Chapter 3

Estimation in the LMCD Assuming Normally Distributed Errors with Unbalanced Designs/Missing Data

In Chapter 2 we considered only designs with equal numbers of observations, n, on each subject. It often happens that $n_i \neq n$ and hence $\text{var}\,(Y_i) = \sum_{i\,(n_i \times n_i)}$; this can arise in different settings. We can have studies with unequal n_i by design: e.g., clustering, sampling households or litters, or family studies. In this case the Y_i are complete responses on the sampling unit, and the LMCD assumes

$$\underset{n_i \times 1}{E\,(Y_i)} = \underset{n_i \times pp \times 1}{X_i \quad \beta} \tag{3.1}$$

and

$$\text{var}\,(Y_i) = \underset{n_i \times n_i}{\Sigma_i}\,, \tag{3.2}$$

for appropriate choices of X_i, β and Σ_i. Here Σ_i depends on i through its dimension, and possibly also X_i, but we will assume a common parameter set for the Σ_i, so that we may write $\Sigma_i(\theta)$, where θ contains all of the variance covariance parameters. For example, if each observation has the same variance σ^2 and any two pairs of Y_{ij}, Y_{ik} have the same covariance, then each Σ_i has the compound symmetry form with different dimension n_i.

Alternately, it may happen that the design calls for n measures per subject, with

$$\underset{n\times 1}{E\,(Y_i)} = \underset{n\times p}{X_i}\ \underset{p\times 1}{\beta}$$

and

$$\text{var}\,(Y_i) = \underset{n\times n}{\Sigma_i}\,,$$

but some observations are missing, so that the vector on n_i observations that are the observed data can be expressed as

$$\underset{n_i\times n}{Y_i^{\text{OBS}}} = \underset{n_i\times n}{I_i}\ \underset{n\times 1}{Y_i}\,.$$

where I_i is obtained from an $n \times n$ identity matrix by removing the rows corresponding to the missing observations. Here again the dimension of each Σ_i will depend upon i; as before we assume a common parameter vector θ which consists of the unique elements of Σ. In general,

$$E\left(Y_i^{\text{OBS}}\right) = I_i\ E(Y_i) = I_i\ X_i\ \beta = X_i^{\text{OBS}}\ \beta \tag{3.3}$$

and

$$\text{var}\left(Y_i^{\text{OBS}}\right) = I_i\ \Sigma_i\ I_i^T = \Sigma_i^{\text{OBS}} \tag{3.4}$$

hold only if the MDM is MCAR. It is important to note that the β and θ are the same for the unequal n_i and equal n case. From a technical point of view, both of the cases (unbalanced by design or because of MCAR data) can be handled in the same way (although the structures for Σ_i will generally be different), but the validity of the estimators will depend upon the design or the MDM. Henceforth we drop the superscript OBS and all Y_i are $n_i \times 1$ and X_i are $n_i \times p$, unless otherwise noted.

3.1 ML and REML Estimation for the Unequal n_i Case

Here we will consider ML and REML estimation based on the multivariate normal distribution for unbalanced data. As discussed above, this can be appropriate for the clustered data setting, or it can arise in the case of missing data. In this chapter we deal with likelihood inference for MAR and/or MCAR MDM's, hence we can assume that each observed Y_i is $N(X_i\beta, \Sigma_i)$. As in the complete data case, we will mainly focus on unstructured Σ in this chapter, i.e., the missing data case where the covariance matrix for each complete data vector is Σ. Chapter 5 takes up a random effects structure which allows Σ to depend on X_i.

Letting θ denote the parameters of Σ_i, we then have that the likelihood based on the observed data is

$$\mathcal{L}(\beta,\theta)=\prod_{i=1}^{N}|\Sigma_i|^{-1/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}(Y_i-X_i\beta)^T\,\Sigma_i^{-1}\,(Y_i-X_i\beta)\right].$$

Again, it is easily seen that for any fixed θ (and Σ_i), the ML estimate of β is given by

$$\widehat{\beta}(\theta)=\left(\sum_{i=1}^{N}X_i^T\Sigma_i^{-1}(\theta)X_i\right)^{-1}\sum_{i=1}^{N}X_i^T\Sigma_i^{-1}(\theta)Y_i. \tag{3.5}$$

Thus, as in the complete data case, the ML estimate of β has a simple form as a function of θ.

