Chapter 2

Estimation in the LMCD Assuming Normally Distributed Errors and Complete Data

In this chapter we consider maximum likelihood inference for the parameters of the Linear Model for Correlated Data (LMCD) of Section 1.2 under the additional assumption that the error vectors are normally distributed. To present the basic ideas, we will assume that the data are obtained from a complete study with each $n_i = n$ and we restrict attention to the setting in which Σ_i is assumed not to depend on covariates, so that $\Sigma_i = \Sigma$ is the same for all i. We will consider both maximum likelihood (ML) and restricted maximum likelihood (REML) estimation when Σ is unstructured. Inference under more restrictive parametric models for Σ is briefly discussed at the end of this chapter. Estimation with unbalanced designs and/or missing data will be taken up in Chapter 3.

2.1 ML Estimation of β and Σ

Suppose that in the LMCD we additionally assume that given X_i, Y_i, $i = 1, \ldots, N$, are N copies from a multivariate normal random vector. Then, under this additional assumption, the likelihood of (β, Σ) is given by

$$\mathcal{L}\left(\beta,\, \Sigma\right) = |\Sigma|^{-N/2}\, e^{-\sum_{i=1}^{N}(Y_i - X_i\beta)^T\, \Sigma^{-1}(Y_i - X_i\beta)/2}.$$

We now characterize the solutions to the score equations

$$\frac{\partial}{\partial\beta}\ln\mathcal{L}(\beta,\,\Sigma)=0 \quad \text{and} \quad \frac{\partial}{\partial\Sigma}\ln\mathcal{L}(\beta,\,\Sigma)=0. \tag{2.1}$$

By convention, for any function $f(M)$ of a $p\times q$ matrix M, we let $\partial f(M)/\partial M$ denote the $p\times q$ matrix with (i,j)th element equal to $\partial f(M)/\partial M_{ij}$. The maximum likelihood estimators of β and Σ, say $\widehat{\beta}_{\mathrm{ML}}$ and $\widehat{\Sigma}_{\mathrm{ML}}$, are solutions of the score equations. The derivatives of $\ln\mathcal{L}(\beta,\,\Sigma)$ are obtained using the following matrix identities (Dwyer, 1967; Harville, 1999). Let x and b be $n\times 1$ vectors and Q be an $n\times n$ symmetric, positive definite matrix, then:

1. $\frac{\partial}{\partial b}(x-b)^T Q^{-1}(x-b)=-2Q^{-1}(x-b)$,
2. $\frac{\partial}{\partial Q^{-1}}\ln\left|Q^{-1}\right|=Q$,
3. $\frac{\partial}{\partial Q^{-1}}(x-b)^T Q^{-1}(x-b)=(x-b)(x-b)^T$.

From

$$\ln\mathcal{L}(\beta,\,\Sigma)=(N/2)\ln|\Sigma^{-1}|-\sum_{i=1}^{N}(Y_i-X_i\beta)^T\,\Sigma^{-1}(Y_i-X_i\beta)/2 \tag{2.2}$$

we use the chain rule, $\partial\ln\mathcal{L}/\partial\beta=\sum_i\left(\partial\mu_i^T/\partial\beta\right)\left(\partial\ln\mathcal{L}/\partial\mu_i\right)$ where $\mu_i=X_i\beta$ to obtain

$$\frac{\partial}{\partial\beta}\ln\mathcal{L}(\beta,\,\Sigma)=-\left\{\sum_{i=1}^{N}X_i^T\Sigma^{-1}(Y_i-X_i\beta)\right\}\Big/\,2$$

and identities 2 and 3 above to obtain

$$\frac{\partial}{\partial\Sigma^{-1}}\ln\mathcal{L}(\beta,\,\Sigma)=\frac{1}{2}\left\{N\Sigma-\sum_{i=1}^{N}(Y_i-X_i\beta)(Y_i-X_i\beta)^T\right\}.$$

Thus, any solution to the score equation $(\widehat{\beta},\,\widehat{\Sigma})$ must satisfy

$$\widehat{\beta}=\left(\sum_{i=1}^{N}X_i^T\widehat{\Sigma}^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N}X_i^T\widehat{\Sigma}^{-1}Y_i\right) \tag{2.3}$$

and

$$\widehat{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}(Y_i-X_i\widehat{\beta})(Y_i-X_i\widehat{\beta})^T. \tag{2.4}$$

Except in the trivial case of $n = 1$, the solutions to the score equations (2.1) do not have closed form analytic expressions when X_i is arbitrary. There are closed form solutions when X_i and Σ have special forms; estimation for these special cases, and those discussed in the examples of Chapter 1, will be discussed at the end of this chapter.

