise for the diligent student) to show that the upper left (2×2) block of the ***observed information matrix*** (3.37) is

$$-\frac{\partial^2 \ell}{\partial\mu\partial\mu^T} = \sum_{i=1}^{N} \left\{ I(C_i = 1)\Sigma^{-1} + I(C_i = 0) \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 0 \end{pmatrix} \right\}, \tag{3.73}$$

Note that (3.73) depends only on the C_i and not the Y_i. Let $\pi = \text{pr}(C = 1)$. It is then straightforward that the upper left (2×2) block of the ***expected information matrix*** $\mathcal{I}(\theta_0)$ is

$$\mathcal{I}_{\mu\mu}(\theta_0) = \pi\Sigma_0^{-1} + (1-\pi) \begin{pmatrix} 1/\sigma_{1,0}^2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.74}$$

which can be estimated in practice by

$$N^{-1}\widehat{I}_{\mu\mu} = (N_1/N)\widehat{\Sigma}^{-1} + \{(N - N_1)/N\} \begin{pmatrix} 1/\widehat{\sigma}_1^2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.75}$$

where $N_1 = \sum_{i=1}^{N} I(C_i = 1)$, and N_1/N is an estimator for π.

Defining

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_{12} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

the upper right (2 × 3) block has columns

$$-\frac{\partial^2 \ell}{\partial \mu \partial \sigma_1^2} = \sum_{i=1}^N \left\{ I(C_i = 1)\Sigma^{-1} E_{11} \Sigma^{-1} \begin{pmatrix} Y_{i1} - \mu_1 \\ Y_{i2} - \mu_2 \end{pmatrix} + I(C_i = 0)(1/\sigma_1^4) \begin{pmatrix} Y_{i1} - \mu_1 \\ 0 \end{pmatrix} \right\} \tag{3.76}$$

$$-\frac{\partial^2 \ell}{\partial \mu \partial \sigma_{12}} = \sum_{i=1}^N \left\{ I(C_i = 1)\Sigma^{-1} E_{12} \Sigma^{-1} \begin{pmatrix} Y_{i1} - \mu_1 \\ Y_{i2} - \mu_2 \end{pmatrix} \right\} \tag{3.77}$$

$$-\frac{\partial^2 \ell}{\partial \mu \partial \sigma_2^2} = \sum_{i=1}^N \left\{ I(C_i = 1)\Sigma^{-1} E_{22} \Sigma^{-1} \begin{pmatrix} Y_{i1} - \mu_1 \\ Y_{i2} - \mu_2 \end{pmatrix}, \right\}. \tag{3.78}$$

and of course the lower left (3 × 2) block has rows that are the transposes of (3.76)-(3.78).

First, suppose that in fact the missingness mechanism is ***MCAR***. Under MCAR, missingness is completely unrelated to Y, so that whether or not Y_2 is missing is effectively ***the same*** as if it were planned in advance to collect Y_2 or not. Under this condition, it is straightforward to demonstrate via a conditioning argument (try it) that the expectations of (3.76)-(3.78) are equal to zero. Thus, the expected information matrix $\mathcal{I}(\theta_0)$ is ***block diagonal***, and it follows that the estimator $\widehat{\mu}$ maximizing (3.72)

$$N^{1/2}(\widehat{\mu} - \mu_0) \xrightarrow{\mathcal{L}} \mathcal{N}[0, \{\mathcal{I}_{\mu\mu}(\theta_0)\}^{-1}], \tag{3.79}$$

and thus an approximate sampling distribution for $\widehat{\mu}$ is

$$\widehat{\mu} \mathrel{\dot\sim} \mathcal{N}\{\mu_0, (\widehat{I}_{\mu\mu})^{-1}\}. \tag{3.80}$$

The results (3.79)-(3.80) are exactly those that arise if the lengths of the vectors Y_i were fixed in advance.

Now suppose that the missingness mechanism is ***MAR***. The form of $\mathcal{I}_{\mu\mu}(\theta_0)$ in (3.74) and thus the estimator $N^{-1}\widehat{I}_{\mu\mu}$ in (3.75) are unchanged. ***However*** , it can be shown by a conditioning argument, left as an exercise for the inquisitive student, that the expectations of (3.76)-(3.78) are ***no longer*** equal to zero. Accordingly, $\mathcal{I}(\theta_0)$ is ***not*** block diagonal, so that the results (3.79)-(3.80) yielding an approximate sampling distribution for $\widehat{\mu}$ are ***no longer valid.*** To obtain the correct approximate sampling distribution, one must derive the upper left (2 × 2) block of the (5 × 5) matrix $\{\mathcal{I}(\theta_0)\}^{-1}$, which can be estimated by the upper left (2 × 2) block of an estimator for this full matrix.