Substituting $\widehat{\beta}(\theta)$ for β in the likelihood gives the profile likelihood

$$\mathcal{L}(\widehat{\beta},\theta)=\prod_{i=1}^{N}|\Sigma_i|^{-1/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}(Y_i-X_i\widehat{\beta}(\theta))^T\Sigma_i^{-1}(Y_i-X_i\widehat{\beta}(\theta))\right].$$

Even when $\Sigma_i = \underset{n_i\times n}{I_i}\ \underset{n\times n}{\Sigma}\ \underset{n\times n_i}{I_i^T}$ for Σ unstructured, the derivation of the score equation is complex and there are not simple closed form solutions to the score equations for Σ. A set of simple iterative computing equations can be derived using the EM algorithm and we can also use the "EM-method" to derive the score equations for Σ. This will be described in Section 3.3; for now, we give the following expression for $\widehat{\Sigma}_{\mathrm{ML}}$ in the case of missing data and $\Sigma_i = I_i\,\Sigma\,I_i^T$:

$$\widehat{\Sigma}_{\mathrm{ML}}=\sum_{i=1}^{N}\widehat{Q}_i/N \tag{3.6}$$

where

$$Q_i=\Sigma-\Sigma\,I_i^T\,\Sigma_i^{-1}\,I_i\,\Sigma+\sum_{i=1}^{N}\Sigma I_i^T\,\Sigma_i^{-1}\,(Y_i-X_i\beta)\,(Y_i-X_i\beta)^T\,\Sigma_i^{-1}\,I_i^T\,\Sigma,$$

$$\beta=\left(\sum_{i=1}^{N}X_i^T\,\Sigma_i^{-1}\,X_i\right)^{-1}\sum_{i=1}^{N}X_i^T\,\Sigma_i^{-1}\,Y_i,$$

$$\Sigma_i=I_i\,\Sigma\,I_i^T,$$

and $\widehat{Q}_i$ is Q_i with $\widehat{\Sigma}$, $\widehat{\Sigma}_i$ and $\widehat{\beta}(\widehat{\Sigma}_i^{-1})$ substituted for Σ, Σ_i and β.

The derivation of the REML estimator of θ proceeds as in the complete data case, by constructing the profile likelihood, with an additional term in the denominator for $\text{var}(\widehat{\beta})$:

$$\mathcal{L}_k(\theta) = \prod_{i=1}^{N} |\Sigma_i|^{-1/2} \left| \sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i \right|^{-1/2} \times \exp\left[-\frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i\widehat{\beta})^T \Sigma_i^{-1} (Y_i - X_i\widehat{\beta}) \right],$$

As before, with unstructured Σ we can use the EM approach explained in (3.3) to derive a solution to the likelihood equations

$$\widehat{\Sigma}_{\text{REML}} = \sum_{i=1}^{N} (\widehat{Q}_i + \widehat{T}_i)/N \tag{3.7}$$

where Q_i is as before,

$$T_i = \Sigma\, I_i^T\, \Sigma_i^{-1}\, X_i \left(\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i \right)^{-1} X_i^T\, \Sigma_i^{-1}\, I_i\, \Sigma,$$

and $\widehat{Q}_i$ and $\widehat{T}_i$ are Q_i and T_i evaluated at $\widehat{\Sigma}_{\text{REML}}$. For both the ML and REML estimators, simple iterative computing algorithms will be derived in Section 3.6.

Notice that setting $I_i = I$ and $\Sigma_i = \Sigma$ for all i (no missing data) gives the corresponding likelihood equations for the complete data case with unstructured Σ. From (3.6)–(3.7) it is clear that $\widehat{\Sigma}_{\text{ML}}$ and $\widehat{\Sigma}_{\text{REML}}$, will be consistent for Σ even if normality does not hold, provided the data are MCAR.