Equations (2.3) and (2.4) suggest the following iterative algorithm for computing solutions to the score equations. Set $\widehat{\Sigma}_0$ equal to any positive definite matrix, for example $\widehat{\Sigma}_0 = I$. Then for $k \geq 0$ set

$$\widehat{\beta}_k = \left(\sum_{i=1}^{N} X_i^T \widehat{\Sigma}_k^{-1} X_i \right)^{-1} \sum_{i=1}^{N} \left(X_i^T \widehat{\Sigma}_k^{-1} Y_i \right)$$

and

$$\widehat{\Sigma}_{k+1} = \sum_{i=1}^{N} \left(Y_i - X_i \widehat{\beta}_k \right) \left(Y_i - X_i \widehat{\beta}_k \right)^T / N.$$

This algorithm is sometimes referred to as Generalized Least Squares (GLS) (Carroll, Wu and Ruppert, 1988). It can be shown that each iteration of the algorithm increases the likelihood function, i.e., $\mathcal{L}(\widehat{\beta}_k, \widehat{\Sigma}_k) < \mathcal{L}(\widehat{\beta}_{k+1}, \widehat{\Sigma}_{k+1})$, because as we show in Section 3.4, this is also an instance of the EM algorithm. Thus, if the likelihood is bounded, then the sequence of likelihoods, $\mathcal{L}(\widehat{\beta}_k, \widehat{\Sigma}_k)$, $k = 1, 2, \ldots$, converges. The sequence of estimates, $(\widehat{\beta}_k, \widehat{\Sigma}_k)$, $k = 1, 2, \ldots$, however, need not converge but if it does, it converges to a solution of the score equation. Unless the likelihood is concave, this solution need not be equal to the maximum likelihood estimator of (β, Σ). See, for example, Wu (1983) or McLachlan and Krishnan (1996, pp. 92–97). We show in the next chapter that this algorithm can also be derived as an EM algorithm.

2.2 Properties

Even though $\widehat{\beta}$ is a function of $\widehat{\theta}$, it has been shown (Kackar and Harville, 1981) that in the setting where $E(Y_i) = X_i\beta$, $\widehat{\beta}(\widehat{\theta})$is unbiased for β, even in small samples, when the ML (or REML) estimate of θ is used. The basic idea of the proof relies on symmetry of the error distribution (not normality necessarily), and the fact that the variance-covariance estimates are translation invariant, even valued functions of the data. In large samples (see, e.g., Newey and McFadden, 1994, Section 2.4,), $\mathrm{vec}(\widehat{\Sigma}_{\mathrm{ML}})$ is a consistent estimator of $\mathrm{vec}(\Sigma)$ for any error distribution. Here, for any symmetric $q \times q$ matrix A, $\mathrm{vec}(A)$ denotes the $q\,(q+1)/2 \times 1$ vector obtained by stacking the columns of the lower diagonal part of A. Furthermore, under additional regularity conditions (Newey and McFadden, 1994, Section 3.2),

$$\begin{bmatrix} \sqrt{N}\left\{\widehat{\beta}_{\mathrm{ML}} - \beta\right\} \\ \sqrt{N}\left\{\mathrm{vec}\left(\widehat{\Sigma}_{\mathrm{ML}}\right) - \mathrm{vec}\left(\Sigma\right)\right\} \end{bmatrix} \xrightarrow{L} N\left(0, \Gamma\right)$$

where

$$\Gamma = -E\left\{\frac{\partial}{\partial\xi\partial\xi^T} \ln \mathcal{L}\left(\beta, \Sigma\right)\right\}^{-1}$$

and $\xi = (\beta^T, \mathrm{vec}(\Sigma)^T)^T$. Interchanging integration and differentation, the expectation of the mixed derivative matrix

$$\partial \ln \mathcal{L}\left(\beta, \Sigma\right) / \partial\beta^T \partial \text{ vec } \left(\Sigma\right)^T$$

is equal to zero, so that Γ is a block diagonal matrix with upper and lower squared blocks of dimensions equal to the number of elements of β and $\mathrm{vec}(\Sigma)$ respectively. Thus, $\widehat{\beta}_{\mathrm{ML}}$ and $\widehat{\Sigma}_{\mathrm{ML}}$ are asymptotically independent.