Although we showed this in this very simple example, the implications hold for more general full data models, such as longitudinal data models as above, as follows:

- In these more complex full data models, standard software implementing maximum likelihood estimation based on the observed data likelihood by default treats the lengths of the Y_i as if they were fixed in advance, and accordingly reports standard errors and confidence intervals for parameters characterizing the mean of Y (possibly conditional on covariates) based on the analogous results to (3.79)-(3.80).
- Thus, if the true missingness mechanism is MAR, the above argument demonstrates that inferences based on the default output will be ***flawed***, as correct assessment of uncertainty in this case should be based on the sampling distribution that acknowledges the more complex form of $\{\mathcal{I}(\theta_0)\}^{-1}$.
- It is thus recommended in practice that, under the assumption of MAR and correct full data model, standard errors should be based instead on the inverse of the ***observed information matrix*** , as this will faithfully reflect the true covariance matrix of the sampling distribution. ***Unfortunately***, many software programs ***do not*** allow this option or do not have an option to output this information matrix.

It has become commonplace in practice (which does not make it correct) to ***ignore*** this issue and to use the ***usual*** approximate sampling distribution analogous to (3.80) for inference as if it were valid. This has the potential to lead to misleading inferences, of course. There have been some empirical studies that suggest that this practice may not be ***too terrible*** in some settings; however, it is critical that the data analyst appreciate this issue.

3.6 Bayesian inference

So far in this course, we have taken a frequentist perspective on inference in the presence of missing data; i.e., we have considered likelihood-based inference and large sample theory approximation to properties of estimators for parameters of interest. It is of course possible to take a Bayesian point of view.

In the Bayesian paradigm, parameters like θ are treated as random quantities, and the basis for inference on θ is the ***posterior distribution*** induced by an overall model consisting of the likelihood for the observed data and a prior distribution for θ. In our context, under MAR, as we have seen, the likelihood for the observed data $(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})$ is

$$p_{\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})}}(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \theta, \psi) \propto \prod_{i=1}^{N} \prod_{r} p_{Z_{(r)}}(Z_{(r)i}; \theta)^{I(R_i=r)}, \tag{3.81}$$

where we ignore the missing data mechanism. Suppose a prior distribution with density $p_\theta(\theta)$ is assumed for θ. Then, using Bayes rule and (3.81), the posterior density for θ given the observed data $(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})$ is given by

$$p_{\theta|\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})}}(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) = \frac{\prod_{i=1}^{N} \prod_{r} p_{Z_{(r)}}(Z_{(r)i}; \theta)^{I(R_i=r)} p_\theta(\theta)}{\int \prod_{i=1}^{N} \prod_{r} p_{Z_{(r)}}(Z_{(r)i}; \theta)^{I(R_i=r)} p_\theta(\theta)\, d\theta}, \tag{3.82}$$

where the density for the missingness mechanism cancels out from the numerator and denominator in (3.82).

If it were possible to obtain an expression for the posterior density in (3.82), then estimators for θ can be obtained as the mean, median, or mode posterior distribution, and Bayesian interval estimates corresponding to these can be obtained from the appropriate quantiles of this distribution. The challenge for Bayesian inference is that, except in simple settings, deriving an analytical expression for the posterior density is not possible owing to the need to carry out what is intractable integration in the denominator of (3.81).

Luckily, computational advances have made it possible to obtain the posterior distribution through simulation. ***Markov chain Monte Carlo*** (***MCMC***) techniques can be used to essentially carry out the integration numerically and result in a sample from the posterior distribution that can then be used to approximate the posterior distribution. The mean, median, or mode of the sample can be used as an estimator for θ, and in fact any functional of the posterior distribution be approximated similarly.

Accordingly, another approach to inference on θ would be to adopt a Bayesian framework as above. Software such as `WinBUGS` or `OpenBUGS` can then be used for implementation. In Chapter 4, we demonstrate in detail such a MCMC formulation.

In the setting of a fully parametric model for the observed data, choosing so-called weak or noninformative priors can yield inference that has good frequentist properties; e.g., the posterior distribution and the asymptotic distribution of the MLE are similar.