Under an MAR assumption on MDM, equations (3.1)–(3.2) do not hold, i.e.,

$$\underset{n_i \times 1}{E\,(Y_i)} \neq \underset{n_i \times p}{X_i}\ \underset{p \times 1}{\beta}$$

and

$$\text{var}\,(Y_i) \neq \Sigma_i = I_i \Sigma I_i^T,$$

where β and Σ determine the moments of the complete data vectors. However, with data MAR, we saw in Section 1.4 that the appropriate likelihood for (β, θ) is proportional to the likelihood based on the ordinary marginal of the observed data, i.e., the likelihood obtained by assuming $Y_i \sim N(X_i\beta, \Sigma_i)$. Thus the likelihood estimators for (β, θ) do not differ

for the MAR and MCAR case even though their properties differ under the different missingness assumptions. So, for $\widehat{\beta}_{\mathrm{ML}}$ defined as

$$\widehat{\beta}_{\mathrm{ML}} = \left(\sum_{i=1}^{N} X_i^T \widehat{\Sigma}_i^{-1} X_i\right)^{-1} \sum_{i=1}^{N} X_i^T \widehat{\Sigma}_i^{-1} Y_i,$$

where $\widehat{\Sigma}_i$ is also based on $\widehat{\theta}_{\mathrm{ML}}$, we have that $\widehat{\beta}_{\mathrm{ML}}$ is a consistent estimate of β and $\widehat{\Sigma}_{\mathrm{ML}}$ and $\widehat{\Sigma}_{\mathrm{REML}}$ are both consistent for Σ in the case of missing data if (a), (b) and (c) or (d) hold:

1. (a) $E\underset{n\times 1}{(Y_i)} = \underset{n\times p}{X_i}\ \underset{p\times 1}{\beta}$ holds for the complete case.

   (b) $\underset{n\times n}{\Sigma} = \mathrm{var}\underset{n\times 1}{(Y_i)}$ holds for the complete case.

   (c) Error terms are normal and the missingness is MAR.

   or

   (d) The missingness is MCAR.

   Note also that:

2. Under (a)–(c), the observed Fisher Information Matrix, but not the expected information, gives a valid estimate of var $(\widehat{\beta}_{\mathrm{ML}})$. So Avar $(\widehat{\beta}_{\mathrm{ML}}) \neq (\sum_{i=1}^{N} X_i^T \widehat{\Sigma}_i^{-1} X_i)^{-1}$ except with MCAR. In practice, however, there is often little difference between the estimated observed and expected information.

3. Under MCAR, $\widehat{\beta}_{\mathrm{ML}}$ is unbiased for β (Karkar and Harville, 1988).

3.2 A General Formulation for Incomplete Data

As mentioned in Section 3.1, using straightforward matrix differentiation techniques to obtain the derivatives of the profile likelihood for θ is tedious, and does not lead to any particularly tractable set of equations. An alternative method for deriving likelihood equations is to rely on a general theory for maximum likelihood in the presence of incomplete data, which includes missingness as well as imbalance by design. The general theory is also the motivation for the EM algorithm.

The EM algorithm is a very general algorithm for computing ML and REML estimates with incomplete data (Dempster, Laird and Rubin, 1977). Indeed, by appropriately defining the complete data, the algorithm can also be applied to solve the ML and REML score equations in a broad class of LMCD models with normal errors (like the random

effects model discussed in the Chapter 5), *even in the absence of missing data.* In this section, we introduce a general notation for describing incomplete data.

Let $\mathcal{Y}$ be an observed data vector which can be represented as an incomplete version of some complete data vector $\mathcal{Z}$ with density $f_{\mathcal{Z}}(z;\Phi)$: $\mathcal{Y}$ may be simply missing components of $\mathcal{Z}$, or it may be a convolution of components of $\mathcal{Z}$. By definition, the density of $\mathcal{Y}$, $f_{\mathcal{Y}}(y;\Phi)$, satisfies

$$f_{\mathcal{Y}}(y;\Phi) = \int f_{\mathcal{Z}}(z;\Phi)\, dz_y, \tag{3.8}$$

where dz_y denotes integration over the missing (or incomplete) data.

We will illustrate the algorithm with four examples. All of the examples are special cases of the LMCD where $\Phi = (\beta, \Sigma)$. In all of these cases, $\widehat{\beta}_{\mathrm{ML}}$ based on the observed data is given by (3.6) for a given Σ. The difficulty is in deriving score equations for Σ and computing $\widehat{\Sigma}_{\mathrm{ML}}$. It helps to keep this in mind when reviewing the examples, since the incomplete data problem is formulated to make ML estimation of Σ easy. The goal is to compute the maximum likelihood estimator of some or all elements of Φ. The observed data $\mathcal{Y}$ may be a pre-specified subset of $\mathcal{Z}$ or it may be a random subset.