The asymptotic variance of $\widehat{\beta}_{\mathrm{ML}}$ can be easily derived by noticing that $\widehat{\beta}_{\mathrm{ML}} = \widehat{\beta}(\widehat{\Sigma}_{\mathrm{ML}})$. Since $\widehat{\Sigma}_{\mathrm{ML}}$ is a consistent estimator of Σ, then the asymptotic distributions of $\widehat{\beta}_{\mathrm{ML}}$ and $\widehat{\beta}(\Sigma)$ are the same (see Section 4.3). Thus the asymptotic variance of $\widehat{\beta}_{\mathrm{ML}}$ is equal to $E\left\{X_i^T\Sigma^{-1}X_i\right\}^{-1}$. This variance can be consistently estimated by

$$\left\{\frac{1}{N}\sum_{i=1}^{N} X_i^T \widehat{\Sigma}_{\mathrm{ML}}^{-1} X_i\right\}^{-1}.$$

2.3 (REML) Restricted Maximum Likelihood Estimation

The use of REML as an alternative to maximum likelihood arises in the context of mixed linear models with multiple variance components and was originally proposed by Patterson and Thompson (1971) for analyzing unbalanced block designs.

The motivating factor of REML is to obtain an inferential procedure for Σ that results in estimators and confidence regions that are approximately centered even when the dimension of the mean parameter β is large. The need for such a procedure is justified because inference based on the profile likelihood for Σ (i.e., the maximum likelihood estimator) is misleading when the dimension of β is large relative to the sample size.

As a simple example of this situation and following Barndorff-Nielsen and Cox (1994, Section 3.5) consider inference about the error variance σ^2 in the univariate linear model with normal errors. This corresponds to the case $n = 1$ with $\Sigma = \sigma^2$. The unbiased estimator of σ^2 is given by $\widetilde{\sigma}^2 = SSD/(N-p)$ where SSD is the residual sum of squares and p is the dimension of β. The MLE of $\sigma^2, \widehat{\sigma}^2$, is equal to SSD/N so that $\widehat{\sigma}^2 = \widetilde{\sigma}^2 (N-p)/N$. Suppose now that the dimension of β grows at the same rate as the sample size, that is p/N converges to a positive constant γ as N goes to ∞ (this is, of course, an idealized scenario that nevertheless helps conceptualize the difficulties that arise when the dimension of β is large). Then, with large N, $\widehat{\sigma}^2$ is approximately equal to $(1-\gamma)\widetilde{\sigma}^2$ so that the profile likelihood for σ^2 is sharply peaked at a value distant from the true value of σ^2.

The key idea of REML is to factorize the likelihood for (β, Σ) into two components, one of which is free of the parameter β and which is maximally informative about Σ. To do this, it is convenient to use the following notation: let $Y = (Y_1^T, \ldots, Y_N^T)^T$, $X = (X_1^T, \ldots, X_N^T)^T$, and $\not\Sigma$ denote an $Nn \times Nn$ block diagonal matrix with blocks equal to Σ. Then Y is MVN$(X\beta, \not\Sigma)$. The analysis proceeds by restricting attention to the likelihood based on a vector comprised of the maximum number of linear combinations of the outcomes, say $C_1^T Y, \ldots, C_q^T Y$, such that the $Nn \times 1$ vectors $C_1, \ldots, C_q$ are linearly independent and the distribution of $H^T Y$ does not depend on β, where $H^T = (C_1, C_2, \ldots, C_q)$. Harville (1977) has called the components of $H^T Y$ "error contrasts." The name arises because if $C^T Y$ has a distribution independent of β, then in particular $E(C^T Y) = C^T X \beta$ does not depend on β. But this occurs for $\beta \neq 0$ if and only if

$$C^T X = 0 \tag{2.5}$$

and so $C^T Y$ has mean zero and can be interpreted as an "error."

If, as we shall assume, X is of full rank p, then the subspace of $Nn \times 1$ vectors satisfying (2.5) is of dimension $q = Nn - p$. Thus, $H^T Y$ is a vector of dimension $q \times 1$ with elements $C_j^T Y$, $j = 1, \ldots, q$, where $C_1, \ldots, C_q$ are any subset of linearly independent $Nn \times 1$ vectors satisfying (2.5). Computation of the likelihood based on $H^T Y$ is simplified if we consider the specific error contrast vector determined by H satisfying $H^T H = I$ and $HH^T = A$ where $A = I - X\left(X^T X\right)^{-1} X^T$, (such H always exists because A is symmetric and idempotent). That H determines an error contrast vector can be readily seen from $H^T X = \left(H^T H\right) H^T X = H^T A X = 0$, where the last identity follows because $AX = 0$. There is no loss of generality in this choice, since any other full rank set of error contrasts may be obtained from H^T as $F^T = P^T H^T$, where P^T is an $(Nn-p)$ dimensional square, orthogonal matrix, and it is easily verified that $F^T F = I$ and $FF^T = A$.

Formally, the REML likelihood is defined as the likelihood of $H^T Y$. We will now show that the REML likelihood can also be written as

$$L_{\text{REML}}(\Sigma) = L_{\text{ML}}(\widehat{\beta}, \Sigma) / \left| X^T \not{\Sigma}^{-1} X \right|^{1/2},$$

where $L_{\text{ML}}(\widehat{\beta}, \Sigma)$ is the profile likelihood for Σ or alternatively,

$$L_{\text{REML}}(\Sigma) \propto f(Y; \beta, \Sigma) / f(\widehat{\beta}; \beta, \Sigma),$$

where for simplicity, we write $\widehat{\beta}(\not{\Sigma}) = \widehat{\beta}$. Since $H^T Y$ is just a linear tranformation on Y, it follows that $H^T Y$ has mean zero, and variance covariance matrix $H^T \not{\Sigma} H$. Hence using the definition of REML,

$$L_{\text{REML}} = e^{-Y^T H (H^T \not{\Sigma} H)^{-1} H^T Y/2} \Big/ \left| H^T \not{\Sigma} H \right|^{1/2}. \tag{2.6}$$

We will use the following identities:

$$H(H^T \not{\Sigma} H)^{-1} H^T = \not{\Sigma}^{-1} - \not{\Sigma}^{-1} X (X^T \not{\Sigma}^{-1} X)^{-1} X^T \not{\Sigma}^{-1} \tag{2.7}$$

(see Searle, Casella and McCulloch, 1992, Appendix M4.f), and

$$\left| H^T \not{\Sigma} H \right|^{1/2} \propto |\not{\Sigma}|^{1/2} \left| X^T \not{\Sigma} X \right|^{1/2}. \tag{2.8}$$

This last identity is proved at the end of this section.

From (2.7) it follows that

$$\begin{aligned} Y^T H (H^T \not{\Sigma} H)^{-1} H^T Y &= Y^T \not{\Sigma}^{-1} Y - Y^T \not{\Sigma}^{-1} X (X^T \not{\Sigma}^{-1} X)^{-1} X^T \not{\Sigma}^{-1} Y \\ &= (Y - X\widehat{\beta})^T \not{\Sigma}^{-1} (Y - X\widehat{\beta}) \\ &= \Sigma_i (Y_i - X_i \beta)^T \Sigma^{-1} (Y_i - X_i \beta). \end{aligned}$$

where

$$\widehat{\beta}(\not{\Sigma}^{-1}) = (X^T \not{\Sigma}^{-1} X)^{-1} X^T \not{\Sigma}^{-1} Y.$$

Thus from (2.6, 2.7, 2.8) and using the block diagonal structure of $\not{\Sigma}$ we have

$$L_{\text{REML}} = e^{-\text{tr}(\not{S} \not{\Sigma}^{-1})/2} \Big/ |\not{\Sigma}|^{1/2} \left| \Sigma_i X_i \Sigma^{-1} X_i \right|^{1/2}$$

for $S = \Sigma_i (Y_i - X_i \widehat{\beta})(Y_i - X_i \widehat{\beta})^T$. It follows immediately that the previous two characterizations of the REML likelihood are also true:

1. The REML likelihood is the profile likelihood, $L(\beta = \widehat{\beta}_{\text{ML}}, \Sigma)$, multiplied by $|\text{var}\widehat{\beta}(\Sigma)|^{1/2}$.

2. The REML likelihood is the likelihood of the full sample Y, conditional on $\widehat{\beta}(\Sigma)$.

These characterizations point to a drawback of REML inference. While the profile likelihood is invariant under reparameterizations, the term $\ln|X^T \; /\Sigma^{-1}X|$, being a function of the design matrix, is not invariant under reparameterizations of the parameter β. This has the consequence that inference about Σ using REML will change with different, but equivalent, formulations of the linear model. This drawback, however, is asymptotically negligible provided the dimension of β is not too large. For further discussion see Bardnorff-Nielsen and Cox (1994, Section 4.4).

The use of the REML likelihood can also be based on marginal sufficiency arguments given in Sprott (1975). That is, H^TY is marginally sufficient because:

1. the distribution of H^TY depends only on Σ,
2. the distribution of $\widehat{\beta}$ contains no information about Σ in the absence of knowledge about β, and
3. together H^TY and $\widehat{\beta}$ are a full rank transformation of the data Y.