Example 1. Suppose each $Z_i \sim N_n(\mu, \Sigma)$, $i = 1, \ldots, N$, and some subjects are missing some of the Z_{ij}'s. Let Y_i denote the $(n_i \times 1)$ subset of the Z_{ij}'s which are observed. Then $\mathcal{Y}^T = (Y_1^T, \ldots, Y_N^T)$ is a $n_+ \times 1$ vector, where n_+ is the summation of the n_i, and we will let the complete data $\mathcal{Z}$ be the $nN \times 1$ vector defined as $\mathcal{Z}^T = (Z_1^T, \ldots, Z_N^T)$. Note this is a special case of the LMCD when $X_i = I$ and $\mu = \beta$.

Example 2. Let $Z_i \sim N_n(X_i\beta, \Sigma)$ where β and Σ are unknown and Σ is an arbitrary symmetrc positive definite matrix. Take $\mathcal{Y}^T = (Z_1^T, \ldots, Z_N^T)$ and $\mathcal{X}^T = (Z_1^T, e_1^T, \ldots, Z_N^T, e_N^T)$. Here the outcomes Z_{ij} are fully observed. However, we regard the unobserved residuals, $e_i = (Y_i - X_i\beta)$, as additional, unobserved data. Example 2 seems a bit contrived, but provides us with a simple way of deriving the likelihood equations for (β, Σ) as given in Chapter 3, which extends easily when some outcomes Z_{ij} are missing, as indicated in the next example.

Example 3. As in Example 2, let $Z_{i\,(n\times 1)} \sim N_n(X_i\beta, \Sigma)$ but now suppose that some Z_{ij} are missing, so only $Y_{i\,(n_i\times 1)}$ is observed. Then take $\mathcal{Y}^T = (Y_1, \ldots, Y_N)^T$ and $\mathcal{Z}^T = (Y_1^T, e_1^T, \ldots, Y_N^T, e_N^T)$ as in example 2. The dimension of $\mathcal{Z}$ is $(nN \times 1)$ and of $\mathcal{Y}$ is $n_+ \times 1$.

Example 4. Suppose $Y_{ij} = X_{ij}^T\beta + b_i + e_{ij}$, $j = 1, \ldots, n_i$, and $i = 1, \ldots, N$ where b_i and $e_{ij}, j = 1, \ldots, n_i$, are all independent normally

distributed with with zero mean, $\text{var}(b_i) = d$ and $\text{var}(e_{ij}) = \sigma^2$. We take $\mathcal{Z}^T = (Y_1^T, b_1, \ldots, Y_N^T, b_N)$ and $\mathcal{Y}^T = (Y_1^T, \ldots, Y_N^T)$. As in Example 2, there are no missing data, and the random effects are considered as missing for convenience. Random effects models will be taken up in detail in Chapter 5.

The first thing to notice is that the definition of the complete data $\mathcal{Z}$ need not be unique. The choice of the observed data $\mathcal{Y}$ is unique because it is actually observed. In general, there may be many choices for $\mathcal{Z}$. We should choose $\mathcal{Z}$ so that computing the MLE of Φ based on $\mathcal{Z}$ is easy. For instance, in example 1 with the complete data $\mathcal{Z} = (Z_1^T, \ldots, Z_N^T)$, the ML estimators of (μ, Σ) have closed form. But in Example 3, where the mean of each Z_i is $X_i\beta$, closed form estimates do not exist except in very special cases. By taking $\mathcal{Z}$ to include the error terms we can get closed form estimates of Σ with complete data. As we will show in Chapter 5, with Example 4, adding only the random effects b_i, means that we can get closed form estimates for (β, d, σ^2) with the complete data.

3.3 Derivatives of the Log-likelihood for the Incomplete Data Model and the EM Algorithm

Derivatives of incomplete data likelihoods can be derived using the following fundamental identity. Suppose that regularity conditions are satisfied so that the order of differentiation and integration can be interchanged in the right hand side of (3.8). Then we have