See also Kalbfleish and Sprott (1970, 1973, 1976).

It is straightforward to show that the REML score equations for Σ are

$$N\,\Sigma - \sum_{i=1}^{N} X_i \left(\sum_{i=1}^{N} X_i^T\,\Sigma^{-1}\,X_i \right)^{-1} X_i^T - S\,(\Sigma) = 0.$$

so that $\widehat{\Sigma}_{\text{REML}}$ satisfies

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

where

$$\widehat{\beta}_{\text{REML}} \equiv \widehat{\beta}\left(\widehat{\Sigma}_{\text{REML}}\right) \quad \text{and} \quad V_{\widehat{\beta}} = \left(\sum_{i=1}^{N} X_i^T\,\widehat{\Sigma}_{\text{REML}}\,X_i\right)^{-1}.$$

The iterative computational algorithm given for ML in Section 2.1 can be readily adapted in this setting. Specifically, at stage k, $\widehat{\beta}_k = \widehat{\beta}(\widehat{\Sigma}_k^{-1})$

where $\widehat{\Sigma}_0$ is an arbitrary positive definite matrix but now

$$\begin{aligned}\widehat{\Sigma}_{k+1} &= \sum_{i=1}^{N}\Bigg[(Y_i - X_i\,\widehat{\beta}_k)\,(Y_i - X_i\,\widehat{\beta}_k)^T \\ &\qquad + X_i \left(\sum_{i=1}^{N} X_i^T\,\widehat{\Sigma}_k^{-1}\,X_i\right)^{-1} X_i^T\Bigg] \Big/ N.\end{aligned}$$

As with ML estimation, the sequence $(\widehat{\beta}_k, \widehat{\Sigma}_k)$ need not converge, but when it does, the limit of $\widehat{\Sigma}_k$, $k = 1, 2, \ldots$, is a solution of the REML score equations.

The question naturally arises as to which of the ML or REML estimators is to be preferred. As indicated in the introduction, the REML analysis is motivated by the desire to adjust the ML estimator of Σ for lost degrees of freedom. More precisely, the hope is to find an estimator that, like the MLE of Σ, is asymptotically efficient, but in contrast to the MLE, is an unbiased estimator of Σ. The REML estimator of Σ is asymptotically efficient (see, e.g., Cressie and Lahiri, 1993). However, quite generally, the REML estimator need not be unbiased. In Section 3.6 we discuss special cases where the REML estimate is unbiased (and ML is not). In addition, we present examples where the REML estimator has a closed analytical expression.

Proof of identity in (2.8).

To show that

$$|\Sigma|\,\left|X^T\Sigma^{-1}X\right| \propto \left|H^T\Sigma H\right|$$

where Σ, X, H and Y were defined in Section 2.3, we will define a full rank transformation on Y as TY, where

$$T = \begin{pmatrix} H^T \\ G \end{pmatrix} \text{ so that } TY = \begin{pmatrix} H^TY \\ GY \end{pmatrix}$$

and we choose $G = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$, so that $GY = \widehat{\beta}(\Sigma) \equiv \widehat{\beta}$.

It is easily checked that $H^T\Sigma G = 0$, so $T\ \Sigma T^T$ is block diagonal, and hence

$$\left|T\Sigma T^T\right| = \left|H^T\Sigma^{-1}H\right|\left|(X^T\Sigma^{-1}X)^{-1}\right|.$$

However we also have that $\left|T\ \Sigma T^T\right| = |\Sigma|\left|TT^T\right|$, and using formulas for determinants of a partitioned matrix (Harville, 1997, pp. 188–189), we have

$$\det\begin{pmatrix} H^TH & H^TG^T \\ GH & GG^T \end{pmatrix} = \left|H^TH\right|\left|GG^T - GH(H^TH)^{-1}H^TG^T\right|.$$

Now recall that $H^T H = I$ and $HH^T = I - X(X^T X)^{-1} X^T$, hence

$$\begin{aligned} \left|TT^T\right| &= \left|G(I - HH^T)G^T\right| = \left|GX(X^T X)^{-1} X^T G^T\right| \\ &= \left|X^T X\right|^{-1}, \end{aligned}$$

whence it follows that

$$\left|H^T \not\Sigma H\right| \left|X^T \Sigma^{-1} X\right|^{-1} = |\not\Sigma| \left|X^T X\right|^{-1},$$

and $\left|H^T \not\Sigma H\right| \propto |\not\Sigma| \left|X^T \Sigma^{-1} X\right|$.

2.4 REML Estimation: A Bayes Approach

The REML estimator of Σ can also be given a Bayesian interpretation (Searle, 1987, Section 9.2) as follows. Suppose we assign both β and Σ uniform (improper) prior distributions over their respective sample spaces. The joint posterior distribution of $(\beta,\, \Sigma)$ is proportional to the likelihood for $(\beta,\, \Sigma)$, where the constant of proportionality depends on the data but not on $(\beta,\, \Sigma)$. Letting $p_s(\beta,\, \Sigma)$ denote the joint posterior distribution of $(\beta,\, \Sigma)$, we thus have $p_s(\beta,\, \Sigma) \propto \mathcal{L}(\beta,\, \Sigma)$. To eliminate the nuisance parameter β, we base inference about Σ on the marginal posterior distribution of Σ, $p_s(\Sigma)$, obtained by integrating $p_s\,(\beta,\, \Sigma)$ over β:

$$p_s\,(\Sigma) \propto \int \mathcal{L}(\beta,\, \Sigma)\, d\beta. \tag{2.9}$$

To show $p_s(\Sigma)$ is proportional to the REML likelihood, first write

$$\begin{aligned} \mathcal{L}(\beta, \Sigma) \propto |\Sigma|^{-N/2} \exp \Bigg\{ -\frac{1}{2} \sum_{i=1}^{N} &(Y_i - X_i\, \widehat{\beta})^T\, \Sigma^{-1} (Y_i - X_i\, \widehat{\beta}) \\ &+ (\widehat{\beta} - \beta)^T \left(\sum_{i=1}^{N} X_i^T \Sigma^{-1} X_i \right) (\widehat{\beta} - \beta) \Bigg\} \end{aligned}$$

where $\widehat{\beta} = \widehat{\beta}(\Sigma^{-1})$. Then use the identity

$$\int \exp \left\{ -\frac{1}{2} t^T \Omega^{-1} t \right\} db = (2\pi)^{n/2}\, |\Omega|^{1/2},$$

where Ω is any $n \times n$ positive definite matrix, to obtain that

$$p_s(\Sigma) \propto |\Sigma|^{-N/2} \exp \left\{ \mathrm{tr} \Sigma^{-1} S\,(\Sigma)\,/2 \right\} \left| \sum_{i=1}^{N} X_i^T \Sigma^{-1} X_i \right|^{-1/2},$$

so that right hand side is precisely the REML likelihood. Thus, inferences based on the REML likelihood and inferences based on $p_s(\Sigma)$ are identical. In particular, the REML estimate of Σ is the posterior mode. Note that we can alternatively view the REML likelihood as an integrated likelihood where we have integrated out the nuisance parameter β.

2.5 Patterned Σ

When Σ is a patterned matrix, that is $\Sigma = \Sigma(\theta)$ is a function of a vector of parameters θ of dimension less than $n(n+1)/2$, both the ML and REML score equations for θ can be obtained using the chain rule to take derivatives with respect to θ. See, for example, Jennrick and Schluchter (1986). There are several computing packages which provide subroutines for computing ML and REML estimators of θ under a variety of model choices for $\Sigma(\theta)$.

2.6 Closed Form Solutions for $\widehat{\beta}_{\mathrm{ML}}$, $\widehat{\Sigma}_{\mathrm{ML}}$ and $\widehat{\Sigma}_{\mathrm{REML}}$.

For certain design matrices, one can obtain closed form solutions for $\widehat{\beta}_{\mathrm{ML}}$ and $\widehat{\Sigma}_{\mathrm{ML}}$ or $\widehat{\Sigma}_{\mathrm{REML}}$ in the setting where Σ is unstructured and all subjects are measured at all n occasions. Studying these cases is instructive since the resulting estimators show when $\widehat{\beta}_{\mathrm{ML}}$ does not depend on Σ and enables us to quantify the bias of $\widehat{\Sigma}_{\mathrm{ML}}$ and $\widehat{\Sigma}_{\mathrm{REML}}$.

Case 1. The simplest case in which closed form solutions exist corresponds to the classical MANOVA or multivariate regression setting discussed in Section 1.2.3. Furthermore, in this setting the REML estimator of Σ is unbiased. Under the MANOVA model of Section 1.2.3, each $X_i = a_i^T \otimes I_n$, where a_i is a $k \times 1$ vector of subject specific covariates, I_n is an $n \times n$ identity matrix, and $\otimes$ is Kroneker matrix product (Harville, 1999). Effectively each of the n variables has a separate regression model with the same k predictors for each outcome. Let A denote the $N \times k$ matrix with rows given by the a_i^T, i.e., A is the usual design matrix for a single response variable.