$$\begin{aligned}
\partial \log f_{\mathcal{Y}}(y;\Phi)/\partial\Phi &= \frac{\int \partial f_{\mathcal{Z}}(z;\Phi)/\partial\Phi \, dz_y}{f_{\mathcal{Y}}(z;\Phi)} \\
&= \frac{\int \partial \log f_{\mathcal{Z}}(z;\Phi)/\partial\Phi \; f_{\mathcal{Z}}(z;\Phi) dz_y}{f_{\mathcal{Y}}(y;\Phi)} \\
&= \int \partial \log f_{\mathcal{Z}}(z;\Phi)/\partial\Phi \; f_{\mathcal{Z}|\mathcal{Y}}(z|y;\Phi) dz_y \\
&= E_{\mathcal{Z}|\mathcal{Y}} \left\{ \partial \log f_{\mathcal{Z}}(z;\Phi)/\partial\Phi | y;\Phi \right\}
\end{aligned}$$

where the subscript $\mathcal{Z}|\mathcal{Y}$ refers to the conditional distribution of $\mathcal{Z}$ given $\mathcal{Y}$. Thus, we see that the score for Φ based on the observed data $\mathcal{Y} = y$ is equal to the expectation of the complete data score for Φ based on

$\mathcal{Z} = z$, conditional on the observed data. This identity provides us with an easy way of deriving the score equations in incomplete data settings.

In the score equations for Φ,

$$E_{\mathcal{Z}|\mathcal{Y}} \{\partial \log f_{\mathcal{Z}}(z; \Phi)/\partial\Phi | y; \Phi\} = 0 \tag{3.9}$$

the parameter Φ enters twice; once as an index of the complete data score $\partial \log f_{\mathcal{Z}}(z; \Phi)/\partial\Phi$ and then as an index of the conditional distribution of $\mathcal{Z}$ given $\mathcal{Y}$. This suggests solving (3.9) by an iterative algorithm that updates the values of Φ separately in its two appearances. Specifically, given Φ_k, stage $k+1$ of the algorithm consists of two steps, namely:

E-step Compute the conditional expectation of the complete data score

$$E \{\partial \log f_{\mathcal{Z}}(z; \Phi)/\partial\Phi | y; \Phi_k\}$$

and

M-step Solve for Φ in the expected complete data score equations

$$E \{\partial \log f_{\mathcal{Z}}(z; \Phi)/\partial\Phi | y; \Phi_k\} = 0$$

and call its solution Φ_{k+1}.

The E-step (E for expectation) essentially imputes the complete data score using its expectation given the observed data. The M-step (M for maximization) solves the complete data likelihood equations using the expected score. One instance in which the algorithm takes a particularly simple form is when the distribution of the complete data has an exponential family distribution, namely

$$f_{\mathcal{Z}}(z; \Phi) = \frac{e^{t(z)^T \theta(\Phi) - b(z)}}{a\{\theta(\Phi)\}},$$

where θ is the vector of natural parameters and $t(x)$ is the corresponding vector of sufficient statistics. Then the complete data score function for Φ is

$$\partial \log f_{\mathcal{Z}}(z; \Phi)/\partial\Phi = \{t(z) - E[t(z); \Phi]\}^T \left\{\frac{\partial\theta(\Phi)}{\partial\Phi}\right\},$$

hence

$$E \{\partial \log f_{\mathcal{Z}}(z; \Phi)/\partial\Phi | y; \Phi\} = \left\{E_{\mathcal{Z}|\mathcal{Y}}[t(z)|y; \Phi] - E[t(z); \Phi]\right\}^T \left\{\frac{\partial\theta(\Phi)}{\partial\Phi}\right\}.$$

Thus, in this case given Φ_k, the $(k+1)$th stage of the EM algorithm is:

E-step Compute $t_{k+1}(y) = E_{\mathcal{Z}|\mathcal{Y}}\{t(z)|y;\Phi_k\}$, and

M-step Solve

$$\{t_{k+1}(y) - E[t(z);\Phi]\}\left\{\frac{\partial\theta(\Phi)}{\partial\Phi}\right\} = 0,$$

or when the transformation from θ to Φ is invertible,

$$t^{(k+1)}(y) = E[t(z);\Phi]. \tag{3.10}$$

Notice that solving for Φ_{k+1} at the M-step is the same as solving the complete data likelihood equations with the sufficient statistics $t(z)$ replaced by its imputation $t_{k+1}(y)$. Also notice that the algorithm assumes the existence of a "complete data vector" $\mathcal{Z}$ but it does not prescribe how to define $\mathcal{Z}$. Ordinarily, one would want to define $\mathcal{X}$ in such a way that the steps of the algorithm can be easily computed. In particular, if $\mathcal{Z}$ has a distribution in an exponential family, one would want to choose $\mathcal{Z}$ so that $E[t(z);\Phi]$ is itself equal to Φ or an easily invertible function of Φ.

We now return to our examples.

Example 1. Deriving the EM equations for Example 1 is very straightforward, and left as an exercise. The M-step has a simple closed form solution for μ_{k+1}, Σ_{k+1} and $E(t(x)|\mathcal{Y},\Phi_k)$ is easily obtained using standard multivariate theory. See, for example, Little and Rubin (1987), Section 8.2.

Example 2. Now we have no missing data (each Z_i is $n \times 1$) but since each subject has a possibly different mean, the ML and REML estimates of Σ do not have closed form. In Section 2.3 we calculated the score equations for β and Σ directly by derivation of the log-likelihood function. Here we show how the "EM Method" can be used to obtain a solution to the score equations avoiding the need of matrix derivations. Our implementation of the EM for this example is unorthodox; because the complete data likelihood is singular (y_i is linear in e_i), we always use the score equations for β obtained from the observed data and use the complete data to derive score equations for Σ. With $(e_1^T,\ldots,e_N^T)$ in the complete data where each e_i is iid $e_i \sim N(0,\Sigma)$, we have that $S = \sum_{i=1}^{N} e_i e_i^T$ is the minimal sufficient statistic for Σ, and the complete data score equations for Σ are

$$N\Sigma - S. \tag{3.11}$$

Further, the score equations for β, based on the observed data, are

$$\sum_{i=1}^{N} X_i^T \Sigma^{-1}(Z_i - X_i\beta).$$

Then we have the following:

E-step Given Σ_k and β_k, set

$$S_k = \sum_{i=1}^{N} E\left(e_i e_i^T | X_i, \Sigma_k, \beta_k\right) = \sum_{i=1}^{N} \left(Z_i - X_i\beta_k\right)\left(Z_i - X_i\beta_k\right)^T,$$

because var $(e_i|Z_i, \Sigma) = 0$ and $E(e_i|Z_i, \Sigma) = (Z_i - X_i\beta)$. Note that the score equation for β does not depend upon the e_i's.

M-step Given S_k, compute Σ_{k+1} as

$$\Sigma_{k+1} = S_k/N = \textstyle\sum_{i=1}^{N}(Z_i - X_i\,\beta_k)\,(Z_i - X_i\,\beta_k)^T/N,$$

and β_{k+1} as

$$\beta_{k+1} = (\Sigma X_i^T \Sigma_{k+1}^{-1} X_i)^{-1} \Sigma X_i^T \Sigma_{k+1}^{-1} Z_i.$$

This is simply iteratively reweighted generalized least squares (IRLS) for the case where we have complete multivariate data. At convergence, $\Sigma_k = \Sigma_{k+1} = \widehat{\Sigma}_{\mathrm{ML}}$ and we obtain the likelihood equations given in Section 3. Because at each iteration, $\widehat{\beta}_{k+1}$ maximizes the observed data likelihood and not the expected complete data likelihood, Liu and Rubin (1994) refer to this as a generalization of the EM.

Example 3. Suppose that in Example 2, some of the Z_{ij}'s are missing. Since $\mathcal{Z}$ is the same, the M-step for Σ remains the same, but now

$$E\left(\underset{n\times 1}{e_i} \mid \underset{n_i\times 1}{Y_i}, \beta, \Sigma\right) = I_i^T\ \Sigma_i^{-1}\ (Y_i - X_i\beta)$$

and

$$\text{var}\ (e_i|Y_i, \beta, \Sigma) = \Sigma - \Sigma\ I_i^T\ \Sigma_i^{-1}\ I_i\ \Sigma,$$

so that

$$E\ (e_i e_i^T|Y_i, \beta, \Sigma) = \Sigma - \Sigma I_i^T I_i \Sigma + \Sigma\ I_i^T \Sigma_i^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^T \Sigma_i^{-1} I_i \Sigma = Q_i.$$