For this design, it straightforward to show (see, e.g., Johnson and

Wichern, 1992) that

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

and hence

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

Since $E(Y_i) = X_i\beta$, it follows that $\widehat{\beta}_{\mathrm{ML}}$ is unbiased. To compute the bias of $\widehat{\Sigma}_{\mathrm{ML}}$, we will use the following matrix identities:

(i) $\left(a_i^T \otimes I_n\right)^T = a_i \otimes I_n$,

(ii) for conformable matrices A, B, C, D, $(A \otimes B)(C \otimes D) = AC \otimes BD$,

and

(iii) for nonsingular matrices A and B $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

Also from Johnson and Wichern (1990) we have

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

thus,

$$\begin{aligned}
\widehat{\Sigma}_{\mathrm{ML}} = & \left[\sum_{i=1}^{N}(Y_i - X_i\beta)(Y_i - X_i\beta)^T + \sum_{i=1}^{N} X_i(\widehat{\beta}_{\mathrm{ML}} - \beta)(\widehat{\beta}_{\mathrm{ML}} - \beta)X_i^T \right.\\
& -\sum_{i=1}^{N}(Y_i - X_i\beta)(\widehat{\beta}_{\mathrm{ML}} - \beta)^T X_i^T \\
& \left. -\sum_{i=1}^{N} X_i(\widehat{\beta}_{\mathrm{ML}} - \beta)(Y_i - X_i\beta)^T \right] \Big/ N.
\end{aligned}$$

Taking expectations and setting $X_i = a_i^T \otimes I$, we find

$$E(\widehat{\Sigma}_{\mathrm{ML}}) = \Sigma - \frac{1}{N} \sum_{i=1}^{N} X_i \operatorname{var}\left(\widehat{\beta}_{\mathrm{ML}}\right) X_i^T. \tag{2.10}$$

Using the matrix identities given above, one can show that

$$\text{var}\,\widehat{\beta}_{\text{ML}} = (A^T A)^{-1} \otimes \Sigma$$

and thus

$$E\left(\widehat{\Sigma}_{\text{ML}}\right) = \Sigma - \frac{1}{N}\left(\sum_{i=1}^{N} a_i^T (A^T A)^{-1} a_i\right) \otimes \Sigma.$$

But

$$\sum_{i=1}^{N} a_i^T (A^T A)^{-1} a_i = \text{tr}\left\{\left(\sum_{i=1}^{N} a_i a_i^T\right)(A^T A)^{-1}\right\} = \text{tr}\,(I) = k,$$

so that

$$E\left(\widehat{\Sigma}_{\text{ML}}\right) = \Sigma(1 - k/N).$$

Thus $\widehat{\Sigma}_{\text{ML}}$ is biased, with bias going to zero as $k/N \to$ zero.

Since $\widehat{\beta}_{\text{ML}}$ does not depend upon $\widehat{\Sigma}_{\text{ML}}$, $\widehat{\beta}_{\text{REML}} = \widehat{\beta}_{\text{ML}} = \widehat{\beta}_{\text{OLS}}$ and

$$\widehat{\Sigma}_{\text{REML}} = \widehat{\Sigma}_{\text{ML}} + \frac{1}{N}\sum_{i=1}^{N} X_i \text{ var}\left(\widehat{\beta}_{\text{ML}}\right) X_i^T.$$

and hence from (2.10) it follows that $E(\widehat{\Sigma}_{\text{REML}}) = \Sigma$.

Case 2. The balanced and complete growth curve model discussed in Example 1.2.4 provides another example where the ML and REML estimators have closed form analytical expressions. In this case,

$$X_i = a_i^T \otimes Z$$

where a_i is as before and Z is some suitably defined $n \times q$ design matrix specifying a "design on time" as discussed in Example 1.2.4. Recall that in this setting we may write:

$$E(Y_i) = X_i \beta = Z \Delta a_i$$

with Δ defined as in Section 1.2.4. Grizzle and Allen (1969) (see also Khatri, 1966) have obtained a closed form solution for Δ in this case as

$$\widehat{\Delta}_{\text{ML}} = (Z^T S^{-1} Z)^{-1} Z^T S^{-1} Y A (A^T A)^{-1} \tag{2.11}$$

where Y is now an $n \times N$ data matrix whose columns are Y_i, and

$$S = Y\left[I - A(A^T A)^{-1} A^T\right] Y^T,$$

where I denotes the identity matrix with dimension n. Notice that we may also write

$$\widehat{\beta}_{\mathrm{ML}} = \mathrm{vec}\ \widehat{\Delta}_{\mathrm{ML}} = \left[(A^T A)^{-1} A^T \otimes (Z^T S^{-1} Z)^{-1} Z^T S^{-1}\right] \ \mathrm{vec}\,(Y)\,.$$