As before, using the observed score equation for β yields

$$\sum_{i=1}^{N} X_i^T \Sigma_i^{-1}(Y_i - X_i\beta). \tag{3.12}$$

Again combining the E- and M-steps, for a given β_k and Σ_k we have

$$\Sigma_{k+1} = \sum_{i=1}^{N} Q_{ik} \Big/ N \tag{3.13}$$

where Q_{ik} is Q_i with (β, Σ) evaluated at (β_k, Σ_k). Solving (3.12) for $\beta_{k+1}, \Sigma_{k+1}$ gives

$$\beta_{k+1} = \left(\sum_{i=1}^{N} X_i^T \Sigma_{ik}^{-1} X_i \right)^{-1} \Sigma X_i^T \Sigma_{ik}^{-1} Y_i, \tag{3.14}$$

where Σ_{ik}^{-1} is $(I_i \Sigma I_i^T)^{-1}$ evaluated at Σ_k. Using (3.13)–(3.14) might be regarded as a variant of multivariate IRLS for missing data. Notice that if we have complete data on all subjects the algorithm reduces to the IRLS algorithm of Example 2. From the definition of Q_i as $E(e_i e_i^T | Y_i, \beta, \Sigma)$ and $S = \sum_{i=1}^{N} e_i e_i^T$, it follows from (3.10) and (3.11) that the score equation for $\widehat{\Sigma}_{\mathrm{ML}}$ is given by (3.6).

The EM approach can also be used to derive the REML likelihood equations and the corresponding iterative algorithms, using the Bayes formulation for EM. Recall that with REML, β and Σ are given flat priors, and the REML likelihood is proportional to the marginal posterior of Σ. Thus using (4.5) where $\Phi = (\beta, \Sigma)$ and $\mathcal{Y}$ is the observed data, the REML likelihood is

$$\mathcal{L}_R(y, \Sigma) \propto \int f_{\mathcal{Y}}(y | \beta, \Sigma) d\beta \propto \int \int f_{\mathcal{Z}}(z | \Sigma, \beta) dz_y d\beta.$$

This implies we can regard β as an additional piece of "missing" data. Hence the sufficient statistic for Σ remains S, but in computing the expectations given $\mathcal{Y}$, we consider the β as missing data. Thus the REML score equations become

$$N\,\Sigma - E(E(S | \mathcal{Y}, \beta, \Sigma))$$

where inner expection is exactly the same as before, and where the outer expectation is with respect to β given $\mathcal{Y} = y$ and Σ.

To evaluate the REML score equations, we use the fact that with linear models, normal likelihoods, and flat priors on β, the posterior of β given Σ and the data is normal, with mean equal to $\widehat{\beta}(\Sigma)$ and variance equal to the variance of $\widehat{\beta}$: $\mathrm{var}(\beta) = (\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i)^{-1}$. Therefore, for Example 3,

$$E(S | \mathcal{Y}, \beta, \Sigma) = \sum_{i=1}^{N} E(Q_i | Y_i, \Sigma)$$

and

$$E(Q_i|Y_i, \Sigma) = Q_i(\widehat{\beta}) + T_i,$$

where Q_i and T_i are defined in Section 4.2 and $Q_i(\widehat{\beta})$ is Q_i with β evaluated at $\widehat{\beta}$. Hence it follows that the REML score equations are

$$\widehat{\Sigma}_{\text{REML}} = \sum_{i=1}^{N} \left(Q_i(\widehat{\beta}) + T_i \right) \Big/ N.$$

Notice that if there is no missing data ($Y_i = Z_i$), $S_i = (Y_i - X_i\,\beta)(Y_i - X_i\,\beta)^T$ and

$$\widehat{\Sigma}_{\text{REML}} = \sum_{i=1}^{N} \Bigg[(Y_i - X_i\,\widehat{\beta})(Y_i - X_i\,\widehat{\beta})^T + X_i \left(\sum_{i=1}^{N} X_i^T\, \widehat{\Sigma}_{\text{REML}}^{-1} X_i \right)^{-1} X_i^T \Bigg] / N$$

where $\widehat{\beta} = \widehat{\beta}(\widehat{\Sigma}_{\text{REML}}^{-1})$. When some Z_{ij} are missing, S_i is replaced by $\mathrm{E}\,(e_i\, e_i^T | Y_i,\, \beta,\, \Sigma) = Q_i$, and we simply add $T_i = X_i(\sum_{i=1}^{N} X_i^T \widehat{\Sigma}_{\text{REML}} X_i)^{-1} X_i^T$ onto Q_i to obtain the corresponding REML estimating equations given above.