The Grizzle and Allen (1969) derivation is instructive and leads to a proof that $\widehat{\beta}_{\mathrm{ML}}$ is unbiased even though $\widehat{\beta}_{\mathrm{ML}}$ depends upon S. The approach is to make a linear transformation from Y_i to B_i of the form

$$\begin{pmatrix} B_{1i} \\ B_{2i} \end{pmatrix} = \begin{pmatrix} Z_1^T \\ \\ Z_2^T \end{pmatrix} Y_i$$

where Z_1 and Z_2 are full rank matrices of dimension $n \times r$ and $n \times n$, with $n = r + q$, selected so that

$$Z_1^T Z = I$$

and

$$Z_2^T Z = 0.$$

Notice that a natural choice for Z_1 is $Z(Z^T Z)^{-1}$, and for Z_2 is a matrix, $\{I - Z(Z^T Z)^{-1} Z^T\}$, with columns where $v_j, j = 1, \ldots, r$, are $n \times 1$ vectors linearly independent with the columns of Z. Since $E(Z_2 Y_i) = 0$, the marginal distribution of $Z_2 Y_i$ does not depend upon β, hence we can write

$$\begin{aligned} \Pi\ f(Y_i|\beta, \Sigma) &\propto \ \Pi\ f(B_i|\beta, \Sigma) \\ &= \Pi\ f(B_{1i}|\beta, \Sigma, B_{2i}) f(B_{2i}|\Sigma) \end{aligned}$$

and the MLE of β is obtained by maximizing the first component only. Because B_{1i} and B_{2i} are jointly multivariate normally distributed, the conditional mean of B_{1i} given B_{2i} is linear in B_{2i}. Thus, letting

$$\begin{aligned} \theta &= \mathrm{cov}(B_{1i}, B_{2i})(\mathrm{var}\ B_{2i})^{-1} \\ &= Z_1^T \Sigma Z_2 (Z_2^T \Sigma Z_2)^{-1}. \end{aligned}$$

we have

$$\begin{aligned} E(B_{1i}|B_{2i}; \Sigma, \beta) &= \Delta a_i + \theta B_{2i} \\ &= [\Delta, \theta] \begin{pmatrix} a_i \\ B_{2i} \end{pmatrix} \\ &= \tau\ W_i, \end{aligned}$$

where Δ is defined as in Section 1.2.4, τ is a $q \times (k + r)$ matrix of parameters, and W_i is a $(k + r) \times 1$ vector of covariates. But this is identical to our model in Case 1 where we write

$$E(B_{1i}) = (W_i^T \otimes I)\Omega$$

and Ω =vec(τ). Hence the ML estimate of τ is OLS, and can be written as

$$\widehat{\tau} = B_1 W (W^T W)^{-1}$$

where B_1 is the $q \times N$ data matrix whose columns are B_{1i} and W is the $N \times (k + r)$ design matrix whose rows are W_i^T. By partitioning $\widehat{\tau}$ and using standard matrix identities, the result for $\widehat{\Delta}$ in (2.11) is obtained. It also follows that $E(\widehat{\Delta}|B_2; \Delta, \Sigma) = \Delta$ where B_2 is the vector with components B_{2i}, $i = 1, \ldots, N$, and therefore $\widehat{\Delta}$ is also unconditionally unbiased for Δ.

Case 3. These results concerning closed form solutions have been extended by Szatrowski (1980), Szatrowski and Miller (1980) and Lange and Laird (1989) for the setting where Σ takes a random effects structure, i.e.,

$$\Sigma = Z_c D Z_c + \sigma^2 I,$$

where Z_c is a subset of the columns of Z of dimension $n \times c$, D is a positive definite matrix and I is the $n \times n$ identity matrix. Here it can be shown that $\widehat{\beta}_{\text{ML}} = \widehat{\beta}_{\text{REML}} = \widehat{\beta}_{\text{OLS}}$, and simple closed form expressions can be derived for the ML and REML estimates of D and σ^2. Lange and Laird (1989) showed also that the REML estimates of Σ are unbiased and the ML estimates have bias which goes to zero as $k/N \to 0$, for fixed q.